Towards a better understanding of protein interaction specificities in cell signalling - PDZ domains in the spotlight of computational and experimental approaches

Katja Luck

UNIVERSITÉ DE STRASBOURG

UNIVERSITÉ DE STRASBOURG

irebs
Institut de recherche
de l'École de biotechnologie de Strasbourg

ÉCOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTE

UMR 7242 Biotechnologie et Signalisation Cellulaire

THÈSE présentée par :

Katja LUCK

soutenue le : 19 octobre 2012

pour obtenir le grade de : **Docteur de l'Université de Strasbourg**

Discipline / Spécialité : Bio-Informatique

Vers une meilleure connaissance de la spécificité des interactions protéiques dans la signalisation cellulaire – les domaines PDZ au centre des approches informatiques et expérimentales

THÈSE dirigée par :

M. TRAVE Gilles	Dr., Université de Strasbourg

RAPPORTEURS :

M. CESARENI Gianni	Prof., Université de Rome Tor Vergata
M. WOLFF Nicolas	Dr., Institut Pasteur, Paris

AUTRES MEMBRES DU JURY :

Mme DEJAEGERE Annick	Dr., IGBMC, Illkirch

If we knew what we were doing,
it would not be called research.

(Albert Einstein)

Acknowledgements - Remerciements - Danksagung

Tout d'abord, je remercie mon directeur de thèse Gilles Travé pour sa motivation "contagieuse", ses idées et son inspiration infinite, sa disposition et patience de discuter à tout moment sur tout sujet avec ses thésards, son effort d'améliorer ensemble des manuscrits de publication, son engagement à la formation des thésards et sa compréhension pour la famille.

Je remercie Sadek Fournane pour sa patience et persévérance à conduire des experiences presque infinites de BIAcore et pour m'avoir appris le traitement des données de BIAcore. Je remercie Sebastian Charbonnier pour sa disposition de partager avec moi à tout moment son expérience scientifique ainsi que ses connaissances sur Strasbourg et la région. Je le remercie également pour son soutien et conseil dans des moments difficiles de thèse. Je remercie Katia Zanier pour m'avoir accueilli chez elle pendant mes premières deux semaines à Strasbourg ainsi que pour son soutien et conseil dans des questions scientifiques et interhumaines.

Je remercie Yves Nominé et Marc-André Delsuc pour leur aide en maths, notamment en statistique, en bio-physique et en programmation. Je remercie Bruno Kieffer qui était toujours prêt à répondre à mes questions. Je le remercie pour des discussions interéssantes sur la dynamique des protéines et ses fêtes d'été qui étaient extrêmement sympathiques. Je remercie Claude Ling et Sylvie Bulot pour leur disponibilité et aide informatique.

Je remercie toute l'équipe onco ainsi que les membres de l'équipe d'Annick Déjaegère et de Bruno Kieffer pour toute aide scientifique et l'ambiance amicale qui m'a constamment entouré pendant mes années de thèse. Je remercie l'équipe d'Olivier Poch pour m'avoir intégré dans leurs réunions et de m'avoir donnée un "refuge" bio-informatique à Strasbourg. Je remercie l'équipe de Danièle Altschuh pour leur assistance pendant l'analyse des données de BIAcore.

Je remercie Renaud Vincentelli et Yves Jacob pour notre collaboration constructive et productive. C'était toujours avec beaucoup de plaisir de travailler avec vous sur des projets communs.

Je remercie la Région Alsace et l'Association de la Recherche contre le Cancer pour avoir financé ma thèse ainsi que le Collège Doctoral Européen pour leur soutien.

I thank all members of my thesis committee including the invited members Danièle Altschuh and Toby Gibson for having made the effort to read and judge my thesis work as well as to come to Strasbourg for the defence.

I thank in particular Toby Gibson for having introduced me to the fascinating world of linear motifs, for having shared with me his ideas and thoughts about cell regulation and signalling processes, and for his ongoing support and supervision.

I thank the previous and current members of Toby Gibson's group for the friendly atmosphere they have always provided to me upon my visits at EMBL as well as for their willingness to assist me whenever I had a question.

I thank the editorial board of the special issue on Protein Modules in the FEBS journal for having provided to Gilles and me the opportunity to write a review on sequence context in the PDZ domain family.

Je remercie Emeline et Matthieu pour avoir partagé avec moi les joies et les peines de la vie en thèse, pour des soirées franco-allemandes délicieuses et inoubliables, pour avoir bien rigolé sur nos marottes allemandes et françaises, et pour des voyages très sympathiques aux capitales de l'Europe. Je remercie Blandine, Emilie, Jessica, Alex, Patty et Jonathan pour de nombreuses fêtes et sorties organisées. J'ai toujours beaucoup aimé de passer du temps avec vous tous.

Je remercie toute l'équipe de la crèche de l'esplanade pour assurer une très haute qualité de la garde des enfants à laquelle j'ai pu donner toute ma confiance. Je remercie particulièrement Sophie pour son énorme motivation et sa façon naturelle et très sympathique de s'occuper des enfants, notamment de Samuel. Sachant que Samuel était dans deux bonnes mains me permettait de me concentrer à 100% sur mon travail.

Je remercie Etienne pour avoir été toujours disponible de s'occuper de notre tortue pendant nos nombreux jours d'absence de Strasbourg.

Ich danke meinem großen Schatz dafür, dass er bei mir ist, mich so nimmt, wie ich bin und liebt, und dafür, dass er für jede meiner verrückten Ideen offen ist. Ich danke meinem kleinen Schatz dafür, dass er mir zeigt, was wirklich wichtig ist, dass das Leben voller Überraschungen und Abenteuer steckt und dass nichts selbstverständlich ist.

Ich danke meinen Eltern für ihre Liebe und Unterstützung jederzeit und in aller Hinsicht, angefangen bei der Ausbildung, die sie mir ermöglicht haben und die letztlich sogar zum Doktortitel geführt hat bis hin zu all den Leckerbissen, die stets unsere Kühltruhe füllten. Insbesondere möchte ich meiner Mutti danken für ihre Mühen diese Doktorarbeit Korrektur gelesen zu haben.

Ich danke Gerald und Gisela für ihre Großzügigkeit und stetige Bereitschaft uns während der Doktorarbeit zu unterstützen sowie für all die zahlreichen gemeinsamen Urlaube und damit verbundene Erholung.

Ich danke allen meinen Angehörigen für deren Zusammenhalt, Interesse und Zuspruch. Ihr gebt mir Sicherheit und Selbstvertrauen.

Summary in French

La signalisation cellulaire comprend tous les processus qui permettent à une cellule de recevoir des signaux externes, de les transformer et propager dans la cellule ainsi que de leur répondre. Ces processus dépendent fondamentalement des protéines et de leurs interactions. Beaucoup de protéines impliquées dans la signalisation cellulaire possèdent une architecture modulaire comprenant des motifs linéaires courts (SLiMs) et des domaines globulaires [1] (voir Figure 0.1). Les domaines globulaires sont des régions de séquence continue capables de se replier indépendamment du reste de la protéine [2]. Des SLiMs sont de courts fragment de séquence ne dépassant généralement pas dix résidus, préférentiellement localisés dans des régions désordonnées, et qui interagissent avec des domaines globulaires en adoptant une structure secondaire [3]. Les interactions entre SLiMs et domaines globulaires constituent une fraction importante des interactions protéiques participant à la signalisation cellulaire.

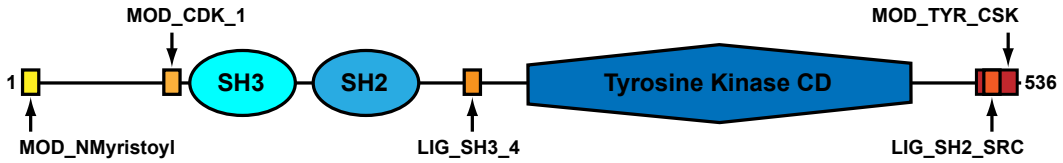


Figure 0.1. Architecture modulaire de la protéine humaine proto-oncogene tyrosine-protein kinase Src. Les domaines globulaires sont illustrés en bleu, les SLiMs en jaune/orange/rouge. Les domaines de la protéine Src ont été prédits avec le serveur web SMART [4], les SLiMs de Src validés expérimentalement ont été extraits de la resource ELM [5]. MOD_NMyristoyl: site de N-myristoylation, MOD_CDK_1: site de phosphorylation de Ser/Thr cyclin dependent protein kinase (CDK), LIG_SH3_4: site de liaison des domaines SH3, LIG_SH2_SRC: site de liaison des domaines SH2, MOD_TYR_CSK: site de phosphorylation de tyrosine des C-Src kinases (CSK). CD=domaine catalytique.

Différents types d'interactions domaine-SLiM ont été identifié. Celles impliquant des domaines PDZ sont parmi les plus étudiées. Les domaines PDZ participent à la polarité cellulaire, au trafic membranaire, et généralement à l'organisation de complexe protéiques [6]. Majoritairement, les PDZs reconnaissent des SLiMs situés à l'extrémité C-terminale de leurs protéines-cibles (voir Figure 0.2). Les motifs de liaison aux PDZs (PBMs) présentent habituellement un résidu hydrophobe à la dernière position (position 0) et peuvent être classés en trois sous-groupes basé sur le résidu en position -2 (Thr/Ser, Asp/Glu, ou des résidus hydrophobes) [6]. Environ 270 domaines PDZ et des milliers de PBMs potentiels ont été identifiés dans le protéome humain. Beaucoup d'études expérimentales et informatiques ont été effectuées pour essayer de déchiffrer les règles de la spécificité qui définissent quel domaine PDZ interagira de préférence

avec quel PBM. Les connaissances obtenues à partir de telles études permettraient en général de mieux comprendre les mécanismes structuraux de la reconnaissance domaine-motif et en particulier de concevoir des prédicteurs fiables des interactions PDZs-motifs.

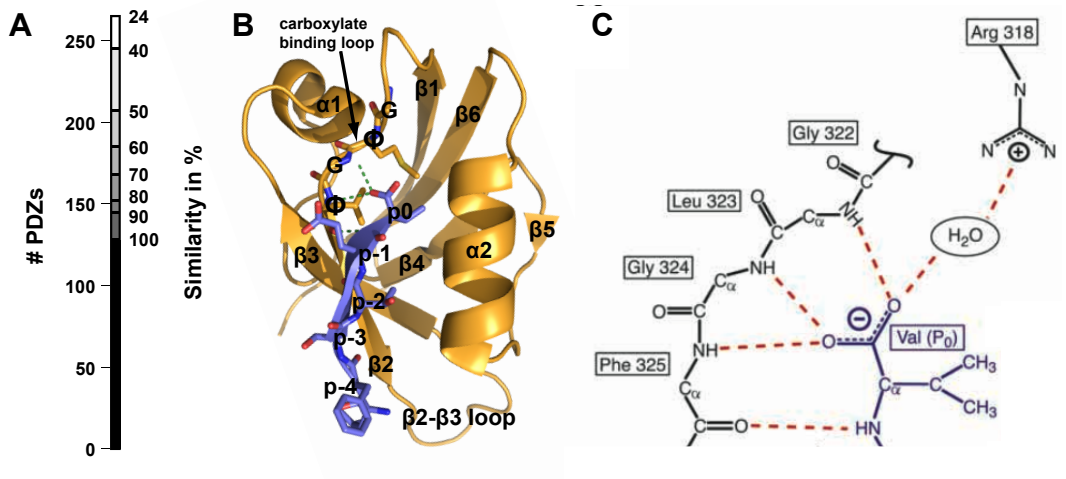


Figure 0.2. Aspects structuraux des domaines PDZ et de la reconnaissance des peptides. A: L'information structurale accessible sur des domaines PDZ. Ce diagramme illustre le nombre des domaines PDZ humains pour lesquels il existe une structure d'un domaine PDZ dans la base de données PDB avec une similarité de séquence de X % ou au-delà, calculé à partir des alignements de séquence locaux des recherches BLAST (par exemple pour 152 PDZs humains il existe une structure d'un PDZ dans le PDB avec au moins 80 % similarité de sequence.) B: Repliement canonique des domaines PDZs et liaison des peptides C-terminaux. Structure du domaine PDZ de la protéine AF6 liée au peptide C-terminal (LFSTEV) dérivé de la protéine Bcr (PDB ID: 2AIN [7]). Les éléments de structure secondaire, les positions de peptide et la signature générale de la boucle de liaison du groupe carboxyle (GϕGϕ, ϕ représente un résidu hydrophobe) sont indiqués. Les lignes hachurées représentent des liaisons hydrogènes qui sont établies entre Val à la position de peptide p0 et la boucle de liaison du groupe carboxyle. C: Reconnaissance d'un peptide C-terminal. Des atomes du domaine PDZ et du peptide sont colorés en noir et bleu, respectivement. Les liaisons hydrogènes entre le résidu Val au C-terminus du peptide et la boucle de liaison du groupe carboxyle du domaine PDZ sont indiquées avec des lignes hachurées. La figure C a été adaptée de [6] et montre une structure de complexe impliquant PDZ3 de PSD-95 [8].

Cette thèse a été centrée sur deux études pionnières à grande échelle, publiées dans des journaux de grand impact, qui avaient combiné des approches expérimentales et informatiques pour explorer la spécificité des interactions aux PDZs. Tonikian *et al.* [9] avaient produit des données de "peptide phage display" établies pour 54 domaines PDZ humains qu'ils avaient ensuite utilisées pour construire des matrices de profils de séquence (PSSMs) afin de cribler le protéome humain pour des partenaires de liaison potentiels à ces PDZs (voir Figure 0.3). L'équipe de MacBeath [10] avait utilisé des "microarrays" combinés avec des mesures de polarisation de fluorescence pour déterminer les constantes de dissociation entre 157 domaines PDZs et 217 peptides C-terminaux dérivés des protéines de la souris. Ils ont utilisé ces données pour

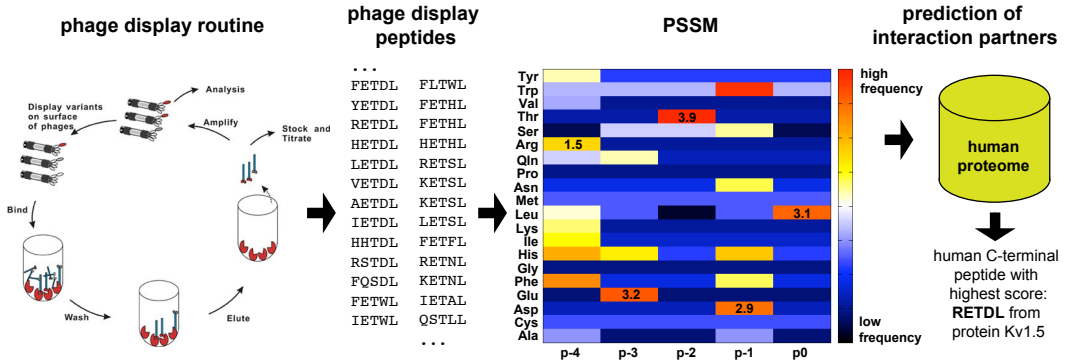


Figure 0.3. Les principes de prédiction des interactions utilisant des PSSMs. De gauche à droite: Pendant les cycles de phage display, des peptides C-terminaux avec une affinité haute pour un domaine PDZ (dans l'exemple PDZ3 de SCRIB) sont sélectionnés, séquencés et alignés. La fréquence de chaque acide aminé à chaque position de peptide (dans cet exemple, seulement, les dernières cinq positions de peptide sont considérées) est calculée et sauvegardée dans une PSSM. Dans l'exemple, la fréquence de chaque acide aminé à chaque position de peptide est illustrée avec un schéma de couleur où les résidus fréquents sont représentés en orange/rouge et les résidus plus rares en bleu foncé. La PSSM peut ensuite être utilisée pour cribler le protéome humain pour des peptides C-terminaux similaires aux peptides issus du phage display en calculant pour chaque peptide du protéome un score qui résulte de la somme des fréquences de ses résidus. Le peptide C-terminal humain qui a obtenu le score le plus élevé dans cet exemple, a la séquence RETDL. Les fréquences de ces résidus sont indiquées sur la PSSM. L'image de la routine de phage display est empruntée à www.creative-biolabs.com.

développer un algorithme de prédiction des interactions PDZ-peptides [11].

En plus de ces approches à grande échelle, des études focalisées sur des interactions PDZ-peptide particulières, ont proposé que le contexte de séquence (c'est-à-dire, des séquences contigües ou non contigües des PDZs et des PBMs, voir Figure 0.4) influencent leurs affinités de liaison. Cependant, peu d'études ont analysé l'impact potentiel du contexte de séquence sur la spécificité des interactions. La spécificité d'interaction ne peut être déterminée qu'en comparant les affinités de liaison d'une protéine (ou d'un fragment d'une protéine) avec ses partenaires d'interaction potentiels et multiples. Une protéine est spécifique, si elle montre pour la majorité de ces partenaires potentiels une affinité plus faible et pour une minorité d'entre eux une affinité plus forte. Donc, la spécificité est une valeur relative en comparaison avec l'affinité, qui est une valeur absolue.

Dans cette thèse, nous avons posé deux questions : Premièrement, les prédicteurs d'interaction PDZ-peptides actuels peuvent-ils être utilisés pour des prédictions des réseaux d'interactions protéiques impliquant les domaines PDZ ? Deuxièmement, le contexte de séquence a-t-il une influence sur la spécificité des interactions aux PDZs, et si oui, quels sont les mécanismes fondamentaux ? En particulier, la spécificité des interactions entre des constructions PDZ et PBM minimales peut-elle être modifiée

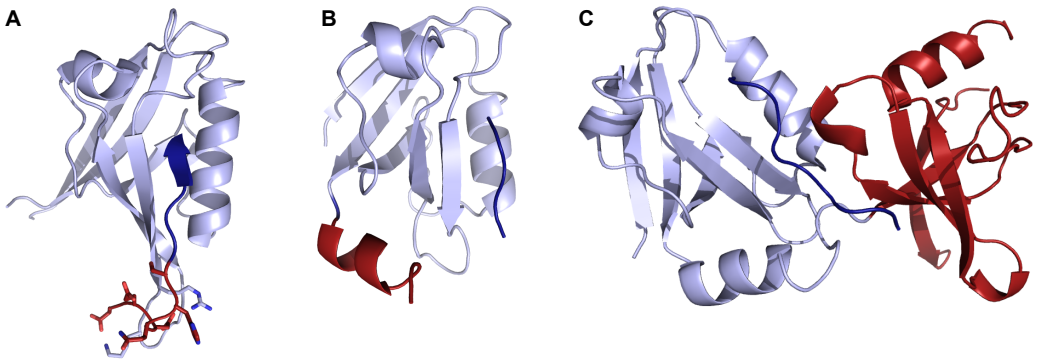


Figure 0.4. Exemples de contexte de séquence chez des PDZs et des PBMs. Les domaines PDZ sont colorés en bleu clair, les peptides C-terminaux en bleu foncé et les parties des structures qui constituent du contexte de séquence en rouge. A: PDZ3 de Par3 lié à un PBM étendu dérivé de PTEN (code PDB: 2K20 [12]). Les résidus du PBM étendu et les résidus cléfs d'interaction du domaine PDZ sont montrés en "sticks". B: Domaine PDZ3 de PSD-95 lié au PBM dérivé de CRIPT (code PDB: 1BE9 [8]). PDZ3 possède une hélice α C-terminale supplémentaire qui influence la liaison des peptides [13,14]. C: Domaine PDZ3 de ZO-1 lié au PBM dérivé de JAM-A (code PDB: 3TSZ [15]). Le domaine SH3 voisin est localisé au C-terminus de PDZ3 et influence la liaison des peptides au PDZ3.

par la présence de séquences flanquantes ?

Pour aborder le problème de la spécificité des interactions aux PDZs et de leur prédiction, nous avons combiné des approches expérimentales et informatiques. Nous nous sommes d'abord concentrés sur les données de phage display et leur utilisation potentielle pour la prédiction des interactions protéiques. Nous avons construit des PSSMs basés sur des données de phage display qui étaient publiées pour 54 PDZs humains [9]. Ensuite, nous avons appliqué ces PSSMs pour prédire des interactions PDZ-peptide et évalué la fiabilité des prédictions en utilisant des outils d'analyse des séquences et de la statistique. Nos résultats montraient que deux tiers des données phage display étaient constituées d'une grande portion des séquences hydrophobes amenant des propriétés de séquence très différentes de celles observées pour des PBMs cellulaires (voir Figure 0.5). La poursuite de l'analyse a mis en évidence que ces caractéristiques non naturelles des peptides dérivés du phage display étaient probablement nuisibles à la qualité des prédictions. Nous spéculons que les propretés différentes des séquences observées entre les PBMs sélectionnées par phage display et celles présentes dans la cellule pourraient provenir des différences dans des procédures de sélection expérimentale et naturelle. La sélection expérimentale est notamment basée sur l'affinité contrairement à la sélection naturelle, qui favorise des interactions domaines-motifs à l'affinité faible mais spécifiques pour la signalisation cellulaire. Nous proposons cependant que le tiers des données de phage display ne montrant pas de biais pour des résidus hydrophobes pourrait être très utile aux prédictions d'interactions aux PDZs. Les résultats de cette étude ont été publiés dans [16].

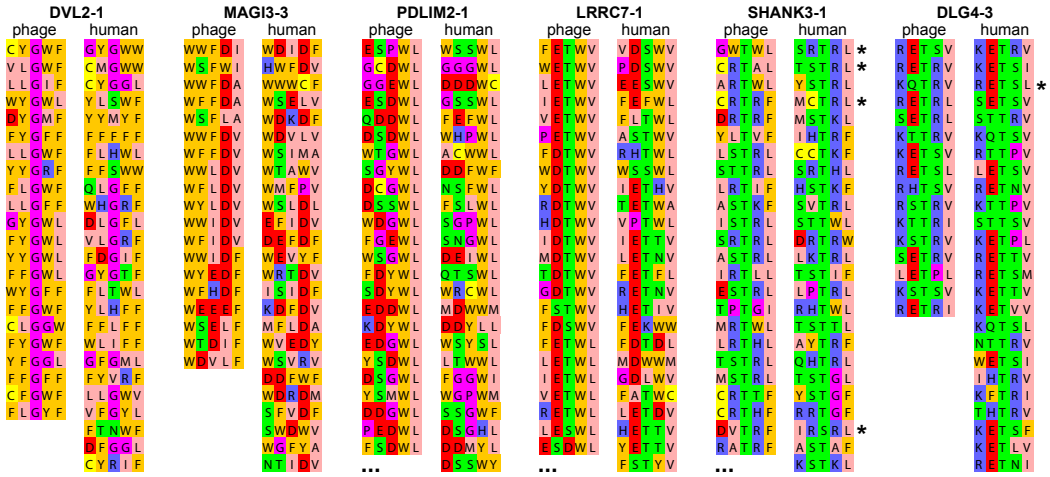


Figure 0.5. Comparaison des peptides phage display [9] avec ceux qui étaient prédits à partir du criblage protéomique. Pour six domaines PDZ humains, les peptides correspondants du phage display sont montrés à côté des peptides C-terminaux humains qui sont issus avec un meilleur score du criblage du protéome humain avec des PSSMs. Les domaines PDZ sont classés du gauche à droite de plus hydrophobe au plus hydrophile selon les peptides de phage display correspondants. Les listes des peptides qui, trop longues pour être montrées entièrement, ont été coupées, comme indiqué par "...". Des astérisques indiquent des C-termini humains qui sont identiques aux peptides de phage display correspondants. Code de couleur: ocre = aromatique, rose = hydrophobe, rose foncé = G ou P, vert = polaire, rouge = acidique, bleu = basique, jaune = C. (La figure a été faite avec Jalview [17].)

Ensuite, nous nous sommes intéressés à l'outil de prédiction publié par l'équipe de MacBeath [11]. Nous avons assemblé une base de données test à partir de la littérature publiée concernant des interactions PDZ-peptide positives et négatives validées expérimentalement. Ces données ont été utilisé pour évaluer objectivement la performance de l'outil de prédiction. Nous avons ensuite développé un protocole pour mesurer des interactions à moyen débit sur des machines BIAcore (basées sur la résonance plasmonique de surface (SPR)) pour valider expérimentalement des interactions PDZ-peptide prédites et pour explorer l'influence des séquences flanquantes des PDZs et des PBMs sur l'affinité et la spécificité de leurs interactions. Nous avons mesuré plus que 200 interactions entre des versions courtes et étendues des PBMs C-terminaux et cinq constructions aux domaines PDZs composées de PDZ2 et PDZ3 de la protéine humaine MAGI1 (membrane-associated guanylate kinase inverted 1), PDZ3 et PDZ4 de la protéine humaine SCRIB ainsi que d'une construction tandem couvrant les PDZ3 et PDZ4 de SCRIB. L'évaluation du prédicteur avec la base de données test a révélé un taux de faux positifs très élevé qui a été confirmé par nos données SPR (voir Figure 0.6). Nous avons pu identifier des points faibles dans la définition du modèle de prédiction ainsi que dans son processus d'entraînement, qui peuvent être à l'origine de la mauvaise performance du prédicteur. Cependant, après l'analyse approfondie des données expérimentales obtenues et des recherches dans la littérature, nous

avons pu proposer 11 nouveaux partenaires potentiels d'interaction pour les protéines à PDZ MAGI1 et SCRIB qui corroborent des suggestions précédentes de l'implication de ces deux protéines dans les réseaux de signalisation des protéines G (voir Figure 0.7).

Les prédicteurs d'interactions PDZs-peptides ont été développés avec la perspective d'être appliqué à la prédiction des réseaux d'interaction protéiques. Au vu des points faibles que nous avons identifiés, l'utilisation de ces outils pour prédire automatiquement des réseaux d'interaction protéiques aboutirait probablement à des résultats très erronés. Cependant, nous avons pu montrer que la prédiction d'interactions PDZs-peptides combinée avec une analyse au cas par cas des résultats et la validation expérimentale peut mener à l'identification de nouvelles interactions potentielles méritant la poursuite des analyses expérimentales.

L'analyse de nos données SPR a de plus relevé que des domaines PDZ montraient des affinités de liaison modifiées pour des séquences PBMs étendues en comparaison avec celles déterminées pour des PBMs courts. Ces altérations d'affinité ont parfois mené à une augmentation de la spécificité des interactions PDZ-PBM. Des études structurales sur des complexes PDZ-peptide conduites dans notre équipe et par d'autres groupes ont mis en évidence des contacts entre des résidus des séquences étendues des PBMs et des résidus de la boucle β2-β3 des domaines PDZ [18]. L'ensemble de ces résultats suggère que la boucle β2-β3 des domaines PDZ, bien que ne faisant pas partie de la poche canonique de liaison, peut jouer un rôle important pour l'affinité et la spécificité des interactions impliquant les PDZs (voir Figure 0.8). De plus, nos données expérimentales ont montré des changements d'affinité entre des constructions contenant un PDZ unique (PDZ3 et PDZ4) de SCRIB et la construction tandem suggérant que ces deux PDZs influencent mutuellement leur liaison aux peptides et qu'ils forment une unité globulaire pouvant être appelée "supramodule".

Beaucoup d'interactions entre domaines PDZ et PBMs minimaux que nous avons analysées par la SPR ont montré des affinités très faibles. Ces affinité faibles ont augmenté (spécifiquement) lorsque nous avons étendu les fragments minimaux d'interaction. Il est intéressant de noter que, bien que nous ayons pu trouvé des cas dans nos données expérimentales où les extensions des fragments protéiques ont modulé l'affinité de liaison, nous n'avons jamais pu trouver des cas où des changements d'affinité étaient aussi importants que des fragments protéiques qui n'ont pas interagit entre eux dans leur version courte commençaient à interagir lorsqu'ils étaient rallongés, ou vice-versa. Donc, nous concluons que des données quantitatives des affinités de liaison obtenues pour des fragments minimaux d'interaction ne sont pas nécessairement valides pour des fragments étendus ou les protéines entières. Néanmoins, ceci pourrait être le cas pour les données d'interaction qualitatives (liaison ou pas liaison). Cette étude a été publiée dans [19].

Nous avons fait une revue extensive de la littérature publiée sur des exemples des séquences flanquantes des domaines PDZ et des PBMs pour pouvoir placer nos

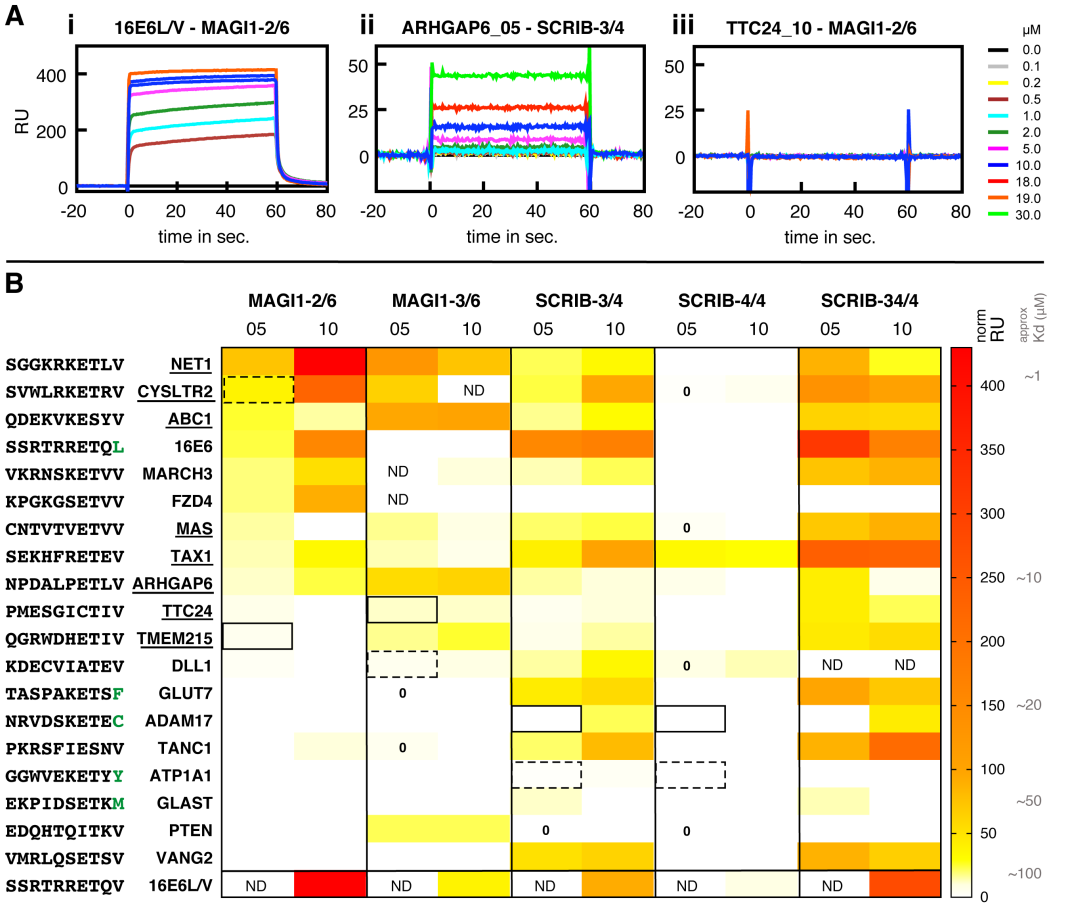


Figure 0.6. Les données expérimentales de SPR obtenues. A: Exemple de sensorgrammes représentatives d'interactions fortes, faibles et de non-interactions. Une augmentation du signal pour l'injection du PDZ indique une liaison. (i) Des concentrations plus élevées du PDZ mènent à un signal plus élevé jusqu'à saturation dans le cas d'une interaction spécifique. (ii) Pour des interactions faibles, la concentration la plus élevée du PDZ injectée, n'a pas mené à une saturation du signal. (iii) Les sensorgrammes pour des non-interactions ne montrent pas de changement de signal. B: Vue d'ensemble des données expérimentales de SPR et comparaison aux prédictions. Des signaux RU normalisés déterminés pour une concentration de 10 µM de PDZ ont été extraits des sensorgrammes et représentés sous forme d'un "heatmap". Une échelle approximative du K_D est indiquée sur la droite. 05 et 10 indiquent des peptides en version courtes et longues. ND = non-déterminé. Les signaux obtenus pour des peptides courts interagissant avec des constructions expérimentales de PDZs uniques ont été comparés aux interactions prédites avec le prédicteur de [11]. Les rectangles normaux et les rectangles hachurés indiquent le premier et le deuxième meilleur peptide prédit pour chaque domaine PDZ, respectivement. Les paires PDZ-peptide pour lesquelles une non-interaction était prédite, ont été labellisées avec zero. Pour toutes les autres paires entre PDZs et peptides minimaux montrées sur le heatmap, une interaction avait été prédite par le programme. Le résultat indique que les non-interactions sont plutôt correctement prédites, tandis que de nombreuses non-interactions ont été prédites comme des interactions.

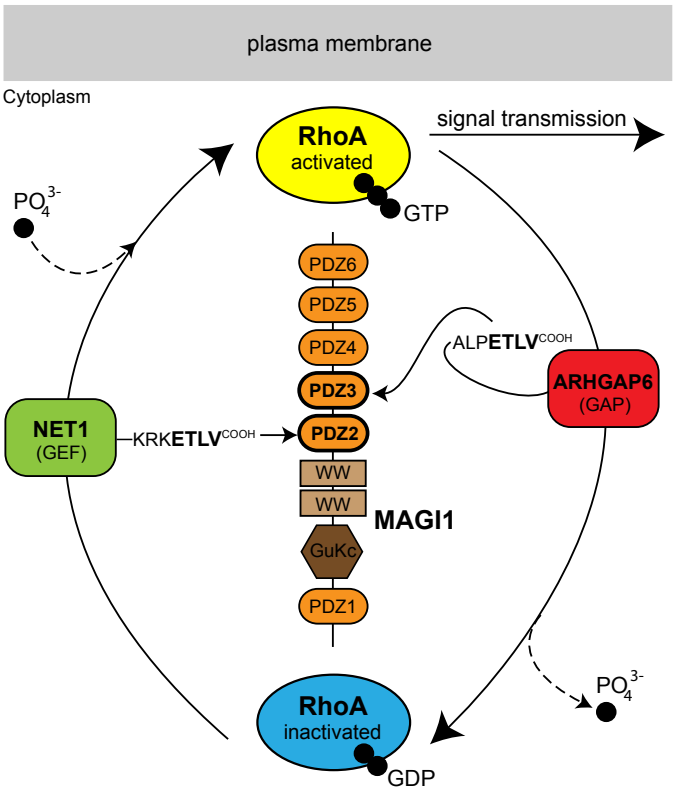


Figure 0.7. Proposition du modèle de la fonction de "scaffolding" de MAGI1 dans la signalisation à l'origine des Rho GTPases. Nos données expérimentales ont montré que PDZ2 et PDZ3 de MAGI1 lient préférentiellement les C-termini des protéines NET1 (vert) et ARHGAP6 (rouge), respectivement. NET1 et un facteur d'échange du nucléotide guanine (GEF), qui transfère un groupement phosphate (PO_4^{3-}) à la "small GTPase" RhoA. Cette protéine, dans sa forme liée au GTP (jaune), est associée principalement à la membrane et stimule des voies de signalisation en aval. ARHGAP6 est une "GTPase-activating protein" (GAP), qui induit RhoA de libérer un groupement phosphate, menant à l'arrêt de la signalisation médiée par RhoA. Dans sa forme liée au GDP, RhoA est inactive (bleu) et principalement localisée au cytoplasme. Ces données suggèrent que MAGI1 recrute, via deux domaines PDZ voisins, un activateur et un inhibiteur de la voie de signalisation de RhoA. Les quatre derniers résidus des deux protéines NET1 et ARHGAP6 sont identiques. Donc, les préférences de liaison des deux domaines PDZ2 et PDZ3 pour ces peptides C-terminaux doivent être déterminées par des résidus en amont.

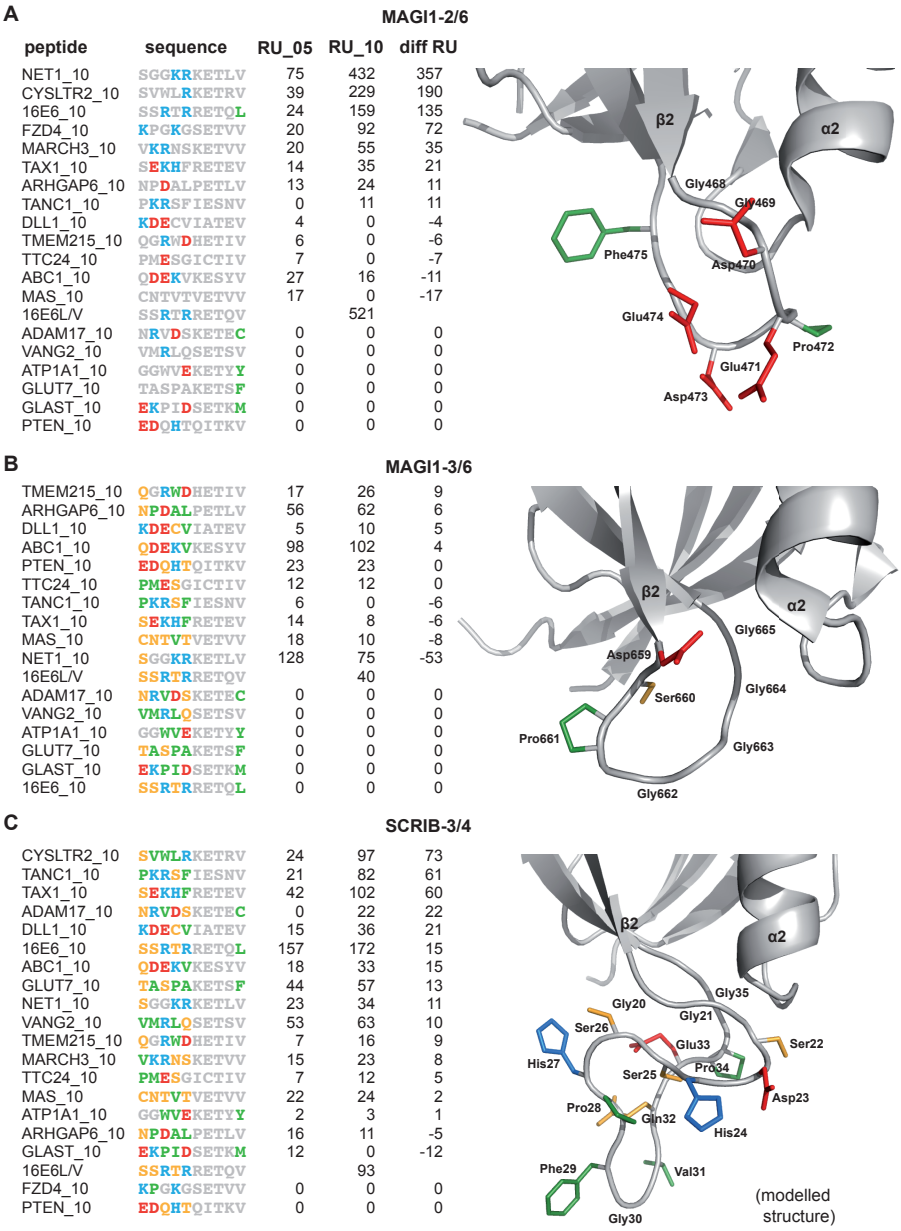


Figure 0.8. L'influence de la boucle β2-β3 des domaines PDZ sur la liaison des peptides. Les colonnes indiquent de gauche à droite les noms des peptides, leurs séquences, les intensités d'interaction en RU pour les peptides avec cinq et dix résidus "wildtype" et la différence de l'intensité d'interaction entre les deux versions de peptides. Pour chaque PDZ la partie de la structure contenant la boucle β2-β3 est montrée avec des résidus représentés en bâtons. Les acides aminés dans les séquences et les structures sont colorés de la façon suivante: rouge = charge negative, bleu = charge positive, jaune = polaire, vert = hydrophobe. A. PDZ2 de MAGI1 lie avec une affinité augmentée les peptides disposant de charges positives en amont de la position p-4 probablement dû au quatre charges négatives dans la boucle (code PDB: 2I04). B. PDZ3 de MAGI1 ne montre pas de différence d'affinité entre des peptides courts et longs, probablement dû au quatres résidus "neutres" (des glycines) dans la boucle (code PDB: 3BPU). C. PDZ3 de SCRIB montre une augmentation aspécifique d'affinité pour des peptides longs. La boucle est particulièrement longue et contient des résidus de chaque type physico-chimique.

résultats dans un contexte plus large. À partir de cette analyse nous avons publié une revue sur l'impact des séquences flanquantes des domaines PDZs et des PBMs (sous presse). D'après cette revue, il apparaît que de nombreuses études ont mis en évidence l'influence des séquences flanquantes sur l'affinité des interactions protéiques. Pourtant, des expériences comparatives nécessaires pour évaluer si elles influencent également la spécificité d'interaction n'ont pas été conduit. En conséquence, le travail présenté dans cette thèse représente une contribution importante pour comprendre le rôle que les séquences flanquantes jouent dans la spécificité des interactions. Cependant, plus d'études de ce type sont requises, qui viseront à déterminer et comparer des affinités de liaison de divers peptides pour un domaine PDZ ou de divers PDZs pour un peptide, en incluant idéalement des constructions de longueurs variées, afin de mieux comprendre les niveaux différents de la spécificité des interactions PDZ-PBM et leur modulation par des séquences flanquantes. Nous espérons que ce travail influencera positivement la conception de futures études expérimentales et informatiques pour explorer la spécificité des interactions PDZ-peptide en particulier, et celle des interactions domaine-motif linéaire en général.

Summary

Cell signalling describes all those biological processes that allow a cell to retrieve external signals, to process and propagate them inside the cell, and to respond to them. These processes substantially depend on proteins and their interactions with each other. It has been found that many proteins that function in cell signalling processes possess a modular architecture including short linear motifs (SLiMs) and globular domains [1]. By definition, globular domains are contiguous sequence regions in proteins that can independently fold into a tertiary structure [2]. SLiMs are short sequence stretches in proteins (usually no more than ten residues in length) that preferentially occur in disordered regions of proteins and adopt a secondary structure upon binding [3]. SLiMs can bind to globular domains and *vice versa* thereby mediating a significant fraction of protein interactions that function in cell signalling.

Many different types of domain–SLiM interactions have been identified, of which PDZ domain-mediated protein interactions are one of the most studied. PDZ domains have been shown to function in cell polarity, membrane trafficking and generally in protein complex organisation [6]. PDZs mainly recognise SLiMs that are situated at the very C-terminus of proteins. Such PDZ-binding motifs (PBMs) usually carry a hydrophobic residue at the last position (position 0). PBMs can roughly be divided into three subgroups based on the residue at position -2 (Thr/Ser, Asp/Glu, or hydrophobic residues) [6]. About 270 PDZ domains and thousands of potential PBMs have been identified in the human proteome. In numerous computational and experimental studies researchers have tried to decipher the specificity rules that define which PDZ domain will preferentially bind to which PBM. Insights gained from such studies would allow for a better understanding of the structural mechanisms of protein recognition in general and the design of valuable PDZ interaction predictors in particular.

Central to this thesis were two groundbreaking large-scale studies, published in high-impact journals, that combined experimental and computational approaches to study PDZ interaction specificities. Tonikian *et al.* [9] published phage display data established for 54 human PDZ domains that they used to build position specific scoring matrices (PSSMs) to screen the human proteome for potential binding partners to these PDZs. MacBeath and co-workers [10] used microarrays combined with fluorescence polarisation to determine binding affinities between 157 mouse PDZ domains and 217 mouse C-terminal peptides. They employed this data for PDZ interaction predictor development [11].

In addition to those large-scale approaches, numerous single case structural studies

on PDZ–peptide interactions have been published that suggest that sequence context, e.g. regions in protein sequences that surround PDZs and PBMs, influence the binding *affinity*. However, much less studies have provided information on whether and how sequence context may impact interaction *specificities*. Interaction specificity can only be assessed by comparing binding affinities of one protein (or protein fragment) towards its numerous potential interaction partners. A protein is specific if it displays for most interaction partners weaker and just for a very few of them higher binding affinities. Thus, specificity is a relative value in comparison to affinity, which is an absolute value.

In this thesis, we have addressed two main questions: first, how useful are state-of-the-art PDZ interaction predictors for the prediction of PDZ-mediated protein-protein interaction (PPI) networks? And second, does sequence context have any influence on PDZ interaction specificity and if yes, what are the underlying mechanisms? In particular we were interested in investigating whether different levels of specificity might exist between minimal interacting fragments, e.g. core PDZ domains and core PBMs, and extended protein fragments (presenting additional flanking sequences).

We combined computational and experimental approaches to address the problem of PDZ interaction specificities and their predictions. First, we focussed on phage display data and its potential to be used for protein interaction predictions. We built PSSMs based on the phage display data that had been published for 54 human PDZs [9], applied these PSSMs to predict PDZ–peptide interactions and assessed the reliability of the predictions via sequence analysis and statistics. This revealed that two thirds of the phage display data displayed high proportions of hydrophobic residues leading to very different sequence properties than those observed for cellular PBMs. Further analysis suggested that the unnatural sequence characteristics of peptides derived from phage display were likely to significantly impair prediction qualities. We speculate that the different sequence properties observed between phage display and cellular PBMs may arise from differences in experimental and natural selection procedures, the former being mainly affinity driven whereas the latter being driven by the need for weak, yet specific linear motif-mediated interactions in cell signalling. We suggest that the remaining one third of the phage display data analysed in this study, which did not display a bias towards hydrophobic sequences may be very promising to be used for PDZ-mediated PPI network predictions. Results of this study have been published in [16].

Next, we turned to the predictor that had been published by MacBeath and coworkers [11]. We assembled test data sets consisting of experimentally validated positive and negative PDZ–peptide interactions from the literature and used them to objectively assess the performance of this predictor. We developed a medium-throughput protocol on a BIAcore instrument (based on surface plasmon resonance (SPR)) to experimentally validate predicted PDZ–peptide interactions and to assess the influence of sequence context of PDZ domains and PBMs on the binding affinity and specificity of their interactions. We measured more than 200 interactions between extended and

core versions of C-terminal PBMs and 5 PDZ domain constructs, consisting of PDZ2 and PDZ3 of the human protein MAGI1 (membrane-associated guanylate kinase inverted 1), PDZ3 and PDZ4 of the human protein SCRIB as well as a tandem construct comprising PDZ3 and PDZ4 of SCRIB. Benchmarking the predictor with the test data sets revealed a high false positive rate (FPR) that was confirmed by our SPR data. We identified weaknesses in the model definition and training process of the predictor that may be responsible for the poor prediction performance. Nevertheless, after careful experimental data analysis and literature searches, we were able to propose 11 new potential binding partners for the PDZ proteins MAGI1 and SCRIB that strengthen previous suggestions for their involvement in G protein signalling pathways.

PDZ–peptide interaction predictors are developed with the prospect to be applied to PPI network predictions. Given the identified weaknesses in PDZ-mediated interaction prediction, fully automatic derivation of PPI networks will be highly error-prone. However, we could show that predictions of PDZ–peptide interactions combined with manual analysis and experimental validation can result in the identification of new potential interactions that are worth further experimental investigation.

Analysis of our SPR data also revealed that PDZ domains displayed altered binding affinities towards extended PBM sequences in comparison to core PBMs and that these alterations could increase the binding specificity of PDZ–peptide interactions. Structural studies on PDZ–peptide complexes performed in our group and by others provided evidence for residue contacts between extended PBM sequences and the β2-β3 loop of PDZ domains [18]. Altogether, this suggests that the β2-β3 loop of PDZs, although not being part of the canonical binding pocket, may be an important player for PDZ interaction affinities and specificities. In addition, our experimental data showed changes in binding affinity between single PDZ3 and PDZ4 of SCRIB and the tandem construct, suggesting that these two PDZs influence the peptide binding of each other and might form one globular unit, which may be called a PDZ "supramodule".

Many interactions that we identified between core PDZ domains and core PBMs (minimal interacting fragments) using SPR displayed very weak binding affinities. These weak binding affinities were shown to (specifically) increase when extending the minimal interacting fragments. Interestingly, the changes in binding affinity that we observed due to fragment extensions were never at a scale where protein fragments that did not bind to each other in their short version started to bind to each other when being extended or *vice versa*. Thus, we conclude that quantitative binding affinity data obtained for minimal interacting fragments is not necessarily valid for extended fragments or the corresponding full length proteins. However, the qualitative results (i.e. binding or not binding) obtained for minimal interacting fragments should in general be transferable to full length proteins. This study has been published in [19].

We intensively surveyed the published literature on instances of sequence context

of PDZ domains and PBMs to put our findings within the wider context of this field. Based on this analysis we published a review on sequence context of both, PDZs and PBMs (in press). From this review, it emerges that many studies provide evidence for influence of sequence context on the binding affinity of protein interactions. Yet researchers do not perform the comparative experiments necessary to assess whether this is also true for interaction specificity. Thus, the work presented in this thesis represents an important contribution to our understanding of the role that sequence context plays for interaction specificities. However, there is a clear need for more studies that systematically compare binding affinities between various peptides to a PDZ domain or various PDZs to one peptide, ideally by varying construct lengths, to better understand the different levels of specificity in PDZ-mediated protein interactions and its modulation by sequence context. We hope that this work will positively influence future design of computational and experimental studies to investigate the specificity of PDZ–peptide interactions in particular and of domain–linear motif interactions in general.

Abstract

PDZ domains recognise C-terminal PDZ-binding motifs (PBMs) thereby mediating many protein interactions that function in cell signalling. About 270 PDZs and thousands of potential C-terminal PBMs have been identified in the human proteome. What are the specificity rules that define, which PDZ will preferentially bind to which PBM? This question has often been addressed by studying interactions between the canonical peptide binding pocket of PDZs and short PBMs leading to the development of numerous PDZ interaction predictors. However, it seems that the sequence context (surrounding regions) of PDZs and PBMs, may contribute to interaction specificities as well. Here, we addressed two questions: First, how reliable are PDZ interaction predictors and second, does sequence context have any influence on PDZ interaction specificity and what are the underlying mechanisms? We assessed two PDZ interaction predictors, the first based on phage display data and the second, trained on mouse PDZ-PBM interactions. We identified a bias towards hydrophobic sequences in the phage display data impairing its application for predictor training. The second predictor displayed a high false positive rate, probably due to incorrect model definition and insufficient training data. We developed a medium throughput protocol using SPR to experimentally validate predicted PDZ-PBM interactions for the human PDZ proteins MAGI1 and SCRIB. We could show that the weak prediction performances can be bypassed with manual inspection and experimental validation resulting in the identification of new potential interactions. Our experimental data also suggests that PDZs can display increased binding specificities towards PBMs when their sequences are extended, probably via involvement of the β2-β3 loop of PDZ domains. Our results contribute to our understanding of the role that sequence context plays for interaction specificities. We hope that this work will positively influence future design of computational and experimental studies to investigate the specificity of PDZ–PBM interactions.

Contents

List of Tables

List of Figures

Part I.

Introduction

1. Protein interactions in cell signalling processes

Cell signalling describes all those biological processes that allow a cell to retrieve external signals, to process and propagate them inside the cell, and to respond [20]. The molecules that organise these processes include extracellular signalling molecules that bind to receptor proteins that in turn activate intracellular signalling proteins. The signal terminates in target or effector proteins, which as a result change themselves or alter the state of the cell (e.g. by changing the transcription of genes). Such target or effector proteins comprise (gene) regulatory proteins, ion channels, components of metabolic pathways, cytoskeleton proteins and many more [20]. The overall process of converting extracellular signals into intracellular responses, as well as the individual steps in this process, is termed signal transduction [21].

In molecular biology reference books, cell signalling is often termed cell communication [20, 21]. Indeed, in multicellular organisms most signals originate from communications between cells with the aim to organise concerted cell division, cell growth, and differentiation, which are essential for tissue organisation [20]. Although cell signalling reached its complexity in multicellular organisms, the basic mechanisms must have evolved in single cell organisms as they also had the need to communicate to other unicellular organisms and to sense and respond to changes in their environment [21]. This is consistent with observations that many entities that mediate interactions between signalling proteins in multicellular organisms can also be found in bacterial genomes and unicellular eukaryotic organisms, such as yeast, although they massively expanded with the evolution of multicellular organisms (see section 2.3) (e.g. tyrosine phosphorylation networks [22]). Interestingly, with increasing organism complexity, the fraction of proteins involved in cell signalling processes increases while the fraction of metabolic enzymes decreases [23]. The importance that cell signalling plays for multicellular organisms is reflected by the estimate that probably more than 40% of the human proteins are involved in signal transduction processes [24].

Almost all cell signalling processes are based on protein interactions. Therefore, investigating cell signalling pathways means first of all to discover and understand the underlying complex network of protein interactions. Researchers try in ongoing efforts to establish a map of the human interactome, e.g. the totality of protein interactions in a human cell, in analogy to the map of the human genome published in 2001 [25]. The total number of different protein interactions in a human cell is likely to exceed the 10 million (following an estimate from Toby Gibson given in a talk in

Garmisch-Partenkirchen, Germany, in 2010) when including post-translational modification (PTM) events such as phosphorylation (more than 170,000 phosphorylation sites are annotated in phosphosite [26]). High-throughput (HTP) methods allow for the rapid detection of thousands of protein interactions. However, such methods often fail to detect transient signalling protein interactions or multiple protein complexes and to differentiate between true and false positive hits [27]. Thus, we still seem very far from obtaining a first draft of the complete human interactome.

Instead of searching for as many as possible interactions between any set of proteins of a particular organism [28–30], a new trend emerged during the last ten years where HTP methods were applied with the aim of obtaining "sub-interactomes" related to particular biological questions, e.g. a particular signalling pathway [31] or a particular type of interaction domain [32]. Such approaches yielded interactomes of manageable size and high accuracy allowing to derive novel insights into the biological functions and properties of the investigated proteins and discovered interactions.

2. Modular architecture of proteins involved in cell signalling

2.1. The discovery of protein modules

Research on protein structures focussed for long time on enzymatic proteins consisting of one single folded domain, which made them amenable for crystallisation studies. This focus of structural biologists dramatically changed with the invention of a new DNA sequencing method by Sanger and co-workers in 1977 [33, 34]. Due to more sophisticated DNA sequencing, the 80s were marked by a rapid increase in available DNA sequences including whole genomes from diverse model organisms [24].

Newly sequenced DNA molecules had to be compared to the growing pool of known DNA sequences for functional prediction purposes and homology searches. This raised the need to be able to rapidly compare biological sequences and to reliably distinguish between evolutionary related sequences and sequence similarity that had been found by chance. Theoretical concepts for automated sequence comparisons were already developed in the 60s with an important contribution from M. Dayhoff. She based her studies on the idea that evolution mainly acts on protein sequences and thus, ancestral relationships (homology) between two genes can be best studied by comparing the corresponding protein sequences with each other. Dayhoff developed the first matrix for amino acid substitution scorings, the PAM matrix, that together with sequence alignment algorithms like Needleman&Wunsch, allowed to calculate the likeliness that two proteins were evolutionary related to each other. PAM and other substitution matrices were later applied in the basic local alignment search tool (BLAST) algorithm, a heuristic strategy employing local sequence alignments, that allowed to rapidly and reliably search for homologous protein sequences in databases [35, 36].

The application of local sequence alignment algorithms for protein homology searches in huge sequence databases turned out to be biologically very meaningful as it led to the discovery that proteins that were overall evolutionary unrelated nevertheless shared regions of high sequence similarity. What was the structure and function of these homologous regions? First, their study turned out to be difficult as these regions were part of large proteins that were impossible to deal with in experimental assays. The idea of splitting up such large proteins into structurally and functionally independent domains largely improved their investigation by methods such as crystallography and nuclear magnetic resonance (NMR) [2, 37]. The identification of more and more of such domains in protein sequences led to the introduction of the term *protein module*

to describe and distinguish them from previously studied enzymatic domains [2].

Thus, the impressive progress in genome/proteome analyses and in structural biology in the 80s had been the basis for the identification and classification of protein modules [1, 24]. Protein modules were first identified in extracellular proteins in the 80s [36]. The majority of intracellular modules were identified in the 90s with a peak in module identification in 1995 and 1996 [36]. Logically, the most frequent modules in proteomes were discovered first (WD40, SH2, SH3, ANK, RING, PH, PDZ, Fbox) [36].

2.2. Defining a protein module

Probably the first definition for a protein module has been given by Campbell and co-workers [2]: *"A domain is probably best defined as a spatially distinct structural unit that usually folds independently. In this definition, the sequence need not be contiguous. Modules are a subset of domains that are contiguous in sequence and that are repeatedly used as building blocks in functionally diverse proteins. They have identifiable amino-acid patterns and can be described by a consensus sequence."* In this definition, the two main characteristics of modules are highlighted:

- widespread occurrence in proteome (genetically very mobile);
- continuous sequence, detectable via sequence similarities.

Additional characteristics of protein modules have later been added and were originally discovered in studies of src homology 2 (SH2) and src homology 3 (SH3) domains [1, 38–42]:

- 40-150 residues in length;
- tertiary structure that folds independently from surrounding sequences and that is unique for each type of module;
- modules frequently bind to short sequence motifs.

Often, the identification of a module next to other modules within one protein sequence increased its likeliness to be a module [2]. Modules were seen as evolutionary independent entities that function in single copies and in this regard clearly differ from repeats [36]. The above cited general definition of a domain is strongly influenced by structural aspects. Other definitions of domains exist, e.g. originating from biochemistry and genetics, where a domain is a *"minimal fragment of a gene that is still able to perform a certain function"* [36] but these definitions are not applied in this thesis.

The existence of different combinations of modules in proteins and their wide distribution in the proteome could not be explained by mechanisms like gene duplication and diversification that had been used to explain evolution of purely enzymatic proteins. Often, module boundaries were found to be consistent with exon boundaries.

Thus, exon shuffling has been proposed as mechanism to explain module rearrangements [2, 37]. However, modules have also been identified in proteins of bacteria that do not possess an intron-exon gene architecture [24]. Overall, module rearrangement has been suggested to be a much more efficient mechanism for protein evolution than gene duplication and subsequent modification [24].

Protein modules were often shown to bind to short stretches of sequences in partner proteins, thereby mediating specific protein interactions. Unfortunately, the characterisation of these short sequences lagged behind the module discovery. In 1998, Sudol [1] suggested the protein recognition code as probably a first attempt in trying to classify and review those sequences that were later called short linear motifs (SLiMs). The name highlights the fact that SLiMs miss any tertiary structural information and thus, from a structural viewpoint are very different from globular protein modules (see section 2.4). However, SLiMs and globular modules actually share the two main characteristics of modules, e.g. continuous in sequence and widespread occurrence in proteomes (see above). Thus, not only globular modules but also SLiMs contribute to a modular architecture of proteins and therefore, the term module has been extended **to include both domains and SLiMs** [3, 43] (see Figure 2.1).

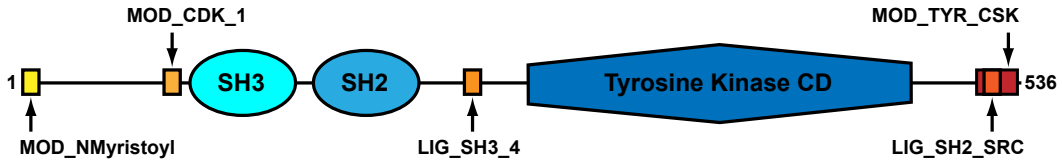


Figure 2.1. Modular architecture of human proto-oncogene tyrosine-protein kinase Src. Globular domains are illustrated in blue tones, SLiMs are illustrated in yellow/red tones. Domains in Src were predicted using SMART [4], experimentally validated SLiMs in Src were extracted from the ELM resource [5]. MOD_NMyristoyl: N-myristoylation site, MOD_CDK_1: Ser/Thr cyclin dependent protein kinase (CDK) phosphorylation site, LIG_SH3_4: SH3 domain binding site, LIG_SH2_SRC: SH2 domain binding site, MOD_TYR_CSK: Tyrosine phosphorylation site for C-Src kinases (CSK). CD=catalytic domain.

2.3. Protein modules and cell signalling

Protein modules have not been identified in phylogenetically "old" proteins, such as many metabolic enzymes [2]. Eukaryotic proteins are on average much longer than prokaryotic proteins owing to their multidomain character [36]. Yeast is a unicellular organism that has more than 20 SH3 domains, several PH and WW domains, and numerous repeats (WD40, ARM, ANK, TPR, ...). However, these modules are much more frequent in multicellular organisms and some other types of modules are completely missing in yeast (e.g. the phosphotyrosine binding domains PTB, SH2) [24]. Studies have shown that modularity is clearly over-represented in regulatory proteins [24]. All these observations indicate that modular architecture of proteins rather

appeared "later" in protein evolution, together with the need for more sophisticated cell communication in multicellular organisms. Protein modularity is a property inherent of proteins involved in cell signalling.

2.4. Current understanding of domain–linear motif interactions

In this thesis, I only refer to domains that are globular modules and thus, for simplification, the term "domain" will be used hereafter instead of the very precise and more technical term globular module. The last ten years have seen an amazing increase of interest for domain–linear motif interactions in various research areas including structural biology, biochemistry, systems biology, and computational biology. Many large-scale studies focussed on the discovery of whole interaction networks that are mediated by domains and SLiMs (see section 6.4.3). The properties of domain–linear motif interactions are defined by the characteristics of the entities, the domains and SLiMs, that establish them. The properties of domains have been introduced in the previous paragraphs. A few more words remain to be said about the characteristics of linear motifs. Davey *et al.* [44] recently reviewed attributes of linear motifs. They systematically analysed the content of the eukaryotic linear motif (ELM) resource, a collection of experimentally validated linear motif instances (see section 6.4.2). This analysis revealed that linear motifs are on average six residues long ranging from just one residue in length to up to 23. On average, four residues of SLiMs show some degree of sequence conservation whereas the remaining positions are completely variable (see Table 2.1). Linear motifs mainly occur in disordered regions of proteins [45] or more rarely in disordered loops of structured regions [46].

name in ELM resource	interacting domain	regular expression
LIG_SH3_1	SH3	[RKY]..P..P
LIG_14-3-3_1	14-3-3	R.[^P]([ST])[^P]P
LIG_PDZ_Class_1	PDZ	...[ST].[ACVILF]$
LIG_SH2_STAT3	SH2 of STAT3	(Y)..Q
LIG_WW_1	WW	PP.Y
MOD_CK2_1	kinase domain of CK2	...([ST])..E

Table 2.1. Examples of SLiMs. A few representative examples of SLiMs that are stored in the ELM resource [5] are given. Interpretation of characters in regular expressions: "[XYZ]": residues X, Y, or Z are possible at this peptide position; ".": any residue possible; "^X": residue X is not allowed at this peptide position; "$": C-terminus; "(X)": residue X is phosphorylated.

Most of the available structures of domain–linear motif complexes in the protein data bank (PDB) [47] reveal that linear motifs adopt a secondary structure upon binding [44]. Linear motifs have been grouped into four categories by the curators of

the ELM resource, those that mediate protein interactions (e.g. SH3-binding motifs), those that are sites for PTMs (e.g. phosphorylation), those that target proteins to subcellular compartments (e.g. to the nucleus) and those that mark cleavage sites in proteins (e.g. for caspases) [3]. Given the short binding interface that linear motifs provide for mediating protein interactions, it is not surprising to see that domain–SLiM interactions are usually of weak micromolar affinity (1 to 150 μM). However, with such weak binding affinities, domain–SLiM interactions fit to the need for transient protein interactions in cell signalling [3, 48, 49].

3. Protein interactions mediated by PDZ domains

Protein interactions that are mediated by PDZ domains are probably one of the most studied types of domain–linear motif interactions to date. More than 1,000 articles, including about 80 reviews, are indexed in Pubmed that have the word PDZ in their title. PDZ domains were first identified in 1992 in the mammalian proteins nitric oxide synthase (nNOS), postsynaptic density protein 95 (PSD-95), membrane protein palmitoylated 1 (MPP1), and the Drosophila protein discs-large (DLG) [50]. They have been termed "GLGF-repeat" following the very conserved sequence motif of the carboxylate binding loop (see paragraph 3.2.2) [50]. One year later, this name had been changed into discs-large homology region (DHR) domain [51]. The breakthrough probably came in 1995 with a first systematic search for this type of domains in public sequence databases using flexible pattern methods [52]. This revealed 27 additional occurrences of DHR domains in intracellular proteins providing first evidence that this domain seemed to be very common in the human proteome. In a comment on this study published the same year, Kennedy [53] suggested the name PDZ *"to better reflect the origin and distribution of the domain"*. The name PDZ had been derived from the initials of the three proteins PSD-95, DLG, zonula occludens protein 1 (ZO-1) that were within the first proteins where PDZ domains had been identified.

PDZ domains were first discovered in invertebrate and vertebrate proteins. Ponting [54] further revealed a few occurrences in bacterial, yeast, and plant proteins. However, the few PDZ domains identified in *S. cerevisiae* and *S. pombe* genomes shared very low sequence similarity with metazoan PDZ domains questioning their overall function as protein recognition modules [6]. Thus, it has been proposed that PDZ domains entered the kingdoms of bacteria and plants via horizontal gene transfer [6, 54]. Together with the widespread occurrence of PDZ domains in metazoa, these observations have led to the assumption that PDZ domains might have evolved with multicellularity [6].

An ongoing disagreement exists about the actual total number of PDZ domains in the human proteome. Published numbers vary from about 250 up to more than 900. Several publications indicate that there are about 270 PDZ domains in the human proteome distributed over about 150 proteins [55, 56]. These numbers are consistent with our own findings (presented in our review on sequence context, chapter 10 and corresponding suppl. data, section D.3).

3.1. Structural aspects of the PDZ domain

The repertoire of structures on PDZ domains deposited in the PDB is immense (see Figure 3.1A) providing us with a unique perspective on the structural diversity that exists within one protein domain family. Many structures from PDZ domains were contributed from structural genomics initiatives, e.g. the Riken Structural Genomics/Proteomics Initiative in Japan (http://www.rsgi.riken.go.jp) or the Structural Genomics Consortium in Great Britain (www.thesgc.org). Structures solved by initiatives represent a source of structural information that is largely unexplored and overlooked because usually no article accompanies publication of those structures in the PDB.

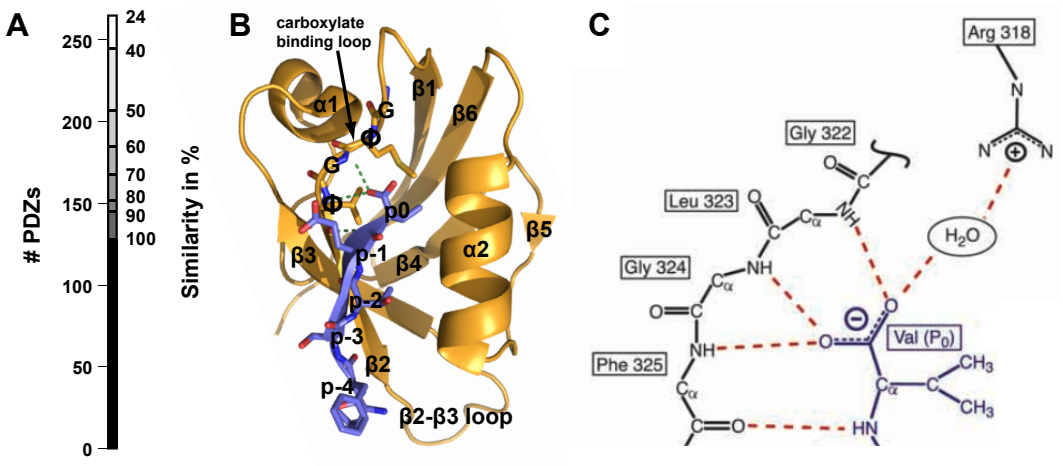


Figure 3.1. Structural aspects of PDZ domains and peptide recognition. A: Available structural information on PDZ domains. Diagram that illustrates the number of human PDZ domains for which there is at least a structure of a PDZ domain in the PDB with X % sequence similarity based on local sequence alignments from BLAST searches (e.g. for 152 human PDZs there is a structure of a PDZ in the PDB with at least 80% sequence similarity.) B: Canonical fold and C-terminal peptide recognition of PDZ domains. Structure of the PDZ domain of AF6 bound to a C-terminal peptide (LFSTEV) derived from the Bcr protein (PDB ID: 2AIN [7]). The secondary structure elements, peptide positions, and the common signature of the carboxylate binding loop (GϕGϕ where ϕ represents a hydrophobic residue) are indicated. Green dashed lines represent hydrogen bonds that are established between Val at peptide position p0 and the carboxylate binding loop. C: Recognition of the C-terminal peptide residue. Domain and peptide atoms are shown in black and blue, respectively. Hydrogen bonds that are established between the C-terminal Val residue of the peptide and the carboxylate binding loop of the PDZ domain are indicated with red dashed lines. Figure C has been adapted from [6], and is based on a complex structure of PDZ3 of PSD-95 [8].

3.1.1. The canonical PDZ fold

The core PDZ fold mostly consists of 80 to 90 residues. The first liganded and unliganded structures of PDZ domains have been published in 1996 [8, 57]. The fold

of PDZ domains has been described as an antiparallel β barrel structure [57] or a β sandwich [8] comprising 5 to 6 β strands and 1 to 2 α helices (see Figure 3.1B). Apart from the secondary structure elements there are two further very conserved structural features in the PDZ fold (see Figure 3.2). The carboxylate binding loop (β1-β2 loop) has the conserved sequence [KR]...GΦGΦ where . denotes any amino acid and Φ a hydrophobic residue, mostly Leu and Phe at the first and second hydrophobic position, respectively. The distance between the conserved positive charge and the GΦGΦ motif is mostly of three residues but can also be of only two residues or much longer. This carboxylate binding loop has important ligand recognition functions (see paragraph 3.2). The second conserved feature is the dipeptide Gly-Asp located N-terminal to the fourth β strand that probably is important for the stability of the fold.

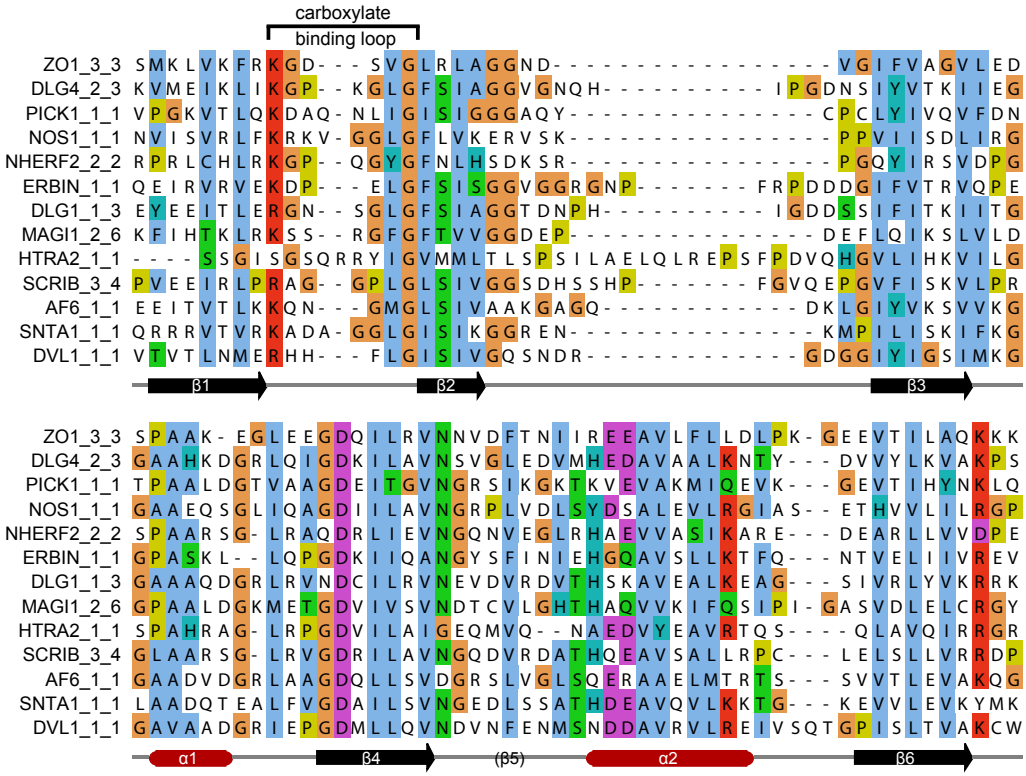


Figure 3.2. Multiple alignment of a few representative human PDZ domain sequences. The names of the proteins containing the PDZ domains as well as the numbers of the PDZ domains are indicated at the beginning of each sequence. Location of secondary structure elements are indicated below the alignment with black arrows representing β sheets and tubes representing α helices. Secondary structure assignment has been based on the structure of PDZ2 of MAGI1 (PDB code: 2KPK). Residues are coloured following the ClustalX colour scheme [58].

3.1.2. PDZ-like folds

A few PDZ domains have been subject to circular permutation, a mechanism by which the order of secondary structure elements of a domain is altered without altering the overall fold of that domain [59]. Circularly permuted PDZ domains have been observed in human in the high-temperature requirement A (HtrA) family of proteases [60], the Golgi-reassembly stacking protein 2 (GORASP2) [61], and eventually in the 26S protasome subunit PSMD9 (own observations based on sequence alignments) as well as in several bacterial PDZ domains [62, 63] and in the plant photosystem II D1 C-terminal processing protease [64] (see Figure 3.3). For a few of those instances it could be demonstrated that they are still able to recognize C-terminal or internal ligands (see paragraph 3.2). The structure of the PDZ domain of the extracellular cytokine interleukin 16 (IL-16) deviates from canonical PDZ domains by displaying a smaller carboxylate binding loop and an occluded peptide binding pocket prohibiting C-terminal ligand recognition [65]. There is no established standard to decide whether domains with such unusual PDZ-like folds and/or altered functions should still be considered as PDZ domains or not. In this respect, the exact number of PDZ domains of the human proteome will be further subject to small adjustments.

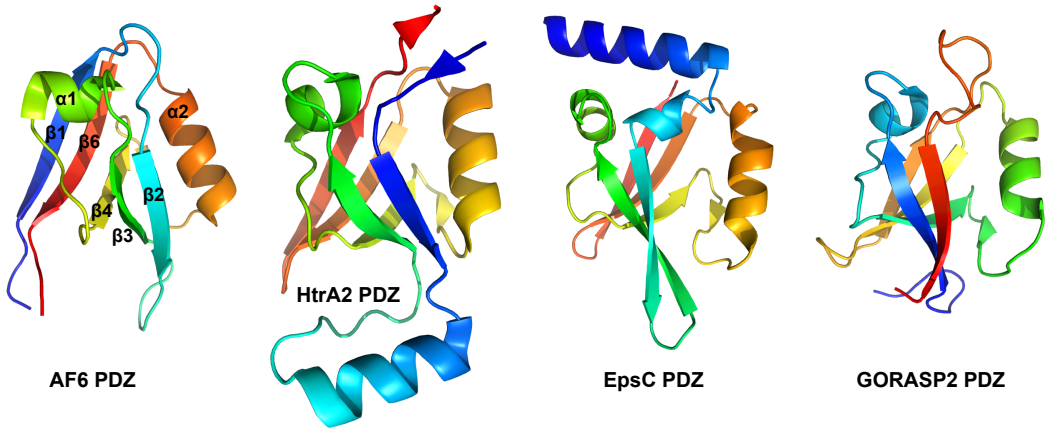


Figure 3.3. Comparison of canonical and circularly permuted PDZ folds. The structure of the human PDZ domain of AF6 (PDB code: 2AIN) serves as an example for a canonical PDZ fold. The PDZ domain of human HtrA2 (PDB code: 2PZD), of bacterial EpsC (PDB code: 2I4S), and of human GORASP2 (PDB code: 3RLE) display circularly permuted PDZ folds. All structures are coloured with a "rainbow" colour scheme, starting with blue at the N-terminus, over cyan, green, yellow, and orange to red at the C-terminus. In addition to the circular permutation, the PDZ domains of HtrA2 and EpsC possess additional secondary structure elements.

3.2. Types of interactions engaged by PDZ domains

3.2.1. Recognition of C-terminal ligands by PDZ domains

Shortly after the discovery of PDZ domains, researchers identified the first sequences that were recognized by PDZs [66, 67]. These were C-terminal sequences, a property that turned out to be common for most of the ligands of PDZ domains identified so far. Hereafter, I will refer to the short sequences that are recognized by PDZ domains as ligands, peptides, or PDZ-binding motifs (PBMs). PBMs bind via β augmentation to PDZ domains, e.g. PBMs adopt a β strand that pairs in an antiparallel manner with the β2 strand of the PDZ domain (see Figure 3.1B). Several hydrogen bonds are established between backbone atoms of the peptide and backbone atoms of the PDZ domain as can be observed for typical β sheets. The peptide binding pocket of PDZ domains is formed by residues from the carboxylate binding loop, the β2 and β3 strand, the β2-β3 loop, and the α2 helix [8]. The structure of PDZ domains change very little upon peptide binding as demonstrated by small root mean square deviation (RMSD) values between the α carbons of apo and holo structures [68].

Based on the recognition of C-terminal SLiMs, peptide positions are numbered starting from the last residue (position 0, p0) going backwards (p-1, p-2, and so forth). The last residue of C-terminal PBMs is almost always a hydrophobic residue, mainly Val, Leu or Ile. This residue inserts into a hydrophobic pocket that is mainly formed by PDZ residues from the carboxylate binding loop, the α2 helix and the β2 strand. The carboxylate group of the C-terminal residue of the PBM is hydrogen-bonded to backbone amides of the GΦGΦ motif (see section 3.1.1) as well as to an ordered water molecule that is stabilised via a hydrogen bond with the conserved positively charged residue of the carboxylate binding loop (see Figure 3.1C). These conserved interactions determine the C-terminal peptide selectivity of PDZ domains [8]. Remarkably, the recognition of the terminal carboxylate group does not involve any direct salt bridges in contrast to other carboxylate recognition domains [6]. Thus, it appears that the chemical recognition of the C-terminus is less important than its spatial recognition. This probably provides two advantages: first, binding affinities remain moderate and second, non-specific recognition of C-termini is hindered [6, 69].

The residue N-terminal to the last residue, at position p-1, is generally very variable. In some PDZ complexes residues at p-1 have been observed to point to the solvent [8], in others they contact neighbouring PDZ domain residues from the α2 helix, β2 strand or β3 strand thereby often establishing electrostatic interactions [70, 71]. Different PDZ domains have been shown to have different preferences for residues at this peptide position. However, the presence of disadvantageous residues at peptide position -1 has usually not been observed to interrupt binding [70, 72]. Very different opinions exist about the eventual role of position p-1 for the specificity of PDZ–peptide interactions.

The residue at peptide position -2 has been found to be key for the classification of all C-terminal PBMs into three different classes [6]. Mutations of residues at p-2 into Ala usually abrogate binding, indicating that like position 0, this is a key position for peptide recognition by PDZ domains [67]. The first PBMs identified carried a Ser or Thr at p-2 [66, 67] that establish with their hydroxyl group a hydrogen bond to a His situated right at the beginning of the α2 helix [8]. It has been proposed that PDZ domains with a His at this helix position are likely to prefer PBMs with a Ser or Thr at p-2 [70]. Thus, researchers tempted to not only group these PBMs but also the "corresponding" PDZ domains into class I (see Figure 3.4A). In addition, Songyang *et al.* [70] observed in a phage display study that Tyr at p-2 might be able to perform the same function as Ser or Thr residues. In the same study, a second class of PBMs had been identified that displayed hydrophobic or aromatic residues at p-2. PDZ domains that prefer peptides with hydrophobic residues at p-2 possessed different residues than His at the first position of the α2 helix [70]. Structures of PDZ domains complexed to class II peptides reveal interaction of hydrophobic/aromatic residues at p-2 with hydrophobic residues or residues with aliphatic portions (e.g. Lys) of the α2 helix (see Figure 3.4B and C).

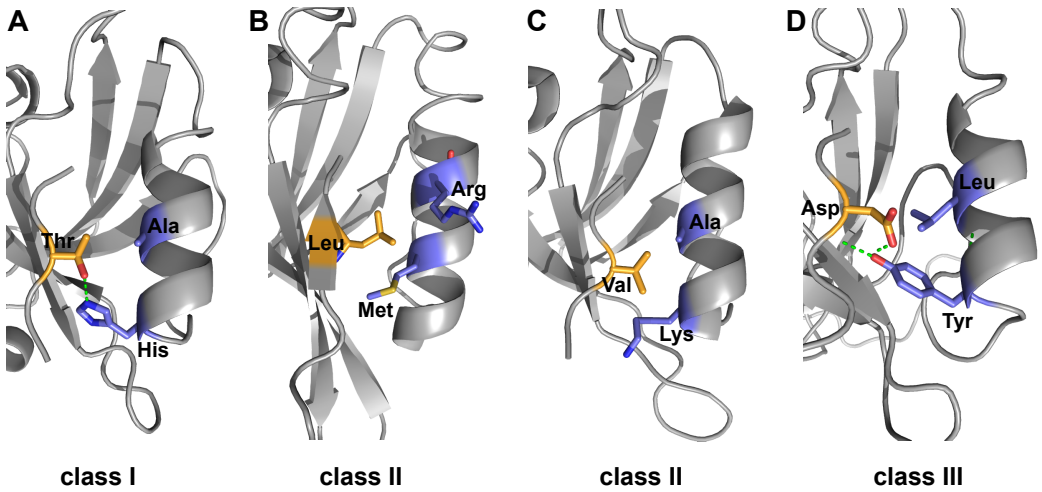


Figure 3.4. Recognition of different subclasses of peptides by PDZ domains. Relevant peptide and domain residues are shown in orange and blue, respectively. Hydrogen bonds are indicated by green dashed lines. A: Recognition of C-terminal peptides of class I having a Ser or Thr at p-2 that establish a hydrogen bond with a His from the α2 helix of the PDZ domain (PDZ3 of PSD-95 complexed to CRIPT peptide, PDB code: 1BE9). B and C: Recognition of C-terminal peptides of class II having a hydrophobic residue at p-2 that interacts with aliphatic portions of side chains from residues of the α2 helix (PDZ3 of Par3 complexed to vascular endothelial cadherin peptide (PDB code: 2KOH) and PDZ of PICK1 complexed to GluR2 peptide (PDB code: 2PKU)). D: Recognition of C-terminal peptides of class III having an Asp or Glu at p-2 that establishes a hydrogen bond to a Tyr from the α2 helix of the PDZ domain (PDZ of nNOS complexed to a synthetic peptide (PDB code: 1B8Q)).

A third class of PBMs had been simultaneously identified by peptide library screens [73]. Class III peptides display negatively charged residues (Asp or Glu) at position -2. The first published structure of a PDZ domain bound to a class III peptide showed that the Asp at p-2 established hydrogen bonds with Tyr at the beginning of the α2 helix (see Figure 3.4D) [74]. Consistently, Songyang *et al.* [70] observed in their phage display study the selection of peptides with Tyr at p-2 by PDZ domains with negatively charged residues at the first position of the α2 helix. Interestingly, mutation of Tyr to His in the α2 helix of the PDZ domain of nNOS switched its preference from class III peptides to class I peptides [73].

A few examples demonstrate that some PDZ domains are of dual specificity, e.g. they can bind PBMs of more than one class. The third PDZ domain of the cell polarity protein Par3 and the PDZ domain of protein interacting with C kinase 1 (PICK1) can bind to class I and class II peptides [75, 76]. The PDZ domain of nNOS has been shown to bind class II and class III peptides [74]. Dual specificity of PDZ domains provides a first glimpse of the plasticity of PDZ domains for ligand recognition and the complexity of this issue. In section 5.6, I introduce more in detail the controversial subject of PDZ interaction specificity.

As it has been observed for position -1, residues at position -3 are very variable. Again, different PDZ domains display different preferences for residues at p-3. Whereas in some structures, residues at p-3 point to the solvent, in other structures often electrostatic contacts could be observed to PDZ domain residues from the β2-β3 sheet [8, 72] or the β2-β3 loop [77].

Up to now, no clear consensus exists between researchers in the PDZ field about the actual length of PBMs. A few structures of PDZ-peptide complexes indicate that only the last three peptide residues might be sufficient for binding to PDZ domains [78, 79]. Clearly, the last four peptide residues insert into the peptide binding pocket between the β2 strand and the α2 helix. In some structures and experimental binding assays residues at peptide position -4 have been observed to contribute to the binding affinity mainly by contacting PDZ domain residues from the β2-β3 loop but sometimes also from the α2 helix or the β2-β3 sheet. A considerable amount of studies demonstrate that some PDZ domains show preferences for certain amino acids at certain peptide positions far upstream position -3 [9, 70, 72] and solved structures provide the molecular details for these observations (see our articles presented in chapter 7 and 9). In our review (see chapter 10), we extensively discussed the published literature presenting cases of such extended PBMs.

The discussion about the binding affinity range of PDZ–peptide interactions is similarly controversial. Early studies often indicated dissociation constants for PDZ–peptide interactions in the nanomolar range [70]. Though, such high affinity values have been observed for artificial peptides, cellular PDZ–peptide interactions were later suggested to be in the low micromolar range (1-10 μM) [6, 80]. More recent studies

indicate that much lower binding affinities (down to 100 μM or even beyond) between PDZ–peptide interactions are actually likely to occur [10, 81, 82] (see also our article in chapter 9). Such low affinity interactions are most probably relevant as they have consistently been observed for other types of domain–SLiM interactions [3] and extensions of core PDZ domain constructs and peptides led to increases in binding affinity suggesting that PDZ domain-mediated interactions of full length proteins may be stronger (see chapters 9 and 10).

3.2.2. Recognition of internal ligands by PDZ domains

In contrast to the clear and well established rules for C-terminal peptide recognition, internal ligands bound by PDZ domains have resisted all attempts of classification. The first identified instance of internal ligand binding by PDZ domains has been highly studied (for a review see [55]). The PDZ domain of nNOS possesses a β hairpin structure in its C-terminal extension that inserts into the binding pockets of the syntrophin PDZ domain and a PDZ domain of PSD-95 [83]. Structural analysis and phage display studies revealed that the sharp turn of the β finger apparently replaces the structural requirements of a C-terminus [83, 84]. Although unexpected, it appears that this remained the only example where the structural context of an internal ligand determined its binding capabilities to PDZ domains [6].

Probably three examples exist where the carboxylate group of an Asp side chain in an internal ligand might replace the function of the carboxylate group of C-terminal ligands. A very detailed structural analysis has been performed by Penkert *et al.* [85] on the cell polarity protein Par6 PDZ domain bound to a C-terminal and internal ligand. Zhang *et al.* [86] assessed the capabilities of dishevelled protein (Dsh) to bind to internal ligands using phage display and crystallography. More ambiguous results have been obtained by Lemaire and McPherson [87] on an interaction between nNOS and Vac14 (scaffolding protein for the phosphatidylinositol 3,5-bisphosphate regulatory complex). It seems that apart from the aspartic residue and maybe a hydrophobic position right before, no other common sequence characteristics are shared by the internal ligands identified in these studies (see Table 3.1).

PDZ	peptide	peptide sequence	PDB code
NOS	Vac14	GDHLDRR	no structure
Par6	Pals1	YPKHREMAVDCP	1X8S
Dsh2	pepN1	WKDYGWIDGK	3CBY
Dsh2	pepN2	SGNEVWIDGP	3CBZ
Dsh2	pepN3	EIVLWSDIP	3CC0

Table 3.1. Internal PBMs with Asp at peptide position 0. The peptide sequences have been aligned based on the Asp that replaces the carboxylate group of C-terminal PBMs.

Yet other examples of internal motif recognition have been observed for the PDZ domains of Dsh (binding to Frizzled) [88] and of the protease HtrA2 [60]. In both cases no negatively charged residues nor loop contexts could be identified. Interestingly, HtrA2 represents a case where sequence similarities were detected between C-terminal and internal ligands derived from phage display libraries, consisting of a patch of three hydrophobic residues that were important for binding [60]. Numerous internal ligands were revealed from a search for binding partners of all *C. elegans* PDZ domains [89]. As has been the case for previous studies, sequence analyses did not reveal any conserved features in these ligands and their biological functions remain to be demonstrated. It also cannot be exluded that there are some cases where identified internal ligands are likely to be artefacts [90, 91].

The recognition of internal ligands by PDZ domains though largely "understudied", is likely to be of biological importance. Comparison of structures of PDZs either complexed to internal or C-terminal peptides [85, 86] and careful analysis of phage display data combined with molecular modelling [92] illustrates the structural plasticity of PDZ domains and their ability to possess multiple specificities.

3.2.3. Binding of lipids by PDZ domains

A more recently discovered feature of PDZ domains is their capability of binding to phospholipids (phosphoinositides) that participate in the regulation of the localisation and active state of protein signalling complexes. Studies on different PDZ domains revealed diverse binding sites and mechanisms by which lipids are recognized by PDZ domains. In general, proteins can either specifically recognize phosphoinositides, or bind unspecifically to membranes via electrostatic interactions (phospholipids are strongly negatively charged) or establish hydrophobic interactions via membrane penetration. These different modes of membrane binding complicate the study of PDZ-lipid interactions [93] and thus, from present studies no clear picture emerged about the influence of lipid binding by PDZs on C-terminal peptide recognition. Most likely, this will strongly depend on the particular PDZ domain under investigation as indicated by Gallardo *et al.* [93] in a review on this topic.

In a very recent large-scale study Chen and co-workers [94] assessed for more than 70 mammalian PDZ domains their lipid-binding capabilities using surface plasmon resonance (SPR). About 40 % of the tested PDZ domains showed lipid binding with an astonishing affinity of better than 1 μM indicating that lipid binding by PDZs might be a more general property. They grouped the lipid-binding PDZ domains into two classes, one with a main cationic patch on their surface without overlap with the peptide binding site, and the other with a cluster of cationic residues that partially overlapped with peptide binding residues. Interestingly, Rhophilin2, a member of the second class, showed more specific binding to a cognate PBM in the presence of lipids.

3.2.4. PDZ–PDZ interactions

The probably best studied example of a PDZ–PDZ interaction involves the PDZ domain of nNOS. As mentioned in section 3.2.2 nNOS has C-terminal to its PDZ domain a β-hairpin that is recognized by PDZ domains from α1-syntrophin and PSD-95 [83, 95]. Although used as a prototype for PDZ–PDZ interactions throughout the "PDZ literature", interactions mediated by the β-hairpin of nNOS do, in my opinion, not strictly correspond to this type of interaction. The β-hairpin of nNOS does not seem to be important for the binding of C-terminal peptides to the nNOS PDZ domain nor for the structural integrity of this PDZ. Thus, the β-hairpin may not be considered as a part of the PDZ domain. Under this regard, interactions between nNOS and α1-syntrophin or PSD-95 should rather be considered as cases of internal ligand recognition than PDZ–PDZ interactions.

Nevertheless, there are many other examples that indicate the potential of some PDZ domains to form homo-dimers, e.g. PDZ domains of sodium-hydrogen exchanger regulatory factor 1 (NHERF1) [96], glutamate receptor-interacting protein 1 (GRIP1) [97], SH3 and multiple ankyrin repeat domains protein 1 (SHANK1) [98,99], ZO-1 and ZO-2 [100–102]. However, the biological significance and structural mechanisms of dimerisation remain to be demonstrated for some of these cases. Based on crystallographic studies, Zhang *et al.* [71] proposed a homo-dimer formed by PDZ2 (out of six PDZs) of membrane-associated guanylate inverted (MAGI)1. However, this dimer is likely to be an experimental artefact as it could not be observed between extended constructs of PDZ2 (see our NMR study in chapter 7).

New insights into PDZ–PDZ interactions were provided from a recent large-scale study performed in the MacBeath lab [103]. About 150 mouse PDZ domains were tested for their abilities to bind other PDZs, leading in total to more than 12,000 interaction tests that had been performed. Positive hits were consequently confirmed with fluorescence polarisation (FP) measurements. About 30 % of all PDZ domains tested displayed at least one interaction with another PDZ domain (including homo-dimers). In total, 37 PDZ–PDZ interactions had been identified and quantified (with binding affinities below 25 μM) and of those 11 were tested and successfully validated in a full length context. Interactions mediated via C-terminal ligands were unlikely to have occurred because all PDZ constructs were checked for C-termini that eventually carried PBMs and in such cases, triple Gly had been added to the construct to prevent C-terminal peptide binding. Contrary, the possibility cannot be excluded that some of these identified interactions are actually interactions between PDZ domains and internal ligands. All PDZ constructs have been designed with a 30 residue C-terminal extension. Such extensions might bear the potential of internal ligand binding as in the case of nNOS.

3.2.5. PDZ–non-PDZ domain interactions

Two interesting single case studies exist that provide evidence for interactions between PDZ and phospho-tyrosine binding (PTB) domains [104, 105]. The PTB domain of Numb, a protein involved in neurogenesis, has been shown to bind to two interaction regions in the E3 ubiquitin ligase LNX1 [105]. The first comprise a phospho-tyrosine motif, the second consists of the first PDZ domain of LNX1. The recognition of both binding sites increased the binding affinity between both proteins and were necessary for the ubiquitination of Numb by LNX1 [105]. The binding interface of the PTB domain for the PDZ domain of LNX1 has been mapped to an 11 residue-long stretch (EFKFFKGFFGK) that constitutes an insertion in two of four isoforms of human Numb. It is unclear whether this region can bind as an internal ligand into the peptide binding pocket of the PDZ domain of LNX1 or whether a distinct binding interface of the PDZ domain is employed.

Richier *et al.* [104] described an interaction between the fourth PDZ domain of the human cell polarity protein SCRIB and the PTB domain of the nitric oxide synthase 1 adaptor protein (NOS1AP). Interestingly, both rho guanine nucleotide exchange factor 7 (ARHGEF7) that binds via a C-terminal PBM to the third PDZ domain of SCRIB [106] and NOS1AP are necessary for the activation of the guanosine triphosphate (GTP)ase Rac1 via SCRIB [104]. Thus, SCRIB could function here as a scaffold protein that brings ARHGEF7 and NOS1AP into close proximity to jointly activate the Rac1 signalling pathway.

3.3. Biological functions of proteins that contain PDZ domains

Often, PDZ domains are mentioned to repeatedly occur within one protein chain. However, it seems that this property is less general than thought as about 75 % of all human PDZ domain-containing proteins (short "PDZ proteins") contain only one PDZ domain (see Figure 3.5). This prevailing view may originate from a few intensely studied multiple PDZ proteins that are recurrently mentioned in the next paragraphs. PDZ domains co-occur with several other types of globular domains in proteins, e.g. the membrane-associated guanylate kinase (MAGUK) family of proteins contains a conserved triplet of domains consisting of a PDZ, an SH3 and a guanylate kinase (GK) domain [107]. This family of proteins has currently 16 members including the DLG proteins and the ZO proteins. The MAGI proteins lack the SH3 domain and present the PDZ and GK domain in inverted order. Thus, they should not be considered as MAGUKs contrary to assertions in the published literature [108]. Other types of domains that frequently co-occur with PDZ domains in proteins are ankyrin, LIM, L27, C2, PH, WW, DEP, and LRR domains [80]. Some enzymatic activities have been observed in PDZ proteins, e.g. serine-threonine kinase, phosphatase, protease, guanine nucleotide exchange factors (GEFs) and GTPase activities [80].

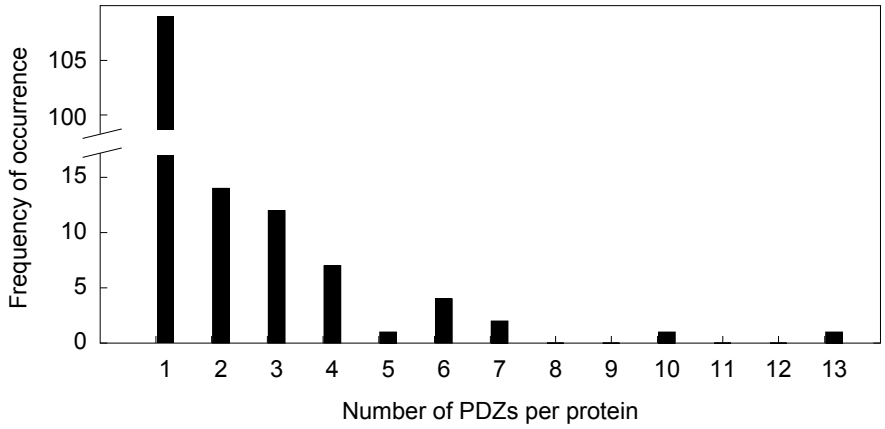


Figure 3.5. Distribution of numbers of PDZ domains per protein in the human proteome. The data used to create the plot has been derived from [56].

Clearly, the multidomain character of PDZ proteins largely determines the functions of these proteins. They have been found to be mainly cytosolic proteins that assist in the assembly and localisation of protein complexes that participate in intracellular signalling pathways [6]. As it seems that these proteins serve as an initial platform for protein complex assembly, the term scaffold has been recurrently used to describe this important functionality of PDZ proteins. In the following, I summarise important biological processes that are regulated by PDZ proteins with a strong focus on cell polarity establishment and maintenance. Names of PDZ proteins are highlighted with bold letters.

One of the most important insights that I obtained during my PhD consisted of the fundamental and diverse roles that cell polarity plays in multicellular organisms. Every cell that is not a completely undifferentiated stem cell is very likely to be polar, e.g. will exhibit an asymmetric distribution of molecules and a shape different from a sphere. Cell polarity in its various forms is important for asymmetric cell division, neuronal transmission, cell migration, immunological responses, and establishment of tissue layers, such as epithelia and endothelia. This non-exhaustive list of functions of cell polarity illustrates its importance and remarkably, in all these processes, PDZ proteins play essential roles.

Establishment and maintenance of cell polarity is highly dependent on the dynamic remodelling of the actin cytoskeleton that is regulated by small Rho GTPases [109]. PDZ proteins have been shown to interact with various members of the G protein cycle (see Figure 3.6), including G-protein coupled receptors (GPCRs), GTPases, and GEFs (see section 3.2.5 and own findings presented in chapter 9). PDZ proteins seem to bring components of the G protein cycle together at precise locations beneath plasma

membranes where they allow for the activation of intracellular signalling pathways in response to incoming stimuli at receptors.

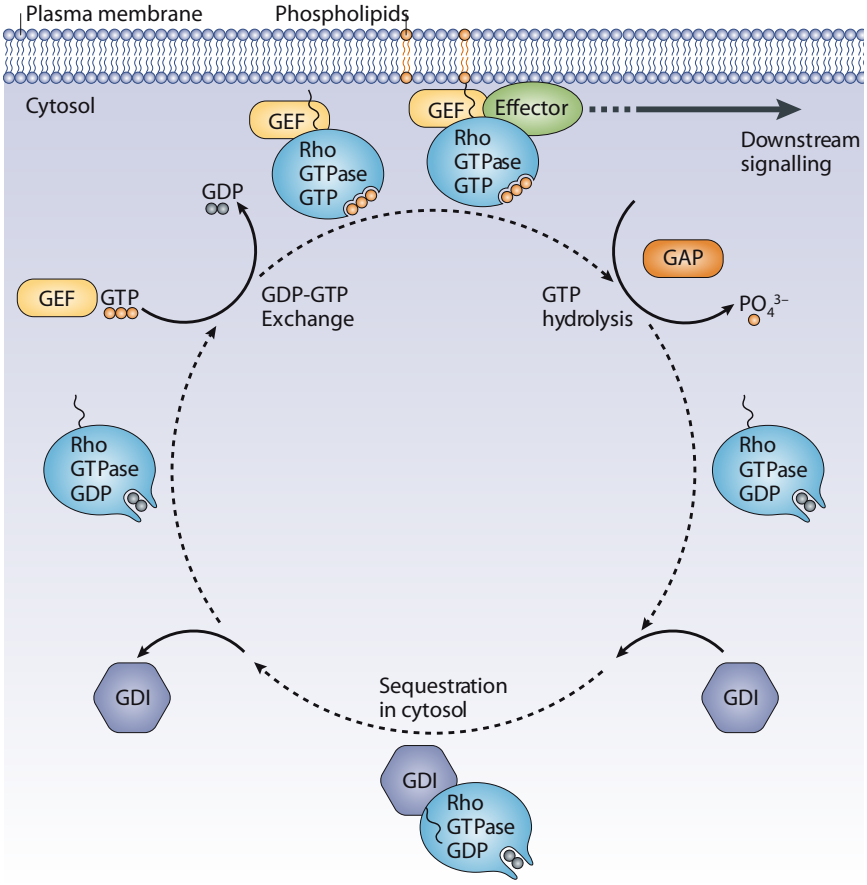


Figure 3.6. The G protein cycle of small Rho GTPases. G proteins are important signalling molecules. They are GTPases that are inactive when bound to GDP (guanosine diphosphate). GEFs (guanine nucleotide exchange factors) activate GTPases by exchanging GDP with GTP (guanine triphosphate). Depending on associated effector proteins, GTPases can function in specific downstream signalling pathways. GTP-activating proteins (GAPs) can accelerate hydrolysation of GTP to GDP by GTPases, which leads to their inactivation. Guanine nucleotide dissociation inhibitors (GDIs) can retain GTPases in their inactive state. The figure has been extracted from [109].

3.3.1. Epithelial apical-basal cell polarity

Epithelial cell layers function as barriers between compartments, e.g. the inside and outside of an organism, and allow for selective transport of molecules from one side to the other [110]. The apical side is oriented towards the outside whereas the basal side is oriented towards the inner side. The basal side of the cell is attached to the

extracellular matrix. The cellular space between the apical and basal part is called the lateral side and is the area of contact with neighbouring cells of the same cell layer (see Figure 3.7). Thus, orientation of the apical–basal axis in an epithelial cell is defined by its environment. Several sites of cell–cell contact ensure the proper anchorage of an epithelial cell within the cell layer. Adherens junctions regulate cell–cell adhesion by providing the mechanical link between cells. They contain cadherins and catenins and are linked to the cytoskeleton via the protein **Afadin (AF6)** [110]. Tight junctions are located above adherens junctions and mark the border between the apical and lateral domain of a cell (see Figure 3.7). Tight junctions create a diffusion barrier for soluble molecules between cells and preclude an intermixture of components of the apical and lateral membrane. They contain occludins, junction adhesion molecules (JAMs) and claudins, and are mainly organised by the **ZO** family of proteins (see Figure 3.8) [111]. MAGI proteins have been shown to be abundant at tight junctions where they link the JAMs to the atypical protein kinase C (aPKC) signalling pathway [112].

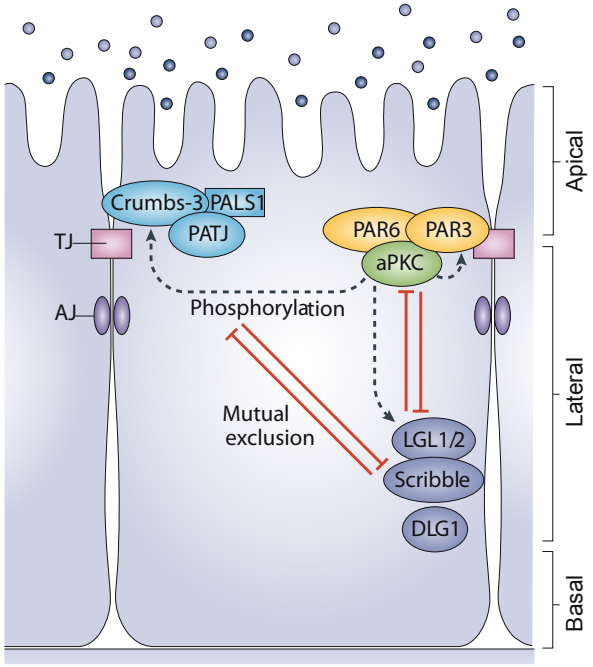


Figure 3.7. Epithelial cell polarity and organisation of polarity complexes. The apical, lateral and basal part of a polar epithelial cell are indicated to the right. TJ = tight junction, AJ = adherens junction. Three polarity complexes are the main regulators of epithelial cell polarity: the Crumbs, Par3, and SCRIB complex. They influence each others cellular localisation. The figure has been adapted from [109].

Establishment of apical–basal cell polarity requires both cadherin-dependent cell-cell adhesion and adhesion to the extracellular matrix [110]. The protein complexes that organise epithelial apical–basal cell polarity are conserved from the fly (where they have been discovered) to worm and human [111]. Asymmetric concentration of

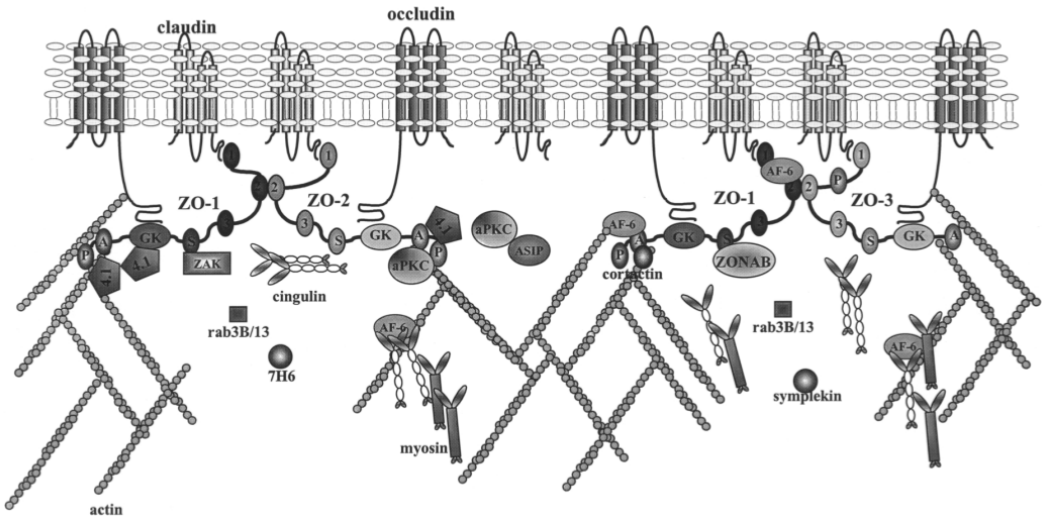


Figure 3.8. Organisation of tight junctions. PDZ proteins like the ZO-family of proteins and AF-6 are mainly responsible for protein complex organisation beneath tight junctions and their connection to the cytoskeleton. The figure has been taken from [113].

phosphatidylinositides may be the initiator of a first localisation of the polarity complexes [110]. The Par3 complex (composed of **Par3**, **Par6**, and aPKC) as well as the Crumbs complex (composed of **MAGUK p55 subfamily member 5 (MPP5)**, **pals1-associated tight junction protein (PATJ)**, and **LIN7**) are localised apically to places where tight junctions will be formed [110]. The SCRIB complex (composed of **SCRIB**, Lgl1/2, and **DLG1**) localises to the lateral membrane (see Figure 3.7). The PDZ protein members of these complexes engage in numerous PDZ domain-mediated protein interactions that organise their localisation, the assembly of the tight and adherens junctions and their link to the cytoskeleton [110, 111]. Knockout experiments of members of the polarity complexes usually revealed only very mild phenotypes making it difficult to study their precise functions. Probably, the high functional redundancy within the polarity complexes confers robustness to the system that regulates cell polarity.

3.3.2. Polarisation in neurons

Neurons are highly polarised cells although it is a very different polarisation from the apical-basal polarity observed in epithelial cells. The asymmetry of neurons is displayed by presynaptic and postsynaptic sites that are formed at the axon terminal and dendrites, respectively [109]. These sites contain distinct ensembles of proteins that ensure the directional transmission of action potentials [111]. The spatially restricted activation of the **Par3** complex together with the protein **T-lymphoma invasion and metastasis-inducing protein 1 (TIAM1)** has been shown to be important for the process of axon specification [109].

PDZ domains have been first discovered in the protein **PSD-95** (see section 3), a main actor in the postsynaptic density. **PSD-95** together with other PDZ proteins such as **GRIP1**, **PICK1**, and **DLG1** regulate the clustering and localisation of receptor channels (e.g. AMPA and N-methyl-D-aspartate (NMDA) receptors) at postsynaptic sites and organise a dense network of protein interactions that link the membrane-anchored protein complexes to downstream signalling pathways [6, 80].

3.3.3. T-cell polarity

T-cells that circulate in blood vessels and lymphatic vessels have a round morphology. Following stimulation by external chemokines, they polarise to be able to migrate towards inflamed tissue or to mediate cell–cell interactions (e.g. with antigen-presenting cells). T-cells polarise along the anterior-posterior axis. **SCRIB** and **DLG** proteins are localised to the rear of the polarising cell to initiate uropod formation (membrane protrusion at rear of T-cells that is involved in their activation and migration) [109]. The **Par3** protein complex is localised to the front of the T-cell where it initiates lamellipodium formation (sheet-like cellular protrusion that is enriched in actin) [109].

An immunological synapse that is formed between a polarised T-cell and an antigen-presenting cell, constitutes a transient and adhesive contact. During immunological synapse formation, polarity proteins such as **DLG** and **SCRIB** have been observed to redistribute from the uropod to the immunological synapse. Knockdown of **SCRIB** led to defects in polarisation indicating the importance of PDZ proteins in immunological synapse formation [109].

3.3.4. Cell migration

Cell migration can be observed in embyonic and adult organisms as well as in pathological situations such as inflammation and cancer. Neurons migrate along glial cells during brain development, epithelial cells migrate during tissue morphogenesis and for maintaining the skin, and tumour cells migrate during metastasis. In order to migrate, cells have to be polarised along a front-rear axis. Rho GTPases and components of the **Par3**, **SCRIB** and **Crumbs** complexes have been demonstrated to jointly regulate front-rear polarisation, chemotactic migration and wound-healing in epithelial cells [109].

3.3.5. Asymmetric cell division

Asymmetric cell division occurs during embryonic development and in adult organisms (e.g. maintenance of stem-cell populations) and requires asymmetric distribution of polarity proteins prior to cell division [109]. In neuroblasts, the apical cortex is enriched for the **Par3** complex whereas the **SCRIB** complex regulates the alignment of the mitotic spindle along the apical-basal axis. Cell-fate determinants accumulate according to the distribution of the polarity complexes and allow for the creation of two different daughter cells [109].

3.3.6. Cell polarity and tumourigenesis

Growth and proliferation of a cell is tightly controlled by its anchorage within a tissue. If a cell looses its cell polarity regulators, it will loose its tight connections to neighbouring cells. Most malignant tumor cells have lost some stages of polarity. Thus, they can escape from normal proliferation control and display increased migratory capacity [109, 114]. Consistently, **SCRIB** has been termed a tumour suppressor for its function in apical-basal cell polarity maintenance [115]. Nevertheless, given the intertwined relationship of the polarity complexes and their implications in a very diverse range of biological processes, their role in tumourigenesis can be oncogenic or suppressive depending on the context [114, 116].

3.3.7. Apart from cell polarity

Of course, PDZ proteins also perform functions that are (at least not directly) linked to cell polarity. **NHERF1** has been in the focus of many studies for its implication in membrane protein activity and trafficking [6]. No SH2 domains nor tyrosine kinase activities have been found in proteins together with PDZ domains [80]. However, PDZ proteins are frequently observed as adaptors for tyrosine kinase receptors, such as the PDZ protein **Erbin** that recognizes a PBM in the receptor tyrosine-protein kinase ERBB2 [80]. The PDZ protein **PATJ** has been highly investigated for its scaffolding function in the phototransduction pathway in the eye of *D. melanogaster* [69,117]. The PDZ protein **Harmonin** is essential for proper mechano-transduction in the inner ear sensory hair cells. Defects in **Harmonin** cause the Usher syndrome, a disease characterised by deafness and blindness [118].

3.4. Hijacking of PDZ domains by viral proteins

Numerous SLiMs have been found in viral proteins where they mediate interactions between viral and host proteins [119]. It has been suggested that SLiMs evolve fast and often by convergent evolution due to their shortness and lack of structural constraints [44,119]. These are properties that are perfect for viruses that have to quickly adapt to changes in their environment, probably explaining the prevalence of SLiMs in viral proteins. Davey *et al.* [119] highlight in their recent review that identified SLiMs in viral proteins are involved in extracellular, cytoplasmic and nuclear processes such as viral entry and exit, protein degradation and transport, immune responses, cell signalling, cell cycle regulation and transcriptional regulation. These are all cellular functions that viruses recurrently hijack for the accomplishment of the viral life cycle.

An impressive number of viral proteins have been shown to possess C-terminal PBMs and for a few examples it has been demonstrated that these PBMs were important for the pathogenicity of the virus [120]. In particular, for the viral proteins E4-ORF1 of Human Adenovirus [121], E6 of human papilloma virus (HPV) [122], and Tax of human T-lymphotropic virus type 1 (HTLV1) [123] it has been suggested that their

PBMs contribute to the oncogenic potential of these viruses. Some HPVs that infect mucosal epithelial tissues have been termed "high-risk", as they have been shown to cause cervical cancer. Interestingly, only high-risk HPV strains, such as HPV16 and HPV18, possess C-terminal PBMs suggesting a major role in tumourigenesis caused by HPV [124]. These three oncogenic viral proteins with PBMs, E4-ORF1, E6, and Tax, target a common ensemble of PDZ proteins that are implicated in the regulation of tight junctions (e.g. **ZO1/2**, **MAGI1/2/3**, **multiple PDZ domain protein (MPDZ)**), of cell polarity (e.g. **SCRIB**, **DLG1**), and of apoptosis (e.g. **SCRIB**, **tyrosine-protein phosphatase non-receptor type 4 (PTPN4) and PTPN3**). Interaction of viral PBMs with cellular PDZ proteins mostly leads to the disruption of their biological functions, either because these interactions promote the sequestration of the PDZ proteins into inactive complexes, their proteasome-mediated degradation or their mislocalisation [120]. Proteasome-mediated degradation of the PDZ targets of E6 depends on the recruitment of the E3 ubiquitin ligase E6AP [125,126]. Unfortunately, degradation experiments prove difficult to clearly identify among PDZ proteins bound by E6 those that are subsequently degraded by the proteasome *in vivo*.

The systematic disruption of the tight junction barrier, disregulation of cell polarity and prevention from apoptosis due to interference with cellular PDZ proteins are obviously beneficial for viral replication. It has been suggested that dismantling tight junctions might facilitate viral spread and transmission as well as increase tissue damage and inflammatory responses [120]. Changes of the polar state of an infected cell as well as interference with apoptosis regulators are likely to promote cell division and lengthen the lifetime of an infected cell, which are clearly favourable for viral replication. Most likely, the perturbation of these important cellular functions can eventually lead to the development of tumours. This destructive process, however, is no longer beneficial for the viral life cycle and therefore should be rather considered as an unfortunate side product of viral infection.

Proteins with C-terminal PBMs from non-oncogenic but still highly pathogenic viruses include non-structural protein 1 (NS1) from avian and human influenca A viruses, the G protein from Rabies virus, and the E protein from severe acute respiratory syndrome (SARS) coronavirus [120]. In an interesting study, Préhaud and co-workers [127] demonstrated that it is a single amino acid change at position -3 of the PBM from Gln in wild type G protein of Rabies virus to Glu in an attenuated strain that is responsible for the observed differences in virulence. Glu at p-3 in attenuated Rabies allows the binding of more PDZ targets than wild type G protein including **PTPN4**. **PTPN4** had been demonstrated to prevent neuronal cells from apoptosis. In the context of infection with attenuated Rabies virus, **PTPN4** is bound and inactivated by the G protein, thus promoting apoptosis.

4. On the binding affinity of protein interactions

4.1. Defining the binding affinity of a protein interaction

The binding affinity of an interaction between two proteins A and B can be defined as the dissociation constant K_D or the change in Gibb's free energy ΔG [128]:

$$K_D = \frac{[A][B]}{[AB]} = \frac{1}{K_A} = \frac{k_{off}}{k_{on}} \tag{4.1}$$

$$\Delta G_{AB} = \Delta H - T\Delta S = -RTlnK_A \tag{4.2}$$

where $[A]$ and $[B]$ are the free concentrations of proteins A and B, $[AB]$ is the concentration of the formed complex, K_A is the association constant, k_{on} the association rate, k_{off} the dissociation rate, H the enthalpy, T the absolute temperature, S the entropy and R the gas constant. Thus, the binding affinity of an interaction depends on the concentration of the formed complex, the association and dissociation rates of the interaction partners or the change in enthalpy and entropy upon reaction. Protein interactions can possess large interaction surfaces that can, in theory, reach very high binding affinities ($K_D > 10^{-15}$ M) [128]. In biological systems when considering a limited range of possible k_{on}, an inversely proportional relationship between the binding intensity and lifetime of a protein interaction can be observed (see Table 4.1). It has been proposed that the lifetime of most protein interactions must be compatible with the duration of the cell cycle, thereby probably defining an upper limit for "acceptable" cellular binding intensities (e.g. for a bacterial cell with 30 min of replication time, $K_D \approx 10^{-10} - 10^{-12}$ M) [128].

K_D range	lifetime
1 Molar	random (microseconds)
1 milli Molar	short lived (milliseconds)
1 micro Molar	transient (seconds)
1 nano Molar	stable (hours)
1 pico Molar	stable (days)

Table 4.1. Affinity and lifetime of protein interactions. This data has been presented in a talk by Joël Janin at the IREBS, Illkirch (F), in 2011.

4.2. Experimental determination of binding affinities

In the following, I will refer to the two interacting proteins A and B as analyte and ligand. Based on SPR terminology, the analyte describes the protein that is free in solution and available in varying concentrations. The ligand describes the protein that is available in limited and stable amounts and depending on the experimental method used, is sometimes fixed on a surface.

The information given in the following paragraphs has been mainly taken from [129], otherwise specified. The determination of the dissociation constant K_D of an interaction between two proteins (often protein fragments) can be readily obtained if the concentrations of the free proteins and the complex are known (see equation 4.1). In most experiments, the concentration of one protein in its free form is unknown necessitating titration experiments and data fitting for K_D estimation. In titration experiments, the ligand is usually titrated with different analyte concentrations. Another possibility consists of titrating with constant analyte concentrations. Here, increases in analyte concentration are obtained by using the sample of the previous injection for the next injection. This latter approach, however, is less accurate because dilution of the ligand concentration after each injection has to be taken into account for K_D determination. For both approaches, a signal is recorded after each injection that is directly proportional to the concentration of the formed complex. The (relative) concentration of the complex can be plotted as a function of the total analyte concentration (see Figure 4.1 and 4.2 for logarithmic scale). This data can be fit to mathematical equations derived from chemico-biological models to obtain an estimation of the K_D.

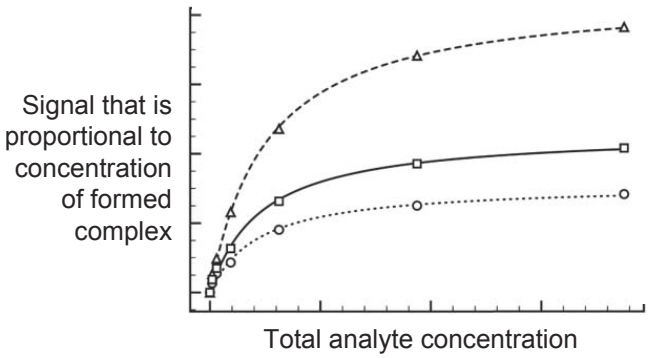


Figure 4.1. Typical diagram obtained during experimental determination of binding affinities. Signals obtained at equilibrium for different analyte concentrations are displayed for three interactions of different binding strengths. Saturation in the binding signal for increasing analyte concentrations is necessary for reliable K_D determination (figure has been modified from [130]).

Numerous methods exist that allow K_D determination. Which of these methods is a good one to choose for a particular setting depends on the strength of the interaction in question and the sensitivity of the method. It appears that it is generally more difficult to reliably determine weak binding affinities between proteins as this prereq-

uisites high concentrations of ligands and/or analytes in order to observe a saturation. Highly concentrated, monomeric protein samples are not always possible to obtain and often bear the risk to form (soluble) aggregates or to precipitate. In the following, I summarise the main characteristics of four methods for dissociation constant determination that have been extensively applied in domain–linear motif biology.

4.2.1. Isothermal titration calorimetry (ITC)

When two proteins form a complex, the free energy, enthalpy, and entropy of the system change (see equation 4.2), and these changes can be measured by highly sensitive calorimetry. Isothermal titration calorimetry (ITC) is special in comparison to other methods presented in this section because analyte concentrations have to be kept constant. The ligand is in solution in a calorimetric cell, and is titrated with consecutive injections of the same analyte concentration. The heat that is released (exothermic reaction) or absorbed (endothermic reaction) upon complex formation is measured. The level of heat of reaction change directly depends on the concentration of free ligand binding sites. Plotting the changes in heat versus the ratio of total analyte and ligand concentration allows an estimation of the K_D. ITC has several advantages: both molecules are in solution, and in addition to the K_D, the enthalpic and entropic contributions to the binding affinity can be directly determined.

4.2.2. Nuclear magnetic resonance (NMR)

The chemical environment of atoms that are part of the binding interface between two molecules change upon complex formation. These changes are visible in changes of chemical shifts that can be observed with NMR. The higher the concentration of one protein, the more complex formed, the stronger the chemical shift perturbations measured by NMR. Differences in chemical shifts for an atom can be plotted versus the varying total analyte concentrations to estimate the K_D. As for ITC, NMR has the advantage that both proteins are in solution but what makes NMR a unique method is the potential to look at the responses of individual atoms upon binding, even allowing to detect site-specific binding constants [131]. In addition, NMR is well suited for determination of very weak binding affinities (e.g. $K_D > 200$ μM). The disadvantage is that the ligand has to be highly concentrated in order to obtain significant signals, and thus, a high affinity K_D (i.e. better than 20 μM) cannot be directly determined (but indirectly via competition experiments for example).

4.2.3. Fluorescence polarisation (FP)

In FP, the ligand is labelled with a fluorophore, e.g. a green fluorescent protein (GFP), and is titrated with varying unlabelled analyte concentrations. The fluorophore will emit light upon stimulation with light. If the incoming light is polarised, the emitted light will be to a certain degree polarised as well. The degree to which the emitted light will be polarised depends on the rotational diffusion rates of the ligand, to which the fluorophore had been attached. The rotational diffusion rate of a molecule depends

primarily on its molecular weight and shape. The smaller the molecule, the higher the rotational diffusion rate, the lower the fluorescence polarisation. If an analyte binds to the labelled ligand, it will increase the overall molecular weight of the ligand (that is now in complex) leading to a decrease of the rotational diffusion rate of the formed complex and thus, increase the degree of polarisation. The degree of polarisation light emitted can be plotted against the total analyte concentration allowing the determination of the K_D. FP has several advantages. The analyte and the ligand are in solution and signals can be obtained under steady state conditions or (more difficult) in real time allowing for the determination of kinetic constants.

4.2.4. Surface plasmon resonance (SPR)

In SPR, the detection principle is based on changes of the optical properties of a surface due to changes in the overall mass of proteins that are bound to it. The ligand is attached to a surface and the analyte is flowed at various concentrations over the surface. In contrast to FP and NMR, SPR does not allow to work with constant analyte concentrations for K_D determination. Depending on the analyte concentration and its binding affinity, a certain amount of analyte will bind to the attached ligands leading to a change of the overall mass on the surface. This change is detected with laser light under total reflection conditions and is translated into response units (RUs). One RU is equivalent to the binding of approximately 1 pg of protein per mm^2 of the SPR chip surface. The time-course of SPR signals are displayed in sensorgrams (see Figure 4.2). The RUs obtained at steady state can be plotted as a function of the total analyte concentration allowing the determination of the K_D. SPR has the advantage of measuring in real time providing the possibility to determine the kinetics of the interaction. Problems can appear when using higher concentrations of analytes (e.g. > 50 μM) that may lead to experimental artefacts (e.g. through blocked flow channels or mass transport effects).

Significant improvements in data quality can be achieved when performing "double referencing" [132]. The first correction consists of substracting non-specific binding signals obtained from analyte injections on a reference surface on which a negative control ligand had been attached. A second correction can be performed by substracting non-specific binding signals obtained from a blank injection (only buffer) that had been flowed over the surface [132]. These two corrections take into account the non-specific contributions to the signal of the analyte and the solvent that are flowed over a surface. Different modes of ligand attachment to the surface exist that can be distinguished into reversible attachments (e.g. via a glutathione S-transferase (GST)-antibody system) and non-reversible attachments (e.g. via a streptavidin-biotin system).

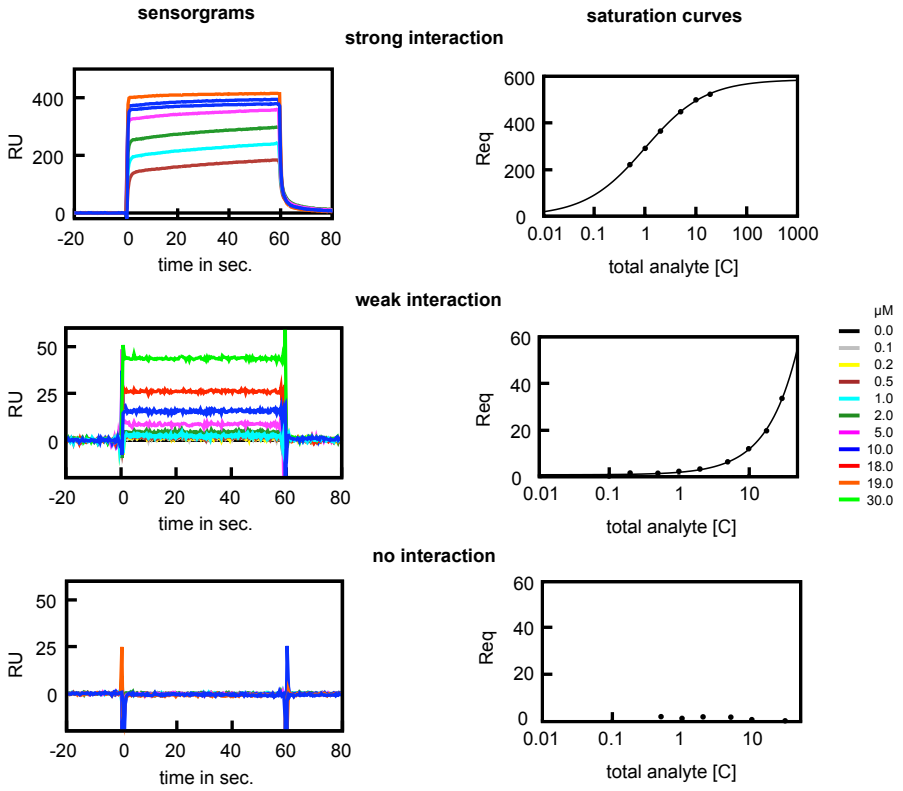


Figure 4.2. Sensorgrams and corresponding saturation curves obtained from SPR experiments. A sensorgram shows the development of the binding signal over time as determined with SPR. Three representative sensorgrams (left) and corresponding saturation curves (right) for a strong (1 µM range), weak (50 - 100 µM range), and no interaction are shown. Different colours represent different concentrations of the analyte that have been injected (see legend to the right). Ideally, for binding affinity determination, RUs are extracted for each run at equilibrium (horizontal signal course). In saturation curves (here with logarithmic x-axis), these RUs at equilibrium (Req) are plotted versus the total analyte concentration. K_D determination is more reliable if saturation of the signal courses can be observed as it is the case for the first saturation plot.

5. On the specificity of protein interactions

5.1. Defining the binding specificity of a protein interaction

The binding specificity of a protein interaction is much more difficult to define than its binding affinity. When is a protein interaction specific? From a qualitative perspective, a protein A binds specifically a protein B (hereafter also called ligand B) if protein A binds with significantly less affinity to other proteins (ligands) B' [133]. However, in order for a protein interaction to be specific, has the reverse case to be fulfilled as well (i.e. protein B binds specifically protein A)? This question has not been addressed by any of the published literature that I have found on binding affinity and specificity. The prevailing view seems to be that the unidirectional way is sufficient for a protein interaction to be specific.

Binding specificity is a measure for the extent of discrimination by a protein between ligand B and other ligands B' [133]. Binding specificity can only be described by assessing the differences in binding affinities between a protein and its ligands. Therefore, binding specificity can only be expressed in *relative* values, in contrast to binding affinity, which can be expressed in *absolute* values. Can binding specificity be quantified? The difference in binding affinity for a protein A towards two ligands B and B' can be described as the difference of the differences in Gibb's free energy $\Delta\Delta G$ of the protein interactions AB and AB' [128, 133]:

$$\Delta\Delta G = \Delta G_{AB} - \Delta G_{AB'} \tag{5.1}$$

Another measure of binding specificity has been proposed by Eaton *et al.* [134]:

$$S_A = \frac{[AB]}{\sum_j [AB_j]} = \frac{K_A[A][B]}{\sum_j K_{A_j}[A][B_j]} = \frac{K_A[B]}{\sum_j K_{A_j}[B_j]} \tag{5.2}$$

where S_A is the specificity of protein A for ligand B in comparison to other ligands B_j. K_A and K_{A_j} are the association constants for the protein interactions AB and AB_j, respectively. S_A will be close to zero for poor binding specificities and greater for better binding specificities. However, specificities of two proteins can only be compared using values of S_A if they have been determined using the same ligand space (comprising the ligands B_j). Is it possible to define a cut-off for $\Delta\Delta G$ or S_A that discriminates between specific and non-specific proteins? Such a cut-off would probably be equally problematic as cut-offs defined for binding affinities that aim to

discriminate between binding and non-binding events (see section 6.2). Thus, as for binding affinities, binding specificities should best be considered at the continuum.

5.2. The intertwined relationship between affinity and specificity of protein interactions

In the previous section, we have seen that binding specificity can only be defined with respect to the binding affinity, pointing to the tight relationship that must exist between both. The two equations 5.1 and 5.2 for binding specificity show that high affinity does not necessarily imply high specificity. Indeed, a low affinity interaction between a protein and a ligand can be specific if interactions between the protein and other ligands are of even weaker binding affinity. Equation 5.2 emphasises that the balance or sum over all binding affinities to all ligands considered finally determines the binding specificity of an interaction.

In the cell, the biological function of a protein interaction determines its requirements for a certain affinity and specificity [128]. Depending on the cellular context, a certain combination of low/high affinity and low/high specificity will be suitable. Accordingly, biological examples for each possible combination have been identified (see Table 5.1 extracted from [128]). This demonstrates that in biological systems affinity and specificity of protein interactions have to be independently modulated and no direct correlation between affinity and specificity can generally exist [128].

	high affinity	**low affinity**
high specificity	antibody/antigen	regulon repressors
low specificity	MHC/peptide	non-specific DNA-protein interactions

Table 5.1. Biological examples for all four possible combinations of high and low affinity and specificity (extracted from [128]).

Greenspan [133] and Szwajkajzer *et al.* [128] further point out that the relationship between affinity and specificity significantly depends on the factors responsible for the affinity differences. Such factors comprise:

1. the structural diversity and number of the ligands available for comparative binding analysis;
2. the structural flexibility of the interaction partners;
3. the kind of non-covalent interaction forces that exist between the interaction partners;
4. the detection ranges of the assays for binding that are employed.

The following examples aim at illustrating these points. If the set of ligands B' considered is structurally very diverse to the ligand B in question, most of the modi-

fications of protein A that lead to an increase in binding affinity for ligand B are less likely to have similar effects for the ligands B'. Here, increases in binding affinity are likely to lead to increases in binding specificity. If a modification of protein A increases the binding site rigidity to better fit the shape of ligand B, there is a certain chance that this modification will lead to an increase in binding specificity as often structural flexibility is important to bind to other ligands B' (see next section) [128, 133]. Electrostatic interactions are highly dependent on the orientation and distance of the functional groups of molecules to each other that establish them. Therefore, electrostatic forces are likely to provide more specific non-covalent contacts than hydrophobic interactions.

All these points raised so far demonstrate that the relationship between affinity and specificity of protein interactions is often complex. Szwajkajzer *et al.* [128] provide an interesting example where affinity and specificity are directly correlated with each other. In host-guest chemistry (see also next section), a rigid host structure is designed to be sterically and electronically complementary to the structure of the guest molecule. Based on the rigidity and complementarity of the molecules, here, affinity and specificity are modulated together where an increase of the former will ultimately lead to an increase of the latter. Such conditions cannot be observed for biological macromolecules such as proteins and DNA that have to maintain a certain degree of flexibility for accomplishing their biological activities [128].

Eaton *et al.* [134] claimed that increasing the affinity of a protein for ligand B is very unlikely to lead to a similar increase in affinity for other ligands B. Hence, increases in affinity will most likely lead to increases in specificity. This view clearly contradicts statements made in the previous paragraphs. Eaton *et al.* worked with the *in vitro* evolution technique systematic evolution of ligands by exponential enrichment (SELEX), a strategy to select high affinity oligonucleotides for target molecules, e.g. peptides. DNA molecules are much more rigid and display much less structural diversity than protein structures [133]. In addition, DNA displays poor hydrophobicity on its surface and is negatively charged. These properties constrain much the structural diversity of the set of ligands when working with protein–DNA interactions [133]. Therefore, the relationship between affinity and specificity may be simpler in the case of protein–DNA interactions [133]. However, Carothers *et al.* [135] have shown that oligonucleotides selected for high-affinity binding do not necessarily bind more specifically to their targets.

In summary, it becomes clear that *"no universal relationship between affinity and specificity can be established"* [136]. Altering the binding affinity of a protein to one of its ligands will most likely impact the affinity to other ligands, in positive or/and negative ways. Thus, especially for drug design, the relationship between affinity and specificity is crucial and it will be necessary to manipulate both affinity and specificity to guarantee a successful therapeutic effect (e.g. by reducing side effects) [133].

5.3. Thermodynamic aspects of the specificity of protein interactions

As we have seen in equation 4.1, affinity can be expressed by the difference in free energy ΔG. A change in ΔG for an interaction *may* thus, in principle, influence its specificity. Hence, mechanisms that lead to changes in ΔG are potential mechanisms to manipulate specificity [128]. Greenspan [133] mentions three major causes for altered affinity, these are changes in: first, shape complementarity; second, chemical complementarity and third, molecular flexibility. In addition, Szwajkajzer *et al.* [128] emphasise the importance of entropy that contributes to the free energy. Changes in entropy upon binding are expressed by changes in the flexibility of the interacting molecules but also by changes in their hydration shells. Upon binding, water molecules or ions can be released or sequestered by the interacting proteins. Thus, binding affinity can be increased by reducing the entropic costs of complex formation, e.g. by locking the interacting molecules into the binding conformation and by controlling the solvation on the surface of the molecules. This is the basic idea of host-guest chemistry (see previous section) [128]. In such a system, optimal complementarity will lead to binding specificity.

Cellular protein complexes are very different from the complexes observed in host-guest chemistry regarding the molecular flexibility. Proteins usually encounter a loss of conformational flexibility and substantial reorganisation of the hydration shell when they bind to other proteins [128]. Disorder to order transitions of protein structures upon binding are common (see section 2.4). Such transitions cost free energy but on the other hand allow for induced-fit interactions that can lead to better fits between protein and ligand. Increases in shape and chemical complementarity between a protein and a ligand do not necessarily lead to an increase in affinity if entropic costs outweigh the enthalpy gain [128]. This illustrates that changes in free energy upon binding are mostly very complex. Attempts to increase the binding specificity between two interacting proteins by studying solely the structure of the complex neglects the thermodynamic nature of protein interactions and are therefore much less likely to be successful [128].

5.4. Biological aspects of the specificity of protein interactions

Protein interaction specificity at the molecular level is often studied between minimal fragments of proteins, e.g. globular domains and SLiMs, that define the *core* binding interface between the two interaction partners. This limitation mostly originates from the fact that larger proteins are more difficult to maintain in stable and active forms for biochemical or biophysical assays. Similar size limitations apply to molecular modelling approaches where bigger molecules exceed the calculation power of available computer clusters. Of course, we can gain important insights on the molecular mecha-

nisms of interaction specificity when focussing on minimal interaction fragments. Yet what can we learn from these results about the specificity of protein interactions in the cell?

5.4.1. The influence of sequence context on the specificity of protein interactions

Numerous studies provide increasing evidence that extended protein fragments or full length proteins display *altered* or/and *more* intermolecular contacts than those observed between the minimal interacting fragments. These changes in molecular contacts were shown to substantially alter the binding affinity and sometimes specificity of the interaction. In addition, it has been shown that the structure and/or dynamics of minimal interacting fragments can be dramatically changed in the extended or full length context (for more details, see our review presented in chapter 10). The term *sequence context* is used to describe these regions in protein sequences that influence the binding interface of the minimal interacting fragments. Sequence context can be extensions of the core interacting fragments, neighbouring domains, or other regions of the sequence that are not contiguous to the fragments under consideration (see Figure 5.1). Intramolecular allostery and cooperative binding can be seen as examples where sequence context influences the binding properties of two interacting proteins. Intramolecular allostery includes cases where the binding of a region of a protein sequence to a globular domain that it carries leads to conformational changes of the binding pocket of this domain that in turn alters binding to its target. Cooperative binding describes cases where several distinct binding interfaces, e.g. formed by several globular domains and linear motifs of the two interaction partners, cooperate to give an overall stronger interaction [43].

5.4.2. Specificity vs. multi-specificity vs. promiscuity

The study of large protein-protein interaction networks revealed that many proteins, called hub proteins, interact with a huge number of other proteins. Given these many interaction partners, it may seem that hub proteins are unspecific [136, 137]. Let us consider the example of the highly studied tumour suppressor protein p53. p53 has more than 230 identified interaction partners in the database STRING (see section 6.3) [138]. p53 is implicated in the regulation of important biological processes such as cell division, cell growth, apoptosis, and transcription [139]. Given these essential functions, it seems impossible that p53 will not specifically bind its targets.

The concept of modular protein architecture (see section 2) may partially serve in resolving these at first sight contradictory observations [43,136,137]. Possessing different modules confers different interaction sites on a protein. Especially disordered regions allow a protein to have different interaction sites (e.g. different classes of SLiMs) located within a relatively short region of a protein. Half of the about 390 residues of p53 are predicted to be disordered [140]. They encode for numerous overlapping

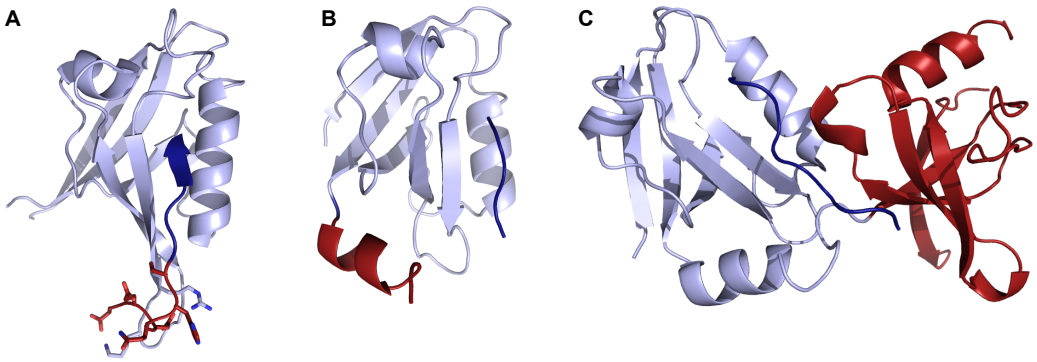


Figure 5.1. Examples of sequence context in the PDZ domain family. PDZ domains, bound C-terminal peptides, and the parts of the structures that constitute the sequence context are coloured in light blue, dark blue, and red, respectively. A: PDZ3 of Par3 bound to an extended PBM derived from PTEN (PDB code: 2K20 [12]). Residues of the extended PBM and key interacting residues of the PDZ domain are shown in sticks. B: PDZ3 domain of PSD-95 bound to a PBM derived from CRIPT (PDB code: 1BE9 [8]). PDZ3 possesses an additional C-terminal α helix that influences peptide binding [13, 14]. C: PDZ3 of ZO-1 bound to a PBM derived from JAM-A (PDB code: 3TSZ [15]). The neighbouring SH3 domain that is located C-terminal to PDZ3 influences peptide binding.

SLiMs including many sites for PTMs [43] (see Figure 5.2).

Using one particular interaction interface, a protein A will bind ligands of a particular type X (e.g. a particular class of SLiMs). If protein A binds a few ligands of type X with much higher affinity than other ligands of type X, then protein A is specific for this type of ligands and the employed interaction interface. At the same time, protein A might be able to bind as well to numerous other proteins using other types of interaction sites (e.g. other globular domains or SLiMs). Thus, the question whether a protein such as p53 is promiscuous or specific cannot solely be answered by looking at the total number of its interaction partners but rather by concentrating on groups of interaction partners that share similar binding interfaces. The term *multi-specificity* might be used to describe such cases where a protein binds specifically to different types of ligands via different interaction interfaces.

5.4.3. The influence of cellular context on the specificity of protein interactions

Pairwise interactions between proteins depend not only on their mutual binding energy but also on other parameters that influence their localisation, concentration and active state in the cell. Proteins are usually expressed at precise moments during the cell cycle or under certain physiological conditions. Proteins are actively transported and localised to certain cellular compartments or sub-locations in the cytosol. Proteins become activated or inhibited via PTMs. Individual proteins often carry out their functions in complexes composed of multiple proteins. Often, an interaction between

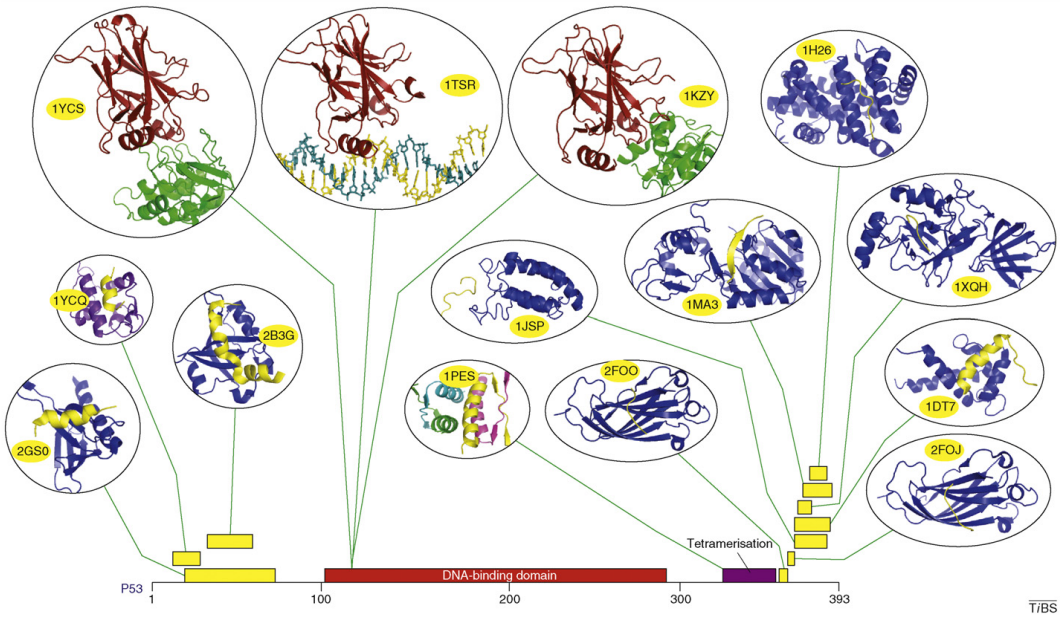


Figure 5.2. Binding interfaces of p53 with structural evidence. Protein modules of p53 are illustrated with red (globular domains) and yellow (SLiMs) boxes at the bottom of the figure. The tetramerisation domain of p53 is shown in purple. Published structures in the PDB involving these interaction sites are displayed in cartoon representation using the same colour code to highlight structure segments of p53. PDB codes of these structures are indicated in yellow ovals. This figure has been extracted from [43].

two proteins prerequisites their assembly into a multiple protein complex via a scaffolding protein. Overlapping interaction sites in one protein (see Figure 5.2) allow for molecular switching, e.g. if one ligand is bound, interactions with other potential ligands may be disabled [141]. All these highly regulated processes point to a discrete and deterministic cellular system controlling that protein interactions are formed at the right moment and place [43]. In this regard, interaction specificity between proteins in biological systems might only be understood when investigating both the molecular and cellular mechanisms.

5.5. Specificity of domain–linear motif interactions

How specific are domain–linear motif interactions? A particular type of SLiM is usually bound by a particular type of domain, e.g. PDZ-binding motifs are recognized by PDZ domains (see Table 2.1). However, it is clear that not every instance of a domain type, e.g. PDZ2 of DLG1, will bind to every instance of the corresponding SLiM type, e.g. all PDZ-binding motifs in the human proteome. A prevailing conception is that there are molecular rules that define, which domain instances will preferentially bind which SLiM instances. In several large-scale studies researchers tried to identify such specificity rules and aimed at assessing how much overlap in ligand space exists

between individual domain instances (see section 6.4.3). Whereas some studies point to substantial overlap in ligand space at least *in vitro* [142], there are also fascinating cases of high specificity in domain–SLiM interactions.

One such example has been published by Zarrinpar *et al.* [143]. They assessed the specificity of the interaction between the two yeast proteins high osmolarity signaling protein 1 (Sho1) (containing an SH3 domain) and polymyxin B resistance protein 2 (Pbs2) (containing an SH3-binding motif). 38 chimeric Sho1 proteins were created by replacing the original SH3 domain either with one of the other 26 known yeast SH3 domains or with one of 12 selected SH3 domains from multicellular organisms. These chimeric proteins were assayed for their capability in activating the high-osmolarity glycerol (HOG)-pathway that is controlled by the Sho1-Pbs2 interaction. Interestingly, only the chimeric proteins with SH3 domains from other organisms than yeast allowed cell growth under high salt concentrations that necessitates activation of the HOG pathway. These results indicated negative selection of Pbs2 to prevent binding to any of the other 26 yeast SH3 domains than Sho1. Mutations in the SH3-binding motif of Pbs2 that either strengthened or weakened the binding affinity to Sho1, were shown to reduce the specificity to the Sho1 SH3 domain suggesting that the Pbs2 motif evolved for an optimal balance between SH3 binding affinity and specificity [144]. Mutations that led to more promiscuous binding behaviours of Pbs2 disfavoured yeast cells in comparison to wild type cells under normal growth conditions, probably due to detrimental interactions of Pbs2 with other SH3 domains.

Even though it has been suggested that negative selection might be a main contributor to binding selectivity in domain-SLiM interactions of unicellular organisms, this mechanism is unlikely to be sufficient for network specificity in more complex organisms [144]. Several authors have pointed to the importance of cellular context in multi- but also unicellular organisms for binding specificities between domains and linear motifs [1, 142, 145]. However, the degree of specificity or promiscuity of a domain–linear motif interaction might actually depend on its biological function [142].

5.6. Specificity of PDZ–peptide interactions

More than a decade of research has been published with the aim of assessing the binding specificities of PDZ domains. As has been discussed in section 5.2, specificity can only be assessed when comparing binding affinities. Thus, systematic determination of binding affinities of several peptides to several PDZ domains combined with structural analysis is likely to significantly contribute to our understanding of PDZ interaction specificities. However, only very few of such studies (including our article presented in chapter 9) have been published so far [10, 81, 82]. Maybe it is because of this lack of data that we still do not clearly understand the rules that define which PBM will bind to which PDZ domain in particular and how specific or promiscuous PDZ–peptide interactions are in general.

5.6.1. Troubles in classification of PDZ domains

As introduced in section 3.2.1, C-terminal PBMs are commonly grouped into three subclasses, based on the residue at peptide position -2. Many studies concentrated on the analysis of which PDZ domain residues define to which of these subclasses of peptides a PDZ domain will bind. Some studies suggest that only very few domain residues, especially at the first and fifth position of the $\alpha 2$ helix (see Figure 3.4), determine the class membership of a PDZ domain. Mutations at these positions have been shown to be sufficient to convert the class specificity of one PDZ into another (see section 3.2.1) [73]. Motivated by these observations, Bezprozvanny and Maximov [146] proposed a classification of PDZ domains into 25 groups based on the chemical properties of the residues at the first and fifth position of the $\alpha 2$ helix. Contradictory opinions have been obtained by other studies, namely from Vaccaro *et al.* [147], who presented evidence that subclass affiliation of PDZ domains is much more complex. In an interesting correspondence, they provide examples like the first PDZ domain of MPDZ that based on its residues at the two helix positions, should bind class I peptides but has been found to bind class II ligands [148].

Overall, researchers agreed on that the classification of PDZ domains and/or ligands into three subclasses is very unlikely to sufficiently explain the binding preferences observed between PDZ domains and their ligands. By now, it is widely accepted that the last four to five peptide residues each contribute to PDZ domain binding. It has been suggested that the number of accepted residues at peptide position -1 may define the level of promiscuity of a given PDZ domain [82]. Based on preferences for residues at the last six peptide positions, 16 distinct specificity classes of PDZ domains have been defined using a large-scale data set on phage display-derived PDZ–peptide interactions [9]. Interesting findings on multiple specificities of PDZ domains have been published by Gfeller *et al.* [92]. They grouped phage display peptides obtained for one PDZ domain into subgroups based on sequence similarities. This revealed that PDZ domains can recognize ligands of different subgroups (others than the three canonical classes described) that are bound by the PDZ domain in structurally very different ways [92]. Velthuis *et al.* [56] applied a published PDZ interaction predictor for screening a whole proteome for potential binders to PDZ domains. Based on the screening results, they concluded that PDZ domains bind with extensive overlap to sets of peptides. However, as discussed in our article presented in chapter 9 the predictor they used suffered from a high false positive rate (FPR), overall questioning their findings.

These various results raise the question about which criteria can be used to classify PDZ domains. Is it sufficient to know one physiological or artificial ligand to be able to assign a PDZ domain to a certain class? How physiologically relevant are overall such class assignments? Quite a few PDZ domains have been shown to be able to bind ligands from at least two peptide subclasses (see section 3.2.1). In a seminal work, Wiedemann *et al.* [81] provided evidence that class assignment of PDZ domains highly depends on the affinity range considered (see Figure 5.3). PDZ domains can be of dual

specificity if low binding affinities are accepted but still might display a clear preference for ligands of one subclass. In line with these findings, their data also suggests that the level of overlap between PDZ domains for sets of bound ligands highly depends on the affinity cuf-off considered with high affinity cut-offs leading to just little overlap and lower affinity boundaries producing substantial overlap [81]. Thus, instead of tempting to group PDZ domains into distinct classes it might be more relevant to consider a continuous specificity space of PDZ domains as it has been suggested after analysis of the first large-scale data set on physiological PDZ–peptide interactions published from MacBeath and co-workers [10].

Kd	<10 μM		<50 μM		<100 μM	
AF6	**11**		**602**		**2,103**	
Class I	2	(18%)	113	(19%)	308	(15%)
Class II	9	(82%)	416	(69%)	1070	(51%)
Non-class I/II	0	(0%)	73	(12%)	725	(34%)
ERBIN	**1**		**48**		**181**	
Class I	1	(100%)	43	(90%)	131	(72%)
Class II	0	(0%)	4	(8%)	25	(14%)
Non-class I/II	0	(0%)	1	(2%)	25	(14%)
SNA1	**30**		**1,622**		**3,527**	
Class I	30	(100%)	1492	(92%)	2003	(57%)
Class II	0	(0%)	4	(0%)	161	(5%)
Non-class I/II	0	(0%)	126	(8%)	1363	(39%)

Figure 5.3. Influence of binding affinity cut-offs on the level of specificity of PDZ domains. Wiedemann *et al.* [81] predicted binding affinities for all possible 4 residue-long peptides towards three human PDZ domains. For three different affinity cut-offs (indicated above the table), the number of predicted binding peptides of class I, class II, and non class I/II for each PDZ were counted and are shown in the table. With weaker binding affinity cut-offs, the PDZ domains become less specific accepting peptides from all three categories. This figure has been extracted from [81].

5.6.2. The role of phage display for studying PDZ binding specificities

Phage display has been recurrently used to study the binding profiles and specificities of PDZ domains. The Sidhu lab developed and refined the protocol that allowed presentation of C-terminal peptides by bacterial phages, and consequently applied it in several single case [60, 63, 82, 149–151] and large-scale studies on PDZ domains [9, 152, 153]. In phage display, billions of different peptides are expressed on the surface of bacteriophages and are presented to an analyte (e.g. a PDZ domain) that is attached to a solid support. In several rounds of binding, washing, and amplification, peptides are selected that bind with high affinity to the analyte. The binding profile of the analyte can be obtained from consequent sequencing and alignment of the selected peptides. Sidhu and co-workers often combined phage display with structural analysis and these studies have provided important contributions to our understanding of binding specificities of PDZ domains.

Nevertheless, the standard phage display procedure has a major drawback that unfortunately, is recurrently disregarded during data analysis. Peptide selection in phage

display is solely based on the binding affinities of the peptides to the presented analyte and their rates of dissociation. In contrast, cellular PBMs are selected by evolution for their functionality, e.g. mediating protein interactions in the signalling context. Here, specific and transient interactions are observed and high binding affinity plays rather a secondary role. As a consequence, phage display peptides have sometimes been shown to bind with higher affinity to PDZ domains and to display different sequences as compared to their natural counterparts [82,149,154] (this has been subject in our article presented in chapter 8). Thus, assessment of binding specificities of PDZ domains using binding profiles derived from phage display data can lead to results that are irrelevant in the biological context. Phage display data is currently the prevailing type of interaction data for PDZ domains that influenced much our understanding of PDZ–peptide interactions and the development of PDZ interaction predictors (see section 6.4.4). Care should be taken to properly interpret the data in its context and to prevent it from biasing our knowledge on PDZ-peptide interaction specificities.

5.6.3. Dynamic aspects of PDZ interaction specificities

Interesting insights in PDZ–peptide binding specificities have also been gained from molecular dynamics simulations. Nonpolar contacts have been observed to substantially contribute to the overall free energy of binding to PDZ domains [155]. It has been hypothesised that this might partially explain their observed promiscuous binding behaviours [155]. The entropic contribution to PDZ binding has often been unfavourable in molecular dynamics simulations [155]. Indeed, the peptide undergoes considerable structuring upon binding. PDZ domains have been observed to sometimes become more rigid, sometimes more flexible upon binding depending on the ligand studied. Thus, entropy might be an important player in binding specificities of PDZ domains [155]. The energetic contribution of water release upon peptide binding to PDZ domains has been studied more in detail by Beuming *et al.* [156]. Calculations suggested that peptides displayed higher affinity to PDZ domains when they had residues, such as Trp, that displaced more water molecules [156]. However, it does not seem that such hydrophobicity-driven affinity changes necessarily lead to higher interaction specificities (see the discussion in our article presented in chapter 8).

5.6.4. Influence of sequence context on PDZ interaction specificities

Most experimental studies on PDZ–peptide binding specificities used constructs comprising the core PDZ domain and peptides of five residues in length. As discussed in chapters 9 and 10, binding affinities and specificities observed for such minimal interacting fragments can change when extending the constructs. Peptide residues upstream of the last five residues can modulate binding affinities and specificities to PDZ domains. Extensions of PDZ domains and neighbouring domains can alter the binding properties of the PDZ towards C-terminal ligands.

PDZ domains actually constitute a prototype for the study of the influence of such sequence context on the structure and function of a globular domain. For many PDZ domains it has been shown that they possess additional secondary structure elements that precede or follow the core PDZ fold and that impact the structure and function of the domain. Fewer examples exist where such extensions are disordered (see our NMR study presented in chapter 7). Neighbouring domains have also been demonstrated to influence significantly the fold and peptide binding properties of PDZ domains. This is true for neighbouring domains that are non-PDZ domains, such as SH3 domains [15, 157] (see Figure 5.1), but has been particular striking for PDZ domains that are separated by just a short linker sequence (see Figure 5.4). Structural studies revealed that PDZ1 and PDZ2 of DLG proteins, PDZ4 and PDZ5 of PATJ, PDZ1 and PDZ2 as well as PDZ4 and PDZ5 of GRIP1, and PDZ1 and PDZ2 of amyloid beta A4 precursor protein-binding family A member 1 (APBA1) display many interdomain contacts and actually form one globular unit, thus the term *supramodule* has been introduced to describe them [158].

PDZ–peptide interactions found to be promiscuous when studied *in vitro* with minimal interacting fragments, might turn out to be more specific in a full length and *in vivo* context.

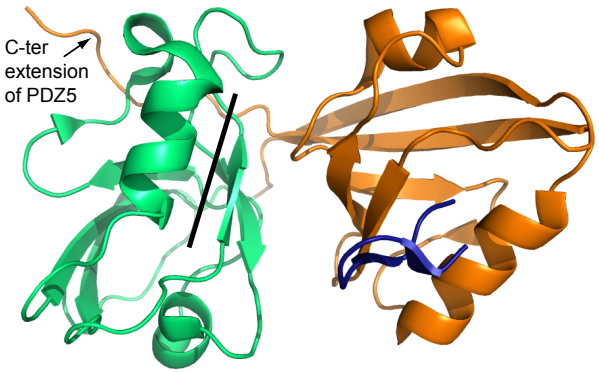


Figure 5.4. The PDZ45 supramodule of PATJ. PDZ4 (green) and PDZ5 (orange) of PATJ form a supramodular structure. PDZ5 is in complex with a C-terminal PBM derived from NG2 (dark blue). Both PDZ domains interact with each other via multiple contacts including an unstructured C-terminal extension of PDZ5. The peptide binding pocket of PDZ4 (indicated by a black line) is inaccessible for PBMs due to the orientation of PDZ4 towards PDZ5 (PDB code: 3R0H [117]).

6. Prediction of protein interactions

The reader will first be introduced to a widely used concept for assessing prediction performances before presenting different approaches for domain–motif interaction predictions in general and PDZ domain-mediated interaction predictions in particular. The terms defined in the following section are recurrently used in this chapter.

6.1. Assessment of prediction performances using ROC statistics

Receiver operating characteristics (ROC) analysis is widely used in computational biology to assess the performance of protein interaction predictors in combination with a gold standard test data set of validated positive and negative protein interactions. With the help of a scoring scheme, the predictor is supposed to correctly classify the positive and negative protein interactions from the test data set. For example, interactions that scored above a certain threshold are classified as positive interactions and otherwise as non-interactions. This leads to four types of entities, true positive, true negative, false positive and false negative interactions that are usually summarised in a "confusion matrix" (see Table 6.1).

	real positive interaction	**real negative interaction**
classified as positive	**t**rue **p**ositive interaction (TP)	**f**alse **p**ositive interaction (FP)
classified as negative	**f**alse **n**egative interaction (FN)	**t**rue **n**egative interaction (TN)

Table 6.1. Confusion matrix for ROC statistics.

Based on the number of TP, FP, FN and TN interactions identified in a certain test run of the predictor, different measures can be calculated that reflect prediction performances.
The **sensitivity** (also called **true positive rate (TPR)** or **recall**) is defined as

$$sensitivity = \frac{TP}{P} \tag{6.1}$$

where P is the total number of positive interactions in the gold standard test data set. It is $P = TP + FN$. The **specificity** (also called **true negative rate (TNR)**) is defined as

$$specificity = \frac{TN}{N} \tag{6.2}$$

where N is the total number of negative interactions in the gold standard test data set. It is $N = TN + FP$. The **false positive rate (FPR)** is defined as

$$FPR = \frac{FP}{N} = 1 - specificity \tag{6.3}$$

Other measures that are frequently used, are **precision** and **accuracy**:

$$precision = \frac{TP}{TP + FP} \tag{6.4}$$

$$accuracy = \frac{TP + TN}{N + P} \tag{6.5}$$

Thus, for a specific threshold, a test run will result in a value for the sensitivity and a value for the FPR. These values vary depending on the threshold employed. Usually, increases in sensitivity lead to increases in the FPR. This relationship is plotted in ROC curves (see Figure 6.1). Perfect classification is represented by 0% FPR and 100% sensitivity, that is point (0,1) in a ROC curve. Thus, the closer the ROC curve passes the upper left corner of the diagram (point (0,1)), the closer the predictor to optimal performance. A predictor that is as good as random would obtain for a certain threshold a sensitivity of 50% and a FPR of 50%, resulting in a bisecting line in the ROC curve. A common way to indicate the quality of a predictor is to measure the area under the ROC curve (AUC). The higher the ROC curve (the closer to point (0,1)), the greater the area under the curve. In theory, the AUC can obtain values between 0 and 1, with 1 for perfect classification. In practice, an AUC can never be below 0.5, e.g. will never be worse than random.

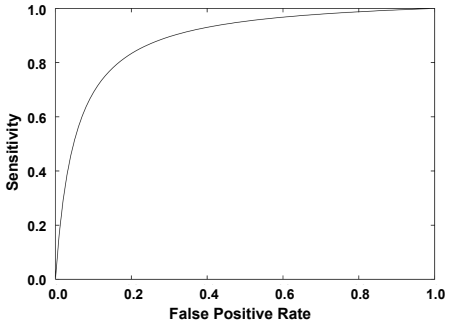


Figure 6.1. Example for a ROC curve. A ROC curve for a predictor with a good relationship between sensitivity and FPR.

6.2. Binary interaction versus binding affinity information

Probably the most elementary simplification that is often made in molecular and computational biology consists in the reduction of protein interactions to be binary, e.g. two proteins are either seen to bind to each other or not. This model is in contradiction to well-established biophysical findings indicating that the binding affinity of protein

interactions is continuous. The detection of a protein interaction largely depends on the experimental method applied. Different experimental methods have different protein interaction detection ranges. This can result in cases where using one method an interaction between two proteins would be detected whereas with another these two proteins would be classified as non-interacting.

Currently, binding affinities can only be determined in *in vitro* systems making it impossible to obtain information about the range of binding affinities that exists in cellular systems. Therefore, every cut-off in binding affinity that is set to discriminate between binding and non-binding is motivated by technical concerns and subjective views and is unlikely to be of any biological significance. Consequently, protein interaction prediction tools that return a score for a given protein pair that correlates with (relative) binding affinities are likely to be biologically more meaningful than predictors that only deliver binary information. However, correct prediction of (relative) binding affinities prerequisites training and test data sets that provide this kind of information. Gathering of binding affinity information from various different sources is likely to result in erroneous data sets because binding affinities obtained from different experimental methods are not necessarily comparable. Thus, attempts of large-scale determination of (relative) binding affinities of protein interactions will be crucial for advances in their predictions (see section 6.4.4).

6.3. Prediction of protein interactions without module information

From the various strategies for protein interaction predictions that exist, I only focus on one publicly available resource, the database STRING, that combines predicted and experimentally validated protein interactions into the probably most complete data set available to date. STRING (search tool for recurring instances of neighbouring genes) has been originally published as a server for the retrieval of genomically associated genes [159]. It has been suggested that genes that repeatedly occur in clusters on the genome are likely to encode for proteins that are functionally associated, e.g. by working in the same metabolic pathway. In the following years, STRING has been consequently extended by integration of direct (physical) and indirect (functional) protein associations from four different sources: HTP experiments, co-expression data, text-mining and other repositories of protein interaction information. By now, STRING covers protein association information for 1,133 organisms [138]. If applicable, protein associations are transferred between organisms based on orthologies. Particular attention has to be paid to the fact that STRING contains a mix of direct and indirect as well as predicted and experimentally validated protein associations. The origin and kind of each protein association is properly tracked and visible to the user.

6.4. Prediction of protein interactions using module information

The prediction of protein interactions has much advanced with our understanding of how protein modules mediate protein interactions. As mentioned in section 5.5, a SLiM is recognized by a specific type of globular domains. Thus, a very rough approximation in protein interaction prediction consists of the prediction of domains and SLiMs in protein sequences and the assumption that if protein A has a domain that in theory can recognize a SLiM that has been identified in protein B, A and B might be able to interact [160]. This approach has a significant advantage in providing in addition to the protein interaction prediction a prediction about the putative binding interfaces. In the following, I provide a quick overview about common methods for domain and SLiM predictions and summarise different approaches for interaction prediction that are based on the protein modules concept. I do not discuss tools that aim at predicting *de novo* domains or SLiMs (e.g. not yet annotated modules). Such newly predicted modules cannot be used for protein interaction prediction because information on the corresponding interacting module is usually missing.

6.4.1. Prediction of globular modules – domains

Several tools have been developed during the last years that predict regions of order and disorder in a given protein sequence (for a performance comparison and review see [161]). These predictors allow to obtain a fairly well overview about whether and where a protein might have globular domains. However, using such predictors no information can be obtained about what type of globular domain the protein might have.

The prediction of occurrences of specific types of globular domains in a given protein sequence is possible with tools like Pfam and SMART. Pfam and SMART have been developed at the end of the 90s, the decade of "protein module discovery" (see section 2.1), as a response to the need to automatically identify domains in protein sequences for protein function prediction as well as resources to manage the knowledge on the growing number of modules discovered [162,163]. Both tools use hidden markov models (HMMs) to capture the signature of a domain and to search for occurrences of domains in protein sequences. An HMM can be defined on the basis of a high quality sequence alignment of several validated instances of a certain type of globular domain. A match of an HMM in a protein sequence reveals a score that indicates how similar the predicted domain sequence is to the domain instances used to define the HMM.

Though Pfam and SMART have many characteristics in common, they also differ in some important aspects. Developers of Pfam (Protein families) originally aimed at classifying proteins into families based on sequence similarities they shared without a particular focus on protein modules [162]. As a consequence, Pfam contains HMMs of protein modules, enzymatic domains that function in metabolic pathways, repeats, structural motifs and signalling peptides. SMART (simple modular architecture re-

search tool) has been designed with the aim to describe domains that occur in cytoplasmic signalling proteins, thus modules [163]. By now, SMART stores 1,009 domain signatures [4] in contrast to Pfam that defines more than 13,000 protein families [164]. Using tools like SMART and Pfam, instances of annotated globular modules in protein sequences can be extremely reliably predicted. Other tools for domain prediction include InterPro, a meta server, that unifies information from secondary databases such as Pfam and SMART [165], Prosite [166], and the conserved domain database (CDD) [167].

6.4.2. Prediction of linear modules – SLiMs

As for prediction of globular domains, linear motif prediction can be seen as a two step process: first, definition/annotation of a type of linear motif with validated instances; second, prediction of SLiM occurrences using the defined signature from the annotation process. Conserved features of SLiMs are often captured in regular expressions that can be manually defined in contrast to HMMs. Given the few instances that are often only available for a certain type of SLiM, manual creation of regular expressions becomes necessary to allow subjective input of expert knowledge from the annotator. Such regular expressions can then be used to screen a given protein sequence for SLiM instances. The ELM resource [5], MiniMotif Miner [168], and ScanSite [169] are databases that store manually annotated types of SLiMs and that allow to search for instances of these annotated SLiMs in protein sequences. In addition, ScanSite offers the possibility to enter one's own regular expression and to query whole protein sequence databases. MiniMotif Miner and the ELM resource contain a diverse set of types of SLiMs whereas ScanSite concentrates on annotation of SLiMs that are linked to phosphorylation. Prosite [166] catalogues profiles of any conserved features in protein sequences, including protein domains, families and functional sites such as linear motifs.

In contrast to well-established and straightforward prediction of globular domain occurrences, linear motif prediction faces many difficulties. SLiMs are much shorter and do not possess a tertiary structure that is well conserved as do domains and thus, SLiMs are much more difficult to detect in multiple sequence alignments [170]. In addition, SLiMs occur in disordered regions of proteins that are very difficult to align due to poor sequence conservations. SLiMs possess only a few (sometimes not more than two) very conserved amino acid positions that can often appear by chance in a protein sequence (e.g. within globular domains where they are likely to be non-functional). As a consequence, developers of linear motif prediction tools have in particular to deal with high false positive rates [3]. The strategy employed in the ELM resource [5] to reduce the false positive rate consists of filter development and application. Here, matches of regular expressions of SLiMs in protein sequences are currently **discarded** if:

1. they occur in globular regions of proteins unless structural information is available that suggests their occurrence in exposed loops of globular domains;
2. they are not conserved in homologous protein sequences;
3. the corresponding ELM class is known to only occur in proteins of specific subcellular compartments that are different from those where the query protein is known to be located;
4. the corresponding ELM class is known to only occur in specific organisms that are different from the organism of the query protein;
5. they are likely to occur by chance (a user-defined probability cut-off can be set).

6.4.3. Prediction of domain–SLiM interactions

As mentioned in section 5.5, it is in most cases insufficient to simply know the presence of a domain instance and corresponding SLiM instance in two proteins for protein interaction prediction. Molecular recognition rules define subsets of SLiM instances that are preferentially bound by a particular domain instance. Thus, in numerous studies researchers tried to identify and describe such specificity rules for their application in domain–SLiM interaction predictions. Some attempts focussed on trying to classify domain instances into subgroups based on some molecular signatures to accordingly split up the ligand space that they recognize. However, more promising predictions are likely to be gained in approaches where binding profiles are defined for each single domain instance. This has been the subject in several studies that combined HTP methods with bioinformatic tools to identify the interactomes of model organisms that are mediated by specific types of globular domains. In the following, a few of such examples are summarised.

Probably the first domain-mediated human interactome that had been tried to map involved WW domains [32]. Based on published instances of WW-binding motifs, sequence patterns were defined and used to retrieve all sequences in the human proteome that matched to them. These peptides were synthesised and probed against all human WW domains that were identified by SMART and Pfam. Interestingly, when considering all WW domains and peptides that were involved in at least one interaction, only 10% of all tested interactions were positive. This nicely illustrates the inherent specificity that can exist within domain–SLiM interactions.

Serrano and co-workers [171] used FoldX, an empirical force field developed by their group, to predict the binding profiles for 9 human SH2 domains of which structures were available. FoldX allows to perform *in silico* mutagenesis, each time calculating the free energy of the resulting complexes. The free energies were transformed into binding profiles that were consequently used to screen the human proteome for binding sites of the 9 SH2 domains. Prediction performances improved when additional filtering with secondary structure predictors and phosphorylation data were employed.

In a collaborative study of the Sidhu, Vidal, and Cesareni lab phage display, yeast-2-hybrid (Y2H), and SPOT array data have been combined for the prediction and experimental validation of most protein interactions in *S. cerevisiae* that are mediated by SH3 domains [172]. The power of the predictive model lies in the combination of experimental data from very different sources (artificial peptide, natural peptide, and full length protein screening) that complement each other.

HTP proteomics studies reveal thousands of phosphorylation sites in proteins but the kinases that recognize and phosphorylate these sites remain unknown. Target site prediction for kinase domains using binding profiles is very difficult even when experimentally determined substrate binding sites are available for profile construction. Many isolated kinase domains have been shown to have very low substrate specificities *in vitro*. Specificity in target recognition is increased *in vivo* by the cellular context, including co-localisation via subcellular compartmentalisation, anchoring and scaffold proteins, cooperative effects via non-catalytic interaction domains that additionally bind substrates or docking motifs in substrates recognized by the kinases themselves as well as temporal and tissue-specific gene expression [173]. Linding *et al.* [173] aimed at increasing prediction quality of phosphorylation networks by incorporating contextual information from the STRING database. They could improve prediction performance by 2.5 fold in comparison to purely sequence-based methods.

SPOT technology has been used to define the binding profiles of the two 14-3-3 domains in *S. cerevisiae* [174]. These binding profiles together with motif conservation information and contextual data (gene ontologies, co-expression, co-localisation) have been combined to predict and rank all potential binding sites recognized by these two domains from the yeast proteome. The predictive model could in part be successfully extended to human 14-3-3 domains. Similar large-scale studies have been published for PDZ domains. They are discussed in detail in the following paragraph.

6.4.4. Prediction of PDZ–peptide interactions

Many studies have developed PDZ–peptide interaction predictors with a wide range of different strategies employed. Table 6.2 summarises 26 articles that deal with PDZ–peptide interaction predictions of which 24 describe the development of new prediction tools. Aside from "pure" PDZ interaction prediction studies, I have also considered studies that aimed at developing general domain–SLiM interaction predictors and that performed tests with PDZ interaction data or that focussed on a small set of types of globular domains including PDZ, SH2, SH3, and WW domains. All of these 26 studies focussed on predictions of interactions involving C-terminal PBMs due to the sparse amount of data available for "non-canonical" types of PDZ-mediated interactions, e.g. involving internal ligands, lipids, or domain–domain interactions.

Interestingly, only four articles about PDZ interaction predictions had been published before publication of the ground-breaking large-scale study on PDZ interactions

from the MacBeath group [10]. This study had been followed one year later by another equally important large-scale study on PDZ domains from the Sidhu group [9]. The huge amount of PDZ interaction data provided by these two studies initiated a "flood" of PDZ interaction predictors that were published two to three years later (see Table 6.2). In eight studies the PDZ interaction data from either or both large-scale studies have been used to train and/or test the predictors. This illustrates the importance of HTP data for the computational biology field. Given the relevance of these two large-scale studies in general for the PDZ domain field and in particular for my PhD project, I shortly summarise their main contributions.

Stiffler *et al.* [10] used microarrays coupled with FP to determine the binding of 157 mouse PDZ domains to 217 mouse peptides. This resulted in 1,301 interactions that were cross-checked by binding affinity determination using solution-phase FP assays. Importantly, this study also revealed a huge amount of negative interaction data that, as will be explained below, is extremely important for predictor development. A first prediction model based on this interaction data had been proposed in the same study (see Table 6.2). A second prediction model has been published one year later by the same group, to which I refer to as the "Chen predictor" [11]. The Chen predictor has been trained with the experimental data from their previous study [10]. The heart of this predictor consists of 38 position pairs of domain and peptide residues that were seen to interact with each other in the crystal structure of the peptide-bound PDZ domain of α1-syntrophin [175] (see Figure 6.2A). The training data was used in a Bayesian approach to obtain sub-scores for the occurrence of all possible combinations of amino acid pairs at these 38 position pairs. These sub-scores quantify the positive, neutral or negative contribution of a pair of amino acids at a certain position pair to the overall interaction between a PDZ domain and a peptide (see Figure 6.2B). The sum of the 38 sub-scores for a given PDZ-peptide pair represents the final score, which was suggested to indicate the relative binding strength.

Tonikian *et al.* [9] applied phage display to determine the binding profiles of 28 *C. elegans* and 54 *H. sapiens* PDZ domains using more than 10 billion random peptides. Phage display data is perfect for the construction of a position specific scoring matrix (PSSM) (also often referred to as position weight matrix (PWM)). A PSSM captures the frequency of occurrence of each amino acid at each position within a list of aligned peptide sequences (see Figure 6.3). A PSSM can be constructed for each PDZ domain representing the sequence profile defined by the phage peptides that bound to it. Such PSSMs can be used to predict PDZ-peptide interactions in the following way. For a given peptide and PDZ domain, the normalised frequencies for each residue of the peptide are extracted and summed up from the corresponding PSSM of the PDZ domain. The resulting score reflects the sequence similarity of that peptide to the phage display peptides obtained for the PDZ domain. The higher the score, the higher the similarity to the phage display peptides, the higher the likeliness that the peptide will be bound by the PDZ. Using such PSSMs, Tonikian *et al.* [9] successfully screened the human proteome and several viral proteomes for C-terminal peptides to find potential

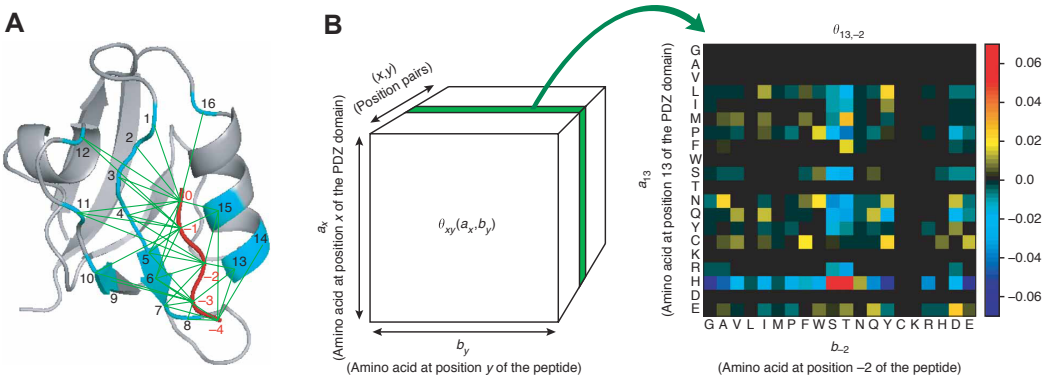


Figure 6.2. Illustration of the prediction model defined by Chen et al. [11]. A: 38 pairs (green lines) were defined based on contacts between domain and peptide residues that have been observed in the crystal complex structure of the PDZ domain of α1-syntrophin [175]. B: Each of these 38 residue pairs can consist of 20x20 possible combinations of amino acids. The PDZ interaction data has been used to derive for each of these amino acid combinations for each of the 38 pairs a value that represents whether this pair of residues at this position is rather favourable or unfavourable for a PDZ–peptide interaction. All 400 values obtained for the residue pair (13,-2), where 13 is the first residue of the α2 helix of the PDZ domain and -2 the peptide position, are illustrated with a heatmap. Orange/red tones represent favourable contribution to binding and blue tones unfavourable contribution. The figure has been adapted from [11].

binding partners for some of the 54 human PDZ domains used in their phage display screen. Unfortunately, phage display does not provide negative interaction data.

The diverse published PDZ interaction predictors can be grouped using different criteria, e.g. kind of input data, main prediction result, and generality of application. Some predictors exclusively use primary sequence data [9, 10, 81, 92, 176–180], others are sequence-based but add information derived from structures, e.g. contact position pairs between PDZ domains and peptides [11, 147, 181, 182]. A third group of predictors is purely structure-based [68, 156, 183–189]. Some predictors aim at assigning to a given PDZ domain the class to which it is likely to belong (e.g. class I, II, III, or dual specificity) [68,177], others aim at predicting a sequence profile, e.g. a PSSM for a given PDZ domain [9, 81, 176, 179, 180, 182, 189] or simply a score for a given PDZ-peptide pair [11, 92, 178, 181]. The most ambitious tools try to predict relative or absolute binding affinities or the free energy of PDZ-peptide complexes [156, 183–188]. Only a few predictors are ready to be applied to any PDZ domain in question [177–180], others prerequisite that certain residue positions of the PDZ domain can be accurately determined (e.g. using sequence alignments) [11, 147]. Frequently, a PDZ domain has to display sufficient sequence similarity to a PDZ domain for which a structure is available [68, 156, 183–189] or a significant amount of binders or/and non-binders has to be known [92, 176]. In very restrictive cases, the predictor can only be applied to PDZ domains that have been used in the training process [81, 182].

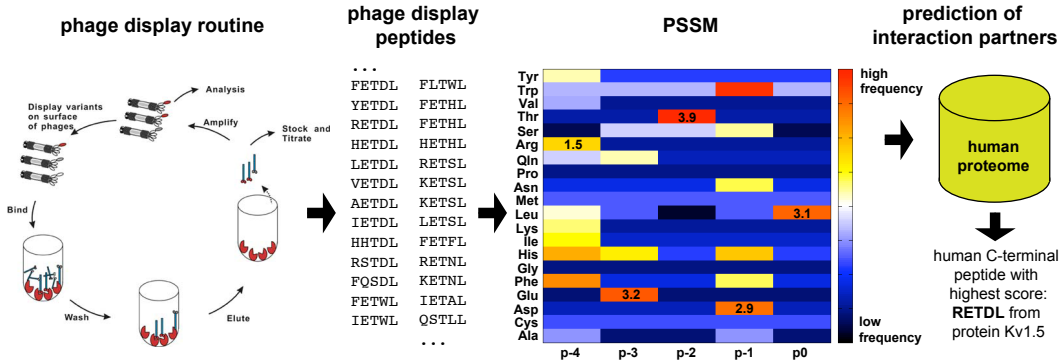


Figure 6.3. Principle of interaction predictions using PSSMs. From left to right: During phage display selection rounds, C-terminal peptides with high affinity for a PDZ domain (here PDZ3 of SCRIB) are selected, sequenced, and aligned. The frequency of each amino acid at each peptide position (in this example only the last five peptide positions are considered) is calculated and stored in a PSSM. In this example, the frequency of each amino acid at each peptide position is illustrated with a colour scheme where very frequent residues at a given peptide position are highlighted with orange tones and very rare residues with blue tones. The PSSM can be used to screen the human proteome for similar C-terminal peptides to the phage display peptides by calculating for each query peptide a score that results from the sum of the frequencies of its residues. The human C-terminal peptide that obtained the highest score out of all available human C-terminal peptides, has the sequence RETDL. The frequencies of these residues are indicated on the PSSM. The phage display routine picture has been taken from www.creative-biolabs.com.

A recurrent finding in several of these studies is that the prediction success for a given PDZ domain depends to a significant degree on its sequence similarity to PDZ domains used in the training process [9, 11, 178, 181]. This observation reflects a frequent problem in computational biology that consists of developing predictors that are biased towards the training data set, often as a result of its limited size [190, 191]. In some cases this can lead to over-optimistic predictions for PDZ domains that are very similar to PDZ domains used in the training data (see our results on the Chen predictor presented in chapter 9).

Controversial findings have been published about whether and to which degree sequence similarity between PDZs results in similar binding specificities. This, of course, depends on whether sequence similarity of the residues of the peptide binding pockets of two PDZ domains is considered or sequence similarity over the whole domain. Whereas some see predictive power using sequence similarities between PDZ domains [9] others are more pessimistic [11].

Very striking is the role that phage display data plays in the field of PDZ–peptide interaction predictions (see Table 6.2). Several HTP phage display studies have been published by Sidhu and co-workers [9, 152, 153]. This data has inspired many predictor developments and led to the formulation of a "DREAM4 Peptide Recognition Domain Specificity Prediction" challenge with phage display data as experimental gold stan-

dard [180,182,189]. However, as has been mentioned in section 5.6 and is demonstrated by our own study presented in chapter 8, high affinity binders selected in phage display can substantially differ in their sequence from natural binders for PDZ domains. Thus, binding profiles derived from phage display data for a given PDZ can highly vary from binding profiles derived from cellular interaction partners. Figure 6.4 shows the binding profiles of mutated Erbin PDZ domains from the DREAM4 prediction challenge. What do they have in common with known properties of cellular PBMs? How useful are tools for the prediction of natural PDZ–peptide interactions that were trained and/or tested on such binding profiles? Hui and Bader [181] took into account these sequence differences between phage display and cellular PBMs when training a predictor by selecting only the genomic-like sequences from the phage display data. To make a test, they screened the human proteome with PSSMs obtained from the unfiltered phage display data. Interestingly, the resulting top predicted binders were more similar in terms of sequence to the genomic-like phage display peptides as compared to the unfiltered phage display data [181].

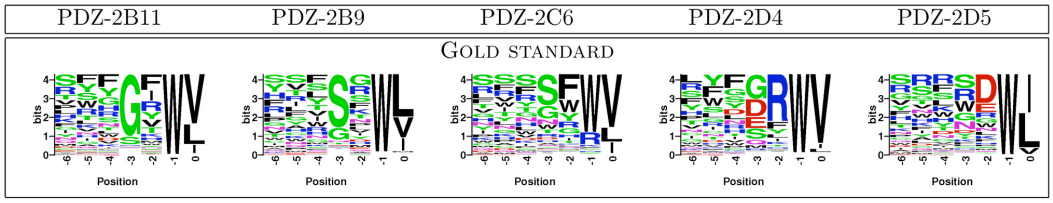


Figure 6.4. The DREAM4 PSSM prediction challenge for PDZs. The gold standard PSSMs (that were to predicted in the DREAM4 prediction challenge) are represented as sequence logos where the size of each letter represents the preference of that residue at a given PBM position. The PSSMs were obtained from peptides that were selected in phage display for binding to mutated Erbin PDZ domains. The figure has been adapted from [182].

The publication of real negative interaction data has probably been the most important contribution of the study of Stiffler *et al.* [10] to the field of PDZ interaction predictions. Interaction predictor development without negative information is very difficult, and the quality and quantity of negative training data determine final prediction performances [178]. Hui and Bader [181] developed a clever method that employs PSSMs derived from phage display data to create negative interaction data that can be used for predictor training. However, artificial negative interaction data, e.g. derived from random assignment of protein pairs or proteins from different cell compartments, has been shown to significantly bias predictors [192]. This problem has been more and more recognized by the scientific community and results in annotation efforts for negative interaction data [193] (see our article presented in chapter 9).

Given the plasticity of PDZ domains to adopt to diverse peptide sequences (as has been strikingly demonstrated for dual specificity PDZs for example) it is questionable to which extent sequence-based prediction approaches will succeed. A yet undiscussed group of PDZ interaction predictors constitutes those that apply molecular dynamics (MD) simulation to this problem. The massive amount of structural data on PDZ

domains encouraged numerous groups to engage in PDZ-peptide complex predictions (10 publications in total, see Table 6.2). An open question is how applicable and successful such approaches will be for the screening of hundreds or thousands of peptide sequences? The developers of PREDIADAN provide a first attempt but inspection of predicted binding profiles questions the biological relevance of the predictions (see Figure 6.5) [185]. Probably the most direct (and most sophisticated) way to predict binding specificities is the prediction of absolute binding affinities from complex structures. Gerek and Ozkan [186] obtained promising results. Nevertheless, an interesting analysis revealed that binding affinity calculations are extremely sensitive to errors in modelled complexes. More than 20% of error is likely as soon as the RMSD of the modelled complex diverges more than 2 Å from the real structure [194].

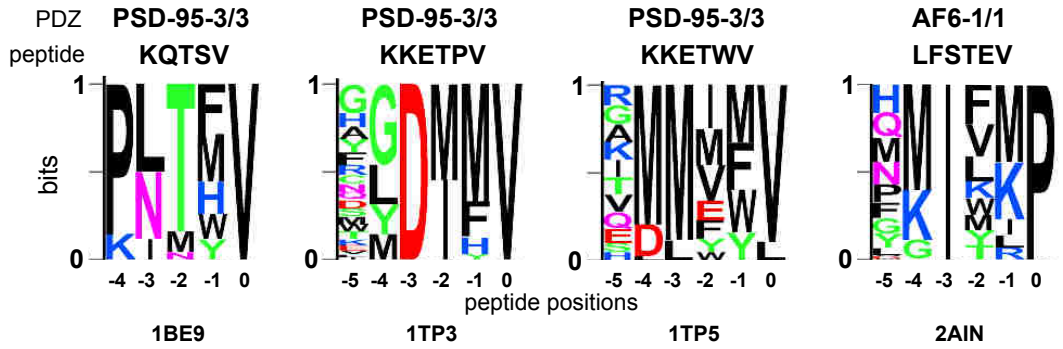


Figure 6.5. Predicted peptide sequence profiles for PDZ domains extracted from PREDIADAN. PREDIADAN has been used to predict for every available structure of a complex between a PDZ domain and peptide, the peptide sequence profile of the respective PDZ domain. Here, the peptide sequence profiles that were predicted for four PDZ complexes are represented with sequence logos. The name of the PDZ domain and the peptide sequence are indicated above, the peptide positions and PDB codes of the respective structures are given below the sequence logos. The first three peptide sequence profiles were predicted for PDZ3 of PSD-95 (PSD-95-3/3) of rat using three different complex structures. In theory, they should be identical, in practice, they vary substantially, even when very similar peptides were bound to the PDZ (second and third profile). The fourth profile had been predicted for the PDZ domain of AF6 (the structure of this complex is shown in Figure 3.1). The pictures of the sequence logos have been obtained from the web server ADAN [185].

In general, it is very difficult to assess the real performance of all these predictors to obtain an overview of the current state-of-the-art in PDZ interaction predictions. Most of the predictors indicate extremely high AUC and accuracy values [92, 176–178, 195] but only very few have been tested on real independent test data or have validated predictions with experiments (see Table 6.2) [9–11, 92, 178]. Even fewer predictors have been tested and applied by other researchers than the developers to obtain independent insights on their performance (see our results when testing the Chen predictor, chapter 9). PDZ interaction predictions are more likely to be successful when biologists and bioinformaticians closely collaborate thereby combining experience in interpretation of experimental data and its application for prediction.

Table 6.2. Published predictors for PDZ-peptide interaction specificities. D=Development of predictor. A=Application of predictor.

D/.	name	method	training data	test data/validation	findings/application	ref
D	iSPOT	contact matrix, 4 peptide and 23 PDZ positions	phage display for 7 human PATJ PDZs	published MPDZ PDZ interaction data	correct predictions but need more training data	[147]
D		analysis of variance	SPOT data with 6,223 human C-termini vs. 3 human PDZs, mutagenesis with SPOT	-	superbinders not necessarily specific, specificity of a PDZ highly depends on K_D cut-off	[81]
D	PDZ-DocScheme	MD with simulated annealing, SCWRL, CHARMM	-	7 PDZ holo structures, cross-docking, 5 homology models	RMSD < 2 Å for holo structures, < 3 Å for most cross-docking	[183]
D	-	co-expression data, localisation data, interologs, literature co-citations	-	human proteome and experimental validation	combination of different data sources reveals new PDZ-peptide interactions	[196]
D		PSSM based	(non)-interactions between 85 mouse PDZs and mouse peptides	85 mouse PDZs and 40 mouse peptides	TPR: 48 %, TNR: 88 %, continuous specificity space	[10]
D	Chen predictor	bayesian model, contact matrix with 38 position pairs	560 interactions, 1,167 non-interactions betw. 82 mouse PDZs and 93 mouse peptides, binary and affinity data	cross-validation, 126 new mouse interactions, worm and fly interactions	performance depends on whether extrapolation to new PDZs and/or peptides	[11]
D		PSSMs	phage display data of 54 human PDZs	published binders of DLG1 PDZs	8 of 11 binders of DLG1 PDZs predicted	[9]
D	-	MD (normal mode analysis (NMA), elastic network model (ENM)), linear discriminant analysis	18 PDZs with apo and holo structure, homology models	cross-validation	prediction of class I, II, dual spec. affiliation based on structural fluctuations of PDZs with TPR of 80 %	[68]

Continued on next page

Table 6.2 – continued from previous page

D/.	name	method	training data	test data	findings	ref
A	WaterMap	MD	-	published Erbin and HtrA-family PDZ data	correlation between predicted and measured binding affinities: corr. coeff. R^2: 0.67, energy contribution of water release upon binding important	[156]
D	-	MD with Monte Carlo	11 PDZ holo structures	9 PDZ holo structures with different peptides, PDZ apo structures	docking successful for holo PDZ structures, difficult for apo PDZ structures	[184]
D	Domain Interaction Footprint	machine learning with correlation-based feature selection	only peptide data (SPOT data from [81])	cross-validation	AUC=0.95, predictor only applicable to PDZs with known binders and non-binders	[176]
D	PREDIADAN	MD, FoldX, PSSMs	apo, holo, modelled structures of PDZs and other domain types	tests only on SH3 and SH2 data	prediction of PSSMs based on *in silico* alanine scanning and energy calculations, prediction of superbinders, proteome scan	[185]
D	-	MD (NMA, all atom minimisation, Rosetta docking, DrugScore affinity calculation)	4 PDZ holo structures	published interaction data for PICK1 PDZ	very complicated procedure but successful binding affinity predictions	[186]
D	-	Rosetta docking, multiple linear regression	published ITC data for PDZ3 of PSD-95	cross-validation	corr. coeff. $R^2 = 0.66$ for calculated vs. measured $\Delta\Delta G$, 0.6 for $\Delta\Delta H$, 0.17 for $\Delta\Delta S$, AUC=0.78 for correct classification prediction	[187]
D	FlexPepDock	Rosetta	14 complexes from peptiDB dataset	89 complexes from peptiDB dataset (many non-PDZ complexes)	80 % and 65 % of holo and apo complexes, respectively, with RMSD < 2 Å	[188]
D	-	triplet residue frequencies, clustering of amino acids, random forest classifiers	dataset of [10]	dataset of [10]	accuracy of 80 % for correct affiliation of PDZs to class I, II, and dual specificity	[177]

Continued on next page

Table 6.2 – continued from previous page

D/.	name	method	training data	test data	findings	ref
D	-	Rosetta, Monte Carlo	-	data from [9], structures of 17 PDZs, Erbin PDZ mutant data	for half of all peptide positions (85) good prediction of PSSM	[189]
D	-	support vector machine (SVM), contact matrix of 38 position pairs	genomic-like phage display data of [9], data of [10], generated negative interactions	proteome scanning and comparison with published PDZ interactions	prediction performance depends on similarity of PDZ to any training PDZ, TPR between 25 and 85 %, PSSM similarity 67 %	[181]
D	SemiSVM	modified support vector regression	data of [10]	cross-validation, exp. validation of predictions for PDZ3 of SCRIB	80 % accuracy, correlation with affinity data: R^2=0.92, negative training data and PDZ sequence similarity important for prediction performances	[178]
D	-	PSSMs, linear regression	phage display with mutated Erbin PDZ	DREAM4 prediction challenge	winner of DREAM4, trained on Erbin PDZ to predict Erbin PDZ specificities, Chen predictor performance poor	[182]
A	Chen predictor	-	-	human PDZome, human C-terminome	intensive overlap in ligand space of PDZs	[56]
D	-	mixture of PSSMs	phage display data for PDZs of [9] and for SH3 and WW domains	SPOT data from [81] and [172], PDZbase	multiple PSSMs perform better than single PSSMs, AUC=0.85	[92]
D	-	MD (perturbation response scanning, ENM)	-	structural and affinity data for PDZ3 of PSD-95 and PDZ2 of hPTP1E	the 2 PDZs have different allosteric pathways, α3 helix of PDZ3 linked to binding site via allosteric pathway and influences binding affinity	[197]
D	DomPep	PSSMs, PDZ sequence similarities	data of [9], data of [10]	measured binding affinities for PDZs of SCRIB	AUC=0.88 %	[195]

Continued on next page

Table 6.2 – continued from previous page

D/.	name	method	training data	test data	findings	ref
D	-	PSSMs, information theory, covariation	-	data of [9], phage display with mutated Erbin PDZ	covariation scores correlate (weakly) with spacial proximity of PDZ and peptide residues, prediction of PSSMs	[179]
D	-	graph invariants for sequence representation, neural networks	PDZbase [198] and data of [9], interaction data from PDB	DREAM4 prediction challenge	second of DREAM4	[180]

Part II.

Results and discussion

7. Solution structure of PDZ2 of MAGI1 bound to a C-terminal peptide derived from HPV16 E6

7.1. Summary

The protein E6 from "high-risk" HPV strains possesses a PBM that has been shown to bind to PDZ2 (out of 6 PDZs) of mammalian MAGI1. This interaction seems to play a role in HPV-mediated tumourigenesis. Previous studies by our group revealed the importance of N- and C-terminal extensions for obtaining monomeric and stable constructs of PDZ2 in solution [199]. The structural details for this observation could not be explained with a published crystal structure of PDZ2 in complex with a C-terminal peptide derived from HPV18 E6 [71].
Approach: We have determined the structure of apo PDZ2 and PDZ2 bound to an 11 residue-long C-terminal peptide derived from HPV16 E6 using NMR. The dynamic responses of PDZ2 upon peptide binding have been analysed by NMR experiments. SPR has been used to determine dissociation constants of wild type and mutant PDZ2 versus the E6 C-terminal peptide.
Findings: We confirmed previously observed ionic interactions between residues upstream of the core PBM of E6 and the β2-β3 loop of PDZ2. PDZ2 has a structured N-terminal and an unstructured C-terminal extension. The N-terminal extension seems to shield from solvent a hydrophobic patch of residues on PDZ2 that comprises a cysteine residue, thereby preventing the formation of soluble aggregates via non-native intermolecular disulfide bonds. The C-terminal extension adopts more restricted conformations upon peptide binding, probably due to atomic contacts that have been observed between residues of the C-terminal extension and R-5 and T-6 of the bound peptide. Mutation of the peptide contacting residues in the C-terminal extension reduced the binding affinity towards the E6 peptide. The hydrogen-bonding pattern of PDZ2 is altered upon peptide binding.
Discussion/Conclusions: Evidence has been provided that the observed dimer in a previously published crystal structure of a PDZ2-E6 complex is likely to be an artefact. The C-terminal extension of PDZ2 plays a role for peptide binding within the fragment context. It remains unknown whether this effect will be similar or completely different in full length MAGI1. Global effects on the hydrogen bonding network upon peptide binding suggests the involvement of distal sites of PDZ2 in peptide binding.
Contribution: I have read and curated the published literature to identify all PDZ domains that have been shown to interact with E6. Results of this analysis have been

used to create the sequence alignment shown in Figure 1a. I have provided information on other structures of PDZ domains with extensions that has been useful for the discussion.

doi:10.1016/j.jmb.2011.01.015 J. Mol. Biol. (2011) 406, 745–763

Contents lists available at www.sciencedirect.com

Journal of Molecular Biology

journal homepage: http://ees.elsevier.com.jmb

The Structural and Dynamic Response of MAGI-1 PDZ1 with Noncanonical Domain Boundaries to the Binding of Human Papillomavirus E6

Sebastian Charbonnier[1], Yves Nominé[1], Juan Ramírez[2], Katja Luck[1], Anne Chapelle[1], Roland H. Stote[2], Gilles Travé[1], Bruno Kieffer[2]* and R. Andrew Atkinson[3]

[1]*Equipe Oncoprotéines, Ecole Supérieure de Biotechnologie de Strasbourg, Boulevard Sébastien Brant, BP 10413, 67412 Illkirch Cedex, France*
[2]*Département de Biologie et de Génomique Structurales, Institut de Génétique et de Biologie Moléculaire et Cellulaire (Université de Strasbourg, CNRS UMR 7104, INSERM U964), 1 rue Laurent Fries, 67404 Illkirch, France*
[3]*Randall Division of Cell and Molecular Biophysics, King's College London, New Hunt's House, Guy's Campus, London SE1 1UL, UK*

Received 19 October 2010; received in revised form 6 January 2011; accepted 7 January 2011
Available online 13 January 2011

Edited by A. G. Palmer III

Keywords:
PDZ domain;
HPV;
E6 oncoprotein;
MAGI-1 PDZ1;
protein dynamics

PDZ domains are protein interaction domains that are found in cytoplasmic proteins involved in signaling pathways and subcellular transport. Their roles in the control of cell growth, cell polarity, and cell adhesion in response to cell contact render this family of proteins targets during the development of cancer. Targeting of these network hubs by the oncoprotein E6 of "high-risk" human papillomaviruses (HPVs) serves to effect the efficient disruption of cellular processes. Using NMR, we have solved the three-dimensional solution structure of an extended construct of the second PDZ domain of MAGI-1 (MAGI-1 PDZ1) alone and bound to a peptide derived from the C-terminus of HPV16 E6, and we have characterized the changes in backbone dynamics and hydrogen bonding that occur upon binding. The binding event induces quenching of high-frequency motions in the C-terminal tail of the PDZ domain, which contacts the peptide upstream of the canonical X-[T/S]-X-[L/V] binding motif. Mutations designed in the C-terminal flanking region of the PDZ domain resulted in a significant decrease in binding affinity for E6 peptides. This detailed analysis supports the notion of a global response of the PDZ domain to the binding event, with effects propagated to distal sites, and reveals unexpected roles for the sequences flanking the canonical PDZ domain boundaries.

Introduction

PDZ domains are protein–protein interaction domains that are commonly found in cytoplasmic proteins involved in signaling pathways or subcellular transport.[1–3] PDZ domains play roles in localizing proteins to the membrane and in acting as molecular scaffoldings or adaptors, but also serve in other functions such as binding to titin Z-repeats

*Corresponding author. E-mail address: kieffer@igbmc.fr.

Abbreviations used: HPV, human papillomavirus; NOE, nuclear Overhauser enhancement; BMRB, BioMagResBank; PDB, Protein Data Bank; SPR, surface plasmon resonance; TOCSY, total correlated spectroscopy; NOESY, NOE spectroscopy; GST, glutathione *S*-transferase; RU, response units; RDC, residual dipolar coupling.

in the Z-disk of sarcomeres[4] or detecting unfolded proteins and activating proteases.[5] Many PDZ-domain-containing proteins are located at the interface between the cytoskeleton and the cellular membrane, where they are implicated in the formation of cellular junctions such as synapses,[6] adherens junctions, or tight junctions.[7] They commonly form multiprotein complexes at the inner interface of the membrane and are associated with the control of cell growth, cell polarity, and cell adhesion in response to cell contact and thus represent a family of proteins targeted during the development of cancer.[8] Multiple PDZ domains are often found within a single protein connected by linkers of varying lengths, and found to be associated with other domains such as WW and SH3 domains, reflecting multiple functions. This further suggests that the composite protein function may be more than the sum of individual domains.

Viral proteins often target PDZ domains, resulting in the disruption of cellular processes for the benefit of the viral life cycle. The targeting of PDZ-domain-containing proteins involved in cellular adhesion and control of polarity has been shown to be a highly important activity in the process of cancer development following infection by "high-risk" human papillomaviruses (HPVs).[9–11] The expression of two HPV oncoproteins, E6 and E7, that cooperate in cell immortalization and transformation has been associated with tumorigenesis.[12,13] E6 has a number of distinct functions in the host cell. It targets the tumor suppressor p53 for degradation through the formation of a trimeric complex with the cellular ubiquitin ligase E6AP[14] and represses p53-dependent cell-cycle control and p53-dependent transcription by inhibiting p300-mediated acetylation.[15] In addition, E6 binds and sometimes drives the proteasome-mediated degradation (*via* PDZ domain recognition) of several PDZ-domain-containing proteins. These include various MAGUKs (*m*embrane-*a*ssociated *gu*anylate *k*inases), such as Dlg-1[16], Dlg-4,[17] and hScrib,[18] and MAGI (*m*embrane-*a*ssociated *g*uanylate kinase with *i*nverted domains) proteins, such as MAGI-1[19], MAGI-2, and MAGI-3[20] (Fig. 1a). The tumorigenic effects of HPV E6 depend partly on interfering with MAGI-1 functions in the living cell.[19] Several non-MAGUK proteins such as CAL,[21] MUPP-1[22], PATJ,[23] PTPN3,[24] Tip1,[25] and Tip2[26] are also targeted by E6. The interaction of high-risk HPV E6 proteins with PDZ domains is mediated by C-terminal peptide sequences matching the X-[T/S]-X-[L/V] motif of "class I" PDZ domains[27–29] and is associated with the development of cervical cancer.[29,30] This consensus sequence is only found in high-risk HPV E6s (such as HPV16 or HPV18) that differ in their C-terminal residues and, as a result, exhibit different affinities for MAGI-1.[8,31,32] A higher affinity seems to correlate with an increased likelihood of recurrence and metastasis in cervical tumors.[20,33] This variability in the C-terminal residue of HPV E6 proteins has been suggested to be important in fine-tuning the affinities of the viral oncoproteins for distinct sets of PDZ domains.[9]

Understanding how the E6 viral protein interferes with the endogenous network of interactions mediated by PDZ domains requires both knowing the general rules governing PDZ–peptide interactions and deciphering the specific strategies used by the virus during infection. In the past few years, many studies attempted to characterize the ligand binding "specificity code" of PDZ domains. Initial analysis of peptide library screens, combined with sequence analysis to identify C-terminal consensus binding sequences, led to the definition of three major PDZ domain classes.[2,34] Later, the mapping of 3100 peptides identified by phage display against 82 PDZ domains from worms and humans allowed a more precise analysis of binding specificity and resulted in a classification of PDZ domains into 16 distinct specificity classes.[35] A study of the binding selectivity of 157 PDZ domains from the mouse proteome using protein microarrays and quantitative fluorescence polarization[36] showed that selectivity is derived from interactions throughout the binding pocket. This led the authors to suggest that, in terms of binding selectivity, PDZ domains constitute a continuum rather than discrete classes.

At the molecular level, a large number of three-dimensional structures of PDZ domains have been determined,[37] with a number of them being available in both unliganded and liganded forms, allowing structural changes induced by peptide binding to be studied. In most cases, the structure of PDZ domains displayed a very small change upon peptide binding.[38,39] However, from a dynamic point of view, computational[40–45] approaches have highlighted the role of regions distal to the peptide binding site in forming dynamic networks within PDZ domains, and a number of studies have characterized changes in dynamics using NMR, thus providing experimental evidence for these networks.[39,46–48]

Initial structural insight into the targeting of PDZ domains by HPV E6 protein was provided by the crystal structures of a short peptide from HPV18 E6 bound to three PDZ domains from MAGI-1 and SAP97/Dlg.[49] This work revealed that peptide residues outside the canonical PDZ binding motif were involved in direct contacts with the canonical core regions of the PDZ domains. In addition, constructs of MAGI-1 PDZ1 [the second of the six PDZ domains in the sequence of human MAGI-1; residues 456–580 of human MAGI-1 (GenBank accession no. AF401656) when referring to the specific polypeptide used in these studies] were seen to form covalent cysteine-bridged dimers both in solution and in the crystal.

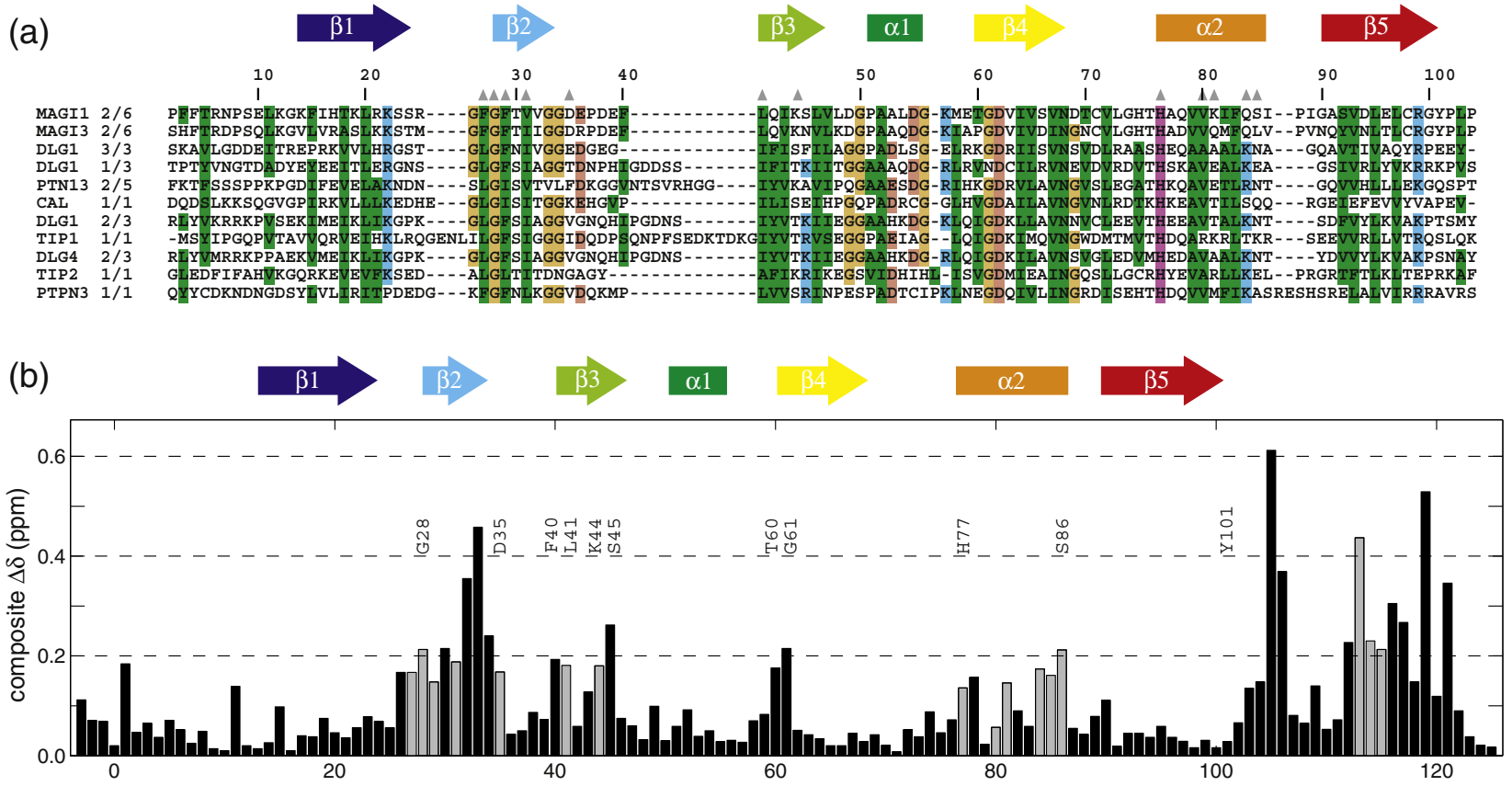


Fig. 1. (a) Multiple sequence alignment of PDZ domains known to bind E6 proteins from "high-risk" HPV types 16 and/or 18. For each domain, its position and the total number of PDZ domains contained in the protein are indicated. Conserved positive and negative charges are shown in blue and red, respectively. Conserved glycine and histidine residues are highlighted in orange and purple, respectively. Residues known or expected to make direct contact with the E6 peptide are indicated by a gray triangle. (b) Composite changes in chemical shift upon the binding of $16E6_{ct}$ L_0/V to MAGI-1 PDZ1 calculated for all assigned ^{1}H, ^{13}C, and ^{15}N resonances. Residues close to the E6 peptide are shown as gray rectangles. For residues Phe40 and Leu41, the major contribution arises from shifts of side-chain resonances. Colored arrows and boxes indicate elements of secondary structure, β-strands, and α-helices, respectively, inferred from chemical shifts and slowly exchanging $^{1}H_N$ resonances. The colors of secondary structure elements are retained in subsequent figures.

Table 1. Experimental restraints and statistics for the sets of the 20 final structures of MAGI-1 PDZ1 and MAGI-1 PDZ1/RSSRTRRETQV

	MAGI-1 PDZ1	MAGI-1 PDZ1/RSSRTRRETQV
Number of experimental restraints		
NOEs		
Intraresidue NOE	480	455
Sequential NOE	663	564
Medium-range NOE	296	207
Long-range NOE	687	543
Intermolecular NOE	—	49
Total NOE	2126	1818
Hydrogen bonds	39	57
TALOS-derived Φ/Ψ pairs	69	74
r.m.s.d. from experimental restraints		
Distance restraints (Å)	0.040 (0.001)	0.036 (0.001)
Dihedral-angle restraints (°)	0.71 (0.07)	0.78 (0.05)
r.m.s.d. from ideal covalent geometry		
Bonds (Å)	0.016 (0.001)	0.015 (0.001)
Angles (°)	1.89 (0.04)	1.73 (0.03)
Impropers (°)	2.15 (0.09)	1.91 (0.07)
Average pairwise r.m.s.d.		
Backbone atoms (residues 4–101) (Å)	0.44 (0.08)	0.58 (0.06)
Heavy atoms (residues 4–101) (Å)	1.12 (0.04)	1.16 (0.03)
Ramachandran plot		
Residues in the most favored regions (%)	77.3	78.0
Residues in additionally allowed regions (%)	18.5	18.3
Residues in generously allowed regions (%)	3.5	2.5
Residues in disallowed regions (%)	0.7	1.3
Equivalent resolution	2.5	2.4
Average pairwise fit between liganded and unliganded structures		
Backbone atoms (residues 4–101) (Å)		1.73 (0.09)

Recently, the extension of domain boundaries for the PDZ1 domain of MAGI-1 enabled us to study its properties in solution by NMR.[50,51] Using NMR, we solved the three-dimensional solution structures of MAGI-1 PDZ1 both unliganded and bound to a peptide derived from the C-terminus of HPV16 E6. In contrast to the crystal structure, we found that the domain was monomeric in both apo and holo forms. Furthermore, we found that peptide binding induced important changes in backbone dynamics and hydrogen bonding. While supporting the notion of a global response of the PDZ domain to the binding event, our study revealed an unexpected contribution to the binding process by sequences flanking the canonical PDZ domain boundaries.

Results

Changes in chemical shift upon E6 peptide binding

Initial studies with C-terminal peptides from HPV16 E6 showed that peptides representing the 11 C-terminal amino acids fully reproduced the spectral changes observed when using the C-terminal domain of HPV16 E6, while shorter C-terminal peptides did not,[50] suggesting the involvement of residues outside the canonical binding motif. Similar observations have been made in other studies.[34,52–54] We previously showed that the mutation of the C-terminal leucine residue into a valine resulted in a significant gain in affinity,[32] which proves beneficial for the detection of intermolecular nuclear Overhauser enhancements (NOEs). Our present studies were therefore performed using a chimeric peptide composed of the 11 C-terminal residues of HPV16 E6 (sequence RSSRTRRETQV, hereafter named $16E6_{ct}$ L_0/V†).

An essentially complete ^{1}H, ^{13}C, and ^{15}N resonance assignment was achieved for MAGI-1 PDZ1 alone and in complex with $16E6_{ct}$ L_0/V. Resonances of the unlabeled peptide could be identified in isotope-filtered experiments performed on the complex. Assignments were completed with the single exception of the side chain of Lys44, which remained unassigned beyond the C^{β} atom due to specific line broadening in both unliganded and liganded forms. Assignments have been deposited at the BioMa-

† The same 11-residue peptide in which the C-terminal leucine residue is replaced by valine (RSSRTRRETQV). Standard three-letter abbreviations for amino acids are used when referring to residues of MAGI-1 PDZ1 (except when space is limited in figures), and standard one-letter abbreviations are used when referring to residues of peptides. The $16E6_{ct}$ L_0/V peptide is numbered backwards from V-0 to R-10.

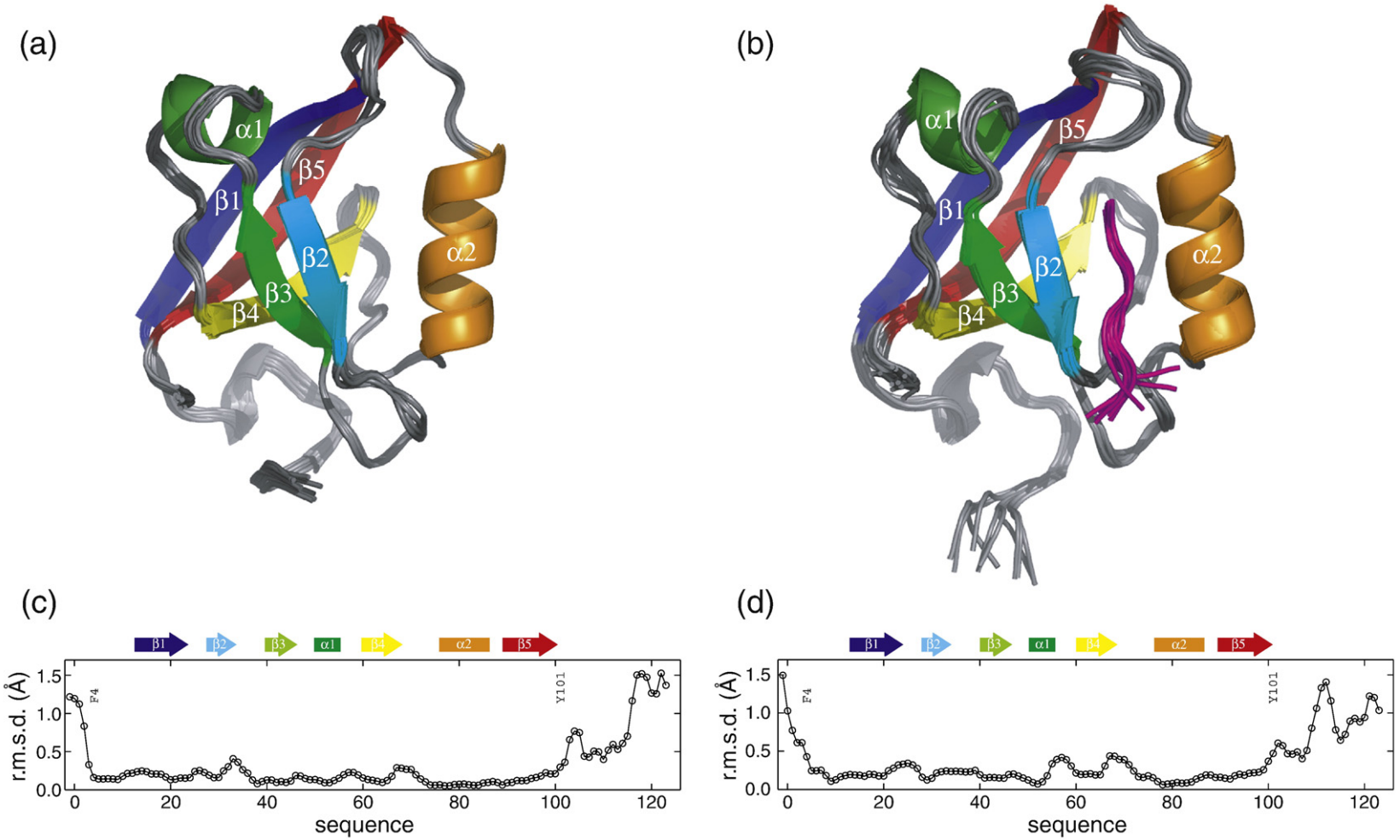


Fig. 2. Bundles of 10 representative structures calculated for (A) MAGI-1 PDZ1 and (b) MAGI-1 PDZ1/16E6$_{ct}$ L_0/V. Secondary structure elements are colored as in Fig. 1 and labeled. For clarity, residues preceding Phe4 and following Tyr101 are not shown. In (b), the backbone of the bound peptide is shown in magenta. A measure of local backbone definition is given by average pairwise r.m.s.d. values calculated over five-residue segments of the primary sequence for 20 calculated structures of (c) MAGI-1 PDZ1 and (d) MAGI-1 PDZ1/16E6$_{ct}$ L_0/V. Residues mentioned in the text are labeled.

gResBank (BMRB)[55] under BMRB ID 16558 (unliganded) and BMRB ID 16559 (liganded).

Two spectroscopic observations suggested a defined structure for the N-terminal portion of the polypeptide chain of MAGI-1 PDZ1 preceding the β1 strand: (i) the resonance of the Thr5 $^{1}H_{N}$ proton is shifted far upfield from the disordered value in both liganded and unliganded domains; and (ii) the $^{15}N^{\varepsilon}$ and $^{1}H^{\varepsilon}$ resonances of the guanidinium group of Arg6 are unexpectedly observed in ^{1}H–^{15}N correlation spectra.

The binding of $16E6_{ct}$ L_0/V to the PDZ domain results in significant spectral changes for a majority of resonances, suggesting a global response of the PDZ domain to peptide binding (Fig. S1; Fig. 1b). In addition to large composite chemical shift changes for residues close to the binding site, two supplementary sets of changes that were less expected are observed: (i) Phe40, Leu41, Lys44, Ser45, Thr60, and Gly61 form a small cluster of affected residues; and (ii) more than half of the residues in the C-terminal extension beyond Tyr101 experienced considerable shifts. This latter observation strongly suggests that this region undergoes a structural and/or dynamic change upon binding of the peptide.

Structures of MAGI-1 PDZ1 and MAGI-1 PDZ1/$16E6_{ct}$ L_0/V

Experimental data, including intramolecular and intermolecular NOEs, hydrogen bonds, and dihedral-angle restraints (Table 1), yielded well-defined structures for both the unliganded PDZ domain and its complex with $16E6_{ct}$ L_0/V (Fig. 2a and b), with no violations of distance restraints greater than 0.5 Å and with no violations of dihedral-angle restraints greater than 5°. Analysis of the distribution of Φ/Ψ angles for the 20 lowest-energy structures of each yielded good statistics, and the PROCHECK-NMR program[56] gave equivalent resolution values of 2.5 Å and 2.4 Å for the unliganded and liganded structures, respectively (Table 1). A high degree of local precision in both sets of structures is reflected in average pairwise r.m.s.d. values (calculated using a five-residue window) below 0.5 Å for all residues (Fig. 2c and d). The average pairwise r.m.s.d. values calculated over the entire backbone between Phe4 and Tyr101 are both close to 0.5 Å, while those calculated with all heavy atoms of the same sequence are both close to 1.1 Å (Table 1).

Both unliganded [Protein Data Bank (PDB) ID 2kpk] and liganded (PDB ID 2kpl) structures adopt the fold expected from studies of other PDZ domains,[38] which is composed of five β-strands and two α-helices (Fig. 2a). The structures are also very similar to the previously reported crystal structure of MAGI-1 bound to the seven last C-terminal residues of HPV18 E6[49] (Fig. S2a and b). In the present structures, the first β-strand (β1) begins at Lys14. Nonetheless, the preceding sequence (Phe4-Gly13) adopts a well-defined structure (Fig. 2a and b), and relaxation measurements confirm that no picoseconds-to-nanoseconds timescale motions (characteristic of unstructured polypeptides) affect this region. This folded N-terminal region effectively masks the side chain of Cys98 and a set of hydrophobic residues (Phe40, Val63, Val65, and Leu73) from the solvent. This observation explains why, in our hands, shorter constructs lacking this N-terminal extension were poorly soluble and that expression of a soluble monomeric form of this domain was only possible upon the addition of an N-terminal extension encompassing Phe3 and Phe4 residues.[50]

Superimposition of the backbones of the core regions of unliganded and liganded structures (between Phe4 and Tyr101) gave an average pairwise r.m.s.d. of 1.7 Å, which is larger than the value of 0.9 Å noted by Doyle *et al.* for PSD-95 PDZ3, but lower than the mean values calculated for other pairs of unliganded and liganded structures (e.g., PDB ID 1GM1/PDB ID 1VJ6: 1.9 Å; PDB ID 3PDZ/PDB ID 1D5G: 1.9 Å; PDB ID 2EV8/PDB ID 2EJY: 2.6 Å).[38]

While the fold of MAGI-1 PDZ1, including the local conformation of many loops, is clearly maintained on complex formation, slight rearrangements occur (Fig. 3a and c). Average pairwise r.m.s.d. values between the two sets of 20 structures, calculated using a five-residue window, reveal those portions of the sequence in which the local structure is altered (Fig. 3b). A small difference in the conformation of Lys12-Lys14 serves to reposition the preceding sequence to a slight degree, although the positions of the side chains are similar; the sequence Ser23-Phe29 (labeled GFGF loop in Fig. 3a) and the end of β2 change upon binding to form interactions with the C-terminal carboxylate and accommodate the peptide chain; the neighboring strand β3 and the loop connecting β3 to α1

Fig. 3. (a) Superposition of the representative structures of MAGI-1 PDZ1 (gray) and MAGI-1 PDZ1/$16E6_{ct}$ L_0/V (deep red), with the C-terminal six residues of the bound peptide shown as sticks color coded by atom type. Secondary structure elements are labeled. For clarity, residues preceding Asn7 and following Pro102 are not shown. (b) Variations in local conformation between MAGI-1 PDZ1 and MAGI-1 PDZ1/$16E6_{ct}$ L_0/V are shown by average pairwise r.m.s.d. values calculated over five-residue segments of the primary sequence between two sets of 20 calculated structures. Residues mentioned in the text are labeled. (c) Changes in average C^{α}–C^{α} distances upon binding greater than 2 Å are shown by blue (increased distance) and red (decreased distance) squares. Regions mentioned in the text are labeled.

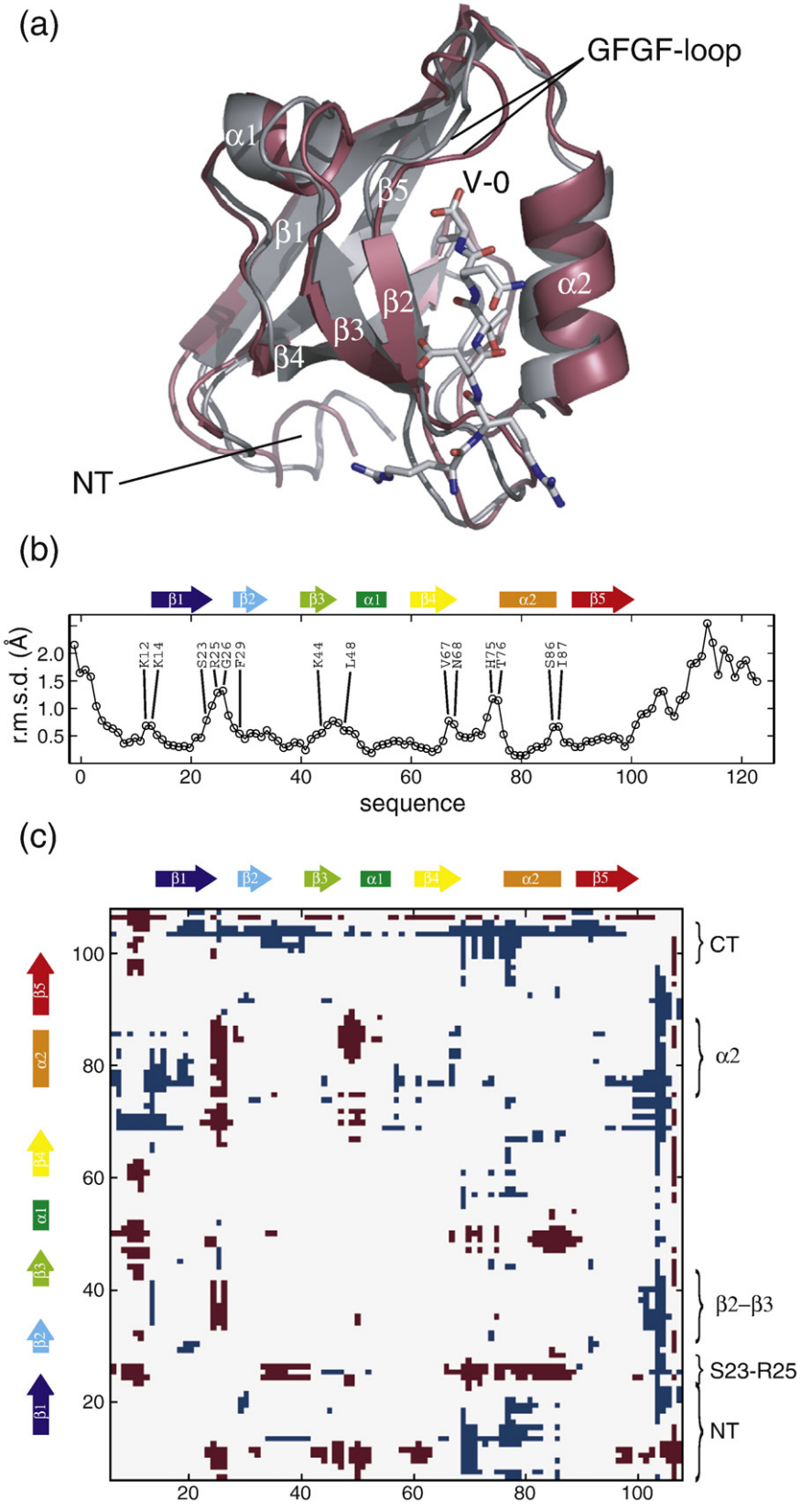


Fig. 3 (*legend on previous page*)

rearrange, and changes in the local conformation of residues at both ends of α2 allow the helix to move away from its position in the unliganded structure to accommodate the bound peptide (Fig. 3a). As a consequence of these changes, helix α2 and strand β2 diverge to accommodate the peptide, with the α-helix undergoing a reorientation of approximately 11°. An overview of global structural changes induced by E6 peptide binding is provided in Fig. 3c, where changes in the mean distances between C^{α} positions in the two sets of structures (apo and holo) are reported as a dot matrix. These distances increase between structural elements located at either side of the β2–β3 loop, whereas they decrease between residues Ser23 and Arg25, which precede the GFGF motif (Gly26-Phe29) and residues of both the β2–β3 loop and helix α2.

The solution structure of the MAGI-1 PDZ1/16E6$_{ct}$ L$_0$/V complex exhibits an extended set of interactions

As in the pioneering work of Doyle *et al.* on PSD-95 PDZ3, the interacting peptide binds in a mode expected for a class I PDZ domain, that is, by β-sheet augmentation with the peptide C-terminal 4 residues running anti-parallel with β2 in the groove between β2 and α2 (Figs. 2b and 3a).[38] The peptide C-terminal carboxylate group interacts with the backbone of the GFGF motif, with additional hydrogen bonds formed between the backbone of β2 and the peptide (Fig. 4). The side chain of the C-terminal valine residue (V-0) of the 16E6$_{ct}$ L$_0$/V peptide lies in a hydrophobic pocket lined by three phenylalanine residues (Phe27, Phe29, and Phe84) and a valine residue (Val31). The side-chain amide group of Gln85 is well positioned to form a hydrogen bond to the backbone carbonyl of Q-1, consistent with changes in chemical shift observed for the nuclei of this residue. The methyl group of T-2 lies in a second hydrophobic pocket formed by Val31, Leu41, Val80, and Val81, while its $O^{\gamma}H$ group lies close to the side chain of His77, thus allowing formation of the hydrogen bond that confers specificity of class I PDZ domains for X-[T/S]-X-[L/V] motifs. In addition to the favorable electrostatic interactions between R-4 and Asp35 observed in the crystal structure,[49] the solution structure shows an additional interaction between E-3 and Lys44. Surface plasmon resonance (SPR) binding measurements performed on several mutants of 16E6$_{ct}$ [an 11-residue peptide derived from the C-terminal sequence of oncoprotein E6 from HPV16 (RSSRTRRETQL)] and on the Lys44/Glu mutant of

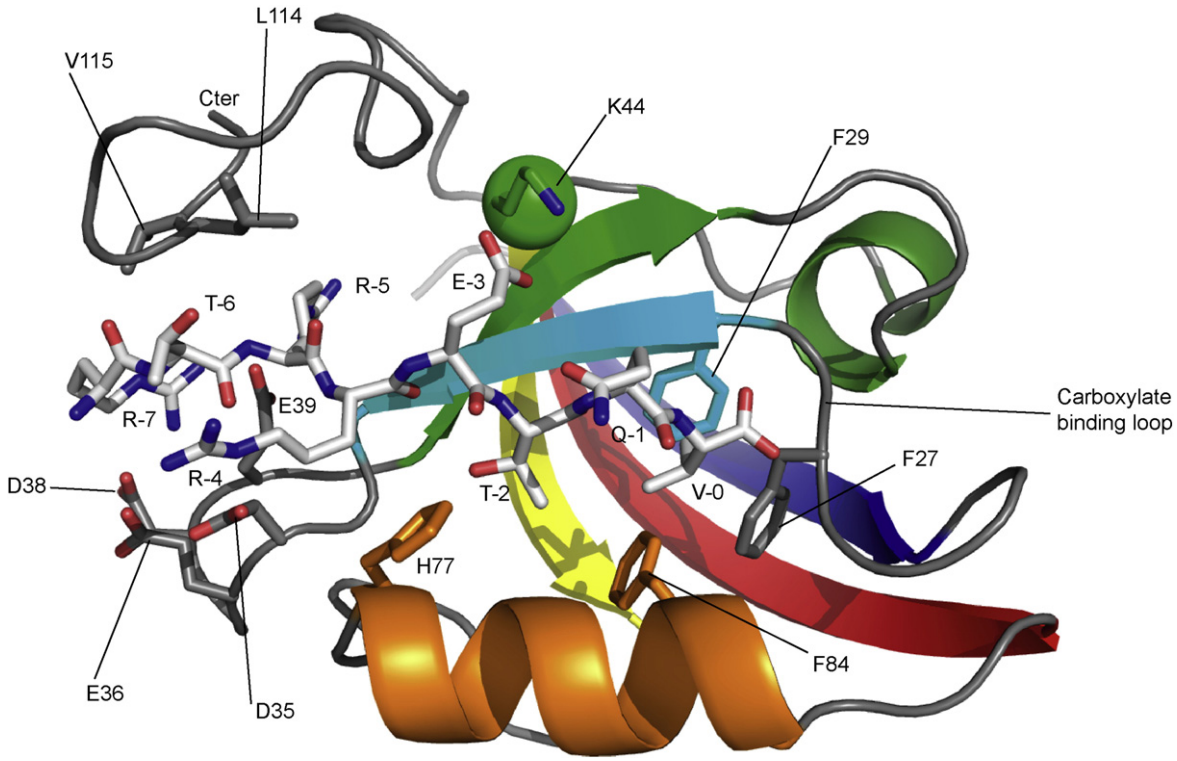


Fig. 4. Schematic representation of the binding of 16E6$_{ct}$ L$_0$/V to MAGI-1 PDZ1 using the lowest-energy model of the ensemble. Only the last seven residues of the 16E6$_{ct}$ L$_0$/V peptide are shown. Most of the MAGI-1 PDZ1 residues involved in canonical interactions—with the exception of V31, L41, V80, and V81, which have been omitted for the sake of clarity—are represented. The position of the C^{α} atom of K44 is indicated with a green sphere.

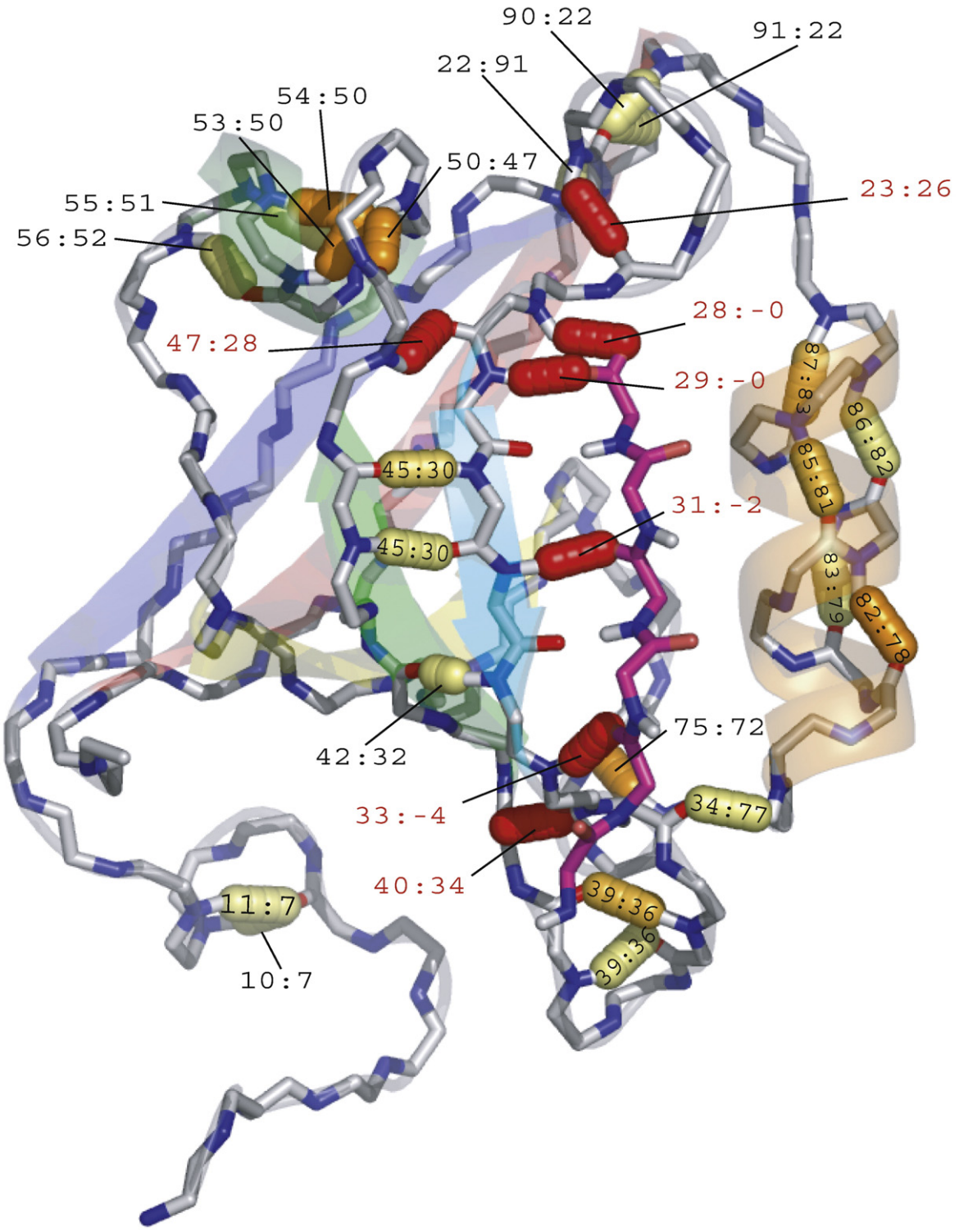


Fig. 5. Changes in hydrogen/deuterium exchange rates upon the binding of $16E6_{ct}$ L_0/V to MAGI-1 PDZ1, mapped onto the tertiary structure of the complex as colored tubes. Decreases in exchange rates (calculated from the values in Table S1) ranged from 0.0 h^{-1} to 2.7 h^{-1}. Values were placed in one of 14 bins of width 0.2 h^{-1} and colored using a linear RGB color scale from white to red. White bars correspond to the smallest detectable changes in exchange rate, while red bars correspond to the largest changes: H_N atoms that exchanged rapidly in the absence of peptide but extremely slowly in the presence of peptide. Hydrogen bonds are labeled with the residue numbers of the donor and acceptors. Where space does not allow labels to be placed on the bars, the first number denotes the donor.

MAGI-1 PDZ1 demonstrated the importance of these electrostatic interactions to binding affinity.[32]

Of particular interest here is a set of intermolecular NOEs between Ser113, Leu114, and Val115 of the PDZ domain, and R-5 and T-6 of the bound peptide, which provide direct experimental evidence for the interaction of the C-terminal extension of MAGI-1 PDZ1 with residues outside the ETQV motif of the bound peptide. This provides a rationale for the differences in heteronuclear single-quantum coherence spectra observed with peptides of different lengths.[50]

Global response to peptide binding involves the hydrogen-bond network

The effect of E6 peptide binding to the PDZ domain was further investigated by measuring the hydrogen/deuterium exchange rates for both the unliganded form and the peptide-bound form of the protein. H_N atoms that are buried within a tertiary fold and/or participate in hydrogen bonding will exchange more slowly with solvent than those in exposed loops, since the probability of accessing an exchange-competent "open" state is reduced in these sites. Amide proton exchange rates therefore provide a measure of overall or local stability on a timescale that is distinct from that upon which NOEs develop. While NOEs may define similar average structures, large differences in exchange rates may be observed under conditions where opening rates are altered. We can define three groups of residues: (i) those that exchange too rapidly for a rate to be measured: no cross-peak is detected at the first time point; (ii) those that exchange too slowly for a rate to be measured: cross-peak intensity does not decrease over the time course of the experiment (more than 20 h); and (iii) those for which a rate can be determined: cross-peak intensity decays over the time course of the experiment. Hydrogen-bond acceptors were identified from initial sets of calculated structures and used to define distance restraints when identification was unambiguous.

In unliganded MAGI-1 PDZ1, a set of H_N atoms that exchange slowly defines the secondary structure of the domain (Table S1). While β1 and β5 pair in a conventional anti-parallel manner, the central strands pair in a rather irregular way: thus, Glu96 on β5 hydrogen bonds to both Val65 and Ser66, resulting in an offset in register in the hydrogen-bonding pattern. Similarly, Thr30 on β2 hydrogen bonds to both Lys44 and Ser45. A second set of slowly exchanging H_N atoms defines turns at key points in the structure (such as Glu36 and Glu39 in the β2–β3 loop, and Thr70 and Val72 in the β4–α2 loop). Interestingly, the hydrogen-bond acceptor for the slowly exchanging H_N atom of Glu59 is a side-chain oxygen atom of Asp62.

This set of structure-defining hydrogen bonds is also found in the complex of MAGI-1 PDZ1 with $16E6_{ct}$ L_0/V, yet a number of rates are decreased, and a set of additional sites is seen to be protected from exchange (Table S1; Fig. 5). A number of these changes result directly from the canonical binding of the peptide to the PDZ domain, such as those observed for residues Gly28, Phe29, Val31, and Gly33. The H_N atom of Ser23 is protected from exchange and forms a hydrogen bond with the backbone carboxyl of Gly26, stabilizing the altered conformation of the carboxylate-binding loop. Elsewhere, in the β2–β3 loop, the rates of exchange for the pair of hydrogen bonds between Glu36 and Glu39 are slowed upon binding, as was that of the H_N atom of Ser45 in the rather irregular β2–β3 pair of strands. The formation of a (weak) hydrogen bond to Thr30 may contribute to stabilizing the altered local conformation of β3 and the β3–α1 loop. Similarly, increased protection from exchange for the H_N atom of His75, which forms a hydrogen bond to Val72, correlates with a change in local conformation in the turn preceding α2. Remarkably, exchange rates in both helices are perturbed upon binding. For residues in α2, exchange rates are slowed upon binding, suggesting a tightening of the helix upon formation of the complex. The behavior of the α1 helix is of particular interest: the fold of this portion of the sequence is well defined in both unliganded and liganded structures, although the backbone H_N atoms were observed to exchange rapidly with solvent in the absence of bound peptide. Upon binding, the H_N atoms of Ala52, Ala53, Leu54, and, to a lesser extent, Asp55 and Gly56 exchange slowly enough for rates to be measured. This is a rather remarkable effect at a site quite remote from the peptide binding site, yet the slowed exchange rates of two further H_N atoms link the two sites: Gly50 forms a hydrogen bond to Val47, which in turn forms a hydrogen bond to Gly28. Identification of affected exchange rates defines a network of interactions leading from the binding groove to distal sites.

Changes in the dynamic properties of MAGI-1 PDZ1 upon binding

Changes in the amplitudes and timescales of molecular motions may be of functional significance for interactions where these affect thermodynamic parameters. For example, the quenching of picoseconds-to-nanoseconds timescale motions is associated with an unfavorable loss of conformational entropy. Since ^{15}N relaxation rates provide sensitive probes of changes in the dynamic properties of the protein backbone, we measured ^{15}N R_1 and R_2 rates and the heteronuclear ^1H–^{15}N NOE for MAGI-1 PDZ1 in the absence and in the presence of $16E6_{ct}$ L_0/V and derived values of the spectral density

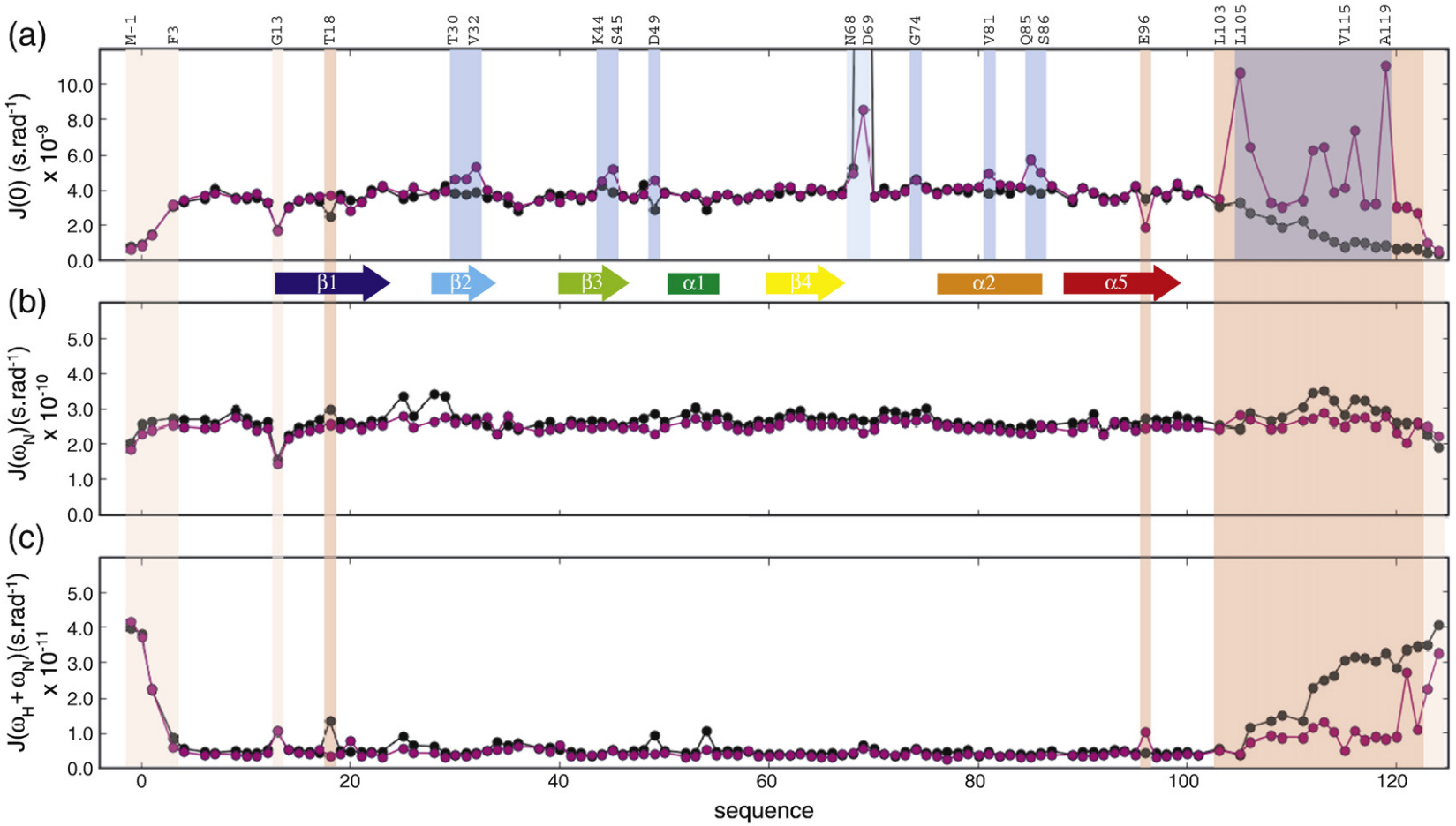


Fig. 6. Spectral density values $J(\omega)$ at frequencies 0 (a), ω_N (b), and $\omega_H+\omega_N$ (c), plotted against sequence. $J(\omega)$ values were determined from the relaxation measurements for backbone ^{15}N nuclei in MAGI-1 PDZ1 (black) and MAGI-1 PDZ1/16E6$_{ct}$ L_0/V (magenta). For clarity, the vertical scale on the upper panel has been adjusted to truncate the value for Asp69, which is affected by conformational exchange. Residues mentioned in the text are labeled above the plots. Light-pink shading indicates residues affected by picoseconds-to-nanoseconds timescale motions in both forms, whereas dark-pink shading indicates residues affected only in one form. Light-blue shading indicates residues affected by microseconds-to-milliseconds timescale conformational exchange in both forms, whereas dark-blue shading indicates residues affected only in the liganded form.

function $J(\omega)$ at frequencies 0, ω_N, and $\omega_H+\omega_N$ (Fig. 6).

Overall rotation

Spectral density values for the core portion of the unliganded domain define a baseline value that reflects overall tumbling in solution (Fig. 6a). The trimmed mean value of $J(0)$ is 3.8×10^{-9} s rad^{-1} and is unchanged upon complex formation. Further insight into the rotational diffusion properties of the two forms was gained by fitting an anisotropic diffusion tensor to the relaxation data using the TENSOR2 program.[57] Similar values of isotropic correlation times were found for the unliganded (9.4 ± 0.1 ns) and liganded (9.7 ± 0.1 ns) proteins. This increase in correlation time is less than that expected from considering the increased molecular weight of the complex (expected value, 10.2 ns), which suggests a more compact form for the complex. Moreover, whereas the eigenvectors of the anisotropic diffusion tensor coincided poorly with the inertia tensor eigenvectors, a strikingly good superposition was observed for the complex (the main axis of the diffusion tensor deviates from that of the inertia tensor by 11°) (data not shown).

Picoseconds-to-nanoseconds timescale motions

For the unliganded domain, values of the spectral density function depart steeply from baseline values (in a manner typical of unrestrained loops and termini) for the N-terminal residues Met1 and Gly0 that result from the cloning strategy, Lys1, and, to a lesser extent, Phe3 (Fig. 6a). This reflects considerable amplitudes of picoseconds-to-nanoseconds timescale motions in the N-terminus. At the C-terminus, on the other hand, the decrease in $J(0)$ (corresponding to a decrease in the order parameter and an increase in the amplitude of picoseconds-to-nanoseconds timescale motions) is quite gradual from Leu103 onwards, attaining values typical of unrestrained polypeptides only around Val115. This suggests a certain degree of restriction in the movement of the chain beyond the end of the last β-strand, which may be due, in part, to proline residues. It should be noted that there was no evidence in the NMR spectra for cis conformations of proline residues nor for cis–trans isomerization. The $J(\omega_H+\omega_N)$ values give a clearer indication of the behavior of the C-terminal 25 residues: uncorrupted by exchange contributions, the slight rise in $J(\omega_H+\omega_N)$ values indicates a limited degree of increased mobility on this timescale (with respect to the core of the domain) for residues beyond Phe105 (Fig. 6c).

Upon binding, there is little change to the motions of the N-terminal residues. At the C-terminus, however, the effect of peptide binding is dramatic: $J(\omega_H+\omega_N)$ values display a significant decrease up to residue 120, while the gradual decrease in $J(0)$ is no longer observed. These observations indicate that picoseconds-to-nanoseconds timescale motions are restricted considerably in this region upon binding, albeit not to the point of defining a single conformation. Subtle changes in picoseconds-to-nanoseconds timescale motions are also observed within the core of the PDZ domain: Thr18 in the center of the β1 strand appears mobile only in the unliganded form, while Glu96 on the opposite β5 strand appears mobile only in the complex.

Microseconds-to-milliseconds timescale conformational exchange

Exchange processes on the microseconds-to-milliseconds timescale cause increases only in $J(0)$ values. While little evidence of contributions from conformational exchange is observed for the unliganded form (with the exception of Asn68 and Asp69), increases in $J(0)$ are evidenced for a small set of residues of the complex (Thr30-Val32, Lys44, Ser45, Asp49, Val81, Gln85, and Ser86), reaching from the peptide binding site across β2 and β3 to the region of α1. A set of C-terminal residues between Phe105 and Ala119 is also affected by conformational exchange, possibly resulting from a defined but transient interaction between the C-terminus and both the peptide and the core domain.

Arginine side chains

Relaxation data were obtained for the N^{ε} nuclei of Arg6 and Arg99 in both unliganded and complexed forms (data not shown). In both cases, the data for Arg99 indicate that the side-chain motions on the picoseconds-to-nanoseconds timescale are essentially unrestricted, while those for Arg6 lie close to the baseline defined by residues of the core of the domain. This is consistent with the observed structured N-terminal portion of the sequence that precedes the canonical PDZ domain and packs tightly against it.

Mutations in C-terminal extension affect peptide binding

In order to probe the possible contributions of the C-terminal PDZ domain extension to peptide binding, we mutated the three residues that display intermolecular NOEs with the peptide: Ser113, Leu114, and Val115 into Arg, Lys, and Arg residues, respectively (S113R-L114K-V115R, hereafter named RKR), or into three Gly residues (S113G-L114G-V115G, hereafter named GGG). The affinities of both mutants for either $16E6_{ct}$ or $16E6_{ct}$ L_0/V peptides were measured using SPR experiments, as previously described.[32] MAGI-1 PDZ1 injection at

concentrations ranging from 50 nM to 10 μM displayed a dose-dependent signal increase (Fig. 7a). The sensorgrams reach a steady-state response (R_{eq}) during the association phase, allowing us to evaluate K_d values by plotting the variations of R_{eq} as a function of the injected analyte concentration (Fig. 7b).

Both RKR and GGG mutations led to a significant reduction of binding affinity, with the largest effect being a fivefold reduction in K_d for the GGG mutation (Fig. 7b and c). This effect is of the same order of magnitude as that previously observed for a mutation altering Arg44,[32] demonstrating unambiguously the contribution of the C-terminal extension. Notably, this effect is also observed for the wild-type 16E6$_{ct}$ peptide, albeit with a reduced amplitude (2.5-fold affinity reduction). The RKR mutation led to a threefold and twofold reduction of the affinity for the 16E6$_{ct}$ L_0/V and 16E6$_{ct}$ peptides, respectively. Independent SPR measurements performed for the same pairs of analyte/ligand interaction showed a reproducibility of K_d determination with less than 10% error, indicating that the observed changes are significant.

Discussion

PDZ domains represent one of the largest families of protein–protein recognition domains. As such, in recent years, they have attracted considerable interest in studies aimed at deciphering recognition specificity mechanisms at the molecular level. Our study of the MAGI-1 PDZ1 domain addresses the specific case of an interaction between a PDZ domain and the HPV E6 viral protein that targets several PDZ domains during the infection phase.[19,58] By comparing NMR measurements on liganded and unliganded MAGI-1 PDZ1, we were able to monitor the global response of a PDZ domain to viral peptide binding in solution. Previous crystallographic studies of MAGI-1 PDZ1 domain and other PDZ domains bound to C-terminal peptides from high-risk genital HPV E6[49,59] have already provided insight into the interaction between the oncoprotein E6 and PDZ domains.

The core PDZ domain region of the solution structure of the MAGI-1 PDZ1 domain in complex with the 11 C-terminal residues of the HPV16 E6 peptide could be superimposed on the crystal

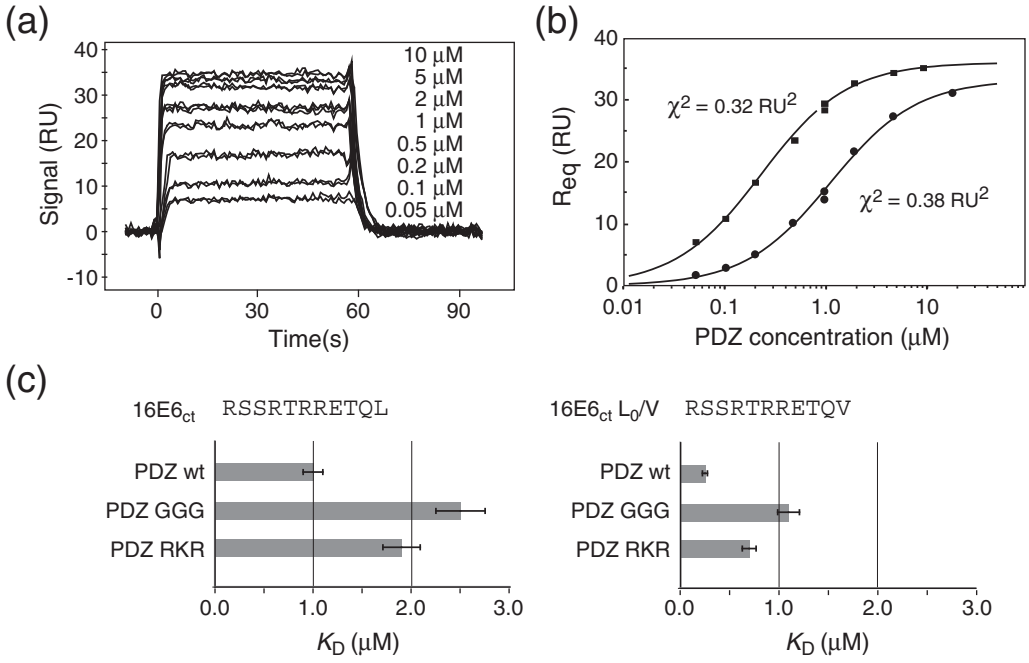


Fig. 7. Steady-state analysis by SPR of the binding of HPV16 E6 peptides to MAGI-1 PDZ1. (a) Representative sensorgrams resulting from GST–16E6$_{ct}$ L_0/V peptide interacting with MAGI-1 PDZ1 injected at different concentrations. The figure displays the superimposition of two independent sets of measurements. The binding curves show the SPR signal (RU) as a function of time. Note the rapid attainment of equilibrium, prohibiting kinetic analysis. (b) Steady-state analysis of the 16E6$_{ct}$ L_0/V peptide/MAGI-1 PDZ1 interaction. Equilibrium responses (R_{eq}) extracted from (a) were plotted as a function of total PDZ1 concentration and fitted with a 1:1 binding model. K_d values of 0.23 ± 0.02 μM and 1.17 ± 0.12 μM were obtained for wild-type MAGI-1 PDZ1 (squares) and PDZ1 GGG mutant (circles), respectively. R_{max} values of 36.4 ± 0.4 RU and 33.6 ± 0.6 RU were found for the wild type and the mutant, respectively. (c) Comparative plot of K_d values obtained for GST–16E6$_{ct}$ and GST–16E6$_{ct}$ L_0/V peptides in contact with wild-type MAGI-1 PDZ1 or mutants. Values are expressed as the arithmetic mean of at least two independent experiments. Error bars indicate a 10% variation.

structure with a backbone r.m.s.d. value of 1.6 Å (Fig. S2a and b). Nevertheless, slight local differences are observed between crystal and solution structures, mostly located in the carboxylate-binding loop. In addition, the orientation of α2 helix differs slightly between the two structures. These rearrangements in solution structure, with respect to that observed in crystallographic studies, were confirmed by the refinement of the structure of the complex using residual dipolar couplings (RDC) measured in samples oriented in strained polyacrylamide gels. These differences may arise from the shorter construct used in the crystal structure, covalent dimerization in the crystal, different peptide sequences, or the effects of crystal packing.

In our earlier study,[50] we found the canonical boundaries of MAGI-1 PDZ1 to be unsuitable for the production of a folded monodisperse unliganded protein. A C-terminal extension of 25 amino acids was required to avoid aggregation during purification, and an additional N-terminal extension of 12 amino acids was needed to remove all traces of heterogeneity from the NMR spectra. From the structural data presented here, it is apparent that the N-terminal extension folds into a well-defined structure at one end of the β-sheet formed by β1, β5, and β4. This folded portion of the N-terminus, involving residues Phe4-Phe15, shields a set of hydrophobic residues (Phe40, Val63, Val65, and Leu73) from the solvent (Fig. 5). It also shields the side chain of Cys98, while Cys71 remains exposed. In the X-ray study of Zhang *et al.*, these hydrophobic residues form a dimer interface, and Cys98 cross-links to Cys71 on the opposite monomer to form the covalent dimer.[49] Our observations strongly suggest that the dimer could not be formed by MAGI-1 PDZ1—as studied here, nor, in all likelihood, *in situ*—but perhaps served to stabilize an otherwise unstable construct during crystallization. We conclude that, in the case of MAGI-1 PDZ1, the N-terminal extension folds to extend the PDZ domain. It is of interest to note that the structure of MAGI-2 PDZ1 (also called atrophin-1 interacting protein 1), which shares a considerable degree of homology with MAGI-1 PDZ1, was also solved by NMR (in the absence of ligand) with N-terminal and C-terminal extensions similar to those used in this study (Zhao *et al.*, unpublished; PDB ID 1UEQ). In both PDZ domains, the N-terminal extension adopts a well-defined fold in a similar position with respect to the core of the PDZ domain (Fig. S3). Few examples of PDZ domains with extended and folded N-terminal tails have been reported in the PDB. Indeed, the N-terminal extensions of several PDZ domains, such as those of the InaD-like protein (PDB IDs 2DB5 and 2DAZ) and the MUPP1 protein (PDB IDs 2O2T and 2IWO), adopt a helical fold that is stabilized by hydrophobic interactions with the surface of the PDZ domain. This observation, which has not yet been reported in the literature, cannot be generalized for all PDZ domains. However, it clearly indicates that the flanking regions of some PDZ domains have been tailored to achieve specific conformational properties at their edges.

The C-terminal extension of MAGI-1 PDZ1 does not adopt a well-defined structure in the absence (or indeed in the presence) of ligand. Experimental evidence shows that this 25-residue tail tends slowly towards greater disorder in the unliganded domain, as expected for an unstructured polypeptide chain attached to a structured domain. Nonetheless, the relaxation data indicate that this decorrelation of the tail occurs rather more progressively than might be expected if it was truly unrestricted (compare N-terminal with C-terminal $J(\omega_H+\omega_N)$ profiles in Fig. 6c). This may result, in part, from the high number of proline residues between Pro102 and Pro110 (four prolines). However, considering the unusual spectral properties of Lys44 and the position of this side chain on the face of the domain (Fig. S4), we suggest that interactions between the positively charged side chain of Lys44 and the negatively charged side chains in the proline-rich sequence between Asp106 and Asp109 may also contribute to restricting the motions and hence the conformations of the C-terminal extension. Indeed, the mutation of Lys44 into an aspartate residue affects the chemical shifts of several residues of the C-terminal extension of the unliganded MAGI-1 PDZ1 domain (data not shown).

The local conformation of the PDZ domain is altered only in a small number of key sites upon peptide binding. The slight changes observed here in the carboxylate-binding loop, the C-terminal end of β2, the β3–α1 loop, and the sequences at both ends of α2 have also been noted for mPTP-BL PDZ2.[46] In addition to the canonical binding interactions shared by a number of PDZ domains, the complex between the MAGI-1 PDZ1 and $16E6_{ct}$ L_0/V peptide revealed additional interactions. As observed in the crystal structure,[49] the solution structure shows interactions between the negatively charged residues located in the loop between β2 and β3 (sequence $_{35}$DEPDE$_{39}$) and the positively charged arginine residues (R-5 and R-4) located upstream of the canonical PDZ binding motif of the E6 peptide. The contribution of these interactions to PDZ–peptide affinity was demonstrated by SPR studies.[32] The observation of a set of NOEs between residues located in the C-terminal extension ($_{113}$SLV$_{115}$) and the R-4 residue, together with the motional restriction induced by peptide binding, suggests that the network of interactions might be more extended than anticipated. Indeed, mutations of the $_{113}$SLV$_{115}$ sequence into either a positively charged stretch of arginines and lysine or a highly mobile stretch of glycines alter the binding affinity, supporting the role of C-terminal extension in the binding process. Surprisingly, the effect of the GGG mutation is more

pronounced than that of the RKR mutation, suggesting that the dynamic properties of the C-terminal sequence, rather than the positive charges, must be preserved for optimal binding. Alternatively, the nonpolar parts of the arginine and lysine side chains may retain, at least partly, the hydrophobic character of the wild-type sequence. To our knowledge, this is the first example where a mutation performed outside the canonical boundaries of a PDZ domain is shown to affect peptide binding.

The question arises as to whether the dynamic changes in the C-terminal extension of MAGI-1 PDZ1 observed upon binding of $16E6_{ct}$ L_0/V reflect a signaling mechanism of the binding event to other domains of the protein. Several studies have reported observations that support the presence of interdomain interactions, such as those for the tandem PDZ domains of GRIP1[60] or PSD-95.[61] In this latter case, the length of the linker between tandem PDZ domains (varied between 5 and 11 residues) was shown to be important. By contrast, a syntenin tandem PDZ domain linked by four residues binds target peptides independently,[62] precluding the generalization of interdomain PDZ communication mechanisms. The extended PDZ1 domain of MAGI-1 is separated by at least 60 residues from both the preceding defined domain and the following defined domain. Even allowing for the possible role of flanking regions for those domains, it remains rather difficult at this stage to assess the consequences of peptide binding on regions of MAGI-1 outside the PDZ1 domain. Studies on multiple-domain constructs from MAGI-1 are currently in progress to address this question.

Of particular note are the dramatic changes in hydrogen/deuterium exchange rates throughout the PDZ domain upon peptide binding. While some of the observed changes were expected from β-strand extension and reduced solvent accessibility, others are more intriguing and may result from subtle changes in dynamics in parts of the PDZ domain. These effects are, for instance, observed for several residues located in helices α1 and α2 (Fig. 5). Interestingly, several hydrogen bonds of β-sheets remote from the binding site undergo a reduction of their exchange rates, indicating that the effect of peptide binding propagates across the PDZ domain through a defined network that couples the binding groove to distal sites. This observation, which indicates a change in the distribution of the open and closed microstates of hydrogen bonds upon binding, supports the model of a global response of PDZ domains to a binding event. Global changes in backbone and side-chain dynamics upon peptide binding have already been reported for a number of PDZ domains, including hPTP1E PDZ2[47] or PSD-95 PDZ3.[63] Dynamic changes are propagated across PDZ domains through dynamic networks characterized by several theoretical studies.[40–43] It is not yet clear whether there is a common mode of response to binding for diverse PDZ domains: indeed, it might be surprising if this were the case. From the limited number of systems studied to date, it seems that the bases of signaling within PDZ domains may differ between domains, and the characterization of a greater number of such domains will be necessary to furnish a more complete understanding of the global response of these entities to peptide binding. Moreover, the binding of E6 peptides to MAGI-1 PDZ domains may be governed by specific constraints aimed at conferring optimal infectivity. This requires the "high-risk" HPV E6 proteins to target a set of PDZ-domain-containing proteins rather than a single one. This would explain why HPV16 E6 binds more strongly to hScrib than to Dlg, while the reverse situation is observed for HPV18 E6.[28,64] Further studies are needed to establish whether molecular events observed in this study reveal either a specific viral strategy aimed at hijacking host cell regulation[65] or some intrinsic properties of the MAGI-1 PDZ domain.

Materials and Methods

NMR experiments

Samples of unlabeled, ^{15}N-labeled, and $^{15}N,^{13}C$-labeled MAGI-1 PDZ1 were prepared in 20 mM or 100 mM phosphate buffer (pH 6.8) with 50 mM NaCl and 2 mM DTT, at protein concentrations between 200 μM and 600 μM. The MAGI-1 PDZ1/$16E6_{ct}$ L_0/V complex was prepared by addition of a 3-fold excess of the synthesized 11-mer $16E6_{ct}$ L_0/V peptide.

NMR experiments were performed (unless otherwise stated) on a Bruker DRX 600-MHz spectrometer equipped with a triple-resonance cryoprobe with z-gradients at 295 K. A set of three-dimensional triple-resonance experiments (HN(CO)CA, HNCA, HN(CO)CACB, HNCACB, and HNCO) was recorded to obtain backbone resonance assignments. Aliphatic side-chain resonance assignments were obtained using HCCH correlated spectroscopy and HCCH total correlated spectroscopy (TOCSY) experiments, and ^{15}N-edited TOCSY and NOE spectroscopy (NOESY) experiments. Aromatic side chains were assigned using two-dimensional homonuclear TOCSY and NOESY experiments recorded at 800 MHz (^1H frequency), and ^{13}C-edited NOESY spectra optimized for aromatic ^{13}C resonances. For the MAGI-1 PDZ1/$16E6_{ct}$ L_0/V complex, ^{12}C-filtered TOCSY and NOESY spectra were used to enable assignment of the bound peptide resonances.[66] Intermolecular contacts were identified from a ^{12}C-filtered and ^{13}C-edited NOESY, in which NOEs from the hydrogen nuclei of atoms not attached to ^{13}C to those attached to ^{13}C are detected. All spectra were processed using NMRPipe[67] and analyzed using CARA[68] and the NEASY module of CARA. Predictions of backbone Φ and Ψ angles were obtained from resonance assignments using the program TALOS.[69] Composite

chemical shift changes were calculated on a per-residue basis using all nuclei of each residue that were assigned in the spectra of both liganded and unliganded forms.[70]

Hydrogen/deuterium exchange was probed by dissolving lyophilized samples of MAGI-1 PDZ1 or MAGI-1 PDZ1/16E6$_{ct}$ L_0/V in ^2H_2O at a concentration of 80 μM and by recording ^1H–^{15}N heteronuclear single-quantum coherence spectra over the following 24 h. Exponential decay rates were obtained from a nonlinear least-squares two-parameter fit using a Levenberg–Marquardt algorithm implemented in MATLAB (The Mathworks, Inc.). Residual dipolar couplings for backbone ^1H–^{15}N pairs were measured in polyacrylamide gels using IPAP pulse sequences.[71] Analysis of RDC values was performed using the MODULE2.0 program.[72]

Structure determination

Three-dimensional structure determination was performed using the semiautomatic ATNOS/CANDID procedure,[73,74] with Xplor-NIH[75] as the molecular dynamics program. A two-dimensional NOESY spectrum recorded at a mixing time of 100 ms, a three-dimensional ^{15}N-edited NOESY spectrum recorded at a mixing time of 150 ms, and three-dimensional ^{13}C-edited NOESY spectra for aliphatic and aromatic ^{13}C nuclei recorded at a mixing time of 150 ms were provided. Hydrogen bonds that could be unambiguously identified from hydrogen/deuterium exchange data and NOE patterns were introduced, together with constraints on dihedral angles predicted with a high degree of confidence by TALOS. For the complex, the peptide chain was attached to the C-terminus of the protein by a 25-residue polyglycine linker. The list of distance constraints generated by CANDID in each cycle was supplemented by a list of intermolecular distances identified in filtered NOESY spectra. The resulting structures were refined in Xplor-NIH using a standard protocol (refine.inp) modified to start from a higher initial temperature (2000 K), with a larger number of cooling steps (10,000 steps) and a longer final energy minimization step (1200 steps). The structures were then refined using parallhdg5.3 parameters, and stereospecific assignments were made, where possible, before the final refinement in explicit solvent.[76] Local and global pairwise r.m.s.d. values were calculated in Xplor-NIH using in-house scripts. Figures were produced using PyMOL.[77]

NMR relaxation measurements

^{15}N relaxation measurements were performed at 295 K and 600 MHz (^1H frequency). ^{15}N R_1 and R_2 relaxation rates were measured using a single-pulse sequence based on those of Farrow *et al.*, in which the lengths of longitudinal or transverse relaxation delays are varied with the other delay set to a minimum value.[78] For ^{15}N R_1 relaxation, intensities were extracted from a set of 14 spectra recorded with relaxation delay values of between 4 ms and 2010 ms, with 180° proton pulses every 2 ms to suppress cross-correlated relaxation.[79] For ^{15}N R_2 relaxation, intensities were extracted from a set of 12 spectra recorded with relaxation delay values of 0 ms and 158 ms, with ^{15}N 180° pulses applied every 1.2 ms at a field strength of 4.2 kHz, and with ^1H 180° pulses applied every 2.4 ms to suppress cross-correlated relaxation.[79] Exponential decay rates were obtained from a nonlinear least-squares two-parameter fit using a Levenberg–Marquardt algorithm implemented in MATLAB (The Mathworks, Inc.).

Surface plasmon resonance

Data were collected on a Biacore 2000 instrument (Biacore AB/GE Healthcare Bio-Sciences Corp. Piscataway, NJ) at 25 °C with the autosampler rack base cooled to 10 °C. We used an optimized protocol published recently[32] for a rapid and accurate estimate of affinity constants between protein domains and glutathione *S*-transferase (GST)-fused peptides. Briefly, GST–peptide fusions were immobilized onto a CM5 chip previously activated with goat anti-GST antibody. Typically, protein densities of 100–200 response units (RU) were used to prevent steric inhibition. The MAGI-1 PDZ1 domain was injected at various concentrations ranging from 50 nM to 10 μM in 20 mM phosphate buffer (pH 6.8) complemented with 200 mM NaCl and P20 (0.005%, vol/vol) at a flow rate of 20 μl min^{-1} at 25 °C. At least two full sets of independent experiments were performed for each series of GST–peptide/PDZ interaction to assess data reproducibility (Fig. 7a). A reference was systematically included on each chip. K_d and R_{max} values were determined using a 1:1 model by fitting the binding isotherms obtained for the interaction between the MAGI-1 PDZ1 domain and various peptides using the BiaEvaluation 3.2 software.

Accession numbers

Assignments have been deposited under BMRB ID 16558 (unliganded) and BMRB ID 16559 (liganded). Structures and experimental data have been deposited under PDB ID 2KPK (unliganded) and PDB ID 2KPL (liganded).

Supplementary materials related to this article can be found online at doi:10.1016/j.jmb.2011.01.015

Acknowledgements

The authors thank Claude Ling (Institut de Génétique et de Biologie Moléculaire et Cellulaire) for technical support and Pascal Eberling (Institut de Génétique et de Biologie Moléculaire et Cellulaire) for peptide synthesis. The authors thank the European NMR Large-Scale Facility Utrecht and Dr. E. Guittet (ICSN, Gif-sur-Yvette, CNRS TGE NMR program) for access to high-field NMR spectrometers, and Dr. T. Herrmann (ENS, Lyon) for advice on the use of software. This work was supported by ANR programs ANR-06-BLAN-0404 and ANR-MIME-2007 (project EPI-HPV-3D), Association pour la Recherche sur le Cancer grants 3127 and 3171, and institute funds from the Centre National de

la Recherche Scientifique, the Institut National de la Santé et de la Recherche Médicale, the University of Strasbourg, and the Fonds National de la Science (GENOPOLE). S.C. was supported by the Association pour la Recherche sur le Cancer and La Ligue Nationale Contre le Cancer. K.L. was supported by the Région Alsace and the Collège Doctoral Européen de Strasbourg.

References

1. Saras, J. & Heldin, C. H. (1996). PDZ domains bind carboxy-terminal sequences of target proteins. *Trends Biochem. Sci.* **21**, 455–458.
2. Schultz, J., Hoffmuller, U., Krause, G., Ashurst, J., Macias, M. J., Schmieder, P. *et al.* (1998). Specific interactions between the syntrophin PDZ domain and voltage-gated sodium channels. *Nat. Struct. Biol.* **5**, 19–24.
3. Jelen, F., Oleksy, A., Smietana, K. & Otlewski, J. (2003). PDZ domains—common players in the cell signaling. *Acta Biochim. Pol.* **50**, 985–1017.
4. Au, Y., Atkinson, R. A., Guerrini, R., Kelly, G., Joseph, C., Martin, S. R. *et al.* (2004). Solution structure of ZASP PDZ domain; implications for sarcomere ultrastructure and enigma family redundancy. *Structure*, **12**, 611–622.
5. Wilken, C., Kitzing, K., Kurzbauer, R., Ehrmann, M. & Clausen, T. (2004). Crystal structure of the DegS stress sensor: how a PDZ domain recognizes misfolded protein and activates a protease. *Cell*, **117**, 483–494.
6. Craven, S. E. & Bredt, D. S. (1998). PDZ proteins organize synaptic signaling pathways. *Cell*, **93**, 495–498.
7. Fanning, A. S. & Anderson, J. M. (1999). Protein modules as organizers of membrane structure. *Curr. Opin. Cell Biol.* **11**, 432–439.
8. Thomas, M., Narayan, N., Pim, D., Tomaic, V., Massimi, P., Nagasaka, K. *et al.* (2008). Human papillomaviruses, cervical cancer and cell polarity. *Oncogene*, **27**, 7018–7030.
9. Thomas, M. C. & Chiang, C. M. (2005). E6 oncoprotein represses p53-dependent gene activation via inhibition of protein acetylation independently of inducing p53 degradation. *Mol. Cell*, **17**, 251–264.
10. James, M. A., Lee, J. H. & Klingelhutz, A. J. (2006). Human papillomavirus type 16 E6 activates NF-kappaB, induces cIAP-2 expression, and protects against apoptosis in a PDZ binding motif-dependent manner. *J. Virol.* **80**, 5301–5307.
11. Shai, A., Brake, T., Somoza, C. & Lambert, P. F. (2007). The human papillomavirus E6 oncogene dysregulates the cell cycle and contributes to cervical carcinogenesis through two independent activities. *Cancer Res.* **67**, 1626–1635.
12. Munger, K. & Howley, P. M. (2002). Human papillomavirus immortalization and transformation functions. *Virus Res.* **89**, 213–228.
13. Ganguly, N. & Parihar, S. P. (2009). Human papillomavirus E6 and E7 oncoproteins as risk factors for tumorigenesis. *J. Biosci.* **34**, 113–123.
14. Huibregtse, J. M., Scheffner, M. & Howley, P. M. (1991). A cellular protein mediates association of p53 with the E6 oncoprotein of human papillomavirus types 16 or 18. *EMBO J.* **10**, 4129–4135.
15. Patel, D., Huang, S. M., Baglia, L. A. & McCance, D. J. (1999). The E6 protein of human papillomavirus type 16 binds to and inhibits co-actiation by CBP and p300. *EMBO J.* **18**, 5061–5072.
16. Grm, H. S. & Banks, L. (2004). Degradation of hDlg and MAGIs by human papillomavirus E6 is E6-AP-independent. *J. Gen. Virol.* **85**, 2815–2819.
17. Handa, K., Yugawa, T., Narisawa-Saito, M., Ohno, S., Fujita, M. & Kiyono, T. (2007). E6AP-dependent degradation of DLG4/PSD95 by high-risk human papillomavirus type 18 E6 protein. *J. Virol.* **81**, 1379–1389.
18. Nakagawa, S. & Huibregtse, J. M. (2000). Human scribble (Vartul) is targeted for ubiquitin-mediated degradation by the high-risk papillomavirus E6 proteins and the E6AP ubiquitin–protein ligase. *Mol. Cell. Biol.* **20**, 8244–8253.
19. Glaunsinger, B. A., Lee, S. S., Thomas, M., Banks, L. & Javier, R. (2000). Interactions of the PDZ-protein MAGI-1 with adenovirus E4-ORF1 and high-risk papillomavirus E6 oncoproteins. *Oncogene*, **19**, 5270–5280.
20. Thomas, M., Laura, R., Hepner, K., Guccione, E., Sawyers, C., Lasky, L. & Banks, L. (2002). Oncogenic human papillomavirus E6 proteins target the MAGI-2 and MAGI-3 proteins for degradation. *Oncogene*, **21**, 5088–5096.
21. Jeong, K. W., Kim, H. Z., Kim, S., Kim, Y. S. & Choe, J. (2007). Human papillomavirus type 16 E6 protein interacts with cystic fibrosis transmembrane regulator-associated ligand and promotes E6-associated protein-mediated ubiquitination and proteasomal degradation. *Oncogene*, **26**, 487–499.
22. Lee, S. S., Glaunsinger, B., Mantovani, F., Banks, L. & Javier, R. T. (2000). Multi-PDZ domain protein MUPP1 is a cellular target for both adenovirus E4-ORF1 and high-risk papillomavirus type 18 E6 oncoproteins. *J. Virol.* **74**, 9680–9693.
23. Storrs, C. H. & Silverstein, S. J. (2007). PATJ, a tight junction-associated PDZ protein, is a novel degradation target of high-risk human papillomavirus E6 and the alternatively spliced isoform 18 E6. *J. Virol.* **81**, 4080–4090.
24. Jing, M., Bohl, J., Brimer, N., Kinter, M. & Vande Pol, S. B. (2007). Degradation of tyrosine phosphatase PTPN3 (PTPH1) by association with oncogenic human papillomavirus E6 proteins. *J. Virol.* **81**, 2231–2239.
25. Hampson, L., Li, C., Oliver, A. W., Kitchener, H. C. & Hampson, I. N. (2004). The PDZ protein Tip-1 is a gain of function target of the HPV16 E6 oncoprotein. *Int. J. Oncol.* **25**, 1249–1256.
26. Favre-Bonvin, A., Reynaud, C., Kretz-Remy, C. & Jalinot, P. (2005). Human papillomavirus type 18 E6 protein binds the cellular PDZ protein TIP-2/GIPC, which is involved in transforming growth factor beta signaling and triggers its degradation by the proteasome. *J. Virol.* **79**, 4229–4237.
27. Simonson, S. J., Difilippantonio, M. J. & Lambert, P. F. (2005). Two distinct activities contribute to human

papillomavirus 16 E6's oncogenic potential. *Cancer Res.* **65**, 8266–8273.
28. Liu, Y., Henry, G. D., Hegde, R. S. & Baleja, J. D. (2007). Solution structure of the hDlg/SAP97 PDZ2 domain and its mechanism of interaction with HPV-18 papillomavirus E6 protein. *Biochemistry*, **46**, 10864–10874.
29. Kiyono, T., Hiraiwa, A., Fujita, M., Hayashi, Y., Akiyama, T. & Ishibashi, M. (1997). Binding of high-risk human papillomavirus E6 oncoproteins to the human homologue of the *Drosophila* discs large tumor suppressor protein. *Proc. Natl Acad. Sci. USA*, **94**, 11612–11616.
30. Lee, S. S., Weiss, R. S. & Javier, R. T. (1997). Binding of human virus oncoproteins to hDlg/SAP97, a mammalian homolog of the *Drosophila* discs large tumor suppressor protein. *Proc. Natl Acad. Sci. USA*, **94**, 6670–6675.
31. Thomas, M., Glaunsinger, B., Pim, D., Javier, R. & Banks, L. (2001). HPV E6 and MAGUK protein interactions: determination of the molecular basis for specific protein recognition and degradation. *Oncogene*, **20**, 5431–5439.
32. Fournane, S., Charbonnier, S., Chapelle, A., Kieffer, B., Orfanoudakis, G., Travé, G., et al. (2010). Surface plasmon resonance analysis of the binding of high-risk mucosal HPV E6 oncoproteins to the PDZ1 domain of the tight junction protein MAGI-1. *J. Mol. Recognit.* in press [E-publication ahead of print].
33. Franco, E. L. (1992). Prognostic value of human papillomavirus in the survival of cervical cancer patients: an overview of the evidence. *Cancer Epidemiol. Biomarkers Prev.* **1**, 499–504.
34. Songyang, Z., Fanning, A. S., Fu, C., Xu, J., Marfatia, S. M., Chishti, A. H. *et al.* (1997). Recognition of unique carboxyl-terminal motifs by distinct PDZ domains. *Science*, **275**, 73–77.
35. Tonikian, R., Zhang, Y., Sazinsky, S. L., Currell, B., Yeh, J. H., Reva, B. *et al.* (2008). A specificity map for the PDZ domain family. *PLoS Biol.* **6**, e239.
36. Stiffler, M. A., Chen, J. R., Grantcharova, V. P., Lei, Y., Fuchs, D., Allen, J. E. *et al.* (2007). PDZ domain binding selectivity is optimized accros the mouse proteome. *Science*, **317**, 364–369.
37. Encinar, J. A., Fernandez-Ballester, G., Sanchez, I. E., Hurtado-Gomez, E., Stricher, F., Beltrao, P. & Serrano, L. (2009). ADAN: a database for prediction of protein–protein interaction of modular domains mediated by linear motifs. *Bioinformatics*, **25**, 2418–2424.
38. Doyle, D. A., Lee, A., Lewis, J., Kim, E., Sheng, M. & MacKinnon, R. (1996). Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ. *Cell*, **85**, 1067–1076.
39. Zhang, J., Sapienza, P. J., Ke, H., Chang, A., Hengel, S. R., Wang, H. *et al.* (2010). Crystallographic and nuclear magnetic resonance evaluation of the impact of peptide binding to the second PDZ domain of PTP1E. *Biochemistry*, **49**, 9191–9280.
40. Sharp, K. & Skinner, J. J. (2006). Pump-probe molecular dynamics as a tool for studying protein motion and long range coupling. *Proteins*, **65**, 347–361.
41. Ota, N. & Agard, D. A. (2005). Intramolecular signaling pathways revealed by modeling anisotropic thermal diffusion. *J. Mol. Biol.* **351**, 345–354.
42. Lockless, S. W. & Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, **286**, 295–299.
43. Kong, Y. & Karplus, M. (2009). Signaling pathways of PDZ2 domain: a molecular dynamics interaction correlation analysis. *Proteins*, **74**, 145–154.
44. De Los Rios, P., Cecconi, F., Pretre, A., Dietler, G., Michielin, O., Piazza, F. & Juanico, B. (2005). Functional dynamics of PDZ binding domains: a normal mode analysis. *Biophys. J.* **89**, 14–21.
45. Gerek, Z. N., Keskin, O. & Ozkan, S. B. (2009). Identification of specificity and promiscuity of PDZ domain interactions through their dynamic behavior. *Proteins*, **77**, 796–811.
46. Gianni, S., Walma, T., Arcovito, A., Calosci, N., Bellelli, A., Engstrom, A. *et al.* (2006). Demonstration of long-range interactions in a PDZ domain by NMR, kinetics, and protein engineering. *Structure*, **14**, 1801–1809.
47. Fuentes, E. J., Gilmore, S. A., Mauldin, R. V. & Lee, A. L. (2006). Evaluation of energetic and dynamic coupling networks in a PDZ domain protein. *J. Mol. Biol.* **364**, 337–351.
48. Dhulesia, A., Gsponer, J. & Vendruscolo, M. (2008). Mapping of two networks of residues that exhibit structural and dynamical changes upon binding in a PDZ domain protein. *J. Am. Chem. Soc.* **130**, 8931–8939.
49. Zhang, Y., Dasgupta, J., Ma, R. Z., Banks, L., Thomas, M. & Chen, X. S. (2007). Structures of a human papillomavirus (HPV) E6 polypeptide bound to MAGUK proteins: mechanisms of targeting tumor suppressors by a high-risk HPV oncoprotein. *J. Virol.* **81**, 3618–3626.
50. Charbonnier, S., Stier, G., Orfanoudakis, G., Kieffer, B., Atkinson, R. A. & Trave, G. (2008). Defining the minimal interacting regions of the tight junction protein MAGI-1 and HPV16 E6 oncoprotein for solution structure studies. *Protein Expression Purif.* **60**, 64–73.
51. Charbonnier, S., Coutouly, M., Kieffer, B., Trave, G. & Atkinson, R. A. (2006). ^{13}C, ^{15}N and ^{1}H resonance assignment of the PDZ1 domain of MAGI-1 using QUASI. *J. Biomol. NMR*, **36**, 33.
52. Zhang, J., Yan, X., Shi, C., Yang, X., Guo, Y., Tian, C. *et al.* (2008). Structural basis of beta-catenin recognition by Tax-interacting protein-1. *J. Mol. Biol.* **384**, 255–263.
53. Tochio, H., Hung, F., Li, M., Bredt, D. S. & Zhang, M. (2000). Solution structure and backbone dynamics of the second PDZ domain of postsynaptic density-95. *J. Mol. Biol.* **295**, 225–237.
54. Kozlov, G., Banville, D., Gehring, K. & Ekiel, I. (2002). Solution structure of the PDZ2 domain from cytosolic human phosphatase hPTP1E complexed with a peptide reveals contribution of the beta2–beta3 loop to PDZ domain–ligand interactions. *J. Mol. Biol.* **320**, 813–820.
55. Ulrich, E. L., Akutsu, H., Doreleijers, J. F., Harano, Y., Ioannidis, Y. E., Lin, J. *et al.* (2008). BioMagResBank. *Nucleic Acids Res.* **36**, D402–D408.

56. Laskowski, R. A., Rullmann, J. A., MacArthur, M. W., Kaptein, R. & Thomton, J. M. (1996). AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR*, **8**, 477–486.
57. Dosset, P., Hus, J. C., Blackledge, M. & Marion, D. (2000). Efficient analysis of macromolecular rotational diffusion from heteronuclear relaxation data. *J. Biomol. NMR*, **16**, 23–28.
58. Narisawa-Saito, M. & Kiyono, T. (2007). Basic mechanisms of high-risk human papillomavirus-induced carcinogenesis: roles of E6 and E7 proteins. *Cancer Sci.* **98**, 1505–1511.
59. Thomas, M., Dasgupta, J., Zhang, Y., Chen, X. & Banks, L. (2008). Analysis of specificity determinants in the interactions of different HPV E6 proteins with their PDZ domain-containing substrates. *Virology*, **376**, 371–378.
60. Long, J., Wei, Z., Feng, W., Yu, C., Zhao, Y. X. & Zhang, M. (2008). Supramodular nature of GRIP1 revealed by the structure of its PDZ12 tandem in complex with the carboxyl tail of Fras1. *J. Mol. Biol.* **375**, 1457–1468.
61. Long, J. F., Tochio, H., Wang, P., Fan, J. S., Sala, C., Niethammer, M. *et al.* (2003). Supramodular structure and synergistic target binding of the N-terminal tandem PDZ domains of PSD-95. *J. Mol. Biol.* **327**, 203–214.
62. Kang, B. S., Cooper, D. R., Jelen, F., Devedjiev, Y., Derewenda, U., Dauter, Z. *et al.* (2003). PDZ tandem of human syntenin: crystal structure and functional properties. *Structure*, **11**, 459–468.
63. Petit, C. M., Zhang, J., Sapienza, P. J., Fuentes, E. J. & Lee, A. L. (2009). Hidden dynamic allostery in a PDZ domain. *Proc. Natl Acad. Sci. USA*, **106**, 18249–18254.
64. Thomas, M., Massimi, P., Navarro, C., Borg, J. P. & Banks, L. (2005). The hScrib/Dlg apico-basal control complex is differentially targeted by HPV-16 and HPV-18 E6 proteins. *Oncogene*, **24**, 6222–6230.
65. Davey, N. E., Trave, G. & Gibson, T. J. (2010). How viruses hijack cell regulation. *Trends Biochem. Sci.* [E-publication ahead of print].
66. Zwahlen, C., Legault, P., Vincent, S. J. F., Greenblatt, J., Konrat, R. & Kay, L. E. (1997). Methods for measurement of intermolecular NOEs by multinuclear NMR spectroscopy: application to a bacteriophage lambda N-peptide/boxB RNA complex. *J. Am. Chem. Soc.* **119**, 6711–6721.
67. Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J. & Bax, A. (1995). NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR*, **6**, 277–293.
68. Keller, R. (2004). *The Computer Aided Resonance Assignment Tutorial*.CANTINA Verlag, Goldau, Switzerland; 3-85600-112-3.
69. Cornilescu, G., Delaglio, F. & Bax, A. (1999). Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR*, **13**, 289–302.
70. van Ingen, H., van Schaik, F. M., Wienk, H., Ballering, J., Rehmann, H., Dechesne, A. C. *et al.* (2008). Structural insight into the recognition of the H3K4me3 mark by the TFIID subunit TAF3. *Structure*, **16**, 1245–1256.
71. Hall, J. B., Daye, K. T. & Fushman, D. (2003). Direct measurement of the N15 CSA/dipolar relaxation interference from coupled HSQC spectra. *J. Biomol. NMR*, **26**, 181–186.
72. Dosset, P., Hus, J. C., Marion, D. & Blackledge, M. (2001). A novel interactive tool for rigid-body modeling of multi-domain macromolecules using residual dipolar couplings. *J. Biomol. NMR*, **20**, 223–231.
73. Herrmann, T., Guntert, P. & Wuthrich, K. (2002). Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. *J. Biomol. NMR*, **24**, 171–189.
74. Herrmann, T., Guntert, P. & Wuthrich, K. (2002). Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J. Mol. Biol.* **319**, 209–227.
75. Schwieters, C. D., Kuszewski, J. J., Tjandra, N. & Clore, G. M. (2003). The Xplor-NIH NMR molecular structure determination package. *J. Magn. Reson.* **160**, 65–73.
76. Linge, J. P., Williams, M. A., Spronk, C. A., Bonvin, A. M. & Nilges, M. (2003). Refinement of protein structures in explicit solvent. *Proteins*, **50**, 496–506.
77. DeLano, W. (2002). *The PyMOL Molecular Graphics System.* DeLano Scientific, San Carlos, CA.
78. Farrow, N. A., Muhandiram, R., Singer, A. U., Pascal, S. M., Kay, C. M., Gish, G. *et al.* (1994). Backbone dynamics of a free and phosphopeptide-complexed Src homology 2 domain studied by ^{15}N NMR relaxation. *Biochemistry*, **33**, 5984–6003.
79. Kay, L. E., Nicholson, L. K., Delaglio, F., Bax, A. & Torchia, D. A. (1992). Pulse sequence for removal of the effects of cross-correlation between dipolar and chemical-shift anisotropy relaxation mechanims on the measurement of heteronuclear T_1 and T_2 values in proteins. *J. Magn. Reson.* **97**, 359–375.

8. Hydrophobic residue bias in phage display data can impair PDZ interaction prediction performances

8.1. Summary

Tonikian *et al.* [9] published in 2008 a large scale phage display study to define the binding profile for 54 human and 28 worm PDZ domains. This data set has been recurrently used for PDZ interaction predictor development (see section 6.4.4). In this article, we have published our results about the evaluation of predictions of cellular PBMs for the 54 human PDZ domains using the phage display data.
Approach: First, we searched the human proteome for C-terminal sequences that are identical to phage display peptides, which have been selected for the 54 human PDZs. Next, we constructed a PSSM for each of the 54 human PDZs based on the phage display peptides to search the human proteome for similar C-terminal sequences to the phage display peptides. We compared the mean hydrophobicity and Trp content of phage display peptides versus those of predicted and experimentally validated PBMs as well as the human C-terminome (entirety of human C-terminal sequences) in general.
Findings: Two third of the 54 human PDZs have hydrophobic phage display peptide lists. Human C-terminal sequences that are identical to selected phage display peptides can almost only be found for the remaining third of the data, which contain hydrophilic phage display peptides. Human C-terminal sequences were found to be more similar to hydrophilic than to hydrophobic phage display peptides. The phage display peptides are much more hydrophobic than experimentally validated cellular PBMs and the human C-terminome in general. Half of the phage display peptides display a Trp at position p-1. This property cannot be found for cellular PBMs.
Discussion/Conclusions: Predictions of cellular PBMs seem to be more reliable when based on hydrophilic phage display peptide lists. SLiMs have a particular sequence composition with hydrophobic or charged residues at conserved positions and non-hydrophobic residues at variable positions [45]. In the phage display data of Tonikian *et al.* [9], hydrophobic residues have been very frequently selected at variable peptide positions. Large aromatic/hydrophobic amino acids contribute much to the binding affinity of PDZ–peptide interactions [155, 156]. Phage display peptide selection is mainly affinity driven. This might explain the selection of hydrophobic residues at flexible peptide positions. SLiM-mediated interactions are of weak binding affinity but specific. Phage display can select peptides with sequence properties that are very different from those of SLiMs due to different selection pressures imposed by

phage display and evolution. Therefore, the application of phage display data for the prediction of cellular SLiMs may in some cases be limited and we recommand sequence analysis of phage display data prior to its application for SLiM-mediated interaction prediction.
Contribution: I have performed the whole programming, prediction, and data analysis. I have contributed most of the ideas how to prove the bias for hydrophobic residues in the phage display data. I have conceived and drafted a first version of the manuscript. Gilles Travé and myself have written together the final version of the manuscript.
(See also supplemental material provided in section D.1.)

BIOINFORMATICS DISCOVERY NOTE Vol. 27 no. 7 2011, pages 899–902
doi:10.1093/bioinformatics/btr060

Sequence analysis Advance Access publication February 7, 2011

Phage display can select over-hydrophobic sequences that may impair prediction of natural domain–peptide interactions

Katja Luck* and Gilles Travé*
Oncoproteins, Unité CNRS-UDS UMR 7242, Ecole Supérieure de Biotechnologie de Strasbourg, 1, Bd Sébastien Brant, BP 10413, 67412 Illkirch - Cedex, France
Associate Editor: Alex Bateman

ABSTRACT

Motivation: The phage display peptide selection approach is widely used for defining binding specificities of globular domains. PDZ domains recognize partner proteins via C-terminal motifs and are often used as a model for interaction predictions. Here, we investigated to which extent phage display data that were recently published for 54 human PDZ domains can be applied to the prediction of human PDZ–peptide interactions.

Results: Promising predictions were obtained for one-third of the 54 PDZ domains. For the other two-thirds, we detected in the phage display peptides an important bias for hydrophobic amino acids that seemed to impair correct predictions. Therefore, phage display-selected peptides may be over-hydrophobic and of high affinity, while natural interaction motifs are rather hydrophilic and mostly combine low affinity with high specificity. We suggest that potential amino acid composition bias should systematically be investigated when applying phage display data to the prediction of specific natural domain–linear motif interactions.

Contact: katja.luck@unistra.fr; gilles.trave@unistra.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 22, 2010; revised on January 28, 2011; accepted on February 1, 2011

1 INTRODUCTION

Many protein complexes that function in cellular regulation and signalling are assembled by multiple linear motif–globular domain interactions, which are mostly specific, yet of low affinity (Diella *et al.*, 2008). One well studied example of such interactions consists of the PDZ domains, which mainly recognize linear motifs at the extreme C-terminus of partner proteins (Doyle *et al.*, 1996). PDZs are implicated in the regulation of cell polarity, tight junctions, intercellular communication and neuronal synapses (Nourry *et al.*, 2003). The last residue (referred to as position p0) in PDZ-binding motifs usually is Val or Leu. The third last peptide residue (position p-2) can be either Thr or Ser, hydrophobic or Glu or Asp, thereby defining three main categories of PDZ-binding motifs (Songyang *et al.*, 1997) (Stricker *et al.*, 1997). These characteristics make PDZ–binding motifs relatively easy to predict. However, the correct prediction of PDZ domain binding specificities, i.e. prediction of which PDZ-binding motif will bind to which PDZ domain, remains challenging and numerous approaches have been proposed to tackle this problem (Brannetti *et al.*, 2001; Chen *et al.*, 2008; Hui and Bader, 2010; Kalyoncu *et al.*, 2010; Schillinger *et al.*, 2009; Smith and Kortemme, 2010).

Most predictors rely on prior experimental knowledge about binding preferences between peptides and globular domains. Phage peptide display has been widely used to provide such information (Sidhu *et al.*, 2003). This approach is based on selecting, out of a library of billions of peptides expressed on the surface of bacteriophages, a limited number of peptides that bind strongly to a given protein attached to a solid support. Several phage display studies have been performed on particular PDZ domains derived from the proteins MAGI1 (Fuh *et al.*, 2000), INADL (Vaccaro *et al.*, 2001), PDZRhoGEF and LARG (Smietana *et al.*, 2008), MUPP1 and DLG4 (Sharma *et al.*, 2009), PTP-BL (van den Berk *et al.*, 2007), Erbin (Skelton *et al.*, 2003), HtrA1 and HtrA3 (Runyon *et al.*, 2007). Tonikian *et al.* (2008) applied phage display in a high-throughput manner to determine and compare binding preferences of 28 *Caenorhabditis elegans* and 54 *Homo sapiens* PDZ domains. The data obtained in this study represent a highly valuable resource that allows to test the general application of phage display data to predictions of natural PDZ–protein interactions using position-specific scoring matrices (PSSMs). This approach was validated on a few PDZ domains in the study of Tonikian *et al.* (2008), and the phage display data were subsequently used in several recent studies for predictions of natural PDZ–peptide interactions (Hui and Bader, 2010; Smith and Kortemme, 2010). PDZ phage display data have also been used as test data in the 'DREAM4 Peptide Recognition Domain Specificity Prediction' challenge (Smith and Kortemme, 2010). Thus, phage display is supposed to capture accurately the binding specificities of domain–linear motif interactions.

Here, we performed and evaluated predictions of human PDZ–peptide interactions using the phage display data of Tonikian *et al.* (2008). Promising predictions were obtained for one-third of the 54 PDZ domains. In contrast, for the other two-thirds of the PDZ domains we detected important bias for hydrophobic amino acids in the phage display peptides that will probably impair the correct prediction of naturally occurring PDZ-binding peptides. We suggest that utilization of phage display data for prediction of natural binders should systematically involve prior analysis of potential sequencial bias in the data.

2 RESULTS

2.1 Prediction of natural PDZ–peptide interactions using phage display data

We searched the human proteome for C-termini of five residues in length that are likely to bind to the 54 human PDZ domains

*To whom correspondence should be addressed.

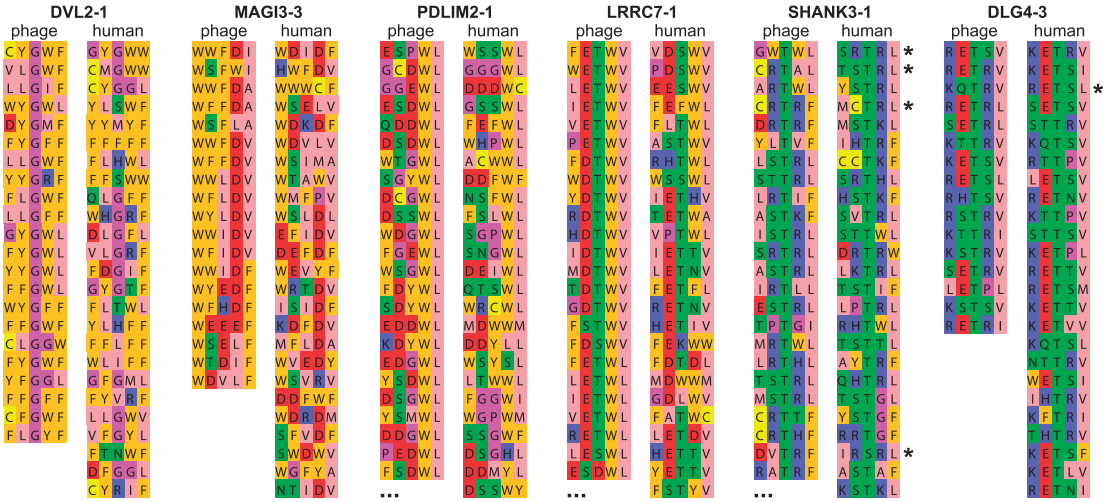


Fig. 1. Prediction of natural binders to PDZ domains using phage display data of Tonikian *et al.* (2008). The last five residues of phage display (PD) peptides together with predicted best-matching human C-terminal peptides are shown for six PDZ domains, ordered from the left to the right from most hydrophobic to most hydrophilic PD peptides. PD lists that were too long for being entirely displayed are indicated by '…'. Asterisks indicate human C-termini that are identical to C-termini of corresponding PD peptides. Colour code: gold = aromatic, light pink = hydrophobic, pink = G or P, green = polar, red = acidic, blue = basic, yellow = C. [Figure made with Jalview (Waterhouse *et al.*, 2009).]

for which Tonikian *et al.* (2008) obtained phage peptides. To this aim, we constructed a PSSM (see Supplementary Material) for each list of peptides selected by the 54 human PDZ domains. A PSSM captures the occurrence of each amino acid at each position within a list of aligned sequences. This allowed us to describe, for each PDZ domain, a sequence profile defined by the phage peptides that bound to it. Using each of the 54 PSSMs obtained in that way, we selected the 25 C-termini of human proteins that matched best to the sequence profile of the corresponding phage peptide list (reported in Supplementary Dataset S1). Within these sets of 25 most similar human C-termini, a number of peptides were actually found to be identical to the corresponding phage peptides (reported in Supplementary Dataset S2). Several instances of this search are shown in Figure 1.

Some of the phage peptide lists seemed to be anomalously enriched in hydrophobic amino acids (such as DVL2-1 and MAGI3-3 in Fig. 1). We used the hydrophobicity index of Kidera *et al.* (1985) to compute the average hydrophobicity (see Supplementary Material) of each list of phage peptides and ranked these lists from the most hydrophobic to the most hydrophilic (Fig. 2A). We observed that more identical human C-termini were returned for the hydrophilic phage peptide lists than for the hydrophobic ones (Fig. 2B, compare left side and right side of the plot, P-value <1.0E-6).

Next, we calculated an additional PSSM for each list of 25 human C-termini and determined its distance to the PSSM of the corresponding phage peptides (see Supplementary Material). The better the 25 human C-termini match to the sequence profile of the phage peptides, the more similar (less distant) the corresponding two PSSMs should be to each other and the more likely the 25 human C-termini would be to bind the corresponding PDZ domain. We observed that the more similar the PSSMs, the more hydrophilic the corresponding phage peptides

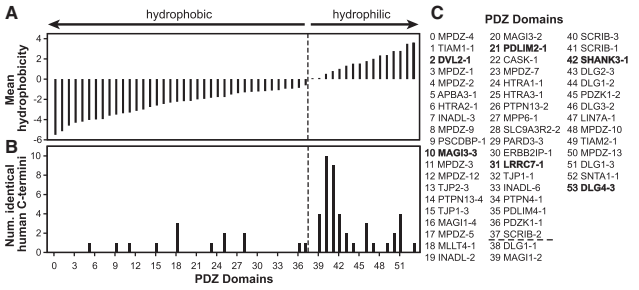


Fig. 2. Analysis of PDZ–peptide interaction predictions. (**A**) The 54 PDZ domains used by Tonikian *et al.* (2008) were ranked based on the mean hydrophobicity of their corresponding phage display (PD) peptides, from the most hydrophobic to the most hydrophilic. This ranking is conserved for plot B. The vertical dashed line separates hydrophobic from hydrophilic peptide lists. (**B**) Numbers of human C-termini that are identical to PD peptides are plotted for each PDZ domain. (**C**) PDZ domains (named as in Tonikian *et al.*) are listed based on the hydrophobicity of the PD peptide lists with numbers that were used in diagram A and B. Names in bold indicate PDZs that are shown in Figure 1.

(Pearson correlation coefficient of −0.51, P-value = 7.5E-5, see Supplementary Figure S1). This analysis indicates that the 25 best-matching human C-termini seem to better reproduce the sequence profile of the corresponding phage peptides when they are hydrophilic (see instances in Fig. 1).

2.2 Analysing the amino acid composition of phage display peptides

The above-mentioned analysis has also revealed that about two-thirds of the human PDZ domains used in the study of Tonikian *et al.* (2008) preferentially selected peptides of rather hydrophobic

Table 1. Comparison of mean hydrophobicity and W content of different peptide datasets

Source	Mean hydrophobicity[a]	% W at p-1[b]	Num peptides[c]
Tonikian *et al.*	−1.41	45.7	1390
C-terminome	0.64	1.9	26 904
PDZbase	0.92	5.6	233
Chen *et al.*	0.93	3.7	108

All peptides were reduced to a length of five residues.
[a]Calculated with index of Kidera *et al.* (1985), value of most hydrophilic peptide of length five: 9.35, most hydrophobic = −7.85.
[b]Percentage of peptides with Trp at peptide position p-1.
[c]Number of peptides.

character (Fig. 2A). We compared the mean hydrophobicity of the phage peptides to that of different peptide sets (Table 1, column 2): the C-terminome (all C-termini from the human proteome, assumed to reflect the general hydrophobicity of human C-terminal sequences, see Supplementary Material); the PDZbase [containing experimentally validated PDZ-binding peptides originating from various proteomes (Beuming *et al.*, 2005)]; and the mouse PDZ-binding peptides published by Chen *et al.* (2008). All three sets, which represent naturally occurring sequences, display a hydrophilic character in contrast to the phage peptides, which are in average markedly hydrophobic. In particular, the natural PDZ-binding peptides (derived both from Chen *et al.* and PDZbase) are significantly more hydrophilic than the phage peptides (2-sample *t*-test, *P*-value = 5E-49).

We further analysed this discrepancy by calculating the frequency of occurrence of each of the 20 amino acids in the phage peptides, the PDZ-binding peptides from the PDZbase and the human C-terminome. The phage sequences are strongly enriched in the aromatic amino acids W and F (Fig. 3). We also computed the amino acid frequencies in these three datasets for each of the five peptide positions separately (Supplementary Figure S2). All positions show an enrichment in hydrophobic amino acids, in particular aromatic residues, for phage peptides. This trend could not be observed for natural PDZ-binding peptides from the PDZbase or in general human C-terminal sequences. For instance, position p-1 is occupied by W in almost 50% of the phage peptides versus only 2% of the human C-terminal peptides, and 6% of the PDZ-binding peptides of the PDZbase (Table 1, column 3, Fisher's exact test *P*-value <2.2E-16). Interestingly, positions p-1 and p-3 and to a lesser extent p-4 seem even to be under-represented for polar or charged residues in phage peptides in contrast to the two other sets (Supplementary Fig. S2).

These results indicate that the hydrophobic character of phage PDZ-binding peptides does not correspond to sequence properties observed in natural PDZ-binding peptides and general human C-terminal sequences. This might explain why our search for best-matching human C-termini to the sequence profile of phage peptides seems to perform better for PDZ domains that preferentially select hydrophilic phage peptides.

3 DISCUSSION

Here, we addressed the problem of predicting natural PDZ–peptide interactions using phage display data. We observed that phage

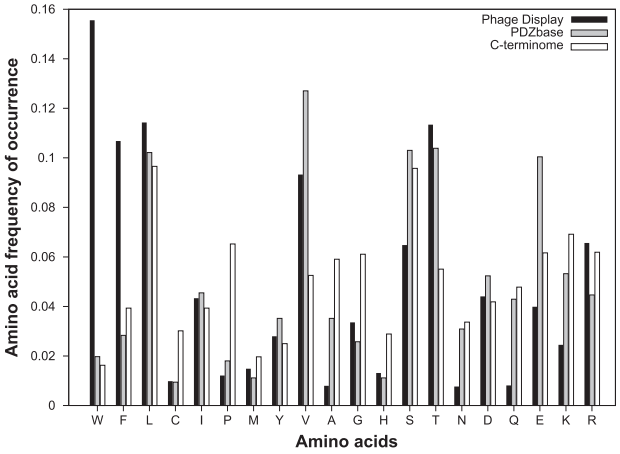


Fig. 3. Amino acid composition of phage peptides versus the human C-terminome and PDZ-binding peptides from the PDZbase. Amino acids are sorted from the most hydrophobic (left) to the most hydrophilic (right) according to the hydrophobicity scale of Kidera *et al.* (1985). All sequences were cut to a length of five residues.

peptide lists of Tonikian *et al.* (2008) can be classified on the basis of their hydrophobic character. Human C-termini matched better to the sequence profiles defined by the phage peptides, when they were hydrophilic. More specifically, human C-termini that are identical to phage peptides could be found more frequently for hydrophilic phage peptides. In addition, we realised that in average the phage peptides were much more hydrophobic than published natural PDZ-binding sequences as well as human C-termini in general. In particular, the phage display data showed a very strong preference for the largest aromatic amino acid Trp at peptide position p-1. All these results indicate that prediction of interactions between PDZs and naturally occurring peptides perform better when based on hydrophilic phage peptides.

It should be noted that short linear interaction motifs (Slims) have been found to display a particular amino acid composition, which distinguishes them from both folded and disordered regions (Fuxreiter *et al.*, 2007). The least conserved positions in Slims are usually non-hydrophobic, whereas the highest conserved positions are very often occupied by hydrophobic and charged amino acids. Indeed, published PDZ ligands generally agree with this trend, since the canonical PDZ-binding motif pattern consists of a hydrophobic (usually not aromatic) amino acid at peptide position p0 and Thr/Ser, hydrophobic (usually not aromatic) or Glu/Asp at position p-2. The phage display procedure of Tonikian *et al.* often selected such characteristics at positions p0 and p-2, but the other less conserved positions (p-1, p-3 and p-4) were, for two-thirds of the PDZ domains tested, very frequently hydrophobic, thereby deviating from sequence characteristics of Slims.

Biological interactions are characterized both by their affinity and specificity. Affinity represents absolute interaction strength, whereas specificity is a relative property derived from the comparison of interaction strengths of different interacting partners. For instance, if a PDZ domain binds with higher (but not necessarily high) affinity to a few peptides than to all others, it will be specific. Molecular dynamic studies (Basdevant *et al.*, 2006) have indicated that hydrophobic interactions are the most important force contributing

to PDZ–peptide affinity, and Beuming *et al.* (2009) have suggested that Trp at p-1 contributes strongly to the affnity of C-terminal peptides to Erbin PDZ domain via hydrophobic effects. In this regard, the phage display procedure, being mainly affinity driven, may have selected hydrophobic and especially aromatic amino acids at the least conserved positions of the PDZ-binding motif. However, transient interactions are required for PDZ-mediated cell signalling. In such a context, PDZ-binding hydrophobic sequences might turn out to be counter-productive due to an excessively high affinity. In addition, interactions involved in signalling also require specificity that might not be conferred by hydrophobic binders. Indeed, by examining SPOT data from Wiedemann *et al.* (2004), we observed that a 'super-binding peptide' with Trp at p-1 displaying high affinity for Erbin PDZ domain seemed to be robust against mutations at other peptide positions indicating a strong contribution of Trp to the binding affinity. Hence, the Trp at p-1, and hydrophobic residues at least conserved positions in general, would probably allow for more putative interaction partners to a PDZ domain and would make specific recognition impossible. In summary, it seems that the phage display approach has a tendency to select high affinity binders presenting artificial sequence features in contrast to evolution rather selecting for specific binders in the context of Slims. We notice that a similar conclusion has independently been drawn in a recent phage display study by Ernst *et al.* (2010). While this property of phage display may be useful for drug design or synthetic biology, it may limit its application for predicting natural domain–motif interactions. Recently, a promising approach was proposed to modify the phage display experimental protocol towards a procedure that will rather select specific than high affinity peptides (Hoffmann *et al.*, 2010).

Our study indicates that PDZ–peptide interaction predictions based on hydrophobic phage peptides should be considered carefully, especially with regard to specific, natural interactions, whereas predictions of interaction networks based on hydrophilic phage peptides are promising. We hypothesize that similar constraints in phage display data might also arise in the context of other types of domain–linear motif interactions. Given the wide use of phage display for the determination of binding specificities of domain–linear motif interactions, the problems addressed here might apply to many other studies as well.

ACKNOWLEDGEMENTS

We thank N. Davey, B. Kieffer and T. Gibson for helpful discussion of the manuscript.

Funding: Région Alsace (K.L.); CDE (K.L.); CNRS; ARC (Grant nr. 3171); and ANR (project ANR-MIME-2007 EPI-HPV-3D).

Conflict of Interest: none declared.

REFERENCES

Basdevant,N. *et al.* (2006) Thermodynamic basis for promiscuity and selectivity in protein-protein interactions: PDZ domains, a case study. *J. Am. Chem. Soc.*, **128**, 12766–12777.

Beuming,T. *et al.* (2005) PDZbase: a protein-protein interaction database for PDZ-domains. *Bioinformatics*, **21**, 827–828.

Beuming,T. *et al.* (2009) High-energy water sites determine peptide binding affinity and specificity of PDZ domains. *Protein Sci.*, **18**, 1609–1619.

Brannetti,B. *et al.* (2001) iSPOT: a web tool for the analysis and recognition of protein domain specificity. *Comp. Funct. Genomics*, **2**, 314–318.

Chen,J.R. *et al.* (2008) Predicting PDZ domain-peptide interactions from primary sequences. *Nat. Biotechnol.*, **26**, 1041–1045.

Diella,F. *et al.* (2008) Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Front Biosci.*, **13**, 6580–6603.

Doyle,D.A. *et al.* (1996) Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ. *Cell*, **85**, 1067–1076.

Ernst,A. *et al.* (2010) Coevolution of PDZ domain-ligand interactions analyzed by high-throughput phage display and deep sequencing. *Mol. Biosyst.*, **6**, 1782–1790.

Fuh,G. *et al.* (2000) Analysis of PDZ domain-ligand interactions using carboxyl-terminal phage display. *J. Biol. Chem.*, **275**, 21486–21491.

Fuxreiter,M. *et al.* (2007) Local structural disorder imparts plasticity on linear motifs. *Bioinformatics*, **23**, 950–956.

Hoffmann,S. *et al.* (2010) Competitively selected protein ligands pay their increase in specificity by a decrease in affinity. *Mol. Biosyst.*, **6**, 126–133.

Hui,S. and Bader,G.D. (2010) Proteome scanning to predict PDZ domain interactions using support vector machines. *BMC Bioinformatics*, **11**, 507.

Kalyoncu,S. *et al.* (2010) Interaction prediction and classification of PDZ domains. *BMC Bioinformatics*, **11**, 357.

Kidera,A. *et al.* (1985) Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *J. Protein Chem.*, **4**, 23–55.

Nourry,C. *et al.* (2003) PDZ domain proteins: plug and play! *Sci. STKE*, **2003**, RE7.

Runyon,S.T. *et al.* (2007) Structural and functional analysis of the PDZ domains of human HtrA1 and HtrA3. *Protein Sci.*, **16**, 2454–2471.

Schillinger,C. *et al.* (2009) Domain interaction footprint: a multi-classification approach to predict domain-peptide interactions. *Bioinformatics*, **25**, 1632–1639.

Sharma,S.C. *et al.* (2009) T7 phage display as a method of peptide ligand discovery for PDZ domain proteins. *Biopolymers*, **92**, 183–193.

Sidhu,S.S. *et al.* (2003) Exploring protein-protein interactions with phage display. *Chembiochem*, **4**, 14–25.

Skelton,N.J. *et al.* (2003) Origins of PDZ domain ligand specificity. Structure determination and mutagenesis of the erbin PDZ domain. *J. Biol. Chem.*, **278**, 7645–7654.

Smietana,K. *et al.* (2008) Degenerate specificity of PDZ domains from RhoA-specific nucleotide exchange factors PDZRhoGEF and LARG. *Acta Biochim. Pol.*, **55**, 269–280.

Smith,C.A. and Kortemme,T. (2010) Structure-based prediction of the peptide sequence space recognized by natural and synthetic PDZ domains. *J. Mol. Biol.*, **402**, 460–474.

Songyang,Z. *et al.* (1997) Recognition of unique carboxyl-terminal motifs by distinct PDZ domains. *Science*, **275**, 73–77.

Stricker,N.L. *et al.* (1997) PDZ domain of neuronal nitric oxide synthase recognizes novel c-terminal peptide sequences. *Nat. Biotechnol.*, **15**, 336–342.

Tonikian,R. *et al.* (2008) A specificity map for the PDZ domain family. *PLoS Biol.*, **6**, e239.

Vaccaro,P. *et al.* (2001) Distinct binding specificity of the multiple PDZ domains of INADL, a human protein with homology to INAD from Drosophila melanogaster. *J. Biol. Chem.*, **276**, 42122–42130.

van den Berk,L.C.J. *et al.* (2007) An allosteric intramolecular PDZ-PDZ interaction modulates PTP-BL PDZ2 binding specificity. *Biochemistry*, **46**, 13629–13637.

Waterhouse,A.M. *et al.* (2009) Jalview version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.

Wiedemann,U. *et al.* (2004) Quantification of PDZ domain specificity, prediction of ligand affinity and rational design of super-binding peptides. *J. Mol. Biol.*, **343**, 703–718.

9. Putting into practice domain–linear motif interaction predictions for exploration of protein networks

9.1. Summary

Stiffler *et al.* [10] published another large-scale study on PDZ–peptide interactions. In contrast to artificial PBMs revealed in the phage display study from Tonikian *et al.* [9], this study involved 217 cellular C-terminal sequences extracted from the mouse proteome that were assayed for binding towards 157 mouse PDZ domains. This interaction data has been used by Chen *et al.* [11] for the development of a PDZ interaction predictor (see section 6.4.4). In this article, we have described the results that we obtained when testing and applying this predictor to PDZ domains from the human proteins MAGI1 and SCRIB followed by experimental validation of predictions using SPR. In addition, we tried to assess changes in binding affinity and specificity when extending minimal interacting fragments.

Approach: Negative PDZ–peptide interaction data has been collected manually from published literature taking advantage of the multiple occurrence of PDZ domains in proteins. The Chen predictor had first to be implemented before tests with assembled and available PDZ–peptide (non)-interaction data sets could be performed. In the following, the Chen predictor has been used to predict out of the human proteome all potential binding partners for the PDZ domains of MAGI1 and SCRIB. Out of these predictions 17 C-terminal human peptides and 2 additional C-terminal viral peptides were selected for experimental validation versus PDZ2 and PDZ3 of MAGI1, PDZ3 and PDZ4 of SCRIB, and the tandem construct PDZ34 of SCRIB. Each peptide has been assessed for binding to each of these five PDZ domain constructs using a long (10 wild type residues) and a short (5 wild type residues) version of the peptide. We developed a medium-throughput protocol to measure these about 200 interactions on a BIAcore machine (SPR). Experimental data obtained has been analysed in light of available structural data.

Findings: Using our data set of real PDZ–peptide non-interactions, we could show that the Chen predictor has a very high FPR of about 50% (double as high as specified by the authors). This high FPR could be confirmed by experimental validation of selected predictions. In addition, initially predicted promiscuous binding behaviour of peptides towards PDZ domains could not be experimentally confirmed. Prediction scores did not correlate with measured binding affinities. The Chen predictor cannot discriminate between C-terminal sequences that carry the class I, II, or III signature

for PBMs and those that do not. The implementation of a filter has been necessary to remove the non-PBM sequences from predictions. The 16 domain residues that are considered by the predictor are not sufficient to describe the set of domain residues that are likely to be implicated in peptide binding. We could demonstrate that the binding affinity and specificity can change when extending the peptide and/or domain constructs. Changes in binding affinity upon peptide extension are likely to result from interactions between residues of extended peptide sequences and residues from the β2-β3 loop of PDZ domains.
Discussion/Conclusions: Inaccuracies in PDZ–peptide interaction predictions observed for the Chen predictor might originate from insufficient training of the underlying prediction model. PDZ4 of SCRIB might display a very specific binding profile due to deviations in the carboxylate binding loop sequence and a distribution of positive charges throughout the binding pocket. New potential binding partners for MAGI1 and SCRIB have been suggested that highlight the scaffolding role of PDZ proteins for G protein-related signalling pathways. Our data indicate that an extrapolation of observed interactions between minimal protein fragments to full length proteins may be possible qualitatively, but not necessarily quantitatively.
Contribution: I have performed the entire work involving the Chen predictor. I have selected predicted C-terminal PBMs for experimental validation. I have performed most of the SPR data treatment and have interpreted the experimental data in conjunction with available structural data. I have searched the published literature on MAGI1, SCRIB and the proteins for which we could experimentally validate binding of their C-terminal sequences to PDZ domains of MAGI1 or SCRIB, to come up with hypotheses about biological functions of these newly identified PDZ–peptide interactions. I have conceived and drafted a first version of the manuscript except of the experimental methods part. Gilles Travé and myself have written together the final version of the manuscript.
(See also supplemental material provided in section D.2.)

OPEN ACCESS Freely available online

PLoS one

Putting into Practice Domain-Linear Motif Interaction Predictions for Exploration of Protein Networks

Katja Luck[1], Sadek Fournane[1], Bruno Kieffer[2], Murielle Masson[1], Yves Nominé[1], Gilles Travé[1]*

1 Group Onco-Proteins, Institut de Recherche de l'Ecole de Biotechnologie de Strasbourg, 1, BP 10413, Illkirch, France, **2** Biomolecular NMR group, Institut de Génétique et de Biologie Moléculaire et Cellulaire, 1, BP 10413, Illkirch, France

Abstract

PDZ domains recognise short sequence motifs at the extreme C-termini of proteins. A model based on microarray data has been recently published for predicting the binding preferences of PDZ domains to five residue long C-terminal sequences. Here we investigated the potential of this predictor for discovering novel protein interactions that involve PDZ domains. When tested on real negative data assembled from published literature, the predictor displayed a high false positive rate (FPR). We predicted and experimentally validated interactions between four PDZ domains derived from the human proteins MAGI1 and SCRIB and 19 peptides derived from human and viral C-termini of proteins. Measured binding intensities did not correlate with prediction scores, and the high FPR of the predictor was confirmed. Results indicate that limitations of the predictor may arise from an incomplete model definition and improper training of the model. Taking into account these limitations, we identified several novel putative interactions between PDZ domains of MAGI1 and SCRIB and the C-termini of the proteins FZD4, ARHGAP6, NET1, TANC1, GLUT7, MARCH3, MAS, ABC1, DLL1, TMEM215 and CYSLTR2. These proteins are localised to the membrane or suggested to act close to it and are often involved in G protein signalling. Furthermore, we showed that, while extension of minimal interacting domains or peptides toward tandem constructs or longer peptides never suppressed their ability to interact, the measured affinities and inferred specificity patterns often changed significantly. This suggests that if protein fragments interact, the full length proteins are also likely to interact, albeit possibly with altered affinities and specificities. Therefore, predictors dealing with protein fragments are promising tools for discovering protein interaction networks but their application to predict binding preferences within networks may be limited.

Citation: Luck K, Fournane S, Kieffer B, Masson M, Nominé Y, et al. (2011) Putting into Practice Domain-Linear Motif Interaction Predictions for Exploration of Protein Networks. PLoS ONE 6(11): e25376. doi:10.1371/journal.pone.0025376

Editor: Anna Tramontano, University of Rome, Italy

Received May 11, 2011; **Accepted** September 2, 2011; **Published** November 1, 2011

Funding: This work was supported by CNRS, University of Strasbourg, Ligue Nationale Contre le Cancer, Association de Recherche contre le Cancer (ARC) grant 3171, Agence Nationale de la Recherche (ANR) programs ANR-06-BLAN-0404 and ANR-MIME-2007 (project EPI-HPV-3D), and National Institutes of Health (NIH) grant R01CA134737. KL was supported by a grant of the "Région Alsace". SF was supported by grants from the Ligue Nationale contre le Cancer. KL and SF were supported by the Collège Doctoral Européen de Strasbourg. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: gilles.trave@unistra.fr

Introduction

Many of the protein interactions that function in cellular regulation and signalling are mediated by linear motifs that bind to globular domains. Such interactions are often specific, yet transient and therefore of low affinity [1]. The efficient prediction of such interactions together with their experimental validation would enormously increase our understanding of the cellular system. The occurrence of specific types of globular domains in protein sequences can mostly be predicted with high accuracy [2] [3] and promising work on linear motif predictions are published [4][5]. However, the correct prediction of which instance of a linear motif will bind to which instance of a type of globular domain, hence the specificity in domain - linear motif interactions, remains one of the hot topics in computational biology.

Approaches for predicting domain-linear motif interactions have very often focussed on PDZ-peptide interactions. PDZs are a very abundant class of globular domains with 267 occurrences in the human proteome [6]. Human proteins often contain several copies of PDZs (up to 13) in their sequence. PDZs bind with a well defined pocket to linear motifs that are mostly situated at the extreme C-termini of proteins. The last residue (referred to as position p0) in PDZ-binding motifs is usually Val or Leu. The third last peptide residue (position p-2) can be either Thr or Ser (class I), hydrophobic (class II), or Glu or Asp (class III), thereby defining three main categories of PDZ-binding motifs [7][8]. 339 experimentally verified PDZ-peptide interactions are currently annotated in the PDZbase [9] and 212 PDZ structures are listed in the ADAN database [10] indicating that PDZs are very well experimentally studied.

PDZs are implicated in the regulation of cell polarity, cell adhesion and intercellular communication [11]. The PDZ-containing proteins MAGI1 (Membrane-associated guanylate kinase inverted 1) and SCRIB (human Scribble) are in the centre of this study. MAGI1, which has six PDZ domains, was found to be located to adherens and tight junctions in epithelial [12] and endothelial cells [13], where it seems to be involved in the maintenance of the junctions and in cell signal propagation. SCRIB, which has four PDZ domains, is known to be involved in the establishment of adherens [14] and tight junctions [15] as well as in the regulation of cell polarity and cell migration [16]. Some data indicate that deregulation of MAGI1 [17] or SCRIB [18] can

promote cell proliferation and tumorigenesis. Interestingly, proteins from different viruses were shown to bind via their C-terminal sequences to MAGI1 or SCRIB and to interfere with their cellular functions for promoting viral replication [19] [20]. For instance, the oncoprotein E6 produced by the human papillomaviruses (HPV) responsible for cervical cancer contains a PDZ-binding motif, which interacts with PDZ domains of MAGI1 and SCRIB [21] [22]. Deletion of this motif in HPV16 E6 impaired its capacity to promote cancer in transgenic mice [23] indicating that binding of E6 to MAGI1 and SCRIB might be implicated in the development of cervical cancer. Therefore, it would be important to better understand the signalling pathways, such as those of cell growth and apoptosis, that are regulated by MAGI1 and SCRIB and that are disrupted upon infection with oncoviruses such as HPV.

Until recently, only specific case studies had been published on the specificity of PDZ-peptide interactions, and the iSPOT tool [24] was for a long time the only attempt to predict PDZ-peptide interactions on a broader scale. In 2007 and 2008, two groups published outstanding large-scale studies on PDZ interactions providing insights into PDZ interaction specificities and strategies for their prediction [25] [26] [27]. Tonikian *et al.* [25] applied phage display to determine the binding profiles of 28 *C. elegans* and 54 *H. sapiens* PDZ domains using 10 billion random peptides. Stiffler *et al.* [26] applied microarrays and fluorescence polarisation to measure binding affinities between 157 mouse PDZ domains and 217 mouse peptides. All interactions and non-interactions (absence of interactions) determined by Stiffler *et al.* were used by Chen *et al.* [27] as training data for a PDZ interaction predictor. The prediction model was defined using the structure of the α1-syntrophin PDZ domain bound to a seven residue-long peptide of which five are visible in the structure [28]. The model consists of 38 position pairs of domain and peptide residues that were seen to interact with each other in this particular structure. The training data was used in a Bayesian approach to obtain sub-scores for the occurrence of all possible combinations of amino acid pairs at these 38 position pairs. These sub-scores quantify the positive, neutral or negative contribution of a pair of amino acids at a certain position to the overall interaction between a PDZ domain and a peptide. The sum of the 38 sub-scores for a given PDZ-peptide pair represents the final score, which was suggested to indicate the binding strength of the potential interaction in question.

A very critical point for the development of protein interaction predictors is the availability of real negative interaction datasets [29]. Stiffler *et al.* [26] provide a negative PDZ interaction dataset, which has already been used to significantly improve PDZ interaction prediction quality [30][31]. However, this negative dataset is the only one existing so far, which implies that PDZ interaction predictors trained with data of Stiffler *et al.* [26], such as the predictor of Chen *et al.* [27], cannot be tested on an independent negative dataset.

The numerous existing predictors for PDZ-peptide interaction specificities focus on the core PDZ domain or binding pocket of the PDZ and mostly on four or five residue long peptides [27] [30] [31] [32] [33] [34] [35]. Generally, it is assumed that interaction specificity predictions based on such protein fragments are also valid in the context of full length protein interactions and hence can be used to predict protein-protein interaction (PPI) networks. However, an increasing amount of biological studies on PDZ domains suggest that peptide residues upstream of the last five residues and domain residues outside of the binding pocket influence binding affinity and specificity [36] [37] [38] [39] [40]. Linker regions flanking the core PDZ domain as well as neighbouring domains, have also been found to influence binding [41] [42]. The term supramodule was introduced for neighbouring PDZs that are separated by particularly short linker sequences and that were shown to significantly influence each other's peptide binding (for a review see [43]).

Based on these observations, several questions are raised: First of all, how correct are PDZ interaction predictors in theory and in practice? Second, to which extent can specificity predictions based on protein fragments be transferred to full length proteins and how much influence do extensions of protein fragments have on affinity and specificity of the corresponding interaction? Third, can existing PDZ interaction predictors be used to extend our knowledge on PPI networks mediated by PDZ-peptide interactions? Here, we attempted to answer these questions by focussing on the well studied predictor published by Chen *et al.* [27]. First, we aimed at assessing its prediction quality *in silico* by using test datasets assembled by ourselves that consisted of real positive and negative interaction data for various PDZ domains. Then, by concentrating on PDZ domains of MAGI1 and SCRIB, we performed proteome-wide interaction predictions and experimentally validated a subset of those, allowing us to also assess the prediction quality *in vitro*. We also assessed how binding was influenced by extended protein fragments, i.e. peptides and PDZ constructs longer than those considered by the predictor. Finally, discovered interactors for MAGI1 and SCRIB were analysed with regard to new biological functions that can be linked to MAGI1 and SCRIB and that might be perturbed in tumours induced by oncoviruses or other factors. In total, this analysis allowed to highlight the power and limits of PPI network predictions involving PDZ domains, to uncover possible ways of improvements, and to obtain further insights into the mechanisms that define affinity and specificity of PDZ-peptide interactions.

Results

Development of real negative test datasets for benchmarking PDZ interaction predictors

We aimed at assessing the performance of the PDZ interaction predictor published by Chen *et al.* [27] with independent datasets of human PDZ-peptide interactions from low-throughput experimental studies. We assembled three test datasets (see Dataset S1) containing interactions and non-interactions involving 95 different human PDZ domains. The first test dataset contained 174 PDZ-ligand interactions including 109 human interactions from PDZbase [9] (a resource of experimentally verified PDZ-ligand interactions) plus 65 interactions that we manually collected from literature, mainly dealing with PDZ domains from MAGI1, 2 and 3. The PDZ domains from MAGI1, 2 and 3 are identical between human, mouse and rat when concentrating on the 16 domain amino acid positions used for predictions by Chen *et al.* Therefore, we included in the datasets interactions that we expect to occur between human proteins although they were originally described in the literature using rat and mouse PDZ domains.

The second and third test dataset contain negative interaction data that were assembled from published literature as follows. We took advantage of the particular characteristic of PDZ domains to occur as repeats within proteins (as illustrated in Figure 1). In order to experimentally determine the PDZ domain to which a peptide will bind out of the PDZ domains of a particular protein, each PDZ domain of the protein is tested separately for binding to the peptide. This approach usually yields one genuine interaction and many non-interactions. These non-interactions were annotated into one negative test set that in total contained 446 human non-interactions involving peptides bearing a PDZ-binding motif. The

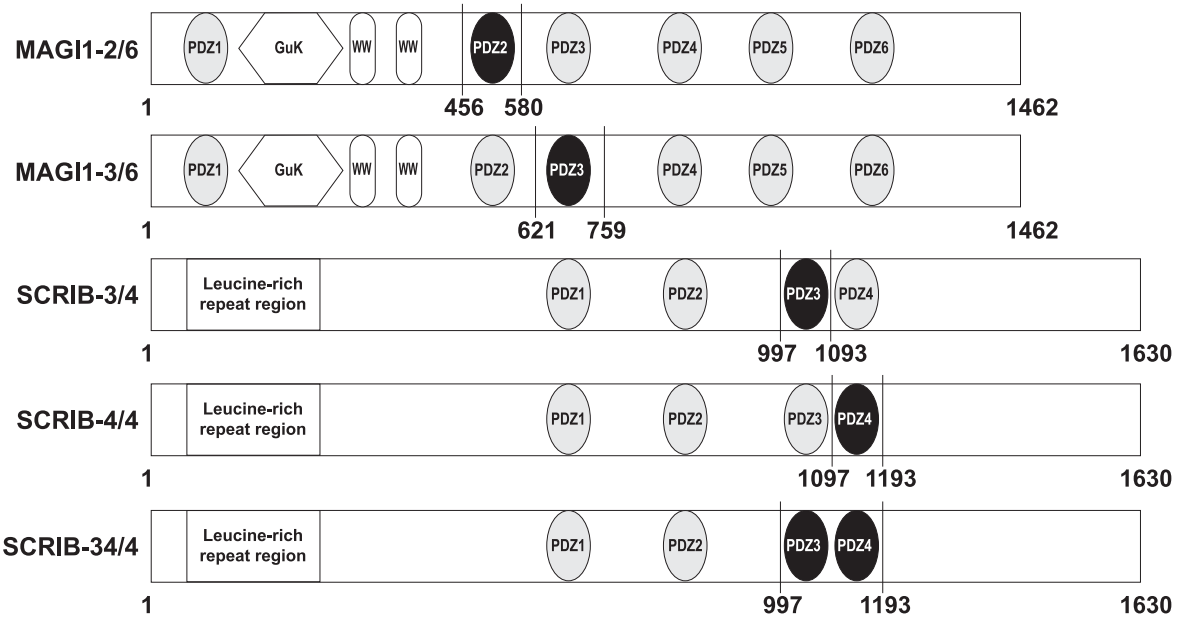


Figure 1. PDZ domains of MAGI1 and SCRIB. MAGI1 has 6 PDZ domains numbered from 1 to 6. SCRIB has 4 PDZ domains numbered from 1 to 4. The PDZ domains that were used for interaction measurements by SPR are highlighted in black and used domain boundaries are indicated.
doi:10.1371/journal.pone.0025376.g001

third test dataset contains 133 human non-interactions collected from the literature where the peptide has a disrupted PDZ-binding motif due to introduced mutations (substitutions or deletions). These real negative experimental data can be expected, as argued by Smialowski *et al.* [29], to outperform artificial negative data (such as randomised protein interactions) in terms of training and test performance.

Benchmarking the PDZ-ligand interaction predictor of Chen *et al.*

When tested on the three established test datasets (Table 1) the predictor of Chen *et al.* obtained a sensitivity of 75.3% in agreement with that indicated by Chen *et al.* (76.5%) [27]. By contrast, the false positive rate (FPR) based on non-interactions with PDZ-binding motifs is about 48%, which is considerably higher than the FPR indicated by Chen *et al.* (24%). Furthermore, the FPR obtained for non-interactions without PDZ-binding motifs is about 26%, which represents a weak performance with regard to the relatively straightforward task to discriminate between peptides that bear a prototypical PDZ-binding motif or not. We then analysed separately, within our test datasets, the data involving human PDZ domains that are either orthologous or not orthologous to the mouse PDZ domains present in the training set of Chen *et al.* Sensitivity and FPR of these subsets show that the predictor tends to be over-optimistic for PDZ domains that are orthologous to domains present in the training data, and over-pessimistic for PDZ domains that are not orthologous to any domain present in the training data (third and fourth column in Table 1).

Our test datasets contain a large portion of interactions and non-interactions involving PDZ domains from MAGI1, 2 and 3. We separately calculated the sensitivity and FPRs of the predictor for subsets of the test datasets consisting only of PDZ domains of MAGI1, 2 and 3 (fifth column in Table 1). The results are considerably different from those obtained with the full datasets, indicating that the MAGI subset does over-influence the calculations.

Prediction of natural PDZ-peptide interactions using the predictor of Chen *et al.*

The predictor of Chen *et al.* [27] was applied to PDZ domains of MAGI1 and SCRIB (see Figure 1 for the domain organisation of these proteins) with the aim of predicting, from the entire human proteome, natural interacting partners for these PDZs. For most domains, the numbers of predicted hits (proteins) were very high (Table 2, second column). An important proportion of these hits might be false positives in relation to the previously observed high FPR (Table 1). Indeed, one third of the C-terminal sequences of the returned hits had a non-hydrophobic amino acid at peptide

Table 1. Performance of predictor of Chen *et al.* for different test data sets.

	complete test data		training[a]		non-training[b]		MAGI1,2,3[c]	
sensitivity[d]	75.3%	(174)	90.7%	(97)	55.8%	(77)	65.9%	(41)
FPR PDZ[e]	48.2%	(446)	53.5%	(213)	43.3%	(233)	17.5%	(240)
FPR NoPDZ[f]	25.6%	(133)	27.6%	(58)	24.0%	(75)	4.0%	(50)

[a]test data containing only (non)-interactions with PDZ domains orthologous to those from the training data of Chen *et al.*
[b]test data containing only (non)-interactions with PDZ domains that were not orthologous to those in the training data of Chen *et al.*
[c]test data containing only (non)-interactions with PDZ domains from MAGI1, 2 and 3 proteins. These subsets were analysed to verify that the overrepresentation of PDZ domains from these proteins did not introduce a bias in calculated sensitivity and specificities.
[d]percentage of interactions that were correctly predicted.
[e]percentage of non-interactions with PDZ-binding motif that were not correctly predicted.
[f]percentage of non-interactions without PDZ-binding motif that were not correctly predicted.
The numbers in brackets represent the total number of items in the respective test data set.
doi:10.1371/journal.pone.0025376.t001

Table 2. Numbers of human proteins predicted to bind to PDZ domains of MAGI1 and SCRIB using the predictor of Chen *et al.*

PDZ domain	unfiltered hits	filtered hits[a]	num. prots. with highest score[b]	num. publ. binders[c]
MAGI1-1/6	0	0	0	1
MAGI1-2/6	457	300	93	4
MAGI1-3/6	160	107	0	1
MAGI1-4/6	43	30	0	3
MAGI1-5/6	1151	623	562	3
MAGI1-6/6	219	179	87	6
SCRIB-1/4	204	89	1	4
SCRIB-2/4	429	203	98	1
SCRIB-3/4	744	293	237	5
SCRIB-4/4	354	113	3	1

[a]proteins without residue C, Y, F, L, I, M, V, W or A at peptide position p0 were filtered out.
[b]numbers of proteins, which were predicted (after filtering) to bind to that domain and scored highest for that domain in comparison to the other domains.
[c]numbers of published mammal binders that we could identify from literature for each PDZ domain.
doi:10.1371/journal.pone.0025376.t002

position p0, in contradiction with most published literature concerning PDZ-binding sequence requirements. We analysed the amino acid composition of the pool of peptide sequences used to train the predictor of Chen *et al.* (Table S1) and observed that this pool of sequences had only V, L, I, F, C or A at position p0. This is due to the fact that the entire training pool of Chen *et al.* contained exclusively peptides that bound at least to one PDZ domain in the experiments of Stiffler *et al.* [26] and hence represent PDZ-binding sequences. In the training process, Chen *et al.* allocated zero (representing a neutral value) to all amino acids that were never seen at particular peptide positions. Whereas this strategy is sound when applying the predictor to peptides matching the general PDZ-binding consensus, it may lead to the selection of irrelevant peptides when querying an entire proteome. To take this issue into account, we applied an additional filter to accept only peptides ending with either C, Y, F, L, I, M, V, W or A, i.e. residues that were observed at position p0 in artificial or natural PDZ-binding peptides. This filter rejected 20 to 60% of the initial hits (Table 2, third column) and was systematically used further on in our study. Detailed information on the predicted interactions is provided in Dataset S2.

As shown in Table 2 (third column), some domains (e.g. MAGI1-5/6 - the fifth out of six PDZ domains of MAGI1) appeared to be very promiscuous as they had a very high number of hits, whereas others (e.g. MAGI1-4/6) had very few hits or even no hit at all (MAGI1-1/6). Within both MAGI1 and SCRIB, the PDZ domains obtaining the highest numbers of hits (MAGI1-5/6, 2/6 and 6/6, and SCRIB-2/4 and 3/4) were also the ones that obtained the highest scores (Table 2, fourth column). This might be correlated with our observation that scores obtained by different domains were distributed over different ranges (Figure 2). While investigating why some domains (e.g. MAGI1-5/6) showed higher scores and higher numbers of hits, we observed that particular peptide residues contributed very high subscores to the overall score for a domain-peptide pair. For instance, the occurrence of a Thr at position p-2 (a characteristic common to all class I PDZ-binding motifs) contributed a value of 0.64 to the prediction score for binding to MAGI1-5/6, while the overall value sufficient for a peptide to be classified as a hit by the preditor is 0.5. This means that any peptide possessing a Thr at position p-2 and residues at other positions that confer a predicted globally neutral effect for binding, would be classified as a binder for the MAGI1-5/6 domain. At present, we do not know whether this characteristic of MAGI1-5/6 is biologically meaningful or whether it just reflects some bias of the predictor's algorithm. Indeed, the predictions differ from published biological data (Table 2, fifth column), which indicate that the PDZ domain of MAGI1 attracting most binders is MAGI1-6/6, rather than MAGI1-5/6.

We also observed (Table S2) that numerous proteins were predicted to bind to more than one PDZ domain of MAGI1 or SCRIB, indicating that not only PDZ domains, but also C-terminal peptides, are considered to be promiscuous by the predictor. This may just originate from the lack of specificity of the predictor as already pointed out before in our analysis (see Table 1). However some PDZ-peptide interactions may indeed be really promiscuous and the predictor may be able to detect this trend.

Structure-based analysis of domain amino acid positions implicated in peptide binding

In the prediction model of Chen *et al.* 16 domain and 5 peptide positions were selected for being implicated in specific binding of peptides to PDZs. This selection was based on one structure, α1-syntrophin [28] (Figure 3). The structural information on PDZs has considerably grown during the last years mainly due to structural genomics initiatives. Here, we comparatively analysed 42 structural complexes of 24 different PDZ domains to get a more general overview about amino acids involved in peptide recognition. Figure 4 shows that the set of domain amino acids found at less than 5 Å from the peptide in the various structures we analysed often differs from the set defined by Chen *et al.* in the structure of α1-syntrophin (these positions are indicated with asterisks above the alignment). For instance, domain positions Leu37 (α1 helix) and Thr74 (α2–β5 loop) in α1-syntrophin (Figure 4), chosen by Chen *et al.*, were only selected once in the 23 other PDZ domains we analysed. Conversely, our approach (see Methods) selected more amino acids on α2 helix. In addition, while Chen *et al.* did not select any amino acid upstream of the GLGF-motif, our approach often selected residues in that region, especially a conserved positively charged position (Arg or Lys) within the β1-β2 loop. The role of this amino acid for peptide binding is discussed in several studies [44] [45] [46]. Finally, our

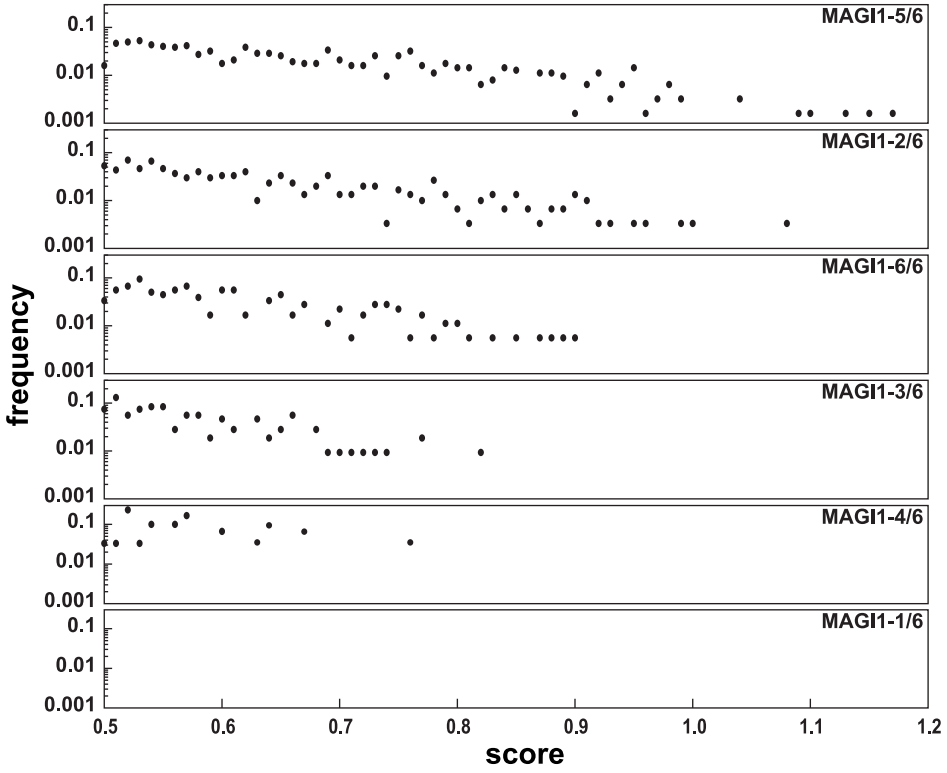


Figure 2. Score distribution of human C-terminal peptides predicted to bind to MAGI1 PDZ domains. Predictions were prefiltered for peptides having either C, Y, F, L, I, M, V, W or A at peptide position p0. Prediction scores were rounded to two decimal places and the frequencies of occurrence of scores within each interval were determined for each PDZ domain of MAGI1.
doi:10.1371/journal.pone.0025376.g002

analysis often selected amino acids of the $\beta2$-$\beta3$ loop, whereas only one residue of that loop was selected in Chen *et al.*'s study. The selection of residues of the $\beta2$-$\beta3$ loop indicates that residues upstream position p-4 are proximal to this loop and therefore may also contribute to binding (Figure 3). Altogether, we suggest that more domain and peptide positions than those defined by Chen *et al.* may influence binding specificity.

Experimental validation of predicted MAGI1-peptide and SCRIB-peptide interactions

From predictions obtained with the predictor of Chen *et al.* we selected 17 human and three viral peptides for interaction measurements against five PDZ constructs: the four single PDZ domains MAGI1-2/6, MAGI1-3/6, SCRIB-3/4, SCRIB-4/4, and the tandem construct SCRIB-34/4 (Figure 1). The 17 human peptides were selected based on different criteria: First, we selected peptides that were predicted to bind promiscuously to all four single PDZ domains. Second, we systematically included the two best predicted hits for each of the four PDZ domains. Third, we preferred proteins already shown to interact with PDZ domains. Further selection criteria were sequence diversity within the set of selected peptides and biological functions related to known functions of MAGI1 and SCRIB. These were inferred from Gene Ontology annotations (Ensembl v52 [47]) and information provided by UniProt [48]. The three viral peptides correspond to the C-terminus of HTLV1 Tax1, HPV16 E6, and a mutated form of HPV16 E6 (further on called 16E6L/V), where Leu at position p0 was mutated to Val. The latter peptide was already assayed against MAGI1 and SCRIB PDZ domains in previous SPR studies performed by our group, and therefore we used it as positive control for the present study. Table S3 provides detailed information about the 19 proteins.

For each of these 19 proteins two peptides were designed, *both of ten amino acids in length*. One peptide, called "long", encompassed the last ten wild type residues of the protein (e.g. VMRLQSETSV for VANG2). The other peptide, called "short", encompassed the last five wild type amino acids of the protein preceded by a GSGAG sequence (e.g. GSGAGSETSV for VANG2). This GSGAG sequence, composed of small neutral residues, was included to prevent the biotin tag N-terminally attached to the peptides to influence the binding to the PDZ domain. The "short" peptides, in which only the last five residues vary and correspond to natural proteins, would allow us to experimentally validate interaction predictions obtained with the predictor of Chen *et al.* that considers the last five residues in the prediction model. The long peptides (as well as the tandem PDZ construct) would allow us to address changes in binding affinity and specificity that might occur when using extended protein fragments.

We opted for the surface plasmon resonance (SPR) method to measure these 190 (19 proteins×2 peptide versions×5 PDZ constructs) interactions. In SPR various concentrations of "analytes" (here, PDZ domains fused to the Maltose Binding Protein (MBP)) flow over surfaces presenting attached "ligands" (here, biotinylated peptides). The amount of analyte interacting with the ligand is measured and quantified in response units (RU). The intensity of this signal is proportional to the binding strength of the assayed interaction (Figure 5A). K_D were obtained using a 1:1 interaction model. However, these calculated K_D were rather inaccurate especially for weak interactions. Therefore, we

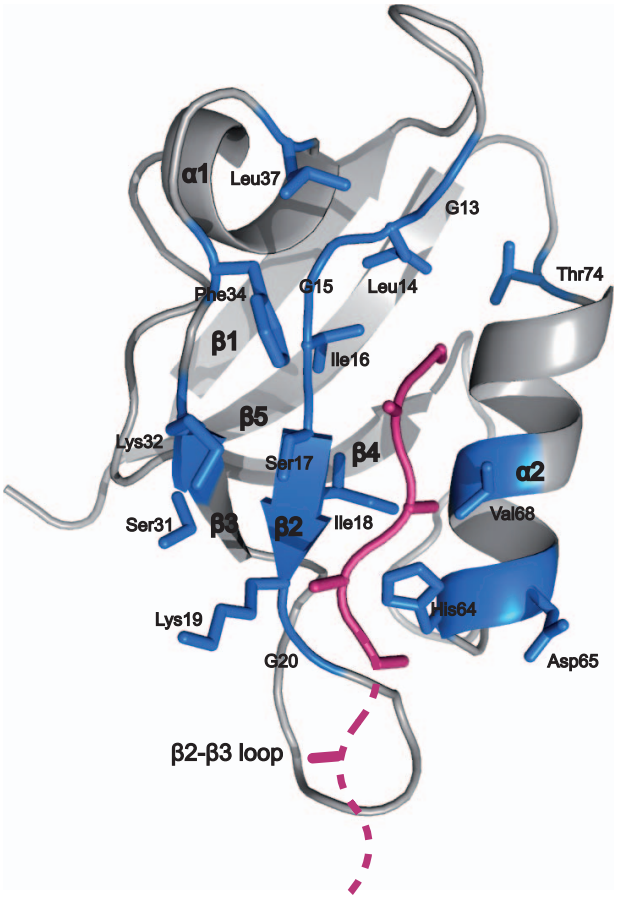


Figure 3. Structure of the PDZ domain of a1-syntrophin used as reference by Chen *et al.* Residues coloured in blue represent the domain positions that are considered in the prediction model of Chen *et al.* The backbone and C_β atoms of the bound peptide are represented as sticks in pink. The pink dashed line indicates where peptide residues upstream position p-4 would be situated in the structure. (PDBcode: 2PDZ).
doi:10.1371/journal.pone.0025376.g003

preferred to rank the binding strengths of the 190 interactions using normalised RU signals at equilibrium (R_{eq}) rather than K_D (see Methods for details). These normalised R_{eq} values were plotted in form of a heat map (Figure 5B). Table S4 contains experimental data for all SPR measurements performed in this study.

Nine out of nine published interactions (including 16E6L/V) were confirmed by our experimental data, of which three out of four published K_D could be confirmed as well, all being high affinity interactions (see Table S3 for more details). This demonstrates the validity of our experimental SPR setup for testing PDZ-peptide interactions.

Peptides do not bind as promiscuously as predicted to PDZ domains

Most tested peptides had been predicted to bind promiscuously to all four single PDZ domains (see Figure 5B, zeros indicate the very few PDZ-peptide pairs predicted not to interact). In practice, the peptides turned out to be much more selective than predicted. Only one peptide, TAX1 (derived from a viral protein), was found to interact with the four PDZ domains, and only at the condition of taking a very weak interaction into account. Even when we discarded the SCRIB-4/4 domain (which bound only one peptide as will be discussed later), we observed that, out of the 16 peptides predicted to bind the remaining three single PDZ domains, only 8 could be confirmed (see Figure 5B, underlined peptide names), again only at the expense of accepting very low interaction signals. This appears to confirm the high false positive rate of the predictor of Chen *et al.* that we have previously noticed (Table 1).

The prediction scores do not correlate with interaction affinities

Chen *et al.* have observed a correlation between prediction scores and binding affinities. In our set of data (19 short peptides vs. 4 single PDZ domains), we did not observe such correlation (for MAGI1-2/6 Pearson correlation coefficient $r = 0.44$ p-value = 0.07, for MAGI1-3/6 $r = 0.13$ p-value = 0.64, for SCRIB-3/4 $r = 0.1$ p-value = 0.69, for SCRIB-4/4 $r = -0.08$ p-value = 0.74) (Figure 6). In particular, the two best predicted hits for each PDZ domain turned out to be non-interactions or very weak interactions in all cases except one (Figure 5B, rectangles).

SCRIB-4/4 may display very specific binding preferences

SCRIB-4/4 was found to significantly bind to only one peptide, TAX1, despite of the fact that SCRIB-4/4 was predicted to bind to 15 out of the 19 peptides tested (Figure 5B). Remarkably, Zhang *et al.* [49] previously noticed that the SCRIB-4/4 domain did not bind any peptide in a phage display experiment. They interpreted this observation by suggesting that recombinant SCRIB-4/4 might be less stable than other PDZ domains. This possibility can be excluded, since we produced highly concentrated folded SCRIB-4/4 for NMR studies (data not shown), and the NMR structure of folded SCRIB-4/4 was solved by the RIKEN Structural Genomics Initiative (PDB code: 1UJU). We suggest that SCRIB-4/4 displays very specific peptide binding preferences, which can be inferred from analysis of available protein structures. We retrieved from the PDB the experimental structures of MAGI1-2/6, MAGI1-3/6 and SCRIB-4/4, and modelled the structure of SCRIB-3/4 (see Methods). The surface electrostatics representations of the four PDZ domains (Figure 7A) show that, in comparison to the other three PDZ domains, SCRIB-4/4 possesses many positive charges surrounding the peptide binding pocket. This should favour peptide sequences with negatively charged residues at position −1 and −3.

The "GLGF-loop", which precedes the β2 strand, coordinates the C-terminal carboxyl group of the peptide and also influences the width of the pocket accomodating the hydrophobic residue at p0 [45]. The first glycine of the "GLGF-loop" is replaced by a bulky arginine residue in SCRIB-4/4 (Figure 7B). This may sterically prevent binding of a peptide presenting a large hydrophobic side chain at p0 and might explain the shallow appearance of the pocket accommodating the peptide residue p0 (Figure 7A). These size and charge constraints may impose sequence properties only found in TAX1 (ETEV) out of the 19 peptides tested.

Different preferences of PDZ domains for residues at peptide position p0

Our interaction data reveal different binding preferences of the PDZ domains for specific hydrophobic amino acids at peptide position p0 (Figure 5B and Figure 8, see green residues at p0 in peptide sequences). SCRIB-3/4 seems to accept larger hydrophobic residues at p0 with a preference of leucine over valine. Indeed, SCRIB-3/4 binds stronger to wild type 16E6 as compared to the

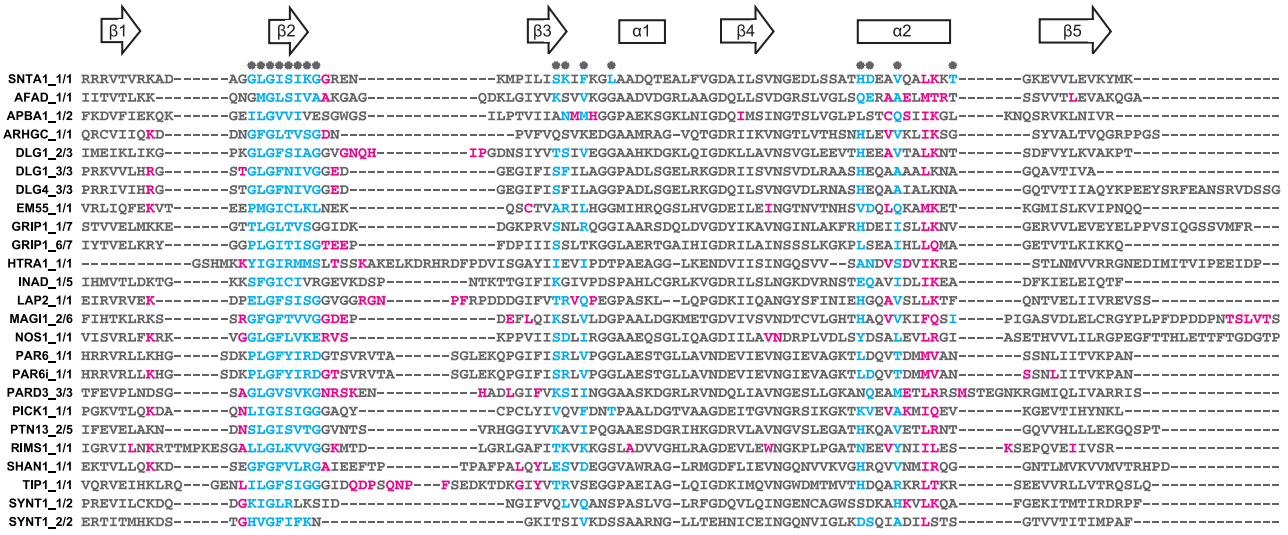


Figure 4. Atomic distance-based selection of peptide-contacting domain positions in different PDZ-peptide structures. For each PDZ domain of the alignment, we extracted from available structural data all domain residues that had at least one atom within a distance of 5 Å to bound peptide atoms. Blue letters indicate residues, which have been selected both, by Chen *et al.* and our approach. Red letters indicate residues, which have been selected by our approach but not by the model of Chen *et al.* Asterisks above the alignment indicate the PDZ residues chosen by Chen *et al.* to be close to peptide residues based on the structure α1-syntrophin (SNTA1, first line of alignment). Arrows and rectangles above the alignment indicate the positions of conserved β-sheets and α-helices, respectively. Note that the sequence of the Par6 PDZ domain occurs twice in the alignment, corresponding to two different structures of Par6, one bound to an internal peptide, the other one bound to a regular C-terminal peptide.
doi:10.1371/journal.pone.0025376.g004

single mutant 16E6L/V, where the last residue of 16E6 has been mutated from leucine to valine. In contrast, MAGI1-2/6 binds stronger 16E6L/V than wild type 16E6, showing that MAGI1-2/6 preferentially accommodates valine in comparison to leucine. This was also observed by Thomas *et al.* [50] using full length E6 proteins. MAGI1-3/6 only accepts valine.

These different preferences for amino acids at p0 might be again correlated with amino acid variations in the conserved "GLGF-loop". The alignment in Figure 7B shows that the two conserved hydrophobic positions of the "GLGF-loop" are occupied by phenylalanine residues in both MAGI1-2/6 and MAGI1-3/6 *vs.* two leucine residues in SCRIB-3/4. This might contribute to a wider pocket in SCRIB-3/4, explaining the preference of this domain for a C-terminal leucine in the bound peptide.

These different preferences for residues at p0 were only partially correctly predicted for MAGI1-2/6 and MAGI1-3/6 by the predictor of Chen *et al.* The predictor failed to predict these amino acid preferences for SCRIB-3/4 (see Dataset S2).

Binding affinities and specificities change for extended interaction fragments

We observed that the tandem construct SCRIB-34/4 bound several peptides with higher affinity as compared to the single domain constructs SCRIB-3/4 and SCRIB-4/4 (Figure 5B). This increase seemed not to depend on the sequence of the peptides.

In addition, we observed that the long peptides often bound PDZ domains with different affinities as compared to the short peptides (Figure 5B). As highlighted in Figure 3, the additional wild type residues present in the long peptides, upstream position p-4, are likely to engage interactions with residues in the $\beta2$-$\beta3$ loop of the PDZ domains. Figure 8 shows part of the structures of the PDZ domains MAGI1-2/6, MAGI1-3/6 and SCRIB-3/4 comprising the region, where the $\beta2$-$\beta3$ loop is situated (see Figure 7B for an alignment). Next to the structures, the differences in RU signals between long and short peptides are ranked from the greatest difference to the lowest. MAGI1-2/6 has four negatively charged residues in the $\beta2$-$\beta3$ loop and shows strong increases in affinity for long peptides having positively charged residues at peptide positions upstream p-4. The closer these positively charged residues are positioned to p-4, the bigger is the increase in affinity for long versions of peptides. By contrast, negative charges at these peptide positions appear to be disadvantageous (Figure 8A). MAGI1-3/6 did not show significant differences in affinity and specificity between short and long peptides. This observation may be explained by the fact that the $\beta2$-$\beta3$ loop contains four consecutive glycine residues unlikely to influence peptide binding (Figure 8B). SCRIB-3/4 shows an unspecific increase in affinity for many long peptide versions. The $\beta2$-$\beta3$ loop of SCRIB-3/4 is twice as long as for the other two PDZ domains and contains amino acids of diverse physico-chemical properties (Figure 8C). This loop might be able to adapt conformationally to many different sequences upstream of peptide position p-4, therefore providing advantageous contacts in most cases.

Discussion

In this study we addressed the problem of predicting naturally occurring protein interactions mediated by PDZ domains and PDZ-binding peptides using the predictor of Chen *et al.* [27]. We analysed the predictor using theoretical and practical approaches. An important step for a fair assessment of prediction qualities is the application of real test datasets independent from the training data. To ensure this, we assembled a novel dataset of real negative PDZ-peptide interactions from the literature, which might turn out to be very useful for further development of PDZ interaction predictors.

Both the *in silico* and *in vitro* tests indicated that prediction accuracies were weak. We could demonstrate that the predictor of

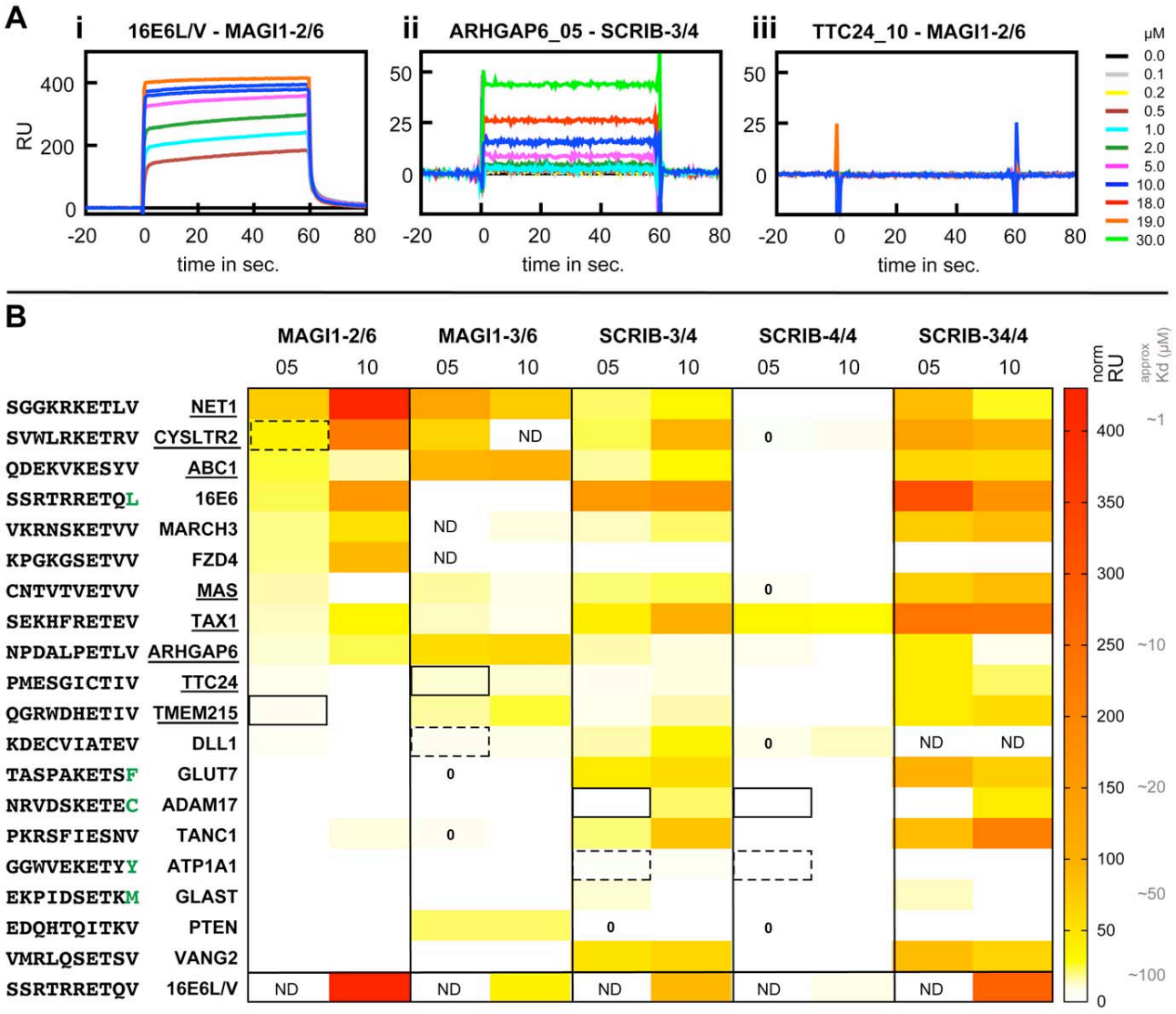


Figure 5. Overview of SPR experimental data. A: Representative sensorgrams for strong and weak interactions as well as non-interactions. An increase of the signal for injection of MBP-PDZ analyte is indicative of binding. (i) The higher the analyte concentration, the higher the R_{eq} up to saturation, indicative of a specific interaction. (ii) For weak interactions the highest analyte concentration, which was injected due to device limitations, did not allow to reach saturation. (iii) Sensorgrams for non-interactions display no change in signal. B: Overview of measured RU signals and comparison to predictions. Normalised RU signals determined for a 10 µM concentration of MBP-PDZ were extracted from SPR sensorgrams and plotted as heatmap for 19 peptides in short and long versions *vs.* the five PDZ constructs MAGI1-2/6, MAGI1-3/6, SCRIB-3/4, SCRIB-4/4 and SCRIB-34/4. An approximate range of K_D is indicated at the right side of the heatmap. 05 and 10 indicate short and long versions of peptides, respectively. ND = not determined. Signals of short peptides interacting with single PDZ constructs were compared to interaction predictions performed with the predictor of Chen *et al.* [27]. Rectangles and dashed rectangles indicate the first and second best hit for each PDZ domain, respectively, out of a proteome-wide screen. PDZ-peptide pairs that were predicted not to interact are labelled with zero. All other pairs of short peptides and single PDZ constructs were predicted to interact. Peptide names that are underlined indicate short peptides that were predicted and confirmed experimentally to bind to at least three of the four single PDZ domains. 16E6L/V served as control.
doi:10.1371/journal.pone.0025376.g005

Chen *et al.* displays a high FPR, as recently suggested by Hui and Bader [30] and that predictions are biased towards the training interaction data. Prediction scores seemed not to correlate with interaction affinities, and amino acid preferences at peptide position p0 were only partially correctly predicted. These limitations may result from both an incomplete model definition and inadequate training of the model. Regarding model definition, we showed that PDZ domains display significant structural variation, so that the model of Chen *et al.*, which is based on a single PDZ-peptide structure, may have excluded residues that are important for peptide binding. Regarding model training, the interaction dataset of Stiffler *et al.* [26] provided values for only about one third of the vast number of the model's parameters ($20 \times 20 \times 38 = 15200$). The other two thirds of the parameters were given by default the value zero, assuming that they are neither positively nor negatively contributing to PDZ-peptide interaction affinities. This allowed in particular for the tolerance of disadvantageous amino acids or over-weighting of advantageous yet non-specific residues in peptides and PDZ domains. This problem was intensified by the fact that the negative training data

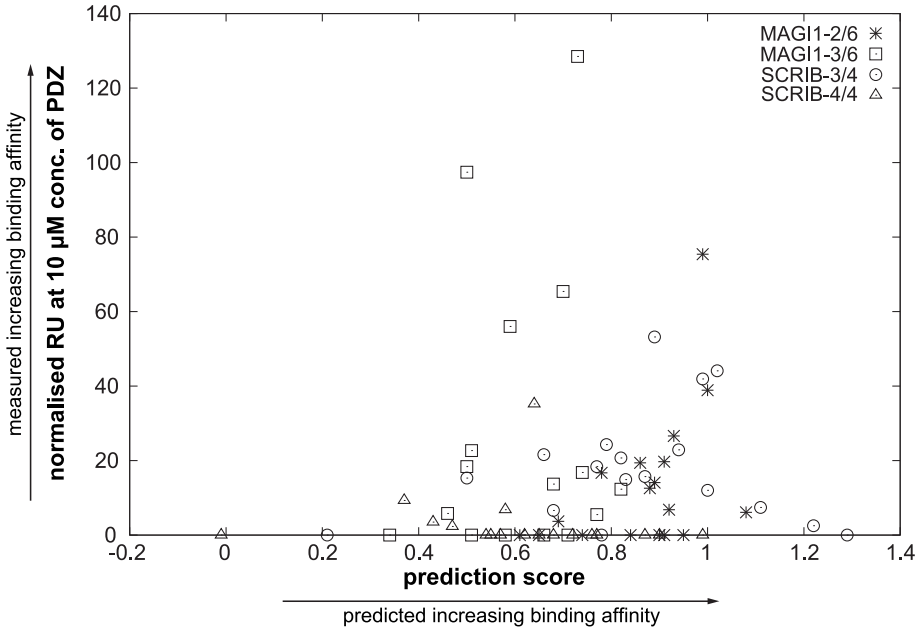


Figure 6. Comparing predicted to measured interaction intensities. The measured interaction intensities (in RU) between short versions of peptides and the PDZ domains MAGI1-2/6, MAGI1-3/6, SCRIB-3/4 and SCRIB-4/4 were plotted against the prediction scores obtained for the PDZ-peptide pairs with the predictor of Chen *et al.* The prediction scores did not correlate with measured signals. Note that SPR measurements were mostly performed for PDZ-peptide pairs that were predicted to bind to each other, explaining why the left region of the graph is empty.
doi:10.1371/journal.pone.0025376.g006

only consisted of peptides that displayed PDZ-binding motifs limiting again the sequence space covered. To turn around these limitations, it might be relevant to reduce the number of parameters that have to be trained by grouping amino acids according to their various physico-chemical properties [51]. Additionally, a filter should be applied that removes all predicted interactions with very unlikely PDZ-binding sequences, as has been done in the present study.

The predictor of Chen *et al.* is based on minimal interacting fragments corresponding to single PDZ domains and five residue-long peptides. We investigated how extensions of these minimal fragments would influence binding. The peptides that showed binding to SCRIB-3/4 generally displayed an increase in binding affinity in the presence of the tandem construct SCRIB-34/4. Since the isolated SCRIB-4/4 domain hardly bound to any peptide, we hypothesise that SCRIB-4/4 contributed indirectly to the increase in affinity of the SCRIB-3/4 domain for its target peptides, maybe by stabilising its structure. Such a long range effect might be favoured by the fact that the linker sequence between the two domains is particularly short (around 10 residues). These observations indicate that SCRIB-34/4 may represent a supramodule as defined by Feng and Zhang [43]. In a recent structure-function study, we have also demonstrated that the affinity of the MAGI1-2/6 PDZ domain to its peptidic target is modulated by the sequence of the C-terminal flanking region of the core structure of the PDZ domain [41].

Analysis of structures of PDZ-peptide complexes from the PDB showed that peptide residues upstream of p-4 are proximal to the $\beta2$-$\beta3$ loop of PDZ domains, and SPR measurements showed that the same residues modulated binding. These observations confirm previous findings [36] [37] [38] [39] [40]. Moreover, we observed that the $\beta2$-$\beta3$ loop of different PDZ domains can display very different effects on affinity and specificity of peptide binding. The observation that flanking sequences surrounding a motif modulate its interactions with the target domain may also account for other classes of domain-peptide complexes [52].

Taken together, our results suggest that extensions of protein fragments may lead to changes in affinity and specificity. However, when comparing binding intensities obtained for long *versus* short peptide constructs or for single *versus* tandem PDZ domains, protein fragment extensions were never found to change an experimentally significant interaction into a non-interaction, nor vice-versa. Therefore, we hypothesise that whenever an interaction is detected between minimal fragments, it is likely that the full length proteins will also interact, albeit possibly with different affinities. Unfortunately, affinity measurements could not be undertaken with full length proteins to provide more evidence for this hypothesis due to experimental limitations in handling large proteins *in vitro*.

Our experimental data showed that many peptides bound weakly, with affinities much weaker than 20 µM, to several of the PDZ domains tested. These observations are consistent with results of Wiedemann *et al.* [53], who predicted that for a K_D cutoff as low as 50 µM, hundreds of ligands would bind to three distinct PDZ domains with largely overlapping specificity ranges. It is often stated that interactions stop to be biologically relevant when their affinity dissociation constants exceed a given threshold (e.g. 100 µM). Such statements may have to be reconsidered when dealing with affinities determined from protein fragments, such as PDZ-peptide interactions, because as our data indicates, weak and promiscuous interactions might become stronger and more specific when moving from short protein fragments towards full length proteins.

Based on the results presented here we suggest FZD4, TMEM215 and ARHGAP6 as new interactors for MAGI1; TANC1, GLUT7, DLL1, MAS and NET1 as new interactors for SCRIB; and ABC1, MARCH3 and CYSLTR2 as new interactors for both MAGI1 and SCRIB. Remarkably, several of these

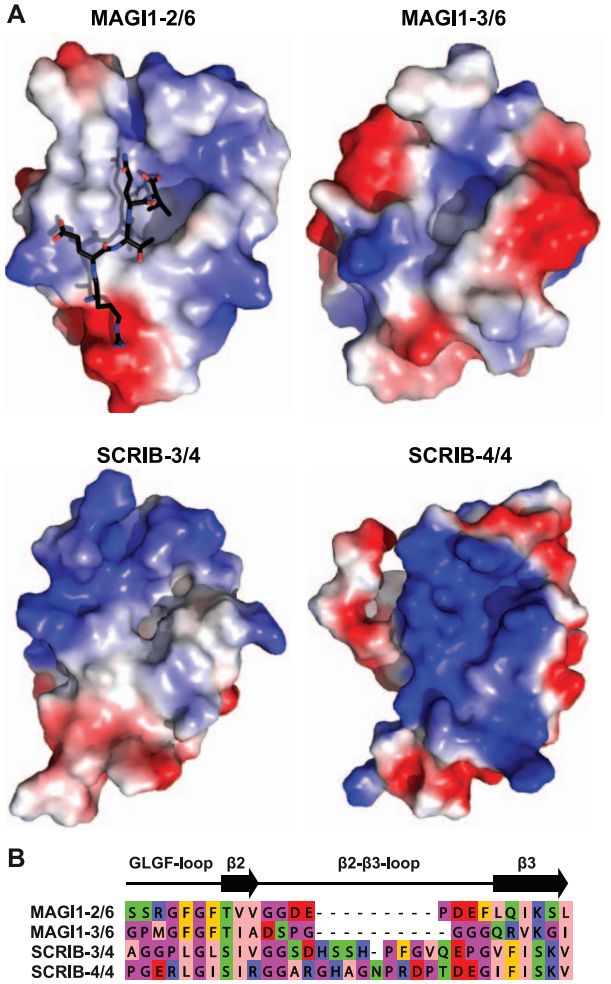


Figure 7. Structural particularities of SCRIB-4/4 in comparison to the three PDZ domains MAGI1-2/6, MAGI1-3/6, and SCRIB-3/4. A. The three experimental structures and one model (SCRIB-3/4) are shown in surface representation with red and blue indicating the electrostatic potentials. The structures are displayed in the same orientation as the PDZ domain in Figure 3. The peptide that was crystallised in complex with MAGI1-2/6 is shown in black. (PDB codes: 2I04, 3BPU, 1UJU for MAGI1-2/6, MAGI1-3/6, and SCRIB-4/4, respectively. The structure of SCRIB-3/4 was modelled from that of DLG4-1/3 (2KA9) using Modeller [79].) SCRIB-4/4 has a particularly positively charged surface around the peptide binding pocket in comparison to the other three domains. In addition, the pocket accommodating the hydrophobic residue at peptide position p0 is particularly shallow in SCRIB-4/4. These characteristics may explain the high ligand specificity displayed by SCRIB-4/4. B. Extract of the sequence alignment of the four PDZs illustrating differences within the GLGF-loop and the β2-β3 loop. SCRIB-4/4 presents a bulky R residue instead of a G in the GLGF-loop probably reducing the available space within the pocket.
doi:10.1371/journal.pone.0025376.g007

proteins are proven or putative membrane proteins (FZD4, TMEM215, GLUT7, ABC1, MARCH3, MAS, CYSLTR2, DLL1) while the three remaining ones (ARHGAP6, TANC1, NET1) are involved in activities localised to the membrane. Indeed, SCRIB and MAGI1 were already known to localise to the membrane where they interact with numerous proteins involved in signal transmission, and more particularly in G protein mediated signalling. On the one hand, MAGI1 had been shown to interact with NET1 [54] [40], a guanine nucleotide exchange factor (GEF) specific for the small G protein RhoA, as well as with PDZ-GEF1 [55], another GEF specific for the small G proteins Rap1A, Rap1B and Rap2B. MAGI3, a close paralog of MAGI1, has been shown to interact with the G protein coupled receptors (GPCRs) FZD4 [56] and LPAR2 [57], and to interact with the integral membrane protein VANG2 leading to the activation of the JNK pathway via the small G protein Rac [56]. On the other hand, SCRIB had been found to interact with two GEFs, βPIX [58] and ARHGEF16 [49], leading to activation of the small RhoA family G proteins Rac1 or Cdc42 [59] [60]. SCRIB has also been shown to interact with TSHR (a GPCR) [61].

In line with these published findings, several of the novel putative interactors of MAGI1 and SCRIB that we identified are also involved in G protein signalling. FZD4, CYSLTR2 and MAS are GPCRs; NET1 is a GEF; ARHGAP6 is a GAP (G protein activating protein); ABC1 is a membrane transporter known to recruit two GEFs (PDZRhoGEF and LARG) involved in Cdc42 and RhoA signalling [62] [63]. Therefore, our data reinforce the view that MAGI1 and SCRIB act as scaffolds that assemble proteins close to membranes to regulate G protein signalling. A remarkable instance is MAGI1 which, as indicated by our data, might be able to recruit simultaneously, via neighbouring PDZ domains, a GEF (NET1) and a GAP (ARHGAP6) that are both specific for the small GTPase RhoA, while possessing inverse enzymatic activities (Figure 9).

MAGI1 and SCRIB are known to participate to the regulation of neuronal synapses via interaction with numerous proteins [64] [60] [65] [66]. Accordingly, TANC1, which was in our hands the strongest cellular binder of SCRIB, is a scaffold component protein in post-synaptic density regions [67]. Some other interactions suggested by our work seem to provide novel links between MAGI1 and SCRIB and pathways in which they were not yet known to participate: Wnt/JNK pathway regulation (FZD4), Notch pathway regulation (DLL1) [68], immune response (CYSLTR2) [69], iron uptake (MARCH3) [70], blood vessel regulation (MAS) [71], glucose transport (GLUT7) [72]. These new interactions can provide interesting starting points for exploration of potential new *in vivo* functions of MAGI1 and SCRIB that might be perturbed upon infection with HPV.

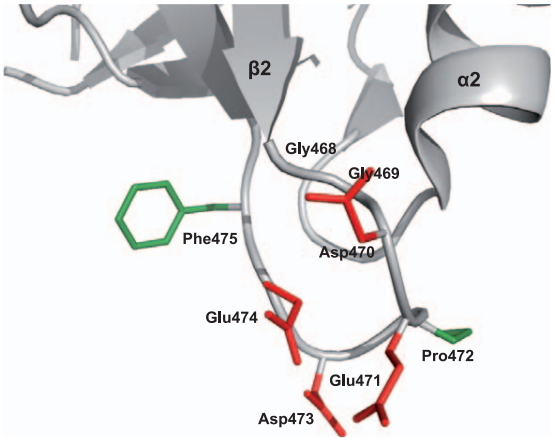
In this work, we showed that inferring protein interaction networks from predictions based on interacting protein fragments should involve at least two very distinct steps. The first step requires accurate prediction of interactions between the isolated protein fragments considered by the predictor. The predictor we used here for completing this step turned out to be rather inaccurate. There is much room for improving this step, in particular by integrating the wealth of structural information recently accumulated about protein domains, especially PDZs. The second step requires correct extrapolation of predicted fragment interactions to interactions between full length proteins. Our data indicate that such an extrapolation may be possible qualitatively, but not necessarily quantitatively. Therefore, while inferring protein interaction networks from minimal interacting fragment predictions appears as a reasonable perspective, more refined predictions addressing binding specificities in these networks remain a challenging, yet fascinating prospect.

Materials and Methods

The programming and data analysis was done using python (www.python.org), biopython [73], gnuplot (www.gnuplot.info) and PyMOL (www.pymol.org). We used the same human proteome as described in Luck *et al.* [74] to perform the proteome-wide screens in this study.

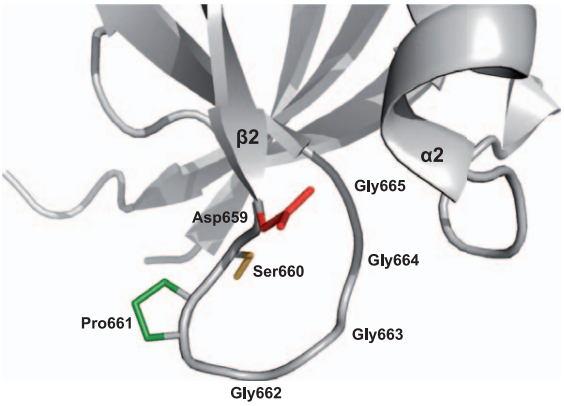
A MAGI1-2/6

peptide	sequence	RU_05	RU_10	diff RU
NET1_10	SGGKRKETLV	75	432	357
CYSLTR2_10	SVWLRKETRV	39	229	190
16E6_10	SSRTRRETQL	24	159	135
FZD4_10	KPGKGSETVV	20	92	72
MARCH3_10	VKRNSKETVV	20	55	35
TAX1_10	SEKHFRETEV	14	35	21
ARHGAP6_10	NPDALPETLV	13	24	11
TANC1_10	PKRSFIESNV	0	11	11
DLL1_10	KDECVIATEV	4	0	-4
TMEM215_10	QGRWDHETIV	6	0	-6
TTC24_10	PMESGICTIV	7	0	-7
ABC1_10	QDEKVKESYV	27	16	-11
MAS_10	CNTVTVETVV	17	0	-17
16E6L/V	SSRTRRETQV		521	
ADAM17_10	NRVDSKETEC	0	0	0
VANG2_10	VMRLQSETSV	0	0	0
ATP1A1_10	GGWVEKETYY	0	0	0
GLUT7_10	TASPAKETSF	0	0	0
GLAST_10	EKPIDSETKM	0	0	0
PTEN_10	EDQHTQITKV	0	0	0



B MAGI1-3/6

TMEM215_10	QGRWDHETIV	17	26	9
ARHGAP6_10	NPDALPETLV	56	62	6
DLL1_10	KDECVIATEV	5	10	5
ABC1_10	QDEKVKESYV	98	102	4
PTEN_10	EDQHTQITKV	23	23	0
TTC24_10	PMESGICTIV	12	12	0
TANC1_10	PKRSFIESNV	6	0	-6
TAX1_10	SEKHFRETEV	14	8	-6
MAS_10	CNTVTVETVV	18	10	-8
NET1_10	SGGKRKETLV	128	75	-53
16E6L/V	SSRTRRETQV		40	
ADAM17_10	NRVDSKETEC	0	0	0
VANG2_10	VMRLQSETSV	0	0	0
ATP1A1_10	GGWVEKETYY	0	0	0
GLUT7_10	TASPAKETSF	0	0	0
GLAST_10	EKPIDSETKM	0	0	0
16E6_10	SSRTRRETQL	0	0	0



C SCRIB-3/4

CYSLTR2_10	SVWLRKETRV	24	97	73
TANC1_10	PKRSFIESNV	21	82	61
TAX1_10	SEKHFRETEV	42	102	60
ADAM17_10	NRVDSKETEC	0	22	22
DLL1_10	KDECVIATEV	15	36	21
16E6_10	SSRTRRETQL	157	172	15
ABC1_10	QDEKVKESYV	18	33	15
GLUT7_10	TASPAKETSF	44	57	13
NET1_10	SGGKRKETLV	23	34	11
VANG2_10	VMRLQSETSV	53	63	10
TMEM215_10	QGRWDHETIV	7	16	9
MARCH3_10	VKRNSKETVV	15	23	8
TTC24_10	PMESGICTIV	7	12	5
MAS_10	CNTVTVETVV	22	24	2
ATP1A1_10	GGWVEKETYY	2	3	1
ARHGAP6_10	NPDALPETLV	16	11	-5
GLAST_10	EKPIDSETKM	12	0	-12
16E6L/V	SSRTRRETQV		93	
FZD4_10	KPGKGSETVV	0	0	0
PTEN_10	EDQHTQITKV	0	0	0

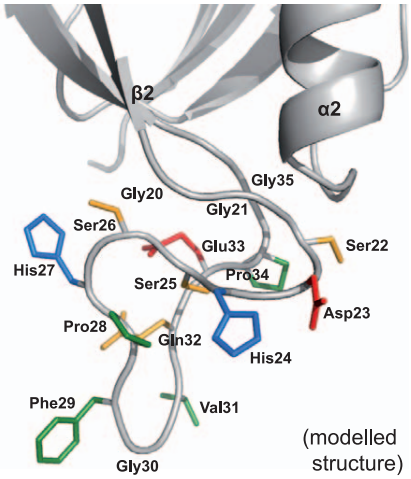


Figure 8. Influence of the β2-β3 loop of PDZ domains on peptide binding. Columns indicate from left to right the names of the peptides, their sequences, the interaction intensities in RU for peptides with five and ten wildtype residues, and the interaction intensity difference between both. Peptides with five wildtype residues had the five N-terminal residues replaced with GSGAG. For each PDZ the part of the structure containing the *β2-β3* loop is shown with loop side chains represented as sticks. Amino acids in the sequences and structures are coloured as follows: red = negative charge, blue = positive charge, yellow = polar, green = hydrophobic. A. MAGI1-2/6 binds with increased affinity to peptides with positive charges upstream p-4 probably due to four negative charges in the loop (pdb code: 2I04). B. MAGI1-3/6 does not show any difference in

affinity to short and long peptides, possibly due to four "neutral" glycines in the loop (pdb code: 3BPU). C. SCRIB-3/4 shows rather an unspecific increase in affinity for long peptides. The loop is very long and contains residues of all physico-chemical types.
doi:10.1371/journal.pone.0025376.g008

Prediction quality assessment

We assessed the performance of the predictor of Chen *et al.* [27] by applying the commonly used measures *Sensitivity* (*SE*) and *False Positive Rate* (*FPR*) of the ROC analysis. Here, the sensitivity is defined as the percentage of PDZ-peptide interactions that were correctly predicted (= True Positives (*TP*)) and is calculated as follows:

$$SE = \frac{TP}{TP+FN} \cdot 100 \quad (1)$$

where *FN* specifies the number of False Negatives (PDZ-peptide interactions not correctly predicted). The False Positive Rate is defined as the percentage of PDZ-peptide non-interactions that were *not* correctly predicted (= False Positives (*FP*)) and is calculated as follows:

$$FPR = \frac{FP}{TN+FP} \cdot 100 \quad (2)$$

where *TN* specifies the number of True Negatives (PDZ-peptide non-interactions correctly predicted).

Implementation, test, and application of the predictor of Chen *et al.*

Chen *et al.* [27] trained the predictor in two different ways, called the binary and affinity mode, of which each of them can be used separately to apply the predictor. For the binary mode the

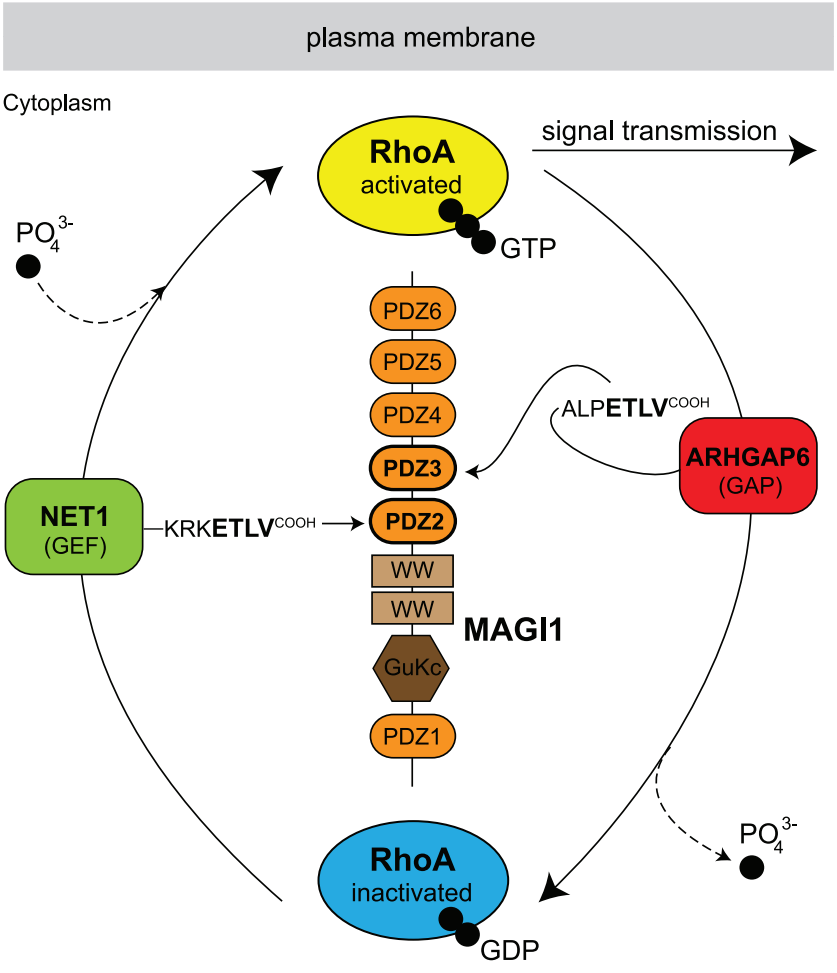


Figure 9. Suggested model for MAGI1 scaffolding function in Rho GTPase mediated signalling. Our data showed that PDZ2 and PDZ3 of MAGI1 bind preferentially to the C-termini of NET1 (green) and ARHGAP6 (red), respectively. NET1 is a guanine nucleotide exchange factor (GEF), which transfers a phosphate group (PO_4^{3-}) to the small GTPase RhoA, which in its GTP-bound form (yellow) is predominantly associated with the membrane and stimulates downstream signalling pathways. ARHGAP6 is a GTPase-activating protein (GAP), which induces RhoA to release a phosphate group, resulting in the shutdown of RhoA signalling. Inactivated GDP-bound RhoA (blue) is mostly present in the cytoplasm. This indicates that MAGI1 recruits, via two adjacent PDZ domains, one activator and one inhibitor of the RhoA signalling pathway. Remarkably, the four last residues of the two proteins NET1 and ARHGAP6 are identical, hence the distinct binding preferences of the two C-terminal peptides for PDZ2 and PDZ3 must be defined by residues upstream.
doi:10.1371/journal.pone.0025376.g009

predictor was trained without consideration of measured binding affinities (e.g. the training data was simply split into interactions and non-interactions). In the affinity mode, binding affinities were directly included in the training process. For all predictions performed in this study, the binary mode was used. No information about performance qualities was provided by Chen *et al.* for the affinity mode. We performed a comparison of both modes that revealed extremely different predictions with the binary mode providing more reliable results (data not shown). The predictor returns a score for each PDZ-peptide pair, which can be used to estimate the likeliness that the PDZ domain will bind the respective peptide. The higher the score, the more likely the interaction. Here, we used a score cutoff of 0.5, which should yield a sensitivity of 76% and FPR of 24% as specified by Chen *et al.*

Each of the 95 human PDZ domains in the test datasets were added to the alignment of mouse PDZ domains provided by Chen *et al.* in order to define the 16 amino acid positions on which predictions are based. Mafft [75] was used to obtain a preliminary alignment, which was corrected manually using Jalview [76] and structural information, if available. The alignment is provided in Dataset S3.

The training set containing 93 peptides of Chen *et al.* was not provided in the publication. The set of peptides from the training data was reconstructed as described by Chen *et al.* taking every peptide that was seen at least once in an interaction with a PDZ domain in the experimental data obtained by Stiffler *et al.* [26]. This revealed 108 peptides.

In Text S1 and Dataset S4 we provide guidelines and programming code, respectively, for users of the predictor of Chen *et al.*, who wish to follow our developed protocol.

PDZ pocket analysis

Available structures of PDZ-ligand complexes were analysed in order to assess important domain residues for ligand recognition. A keyword search with "PDZ" in the PDB [77] revealed 267 structures. Crystal structures were excluded, if the PDB files did not contain coordinates of the full complex but just of one chain (e.g. PDB code 2EGN). After manual inspection, a final set of 42 structures with PDZ-peptide complexes was retained for further analysis representing 24 unique PDZ domains. For each PDZ domain all structural models obtained by NMR and all complexes shown in the crystal obtained by X-ray were taken into consideration for the determination of all domain residues that are in close proximity to bound peptides. A domain-peptide residue pair was only accepted, if in all complexes of this particular PDZ domain the distance between the two amino acids was in average below a defined threshold. Three different distance measures were implemented: C_α distances, distances between residue's centre of mass, and minimal atom distances between residues. Different thresholds were tested from 0 to 40 Å. The distance measure and cutoff that represented best the selection of the 16 domain amino acids in α-syntrophin of Chen *et al.* [27] was chosen: minimal atom distance with a threshold of 5 Å.

The PDZ sequences shown in Figure 4 were extracted from the following PDB entries and chains: SNTA1_1/1 (2PDZ A), AFAD_1/1 (2AIN A), APBA1_1/2 (1U38 A), ARHGC_1/1 (2OS6 A), DLG1_2/3 (2AWW A), DLG1_3/3 (2I0I C), DLG4_3/3 (1TP5 A), EM55_1/1 (2EJY A), GRIP1_1/7 (2QT5 A), GRIP1_6/7 (1N7F B), HTRA1_1/1 (2JOA A), INAD_1/5 (1IHJ A), LAP2_1/1 (1N7T A), MAGI1_2/6 (2KPL A), NOS1_1/1 (1B8Q A), PAR6_1/1 (1RZX A), PAR6i_1/1 (1X8S A), PARD3_3/3 (2K20 A), PICK1_1/1 (2PKU A), PTN13_2/5 (1D5G A), RIMS1_1/1 (1ZUB A), SHAN1_1/1 (1Q3P B), TIP1_1/1 (3DIW A), SYNT1_1/2 (1W9E A), SYNT1_2/2 (1V1T A).

Structure modelling

The structure of the PDZ domain SCRIB-3/4 was modelled using the program Modeller 9v7. The structure template was obtained by querying the PDB with the sequence of SCRIB-3/4 (using the BLAST option) and choosing the structure with the best sequence match (PDZ domain DLG4-1/3, PDB-code 2KA9, 45% sequence identity, e-value 1.0E-11). Modeller was run using the automodel routine and default options. Model quality was assessed using the output information of Modeller and visual inspection. A model of SCRIB-3/4 of intermediate quality was sufficient for the purpose of this study.

cDNA constructs

The cDNA encoding residues 448–572 and 613–752 of mouse MAGI-1 (UniProt acc.: Q6RHR9-1) encoding for MAGI1-2/6 (100% identical to human MAGI1-2/6) and MAGI1-3/6 (99% identical to human MAGI1-3/6) PDZ domains, respectively, were inserted into the NcoI/KpnI sites of the pETM-41 expression vector (EMBL) containing a 6×His-MBP tag followed by a TEV protease cleavage site. A similar cloning strategy was adopted for cDNA bearing residues 997–1093, 1097–1193 and 997–1193 of human SCRIB (Uniprot acc.: Q14160-1) encoding for SCRIB-3/4, SCRIB-4/4 PDZ domains and SCRIB-34/4 tandem PDZ construct, respectively.

Protein sample production

Bacterial over-expression of PDZ domains was performed using BL21 DE3 *Escherichia coli* cells in 300 ml of M9 minimal medium supplemented with $^{15}NH_4Cl$ at 37°C until an OD_{600} of 0.6 was reached. Cultures were then adjusted to 0.5 mM isopropyl-D-thio-galactopyranoside (IPTG) and transferred to 15°C overnight. Plasmid loss was suppressed by adding 15 μg/ml of kanamycin to the expression media. Expression cultures were harvested by centrifugation. The pellets were stored at −20°C.

MBP-PDZ domains purification

Bacterial expression of ^{15}N-labeled 6×His-MBP-PDZ constructs were sonicated in buffer A (50 mM Tris-HCl at pH 6.8, 200 mM NaCl, 1 mM DTT) supplemented with 1 μg/ml DNase I and RNase A and EDTA-free anti-protease cocktail inhibitor (Roche), cleared by ultracentrifugation at 60000·g and filtered (Millipore 0.22 μm). MBP-PDZ extracts were loaded on an amylose column (New England Biolabs) pre-equilibrated with buffer A. Protein was eluted with buffer A supplemented with 10 mM maltose. MBP-PDZ samples were then subjected to a 15 hour ultracentrifugation at 130000·g prior to loading on a Hiload 16/60 Superdex 75 gel-filtration column (Amersham Biosciences) pre-equilibrated with buffer B (20 mM sodium phosphate at pH 6.8, 200 mM NaCl) resulting in pure and mono-disperse protein samples according to the column calibration. The concentration of purified MBP-PDZ fusion samples was evaluated from UV absorption measurements at 280 *nm*. After SPR experiments MBP-PDZ fusions were cleaved by TEV and PDZ domains were separated from MBP by gel size exclusion chromatography. Subsequently, ^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) spectra were recorded on a 600 MHz Bruker instrument in order to verify structural integrity of the domains.

Synthetic peptides

The synthetic peptide 16E6L/V (RSSRTRRETQV), corresponding to the last 11 C-terminal residues of HPV16 E6 with the

last residue L mutated to V, was synthesised by the Chemical Peptide Synthesis Service, IGBMC, France. Lyophilised peptide was re-suspended in water, passed on a NAP-5 desalting column (GE Healthcare) in order to remove residual contaminants. The desalted peptide was lyophilised prior to its dilution into buffer A. The peptide was checked by homonuclear 2D NMR experiments and its concentration estimated to be at 6 mM by measuring the peptide bond absorption at 205 nm as described previously [40]. All other synthetic peptides with biotin at N-terminus that were used as ligand in surface plasmon resonance experiments were synthesised by JPT Peptide Technologies GmbH, Berlin, Germany. Lyophilised peptides were re-suspended in water at a final concentration at 10 mM. The pH of peptide solution was adjusted to 6.8.

Surface plasmon resonance (SPR) measurements

Data were collected on a Biacore 2000 instrument (Biacore AB/GE Healthcare Bio-Sciences Corp., Piscataway, NJ, USA) at 25°C. SPR experiments (ligand immobilisation and binding measurements) have been performed as described in Fournane *et al.* [40]. Briefly, biotinylated peptides (instead of GST-fused recombinant peptides) were immobilised on CM5 sensorchips on which Neutravidin was previously attached. The MBP-PDZ domain analyte was injected at 8 to 10 different concentrations ranging from 0 up to 30 μM. Data were processed using the BiaEvaluation 3.2 software (Biacore AB/GE Healthcare Bio-Sciences Corp.) using "double referencing" [78] in which sensorgrams were corrected for buffer effects and bulk refractive index changes. Representative sensorgrams are shown in Figure 5A.

The steady-state binding signal (R_{eq}) was derived by averaging the signals in a five second window at equilibrium. Steady-state analysis was performed by fitting the average signal R_{eq} as a function of total MBP-PDZ concentrations, assuming a simple 1:1 interaction binding isotherm model. For many weak interactions we observed calculated binding affinities (K_D) with fits that produced high χ^2 suggesting that the K_D were likely to be inaccurate (see Table S4). Reasons for this inaccuracy are likely to be the following: 1. As previously described [40], several repetitions of all the measurements are required to determine accurate K_D. In our case, such repetitions were not achievable in reasonable time due to the large amount of interactions measured in this study. 2. The highest injected analyte concentration restricts the maximal K_D (weakest interaction) that can be accurately obtained. 3. A K_D is estimated based on a mathematical extrapolation of observed R_{eq} signals leading to additional uncertainty. Based on these reasons, we considered the calculated K_D not as accurate enough to be used for absolute binding strength comparison in this study. We rather performed a relative analysis of binding strengths using directly R_{eq} signals which are not biased by any mathematical assumption. We focussed on R_{eq} signals obtained at 10 μM MBP-PDZ concentration, which have been systematically measured in duplicate. The R_{eq} signal is directly proportional to the molecular weight of the analyte and the amount of immobilised ligand. Therefore, the R_{eq} signals were normalised taking those into account before being used for binding strength comparison. The large amount of raw experimental data, which have been collected and the methodological approach that we have developed for their exploitation will be presented and discussed in detail in a separate, SPR-oriented paper.

Supporting Information

Dataset S1 PDZ interaction and non-interaction test datasets. The archive contains three files, one for each test dataset established: interactions, non-interactions with PDZ-binding motif, and non-interactions without PDZ-binding motif. First column: PDZ domain, second column: name of binder, third column: C-terminus of binder.
(BZ2)

Dataset S2 Prediction results of proteome-wide screen for MAGI1 and SCRIB PDZ-binding ligands using the predictor of Chen *et al.* [27]. The prediction results were performed in binary mode using a cutoff of 0.5 and are provided without any additional filtering. No result file is provided for the PDZ domain MAGI1-1/6 because the screen did not reveal any peptides for this domain.
(BZ2)

Dataset S3 Alignment of human PDZ domains. The archive contains an alignment in fasta format of 95 PDZ domains. These include all PDZ domains that occur in the three test datasets as well as all MAGI1 and SCRIB PDZ domains. Additionally, a file is provided containing a translation between the PDZ domain names used in the test datasets and the PDZ domain names used in the alignment.
(BZ2)

Dataset S4 Implementation of the predictor of Chen *et al.* [27]. The archive contains data files and python scripts necessary to launch the predictor. The only prerequisite for running the program is an installed python version. Check the README.txt for more information.
(BZ2)

Table S1 Diversity of amino acids at last five positions of PDZ-binding peptides in the training data of Chen *et al.* [27].
(PDF)

Table S2 Filtered numbers of proteins predicted to bind to 1, 2, 3, … or all PDZ domains of MAGI1 (6 PDZs) or SCRIB (4 PDZs).
(PDF)

Table S3 Annotations for all proteins tested experimentally in this work for interaction to MAGI1 and SCRIB. The table contains UniProt IDs and information about biological functions of the proteins with regard to PDZ domain binding as well as published information on interactions with PDZ domain-containing proteins.
(PDF)

Table S4 Experimental data for all interactions measured. The table contains "double referenced" and normalised R_{eq} signals obtained for a 10 μM analyte concentration as well as tentative calculated K_D assuming a simple 1:1 interaction binding isotherm model. These K_D have to be considered with caution, especially for interactions for which weak RU signals were obtained.
(PDF)

Text S1 Recommendations for application of the predictor of Chen *et al.* [27].
(TXT)

Acknowledgments

We thank the members of the BIAcore platform at IREBS, MA Delsuc, N Davey, T Gibson and R Vincentelli for helpful discussions, A Chapelle for help in protein production and P. Eberling (IGBMC) for peptide synthesis.

Author Contributions

Conceived and designed the experiments: GT KL YN BK MM. Performed the experiments: KL SF YN. Analyzed the data: KL SF YN MM. Wrote the paper: KL GT YN SF.

References

1. Diella F, Haslam N, Chica C, Budd A, Michael S, et al. (2008) Understanding eukaryotic linear motifs and their role in cell signaling and regulation. Front Biosci 13: 6580–6603.
2. Finn RD, Mistry J, Tate J, Coggill P, Heger A, et al. (2010) The Pfam protein families database. Nucleic Acids Res 38: D211–D222.
3. Letunic I, Doerks T, Bork P (2009) SMART 6: recent updates and new developments. Nucleic Acids Res 37: D229–D232.
4. Gould CM, Diella F, Via A, Puntervoll P, Gemünd C, et al. (2010) ELM: the status of the 2010 eukaryotic linear motif resource. Nucleic Acids Res 38: D167–D180.
5. Edwards RJ, Davey NE, Shields DC (2007) SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. PLoS ONE 2: e967.
6. Velthuis AJWT, Sakalis PA, Fowler DA, Bagowski CP (2011) Genome-wide analysis of PDZ domain binding reveals inherent functional overlap within the PDZ interaction network. PLoS One 6: e16047.
7. Songyang Z, Fanning AS, Fu C, Xu J, Marfatia SM, et al. (1997) Recognition of unique carboxylterminal motifs by distinct PDZ domains. Science 275: 73–77.
8. Stricker NL, Christopherson KS, Yi BA, Schatz PJ, Raab RW, et al. (1997) PDZ domain of neuronal nitric oxide synthase recognizes novel C-terminal peptide sequences. Nat Biotechnol 15: 336–342.
9. Beuming T, Skrabanek L, Niv MY, Mukherjee P, Weinstein H (2005) PDZBase: a protein-protein interaction database for PDZ-domains. Bioinformatics 21: 827–828.
10. Encinar JA, Fernandez-Ballester G, Sánchez IE, Hurtado-Gomez E, Stricher F, et al. (2009) ADAN: a database for prediction of protein-protein interaction of modular domains mediated by linear motifs. Bioinformatics 25: 2418–2424.
11. Roh MH, Margolis B (2003) Composition and function of PDZ protein complexes during cell polarization. Am J Physiol Renal Physiol 285: F377–F387.
12. Ide N, Hata Y, Nishioka H, Hirao K, Yao I, et al. (1999) Localization of membrane-associated guanylate kinase (MAGI)-1/BAI-associated protein (BAP) 1 at tight junctions of epithelial cells. Oncogene 18: 7810–7815.
13. Wegmann F, Ebnet K, Pasquier LD, Vestweber D, Butz S (2004) Endothelial adhesion molecule ESAM binds directly to the multidomain adaptor MAGI-1 and recruits it to cell contacts. Exp Cell Res 300: 121–133.
14. Yoshihara K, Ikenouchi J, Izumi Y, Akashi M, Tsukita S, et al. (2011) Phosphorylation state regulates the localization of Scribble at adherens junctions and its association with E-cadherincatenin complexes. Exp Cell Res 317: 413–422.
15. Ivanov AI, Young C, Beste KD, Capaldo CT, Humbert PO, et al. (2010) Tumor suppressor scribble regulates assembly of tight junctions in the intestinal epithelium. Am J Pathol 176: 134–145.
16. Humbert PO, Dow LE, Russell SM (2006) The Scribble and Par complexes in polarity and migration: friends or foes? Trends Cell Biol 16: 622–630.
17. Kotelevets L, van Hengel J, Bruyneel E, Mareel M, van Roy F, et al. (2005) Implication of the MAGI-1b/PTEN signalosome in stabilization of adherens junctions and suppression of invasiveness. FASEB J 19: 115–117.
18. Zhan L, Rosenberg A, Bergami KC, Yu M, Xuan Z, et al. (2008) Deregulation of scribble promotes mammary tumorigenesis and reveals a role for cell polarity in carcinoma. Cell 135: 865–878.
19. Javier RT (2008) Cell polarity proteins: common targets for tumorigenic human viruses. Oncogene 27: 7031–7046.
20. Liu H, Golebiewski L, Dow EC, Krug RM, Javier RT, et al. (2010) The ESEV PDZ-binding motif of the avian inuenza A virus NS1 protein protects infected cells from apoptosis by directly targeting Scribble. J Virol 84: 11164–11174.
21. Glaunsinger BA, Lee SS, Thomas M, Banks L, Javier R (2000) Interactions of the PDZ-protein MAGI-1 with adenovirus E4-ORF1 and high-risk papillomavirus E6 oncoproteins. Oncogene 19: 5270–5280.
22. Nakagawa S, Huibregtse JM (2000) Human scribble (Vartul) is targeted for ubiquitin-mediated degradation by the high-risk papillomavirus E6 proteins and the E6AP ubiquitin-protein ligase. Mol Cell Biol 20: 8244–8253.
23. Simonson SJS, Difilippantonio MJ, Lambert PF (2005) Two distinct activities contribute to human papillomavirus 16 E6's oncogenic potential. Cancer Res 65: 8266–8273.
24. Brannetti B, Zanzoni A, Montecchi-Palazzi L, Cesareni G, Helmer-Citterich M (2001) iSPOT: A web tool for the analysis and recognition of protein domain specificity. Comp Funct Genomics 2: 314–318.
25. Tonikian R, Zhang Y, Sazinsky SL, Currell B, Yeh JH, et al. (2008) A specificity map for the PDZ domain family. PLoS Biol 6: e239.
26. Stiffler MA, Chen JR, Grantcharova VP, Lei Y, Fuchs D, et al. (2007) PDZ domain binding selectivity is optimized across the mouse proteome. Science 317: 364–369.
27. Chen JR, Chang BH, Allen JE, Stiffler MA, MacBeath G (2008) Predicting PDZ domain-peptide interactions from primary sequences. Nat Biotechnol 26: 1041–1045.
28. Schultz J, Hoffmüller U, Krause G, Ashurst J, Macias MJ, et al. (1998) Specific interactions between the syntrophin PDZ domain and voltage-gated sodium channels. Nat Struct Biol 5: 19–24.
29. Smialowski P, Pagel P, Wong P, Brauner B, Dunger I, et al. (2010) The Negatome database: a reference set of non-interacting protein pairs. Nucleic Acids Res 38: D540–D544.
30. Hui S, Bader GD (2010) Proteome scanning to predict PDZ domain interactions using support vector machines. BMC Bioinformatics 11: 507.
31. Shao X, Tan CSH, Voss C, Li SSC, Deng N, et al. (2011) A regression framework incorporating quantitative and negative interaction data improves quantitative prediction of PDZ domain-peptide interaction from primary sequence. Bioinformatics 27: 383–390.
32. Gerek ZN, Keskin O, Ozkan SB (2009) Identification of specificity and promiscuity of PDZ domain interactions through their dynamic behavior. Proteins 77: 796–811.
33. Kalyoncu S, Keskin O, Gursoy A (2010) Interaction prediction and classification of PDZ domains. BMC Bioinformatics 11: 357.
34. Smith CA, Kortemme T (2010) Structure-based prediction of the peptide sequence space recognized by natural and synthetic PDZ domains. J Mol Biol 402: 460–474.
35. Gerek ZN, Ozkan SB (2010) A exible docking scheme to explore the binding selectivity of PDZ domains. Protein Sci 19: 914–928.
36. Imamura F, Maeda S, Doi T, Fujiyoshi Y (2002) Ligand binding of the second PDZ domain regulates clustering of PSD-95 with the Kv1.4 potassium channel. J Biol Chem 277: 3640–3646.
37. Wang L, Piserchio A, Mierke DF (2005) Structural characterization of the intermolecular interactions of synapse-associated protein-97 with the NR2B subunit of N-methyl-D-aspartate receptors. J Biol Chem 280: 26992–26996.
38. Birrane G, Chung J, Ladias JAA (2003) Novel mode of ligand recognition by the Erbin PDZ domain. J Biol Chem 278: 1399–1402.
39. Kachel N, Erdmann KS, Kremer W, Wolff P, Gronwald W, et al. (2003) Structure determination and ligand interactions of the PDZ2b domain of PTP-Bas (hPTP1E): splicing-induced modulation of ligand specificity. J Mol Biol 334: 143–155.
40. Fournane S, Charbonnier S, Chapelle A, Kieffer B, Orfanoudakis G, et al. (2010) Surface plasmon resonance analysis of the binding of high-risk mucosal HPV E6 oncoproteins to the PDZ1 domain of the tight junction protein MAGI-1. J Mol Recognit.
41. Charbonnier S, Nominé Y, Ramírez J, Luck K, Chapelle A, et al. (2011) The structural and dynamic response of MAGI-1 PDZ1 with noncanonical domain boundaries to the binding of human papillomavirus E6. J Mol Biol 406: 745–763.
42. Wang CK, Pan L, Chen J, Zhang M (2010) Extensions of PDZ domains as important structural and functional elements. Protein Cell 1: 737–751.
43. Feng W, Zhang M (2009) Organization and dynamics of PDZ-domain-related supramodules in the postsynaptic density. Nat Rev Neurosci 10: 87–99.
44. Harris BZ, Lau FW, Fujii N, Guy RK, Lim WA (2003) Role of electrostatic interactions in PDZ domain ligand recognition. Biochemistry 42: 2797–2805.
45. Doyle DA, Lee A, Lewis J, Kim E, Sheng M, et al. (1996) Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ. Cell 85: 1067–1076.
46. Dev KK, Nakanishi S, Henley JM (2004) The PDZ domain of PICK1 differentially accepts protein kinase C-alpha and GluR2 as interacting ligands. J Biol Chem 279: 41393–41397.
47. Hubbard TJP, Aken BL, Ayling S, Ballester B, Beal K, et al. (2009) Ensembl 2009. Nucleic Acids Res 37: D690–D697.
48. Consortium U (2010) The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res 38: D142–D148.
49. Zhang Y, Yeh S, Appleton BA, Held HA, Kausalya PJ, et al. (2006) Convergent and divergent ligand specificity among PDZ domains of the LAP and zonula occludens (ZO) families. J Biol Chem 281: 22299–22311.
50. Thomas M, Glaunsinger B, Pim D, Javier R, Banks L (2001) HPV E6 and MAGUK protein interactions: determination of the molecular basis for specific protein recognition and degradation. Oncogene 20: 5431–5439.
51. Schillinger C, Boisguerin P, Krause G (2009) Domain Interaction Footprint: a multi-classification approach to predict domain-peptide interactions. Bioinformatics 25: 1632–1639.
52. Stein A, Aloy P (2008) Contextual specificity in peptide-mediated protein interactions. PLoS One 3: e2524.
53. Wiedemann U, Boisguerin P, Leben R, Leitner D, Krause G, et al. (2004) Quantification of PDZ domain specificity, prediction of ligand affinity and rational design of super-binding peptides. J Mol Biol 343: 703–718.
54. Dobrosotskaya IY (2001) Identification of mNET1 as a candidate ligand for the first PDZ domain of MAGI-1. Biochem Biophys Res Commun 283: 969–975.

55. Mino A, Ohtsuka T, Inoue E, Takai Y (2000) Membrane-associated guanylate kinase with inverted orientation (MAGI)-1/brain angiogenesis inhibitor 1-associated protein (BAP1) as a scaffolding molecule for Rap small G protein GDP/GTP exchange protein at tight junctions. Genes Cells 5: 1009–1016.
56. Yao R, Natsume Y, Noda T (2004) MAGI-3 is involved in the regulation of the JNK signaling pathway as a scaffold protein for frizzled and Ltap. Oncogene 23: 6023–6030.
57. Zhang H, Wang D, Sun H, Hall RA, Yun CC (2007) MAGI-3 regulates LPA-induced activation of Erk and RhoA. Cell Signal 19: 261–268.
58. Audebert S, Navarro C, Nourry C, Chasserot-Golaz S, Lécine P, et al. (2004) Mammalian Scribble forms a tight complex with the betaPIX exchange factor. Curr Biol 14: 987–995.
59. Nola S, Sebbagh M, Marchetto S, Osmani N, Nourry C, et al. (2008) Scrib regulates PAK activity during the cell migration process. Hum Mol Genet 17: 3552–3565.
60. Momboisse F, Lonchamp E, Calco V, Ceridono M, Vitale N, et al. (2009) betaPIX-activated Rac1 stimulates the activation of phospholipase D, which is associated with exocytosis in neuroendocrine cells. J Cell Sci 122: 798–806.
61. Lahuna O, Quellari M, Achard C, Nola S, Méduri G, et al. (2005) Thyrotropin receptor trafficking relies on the hScrib-betaPIX-GIT1-ARF6 pathway. EMBO J 24: 1364–1374.
62. Tsukamoto K, Hirano K, Tsujii K, Ikegami C, Zhongyan Z, et al. (2001) ATP-binding cassette transporter-1 induces rearrangement of actin cytoskeletons possibly through Cdc42/N-WASP. Biochem Biophys Res Commun 287: 757–765.
63. Okuhira K, Fitzgerald ML, Tamehiro N, Ohoka N, Suzuki K, et al. (2010) Binding of PDZ-RhoGEF to ATP-binding cassette transporter A1 (ABCA1) induces cholesterol effux through RhoA activation and prevention of transporter degradation. J Biol Chem 285: 16369–16377.
64. Stetak A, Hörndli F, Maricq AV, van den Heuvel S, Hajnal A (2009) Neuron-specific regulation of associative learning and memory by MAGI-1 in C. elegans. PLoS One 4: e6019.
65. Sun Y, Aiga M, Yoshida E, Humbert PO, Bamji SX (2009) Scribble interacts with beta-catenin to localize synaptic vesicles to synapses. Mol Biol Cell 20: 3390–3400.
66. Moreau MM, Piguel N, Papouin T, Koehl M, Durand CM, et al. (2010) The planar polarity protein Scribble1 is essential for neuronal plasticity and brain function. J Neurosci 30: 9738–9752.
67. Nonaka H, Takei K, Umikawa M, Oshiro M, Kuninaka K, et al. (2008) MINK is a Rap2 effector for phosphorylation of the postsynaptic scaffold protein TANC1. Biochem Biophys Res Commun 377: 573–578.
68. Jaleco AC, Neves H, Hooijberg E, Gameiro P, Clode N, et al. (2001) Differential effects of Notch ligands Delta-1 and Jagged-1 in human lymphoid differentiation. J Exp Med 194: 991–1002.
69. Woszczek G, Chen LY, Nagineni S, Alsaaty S, Harry A, et al. (2007) IFN-gamma induces cysteinyl leukotriene receptor 2 expression and enhances the responsiveness of human endothelial cells to cysteinyl leukotrienes. J Immunol 178: 5262–5270.
70. Fukuda H, Nakamura N, Hirose S (2006) MARCH-III is a novel component of endosomes with properties similar to those of MARCH-II. J Biochem 139: 137–145.
71. Canals M, Jenkins L, Kellett E, Milligan G (2006) Up-regulation of the angiotensin II type 1 receptor by the MAS proto-oncogene is due to constitutive activation of Gq/G11 by MAS. J Biol Chem 281: 16757–16767.
72. Li Q, Manolescu A, Ritzel M, Yao S, Slugoski M, et al. (2004) Cloning and functional characterization of the human GLUT7 isoform SLC2A7 from the small intestine. Am J Physiol Gastrointest Liver Physiol 287: G236–G242.
73. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, et al. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25: 1422–1423.
74. Luck K, Travé G (2011) Phage display can select over-hydrophobic sequences that may impair prediction of natural domain-peptide interactions. Bioinformatics 27: 899–902.
75. Katoh K, Toh H (2007) PartTree: an algorithm to build an approximate tree from a large number of unaligned sequences. Bioinformatics 23: 372–374.
76. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009) Jalview version 2–a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189–1191.
77. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The protein data bank. Nucleic Acids Res 28: 235–242.
78. Myszka DG (1999) Improving biosensor analysis. J Mol Recognit 12: 279–284.
79. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234: 779–815.

10. The emerging contribution of sequence context to the specificity of protein interactions mediated by PDZ domains

10.1. Summary

Content: Numerous structural studies have been published on PDZ domains that illustrate the influence of sequence context on the binding affinity and specificity of protein interactions. This review discusses sequence context of both PBMs and PDZs. Extensions constitute so far the only known cases of sequence context of PBMs, whereas known instances of sequence context of PDZ domains can be divided into extensions and neighbouring domains. Studies on both, extensions and neighbouring domains of PDZs have already previously been reviewed [55,158]. Thus, we focussed on the more recently published instances of sequence context of PDZ domains. On the other hand, to our knowledge, this article is the first to review the numerous instances of sequence context of PBMs.

The study of sequence context of globular domains is tightly linked to the definition of domain boundaries. Domain boundaries can be theoretically defined by the beginning and end of the first and last secondary structure element. Often, boundaries that are defined this way lead to non-functional molecules when transferred to experimental domain constructs. We dedicated a section on this topic and suggest manually curated domain boundaries for the complete human PDZ domain family that are likely to result in functional protein fragments upon expression. These annotations were based on available PDZ domain structures.

Finally, a small section has been added on the ongoing controversy about the total number of PDZ domains in the human proteome. We hope that this analysis will lead to better estimates of the size of the human PDZome in future published PDZ domain-related studies.

This article is an invited review for the special issue: "Modular Protein Domains" in FEBS Letters, edited by Gianni Cesareni, Wilhelm Just, Giulio Superti-Furga, and Marius Sudol.

Contribution: I have red the published literature on sequence context of PBMs and PDZ domains. I have conceived the manuscript and have written most of it. I performed most of the work on the manual domain boundary definition of the human PDZome. (See also supplemental material provided in section D.3.)

FEBS Letters xxx (2012) xxx–xxx

ELSEVIER

FEBS Letters

journal homepage: www.FEBSLetters.org

Review

The emerging contribution of sequence context to the specificity of protein interactions mediated by PDZ domains

Katja Luck *, Sebastian Charbonnier, Gilles Travé *

UMR 7242, Institut de Recherche de l'Ecole de Biotechnologie de Strasbourg, Bd Sébastien Brant, BP 10413, 67412 Illkirch, Cedex, France

ARTICLE INFO

Article history:
Received 29 February 2012
Revised 26 March 2012
Accepted 27 March 2012
Available online xxxx

Edited by Marius Sudol, Gianni Cesareni, Giulio Superti-Furga and Wilhelm Just

Keywords:
PDZ
Specificity
Sequence context
Extension
β2–β3 Loop
Linear motifs

ABSTRACT

The canonical binding mode of PDZ domains to target motifs involves a small interface, unlikely to fully account for PDZ-target interaction specificities. Here, we review recent work on sequence context, defined as the regions surrounding not only the PDZ domains but also their target motifs. We also address the theoretical problem of defining the core of PDZ domains and the practical issue of designing PDZ constructs. Sequence context is found to introduce structural diversity, to impact the stability and solubility of constructs, and to deeply influence binding affinity and specificity, thereby increasing the difficulty of predicting PDZ-motif interactions. We expect that sequence context will have similar importance for other protein interactions mediated by globular domains binding to short linear motifs.

1. Introduction

Interactions between proteins are essential for most of the processes that happen in living cells. Many protein interactions in cell signalling are mediated by interactions between globular domains and short linear motifs (SLiMs) [1]. SLiMs are short disordered protein sequence segments that often become folded in their bound state [1,2]. Important insights on domain–SLiM interactions have been gained from studies on PDZ (PSD95-DLG1-ZO1) domains. PDZ domains constitute a large family of globular domains found in prokaryotes and eukaryotes [3] with about 270 occurrences in the human proteome (see next section). PDZ-domain containing proteins are implicated in diverse cellular functions such as establishment and maintenance of cell polarity [4], signal transmission in neurons [5] or in visual and auditive processes in the eye and ear [6,7], cell migration [8], and regulation of cell junctions [9] (for reviews see [10,11]).

The core PDZ fold adopts an antiparallel β barrel structure [12] comprising 5–6 β strands and 1–2 α helices (Fig. 1A). PDZ domains mainly recognize PDZ-binding motifs (PBMs) that are situated at the very C-terminus of proteins. Some PDZ domains may also bind internal (i.e. non-C-terminal) PBMs [13,14] or lipids [15]. PBMs bind via β augmentation to PDZ domains, e.g. PBMs adopt a β strand that pairs in an antiparallel manner with the β2 strand of the PDZ domain (Fig. 1A). The carboxylate group of the last residue of the SLiM (here, the term peptide will be equally used) is hydrogen-bonded to backbone amides of residues from the carboxylate binding loop (β1–β2 loop), thereby determining the C-terminal peptide selectivity of PDZ domains (Fig. 1A). Based on the recognition of C-terminal SLiMs, peptide positions are numbered starting from the last residue (position 0, p0) going backwards (p−1, p−2, and so forth). The last residue is almost always a hydrophobic residue, mainly Val, Leu or Ile. The third last peptide residue (p−2) can be either Thr or Ser (class I), hydrophobic (class II), or Glu or Asp (class III), thereby defining three main categories of PDZ-binding motifs [16,17]. Thus, recognition of SLiMs by PDZ domains is based on residues of two key peptide positions, p0 and p−2.

Indeed, it has been generally observed that SLiMs have on average less than four defined positions [2]. Given this small binding interface, numerous studies addressed the question about how SLiMs can fulfill the need for specific protein interactions in cell signalling [18–22]. An increasing number of studies now suggests that protein interactions in cell signalling are not only determined by their minimal interacting fragments (e.g. core globular domain

* Corresponding authors.
E-mail addresses: katja.luck@unistra.fr (K. Luck), gilles.trave@unistra.fr (G. Travé).

0014-5793/$36.00 © 2012 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.febslet.2012.03.056

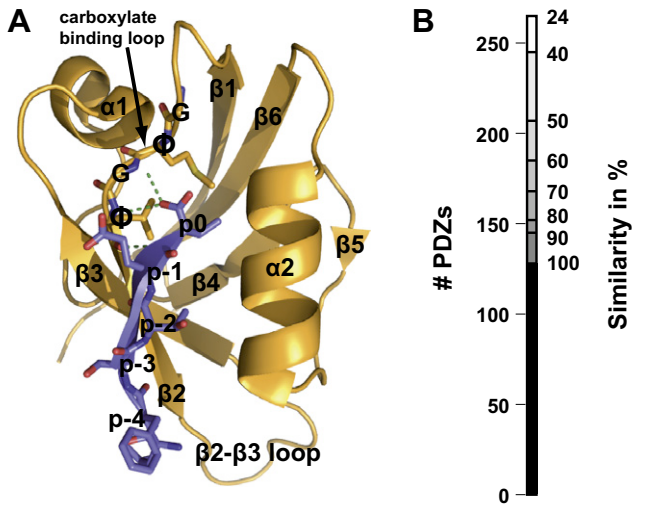


Fig. 1. Available structural information on PDZ domains. (A) Structure of the PDZ domain of AF6 (Afadin) bound to a C-terminal peptide (LFSTEV) derived from Bcr (PDB ID: 2AIN [112]). The secondary structure elements, peptide positions, and the common signature of the carboxylate binding loop (GϕGϕ where ϕ represents a hydrophobic residue) are indicated. Green dashed lines represent hydrogen bonds that are established between Val at p0 and the carboxylate binding loop. Figure was created with Pymol [113]. (B) Diagram that illustrates the number of human PDZ domains for which there is at least a structure of a PDZ domain in the PDB with X % sequence similarity based on local sequence alignments from BLAST searches (e.g. for 152 human PDZs there is a structure of a PDZ in the PDB with at least 80% sequence similarity.)

and SLiM) but also by their *context* [23–26]. The context can be either encoded within the *sequences* of the two interacting proteins; or defined by the *cellular environment*. In our definition, cellular context comprises factors that influence the temporal and spacial distribution of proteins, thereby determining when and under which conditions (e.g. local concentration) two proteins will meet in order to bind each other [25]. Sequence context comprises the regions in proteins that surround SLiMs or globular domains and that were shown to have an impact on the domain–SLiM interactions. Here, we refer to sequence context as *extensions*, if they occur directly upstream or downstream of the SLiM or domain and if they are not part of other domains. Sequence context that is not considered as extensions consists of neighbouring domains and regions that are not in the neighbourhood of the SLiM or domain.

The repertoire of structures on PDZ domains deposited in the Protein Data Bank (PDB) [27] is immense (Fig. 1B) providing us with a unique perspective on the structural diversity that exists within one protein domain family. Given the accumulated knowledge about PDZ–peptide interactions, they can serve as a model system to understand how domain–SLiM interactions are influenced by their sequence context. In the following, we will review studies that provide insights on extensions of PBMs and PDZs as well as studies that investigated the interplay between PDZ domains and their neighbouring domains.

2. Precision of the total number of PDZ domains in the human proteome

There has been considerable confusion about the total number of PDZ domains in the human proteome (hereafter called human PDZome). Numbers that are frequently referred to in the "PDZ literature" range from about 250 up to 900. Based on the articles that justify the number of human PDZs claimed in their text, we identified three different sources. Articles that claim a total number of 450 human PDZ domains (see suppl. data for references), refer to the database SMART [28]. Indeed, SMART provides a list of the human PDZome. However, upon closer inspection, this list turned out to be highly redundant, probably because it has been based on a set of protein sequences containing several isoforms for the same gene that have identical PDZ sequences. By now, SMART decreased the original number from 450 down to 364 (as of February 2012). Spaller [29] represents the second source, which claimed a total number of 918 human PDZ domains. In an erratum published four years later, Spaller corrected this number down to 234 and explained that he had been misled by erroneous numbers that were present in the preprint (but not the final version) of Bhattacharyya et al. [30]. Nonetheless, articles are still being published that seem to ignore the erratum [31]. Finally, several independent studies [32–34] that aimed at finding all human PDZ domains for bioinformatic analyses converge on a total number of about 270 PDZs in the human proteome without counting alternatively spliced forms of PDZs. This latter number, which we confirmed in our own investigations, is most probably the best estimate of the size of the human PDZome.

3. Extensions of PBMs

In numerous low-scale and large-scale studies as well as biocomputational analyses researchers sought to decipher the specificity rules of PDZ–peptide interactions [19,35–38]. Apart from peptide residues at p0 and p−2, which are the hallmarks for recognition by PDZs, the role of other residues of the PBM is more variable. Some studies suggested that the contribution of residues at p−1 and p−3 to the overall binding event is minor whereas in other studies PDZ domains exhibited clear preferences for certain residues at these sites over others [39–43]. More and more studies now converge on the idea that PBMs should at least be extended to p−4 as residues at this position have also been observed to significantly contribute to the binding to PDZs [19,35]. By extending the PBM further and further, a few interesting studies indicate that peptide residues up to p−10 are also implicated in PDZ binding. In the following, we define the core PBM as consisting of the last four residues (these residues clearly bind in the binding pocket of PDZ domains between the β2 strand and the α2 helix) and any longer PBMs will be considered as being extended.

3.1. Residues upstream of the core PBM modulate the binding affinity to PDZ domains

Interesting observations about PDZ–peptide recognition involving an extended PBM were obtained when comparing the binding of peptides derived from Wnt-signalling protein βCatenin and inward rectifier K(+) channel protein Kir2.3 to the PDZ domain of TIP1 (tax-interacting protein 1). A long βCatenin peptide bound stronger than a short peptide to the TIP1 PDZ [44,45]. The main contributor to this difference in affinity is most probably Trp at peptide position p−5 (in the following, we will write W_{-5}). Mutation of W_{-5} to Ala significantly weakened binding, as did mutation of Pro to Ala or Ser in the β2–β3 loop of the PDZ domain facing W_{-5} [44]. Interestingly, mutation of R_{-5} to Trp in Kir2.3 peptide led to an astonishing increase in binding affinity from 6.4 μM to 8.5 nM [46] (Table 1). As for the TIP1-βCatenin interaction, W_{-5} in the βPix C-terminus (Rho guanine nucleotide exchange factor 7) contributes to the binding to PDZ1 of SHANK1 (SH3 and multiple ankyrin repeat domains protein 1) as its mutation to Ala reduced binding affinity, too [47] (Table 1).

3.2. Electrostatic interactions between residues of extended peptides and of PDZ domains

The charge of the residue at position p−5 of the C-terminus of two inward rectifier K(+) channel proteins GIRK3 and IRK1

Please cite this article in press as: Luck, K., et al. The emerging contribution of sequence context to the specificity of protein interactions mediated by PDZ domains. FEBS Lett. (2012), http://dx.doi.org/10.1016/j.febslet.2012.03.056

Table 1
Mutagenesis performed to study extended PBMs. The table summarizes the mutational data obtained from studies that analysed interactions between residues from extended PBMs and PDZ domain residues. If available, measured binding affinities are indicated. wt = wild type, Cter = C-terminal extension of PDZ, alt.spl. = alternatively spliced, Nter-PDZ = construct comprising the PDZ extended at its N-terminus, AA = amino acids, CC = coiled-coil.

PDZ	peptide	peptide sequence	PDZ modification	peptide modification	affinity in μM	ref
PAR3-3/3	PTEN	DEDQHSQITKV	wt	11AA, wt	19	[58]
			wt	8AA, wt	/	
			wt	D-10A/E-9A/D-8A	weaker	
			β2-β3:K609/R611A	11AA, wt	240	
			β2:K606/β2-β3:K609/R611A	11AA, wt	550	
			β2:K606A	11AA, wt	200	
	Cadherin	YGSDPQEELII	wt	12AA, wt	6	[57]
			wt	6AA, wt	28	
			wt	D-7A	42	
			wt	Y-10A	15	
			β2-β3:R609A	12AA, wt	25	
			wt	phosphoS-8	2	
Erbin-1/1	ErbB2	EYLGLDVPV	wt	9AA, wt	50	[61]
			wt	phosphoY-7	128	
TIP1-1/1	Catenin	QLAWFDTDL	wt	9AA, wt	0.19	[44]
			β2-β3:P45A/S	wt	3	
			wt	W-5A	20	
			wt	5AA, wt	21	[45]
	kir2.3	ISYRRESAI	wt	9AA, wt	6.4	[46]
			wt	10AA, wt	45	[114]
			wt	R-5W	0.0085	[46]
			wt	S-2D	/	
			wt	R-5W/S-2D	binding	
			α2:D91A	9AA, wt	29	
SHANK1-1/1	βPix	AWDETNL	wt	CC domain	2.6	[47]
			wt	W-5A	weaker	
DLG1-2/3	E6	RRETQV	wt	6AA, wt	1	[54]
DLG1-2/3	APC	RHSGSYLVTSV	wt	11AA, wt	1	[53]
			β2-β3:Q339P	11AA, wt	10	
DLG1-1/3	APC	RHSGSYLVTSV	wt	11AA, wt	18	
DLG1-2/3	NR2B	KLSSIESDV	wt	9(14?)AA, wt	1	[55]
DLG1-1/3	NR2B	KLSSIESDV	wt	9(14?)AA, wt	10	
			β2-β3:P245Q	9(14?)AA, wt	5	
			β2-β3:P245A	9(14?)AA, wt	10	
MAGI1-2/6	E6	SSRTRRETQL	wt	10AA, wt	3	[51]
			wt	R-7A/R-5A/R-4A	7	[51]
			wt	L/V at p0	0.25	[52]
			β2-β3:DEPDE/NQPNQ	L/V at p0	35	[51]
			β2-β3:DEPDE/AAPAA	L/V at p0	60	[51]
			Cter:S113R/L114K/V115R	L/V at p0	0.75	[52]
			Cter:S113G/L114G/V115G	L/V at p0	1.25	[52]
hPTP1E-2/5	APC	HSGSYLVTSV	wt	10AA, wt	200	[64]
			alt.spl.	10AA, wt	/	
	RIL	VAVYPNAKVELV	wt	12AA, wt	1100	
			alt.spl.	12AA, wt	19000	
Harmonin-1/3	Sans	PALEDTEL	Nter-PDZ	SAM-PBM	0.001	[66]
			Nter-PDZ	PBM alone	1	
			Nter-PDZ	SAM alone	40	
ZO1-2/3	Connexin43	ASSRPRPDDLEI	dimer	9AA, wt	17	[69]
			dimer	12AA, wt	7	
			β2-β3: GGGAinsertion	9AA, wt	/	
			dimer	phosphoS-9	>100	

determined the binding of these two proteins to either the PDZ domain of SNX27 (sorting nexin 27) or the first two PDZ domains of PSD95 (postsynaptic density protein 95, DLG4). Exchanging Glu at position p−5 of GIRK3 with Arg, as observed in IRK1, was sufficient to induce the binding of GIRK3 to PSD95 and to disrupt its interaction with SNX27 [48].

Please cite this article in press as: Luck, K., et al. The emerging contribution of sequence context to the specificity of protein interactions mediated by PDZ domains. FEBS Lett. (2012), http://dx.doi.org/10.1016/j.febslet.2012.03.056

In a recent study [49], molecular dynamics simulations suggested ionic contacts between the β2–β3 loop of PDZ3 of DLG4 (disks large homolog 4) and of a C-terminal peptide derived from CRIPT (cysteine-rich interactor of PDZ3) that seemed to be important for peptide binding.

Structural and mutagenesis studies provided evidence for electrostatic interactions between PDZ2 (also referred to as PDZ1) of MAGI1 (membrane-associated guanylate kinase inverted 1) and an extended C-terminal peptide derived from Human Papillomavirus (HPV) 16 E6 [50–52]. Mutation of the negative charges in the β2–β3 loop of PDZ2 to Gln and Asn or Ala significantly reduced the binding affinity as did mutation of the positive charges at position -4, -5 and -7 to Ala in the E6 peptide [51] (Table 1). In another study, the contribution of upstream peptide residues to the binding affinity to PDZ2 of MAGI1 has been more generally assessed. By measuring binding affinity of several peptides of different length, the preference of PDZ2 of MAGI1 for peptides with positive charges upstream the core PBM has been confirmed [26]. This is one clear case where an increase in specificity is driven by interactions between residues of extended PBMs and residues of the β2–β3 loop. In contrast, the β2–β3 loop of PDZ3 of MAGI1 did not seem to contribute at all to peptide binding [26]. The same peptides were also assayed for binding to PDZ3 of the cell polarity protein hScrib revealing that extended peptides generally bound with higher affinity to PDZ3. This study demonstrates that depending on the PDZ in question, interactions between upstream peptide residues and the β2–β3 loop can have different implications on peptide binding [26].

3.3. The sequence of the β2–β3 loop influences peptide binding

C-terminal peptides of APC (adenomatous polyposis coli protein), the glutamate receptor subunit NR2B, and HPV18 E6 bind stronger to PDZ2 of DLG1 (disks large homolog 1) as compared to PDZ1 of the same protein [53–55] (Table 1). Several studies suggest that amino acid differences in the β2–β3 loop of these two PDZ domains mainly account for the different binding preferences via interaction with residues of extended PBMs. Mutation of a Gln in the β2–β3 loop of PDZ2 to Pro (the corresponding residue in PDZ1) decreased binding affinity for APC [53]. The converse mutation (Pro to Gln) directed at the equivalent position of the β2–β3 loop of PDZ1 increased binding affinity of PDZ1 for NR2B whereas mutation of the same Pro of PDZ1 to Ala did not alter the binding affinity [55]. The Pro in the β2–β3 loop of PDZ1 might also be responsible for weaker binding to E6 in comparison to PDZ2. Liu et al. [54] noticed that the neighbouring residues of the Gln in PDZ2 adopt a particular conformation that contributes to peptide binding, and suggested that the Pro in PDZ1 did not allow its neighbouring residues to adopt a similar favorable conformation. One reason for the weaker binding displayed by PDZ3 to these peptides might be its shorter β2–β3 loop, which does not provide such a platform for extended peptide binding as does PDZ2 [53–55].

3.4. Extended PBMs confer dual binding specificity to PDZ domains

A few PDZ domains were shown to have dual specificity, e.g. binding to PBMs of class I and class II. The dual specificity of PICK1 (protein interacting with C kinase 1) has been mainly attributed to specific residues of the α2 helix [56]. In contrast, the dual specificity of PDZ3 of the cell polarity protein Par3 can be attributed to an extended binding pocket [57]. PDZ3 of Par3 binds to long C-terminal peptides derived from both the vascular endothelial Cadherin (class II PBM, 12 residues) and phosphatase PTEN (class I PBM, 11 residues). However, it significantly binds weaker to a shorter (6-residue long) Cadherin peptide and does not detectably bind anymore to a shorter (8-residue long) PTEN peptide [58,57] (Table 1).

Analysis of the two available NMR structures of these two complexes revealed several contacts between negatively charged residues in the extended peptides and positively charged residues of the β2 and β3 strand as well as the β2–β3 loop of PDZ3 [58,57] (Fig. 2A and B). Whereas the Cadherin peptide mediated more favourable interactions by means of its last four peptide residues, the PTEN peptide established more favourable contacts to PDZ3 with its upstream residues. At first sight this might be interpreted as an example where a longer binding pocket leads to more promiscuous binding behaviour. Yet, class I peptides such as PTEN have to fit very well to the extended binding site with their upstream residues in order to be bound by PDZ3, a constraint that might only be fulfilled by a few peptide sequences.

3.5. Conformational changes of the β2–β3 loop upon peptide binding

As indicated by the previous examples, residues from extended PBMs mainly modulate binding affinity to PDZ domains via interaction with residues of the β2–β3 loop. Comparison of available apo (unbound) and holo (bound) NMR and crystal PDZ structures revealed that the β2–β3 loop either changes conformation upon peptide binding or remains unchanged. Two examples of the latter case are represented by PDZ2 of DLG1 and Erbin (Erbb2-interacting protein) PDZ of which the β2–β3 loops exist in a stable conformation when no peptide is bound and this conformation remains unchanged upon peptide binding, also when complexed to different peptides. In Erbin PDZ, there is a chain of aliphatic contacts from the β2–β3 loop to the β3 strand and the β4 strand (Fig. 2C). Additionally, N1345 of the β2–β3 loop seems to establish a hydrogen bond to the backbone of the loop. These interactions between residues of the Erbin PDZ are probably the driving forces that keep that loop very rigid providing a stable platform for peptide binding. The β2–β3 loop of PDZ1 of ZO1 (Zonula occludens protein 1) is equally long as that of Erbin PDZ, and it adopts a similar conformation upon peptide binding, but it displays a different conformation in its unbound form. Here, the loop seems to restructure upon peptide binding allowing for accommodation of upstream peptide residues (Fig. 2D). Similar observations were obtained for the β2–β3 loop of the TIP1 PDZ domain (Fig. 2E2) and the PDZ of SHANK1 [47].

3.6. The β2–β3 loop can form an additional peptide binding pocket that accommodates upstream peptide residues

In many of the examples mentioned in the previous subsections, the β2–β3 loop contributed to the formation of an additional peptide binding pocket together with residues from strands β2, β3, and sometimes β4. In particular, an aromatic residue (mainly F or Y), located right at the beginning of the β3 strand, is often involved in the formation of this additional pocket. This residue contributed to the binding of peptide residues upstream position -3 and seemed to serve as an anchoring point for the structuring of the β2–β3 loop. The conserved aromatic character of this position in the family of PDZ domains suggests that residues at this position might be of more general importance for the structure and function of PDZs [49].

In general, this additional pocket was frequently observed to have hydrophobic character being occupied by upstream peptide residues with large aliphatic side chains such as Trp, Tyr, or Arg (Erbin PDZ, PDZ1 of ZO1, TIP1 PDZ, Fig. 2C–E1, respectively). However, in the case of the PDZ of SNX27, this additional pocket is rather of hydrophilic character being formed by three arginines

Please cite this article in press as: Luck, K., et al. The emerging contribution of sequence context to the specificity of protein interactions mediated by PDZ domains. FEBS Lett. (2012), http://dx.doi.org/10.1016/j.febslet.2012.03.056

	PDZ	peptide	peptide sequence	pep. pos.	domain position
A	PAR3-3/3	PTEN	DEDQHSQITKV	-8	β2-β3:R609
				-9	β2:K606, β3:K611
				-10	β2-β3:K611
B	PAR3-3/3	Cadherin	YGSDPQEELII	-4	β2-β3:R609
				-7	β2-β3:R609, β2-β3:K611
				-9	β2-β3:K611
				-10	β2-β3:K611
C	Erbin-1/1	ErbB2	EYLGLDVPV	-7	β2-β3:P1305
	Erbin-1/1	artificial peptide	TGWETWV	-4	β2:S28, β3:T48, β2-β3:P37, β2-β3:F38
D	ZO1-1/3	artificial peptides	WRRTT[WY]L	-4	β2-β3:N44, α2:H88
				-5	β2-β3:N44, β2-β3:P45, β2-β3:F47
				-6	β3:V55, β2:S39, β2-β3:H46, β3:S57, β2-β3:E51
E	TIP1-1/1	Catenin	QLAWFDTDL	-4	α2:H90, β2-β3:Q39
				-5	β2:G34, β2-β3:N44, β2-β3:P45, β2-β3:F46, β3:Y56, β3:T58
	TIP1-1/1	Kir2.3	ISYRRESAI	-4	α2:H90, α2:D91
				-5	β2-β3:Q43, β2-β3:N44, β2-β3:P45
F	SHANK1-1/1	βPix	AWDETNL	-4	β2-β3:K682, α2:R736
				-5	β2:R679, β2-β3:F696, β3:Y701
				-8	β2-β3:K682
G	SHANK1-1/1	GKAP1a	EAQTRL	-5	β2:R679
	SNX27-1/1	GIRK	ESESKV	-5	β2:R56, β2-β3:R66, β3:H78, β4:R98
H	DLG1-2/3	E6	RRETQV	-4	β2-β3:G334, β2-β3:N338, α2:E384
				-5	β2-β3:H340
	MAGI1-2/6	E6	RSSRTRRETQL	-4	β2-β3:D35
				-5	β2-β3:E39, C-ter:S113, C-ter:L114, C-ter:V115
				-6	C-ter:S113, C-ter:L114, C-ter:V115
				-7	β2-β3:E36, β2-β3:D38
	hPTP1E-2/5	RA-GEF2	EADENEQVSAV	-5	β2-β3:V26, β2-β3:N27
I	Harmonin-1/3	Sans	PALEDTEL	-5	β2:R103, β2-β3:F108, β3:C110, C-ter:V172, C-ter:K173, C-ter:L180
				-6	C-ter:K173
				-7	β2-β3:F108
J	Harmonin-2/3	Cadherin23	TPLEITEL	-4	α2:K73
				-5	β2-β3:P31, β3:F38, C-ter:L97, C-ter:F98
K	ZO1-2/3	Connexin43	RPRPDDLEI	-4	β5-α2:N239
				-5	allows for kink to avoid clashes with β5-α2 loop
				-6	β5-α2:E238

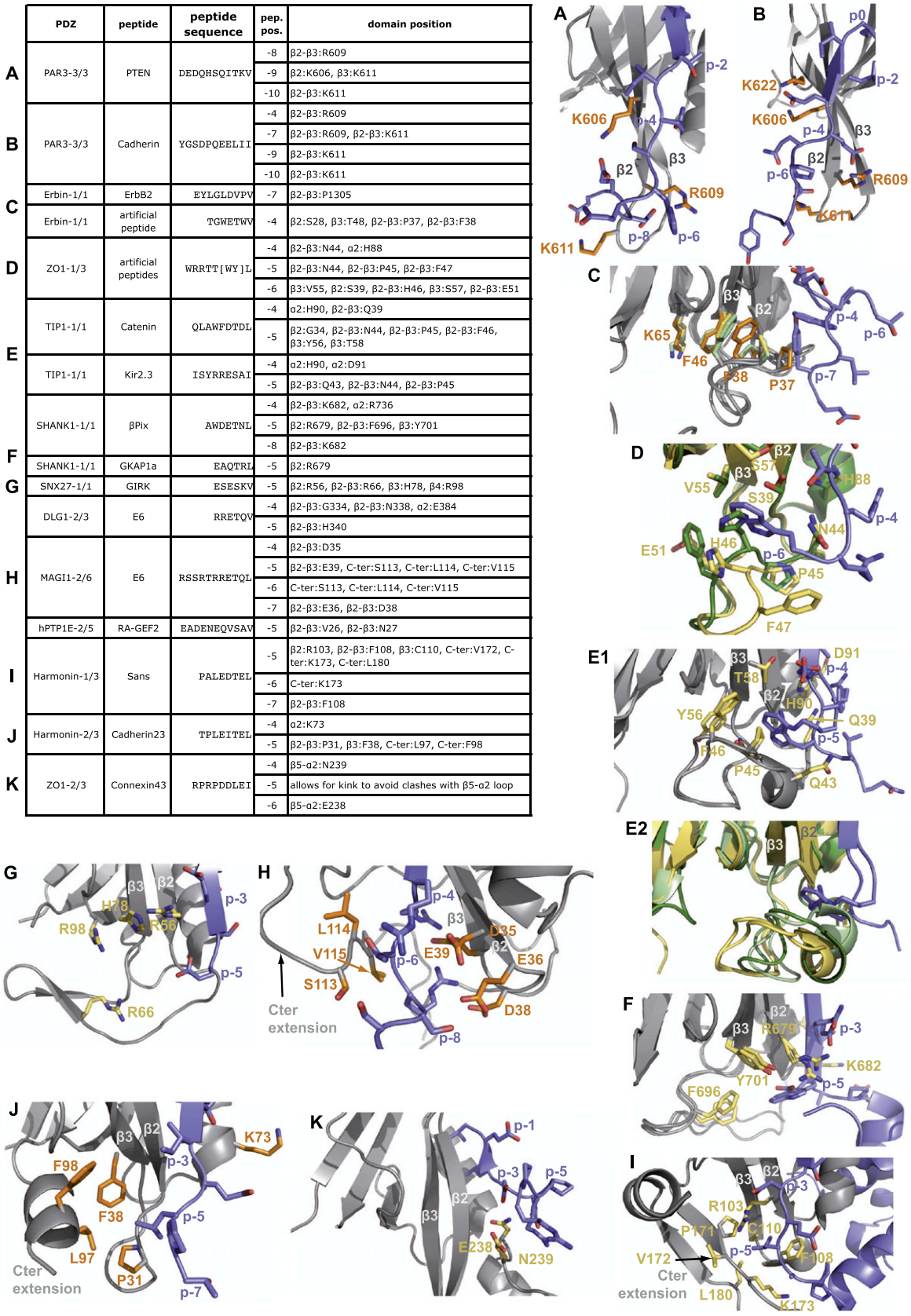


Fig. 2. Contacts between residues of extended PBMs and PDZ residues. The table summarizes the contacts that were observed between residues in extensions of PBMs and residues of PDZ domains or their extensions. A structural representation is provided for most of these PDZ–peptide complexes. Colour code: peptide residues in blue, PDZ residues from holo NMR structures in orange, from apo NMR structures in dark-green, from holo crystal structures in yellow, and from apo crystal structures in light-green. Structural information was used from the following PDB entries: (A) 2K20 [58]; (B) 2KOH [57]; (C) 1MFG [61], 1N7T [59], 2H3L [60]; (D) 2H2B, 2H3M [60]; E1: 3DIW [44], 3GJ9 [46]; E2: 3DJ1, 3DIW [44], 3GJ9 [46], 2KG2 [114]; (F) 3L4F [47], 1Q3P [115]; (G) 3QGL [48]; (H) 2KPL [52], 2I04 [50]; (I) 3K1R [66]; (J) 2KBS [67]; (K) 3CYY [69]. Figures were created with Pymol [113] (see suppl. data for the pymol session files).

Please cite this article in press as: Luck, K., et al. The emerging contribution of sequence context to the specificity of protein interactions mediated by PDZ domains. FEBS Lett. (2012), http://dx.doi.org/10.1016/j.febslet.2012.03.056

from the β2 strand, β2–β3 loop, and β4 strand, and accommodates E_{-5} of the C-terminal peptide of GIRK3 (Fig. 2G). Another interesting example is that of the trimer of βPix bound to the SHANK1 PDZ. The homotrimer is formed by the coiled-coil domain of βPix leading to three closely located C-termini all carrying the PBM of βPix [47]. However, due to steric hindrance, only one of the three C-termini is accessible for binding to SHANK1 PDZ [47]. Here, the additional pocket formed by the β2–β3 loop, β2, and β3 strand adopts a hemispheric shape and interacts with one side of the ring of W_{-5} from βPix (Fig. 2F). The other side seems to be covered from solvent by residues from the coiled-coil domains. In that way, W_{-5} may contribute to the stabilisation of the whole complex.

In most cases, this additional pocket is occupied by peptide residues from position -5. However, it can sometimes be occupied by residues at different positions, such as W_{-4} for a phage display-derived peptide bound to Erbin PDZ (Fig. 2C) [59], W_{-6} for a phage display-derived peptide bound to PDZ1 of ZO1 (Fig. 2D) [60], and Y_{-7} for ErbB2 peptide bound to Erbin PDZ (Fig. 2C) [61]. Together with observations that peptide residues at position -4 can either contact domain residues from the β2 or β3 strand, from α2 helix or from the β2–β3 loop (see table in Fig. 2), this demonstrates to which extent peptides can adapt to PDZ domains. A rigid definition of pairs of domain and peptide residues as often considered in PDZ–peptide interaction predictors, would not reflect this adaptability and is therefore likely to be an inappropriate model.

3.7. Alternative splicing of the β2–β3 loop can modulate the peptide binding properties of PDZ domains

A very different example of influence of the β2–3 loop on peptide binding is provided by PDZ2 of the tyrosine protein phosphatase hPTP1E (or its mouse homolog PTP-BL). This PDZ domain represents one of the few PDZs subject to alternative splicing, and to our knowledge is the only one that was studied in detail (with in total 8 structures deposited in the PDB). The alternatively spliced form (PDZ2as) exhibits an insertion of 5 residues (VLFDK) at the start of the β2–β3 loop that abrogates binding to the C-terminal peptide of APC and RIL (reversion-induced LIM protein) [62–64] (Table 1). Two NMR structures of PDZ2as exist but the fine structural interpretation of how the insertion negatively influences peptide binding remains a matter of debate as the two structures considerably diverge from each other (Fig. S1) [63,64]. The alternative splicing event leads to replacement of the GG hinge region at the beginning of the β2–β3 loop (a conserved feature in the PDZ fold) by the more conformationally restricted VL sequence [64]. This appears to induce a displacement of the loop, which, together with a global destabilisation of the domain, seems to be the main cause for the observed inability of PDZ2 as to bind to C-terminal peptides [64]. It will be interesting to see how alternative splicing might modulate the binding behaviour of other PDZ domains and whether these are used in vivo to regulate PDZ–peptide interactions [65].

3.8. Upstream residues of a PDZ-bound peptide can interact with residues that do not belong to the core of that PDZ domain

In most of the cases upstream peptide residues interacted with residues of the core PDZ domain but there are also a few very interesting instances where they interact with residues from PDZ extensions or from other proteins. Atomic contacts could be observed by NMR between residues (Ser, Leu, and Val) of the C-terminal extension of PDZ2 (PDZ1) of MAGI1 and R_{-5} and T_{-6} of a bound C-terminal peptide derived from HPV16 E6 [52] (Fig. 2H). This significantly restricted the disordered conformation of the C-terminal extension of PDZ2 in the presence of bound peptide [52]. Mutation of the three residues SLV to either RKR or GGG resulted in threefold and fivefold reduced binding affinity in comparison to the wild type, respectively [52] (Table 1).

Harmonin and Sans are two proteins that are implicated in the Usher syndrome. The N-domain, PDZ1, and a C-terminal extension of PDZ1 of Harmonin form an integral domain that binds to a PBM located at the C-terminus of Sans [66]. This interaction is further stabilised by interaction between the SAM (sterile alpha motif) domain of Sans and PDZ1 of Harmonin leading to an extremely tight complex (reviewed in detail by Wang et al. [33]) (Table 1). L_{-5} inserts into a hydrophobic pocket formed by residues from the β2 and β3 strand, from the β2–β3 loop, and from the C-terminal extension (Fig. 2I). Additionally, hydrogen bonds were observed between Lys from the C-terminal extension and the backbone of A_{-6} [66]. The PDZ2 domain of Harmonin, complexed to the C-terminal peptide of the cell adhesion protein Cadherin23, exhibits an additional C-terminal α helix [67]. Interestingly, the C-terminal tails of Cadherin23 and of Sans are very similar with identical residues at p−5, −4, −2, −1, 0. L-5 of Cadherin23 interacts with residues of the extension and the core PDZ domain in a similar way as L-5 of Sans, although in the case of PDZ2 it is a hydrophobic patch that is formed instead of the pocket reported for PDZ1 (Fig. 2J). Comparing the structure of PDZ2 of Harmonin with the one from PDZ3 of ZO1 revealed a very similar hydrophobic patch formed from the additional C-terminal α helix of PDZ3 and residues from the β2–β3 loop and the β2–β3 sheet. No difference in binding affinity could be observed to the C-terminal peptide of the gap junction protein Connexin45 when the helical extension of PDZ3 of ZO1 was removed [68]. Yet we speculate that an increase in affinity might have been observed for peptides possessing residues at upstream peptide positions that would have been capable to interact with this hydrophobic surface.

ZO1 and ZO2 PDZ2 were shown to exist as homodimers [69,70]. The complex of ZO1 PDZ2 bound to the PBM of Connexin43, revealed that due to dimerization the extended peptide binding pocket is altered. This has been reviewed more in detail by Wang et al. [33]. The β5–α2 loop of the other PDZ is placed in front of the end of the binding pocket forcing the peptide to make a bend at position -5 that is occupied by a Pro in Cx43. D_{-4} and R_{-6} are involved in several interactions with residues from the β5–α2 loop (Fig. 2K, Table 1) [69].

Unfortunately, the Harmonin-Sans/Cadherin23 and ZO1-Cx43 complexes have neither been mutated nor compared to complexes of Harmonin and ZO1 with PBMs of other known binding partners. Such analyses would have helped to better understand the contribution of observed interactions between residues from PBM and PDZ extensions to complex formation. However, given the numerous precise residue contacts observed, it is tempting to speculate that these additional interactions between PBM and PDZ might increase the specificity of the PDZ domains for their peptide targets.

3.9. General remarks

It clearly emerges that residues of PBM extensions influence the affinity of peptides to PDZ domains and some studies provide clear evidence that this increases the specificity of PDZ–peptide interactions. The β2–β3 loop contributes to an extended binding pocket. Owing to the huge variability of this loop in length (up to 36 residues, PDZ of ARHGAP21, PDB ID: 2YUY [71]) and sequence composition, observed interactions between residues of the loop and of the peptide were of impressive diversity. It seems rather impossible to derive any rule about how the β2–β3 loop and the peptide extensions contribute to PDZ–peptide binding. Most likely, this complex system of interactions is at the moment unpredictable.

Interestingly, the contribution of upstream peptide residues for the binding to PDZ domains seems to be subject to regulation as

Please cite this article in press as: Luck, K., et al. The emerging contribution of sequence context to the specificity of protein interactions mediated by PDZ domains. FEBS Lett. (2012), http://dx.doi.org/10.1016/j.febslet.2012.03.056

they were observed to overlap with phosphorylation sites. Reported phosphorylation of upstream PBM residues led either to an increase in binding affinity (S_{-8} of Cadherin bound to PDZ3 of Par3 [57]) or to a decrease (Y-7 of ErbB2 bound to Erbin PDZ [61], S-9 of Cx43 bound to PDZ2 of ZO1 [69]) [72] (Table 1). It appears that experimental studies on PDZ–peptide interaction specificities should systematically include peptides of different length (e.g. core PBMs and extended PBMs) to account for the role that upstream peptide residues play for PDZ binding.

In principle, extending the binding interface between PDZs and peptides is likely to increase affinity between both. However, none of the cellular peptides that we have described above possessed an optimal sequence for the entire binding site. This observation is reflected by the fact that three artificial peptides (a phage display peptide, mutated Kir2.3 peptide, mutated HPV16 E6 peptide) displayed much higher affinity to PDZ domains than the identified natural binding partners (Table 1). Noteworthy enough, a high affinity interaction is not necessarily specific, as one or both of the interaction partners might bind many other proteins with similarly high affinity. Conversely, a low affinity interaction can be specific, if most other interactions engaged by the two partners are of even lower affinity. As mentioned in the introduction, SLiMs seem too short to fully account for specific protein interactions. It is tempting to speculate that the extended binding interfaces reviewed herein serve to provide a balanced mix of advantageous and disadvantageous interface contacts, which guarantee the weak yet specific interactions typically observed in cell signalling.

4. Sequence context of PDZ domains

4.1. Defining the core and boundaries of PDZ domains: theory and practise

The canonical core PDZ fold is defined by the arrangement of its secondary structure elements, namely six β strands and two α helices (Fig. 1A). In theory, the boundaries of a PDZ domain are thus defined by the first residue of the first β strand and the last residue of the last β strand of the canonical PDZ fold. The web servers SMART [28] and Pfam [73], both based on Hidden Markov Models, efficiently detect PDZ domains, including quite divergent instances presenting structural rearrangements (such as the PDZ-like domains from HtrA and EpsC (see below) [74,75]). However, these programs have not been developed for accurately predicting the boundaries of the core domain structures.

In practise, the boundaries of most experimental PDZ constructs that have been used for structural studies or interaction assays, generally extend beyond the strict boundaries of the core PDZ structure. Such extensions vary from a few residues to longer stretches, which may be structured. Whereas Bhattacharya et al. [76] and Wang et al. [33] proposed approaches for predicting some secondary structure elements within PDZ extensions, the conformation of extensions and their impact on the solubility, stability and peptide binding properties of PDZ constructs (see following paragraphs) remain largely unpredictable.

4.2. A PDZome-wide database of suggested PDZ construct boundaries derived from structural information

The lack of a general rule for designing boundaries of PDZ domain constructs raises a practical issue concerning past and future experimental studies aimed at producing and comparing the functional properties of large numbers of PDZ domains. How can hundreds of PDZ constructs be properly designed? One solution to this problem may be to take advantage of the impressive structural data already available on PDZ domains. Indeed, successful structure determination probably represents the most relevant *a posteriori* quality proof for construct boundaries. Based on this principle, we propose here a list of manually curated domain boundaries sequences for the 266 PDZs that constitute to our knowledge the human PDZome (Table S1). For each known PDZ domain, we identified the three most similar PDZ domains for which structures were available and applied a set of hierarchized criteria (see suppl. material for protocol) to finally define the construct boundaries of the considered PDZ domain.

Noteworthy enough, the structures of about 120 human PDZs have been solved, and for most of the other human PDZs, the structure of one or several very closely related orthologous PDZ domains is available in the Protein Data Bank (PDB) (Fig. 1B). Therefore, the majority of constructs proposed in our curated list are either identical or highly similar to constructs for which a structure has been solved. On average, the constructs proposed in our list, including those of known structure, are 16 and 5 residues longer at the N- and C-terminus, respectively, than the start and end of the sequence predicted by SMART. In particular, SMART-predicted boundaries consistently excluded the first β strand of PDZ domains. Therefore, the raw SMART output is not sufficient for domain boundary prediction. This database provides a first suggestion of boundaries for the cloning of PDZ constructs from human or related proteomes. Of course, semi-empirical optimization of construct solubility and/or stability may still remain necessary in some cases [77].

4.3. PDZ extensions

The emerging roles of extensions of PDZ domains was addressed in a review by Wang et al. [33]. Here, we aim at complementing this review by focussing on recently published studies and by putting emphasis on the implications that extensions have on peptide binding.

4.3.1. Extensions that influence the dynamics, stability and solubility of PDZ domains

PDZ2 of NHERF1 (Na(+)/H(+) exchange regulatory cofactor 1) and PDZ3 of DLG4 (disks large homolog 4) have helical C-terminal extensions that were shown to be important for the stability and dynamics of the PDZ domain, respectively [76,78]. In both cases, removal of the extension did not alter the fold of the core PDZ domain but led to significantly decreased affinity to C-terminal peptides [76,78] (Table 2). Petit et al. [78] showed that a construct of PDZ3 of DLG4 lacking the extension, displays more side chain flexibility leading to increased entropy that makes peptide binding energetically less favourable. Both studies were reviewed more in detail by Wang et al. [33]. A recent molecular dynamics study [49] suggested that the helical extension of PDZ3 of DLG4 establishes ionic contacts with core PDZ residues. These contacts seemed to restrain the backbone flexibility of the β2–β3 loop and the carboxylate binding loop, thereby facilitating peptide binding.

The PDZ2 (PDZ1) domain of MAGI1 was shown to have a structured N-terminal and an unstructured C-terminal extension [52]. The C-terminal extension seemed to be essential for obtaining soluble and stable constructs. The N-terminal extension was shown to shield from solvent a hydrophobic patch that comprised a cysteine residue, thereby preventing the formation of soluble aggregates via non-native intermolecular disulfide bonds [77,52]. Mutations in the C-terminal (see previous section dealing with PBM extensions) and N-terminal (unpublished data) extensions weakened the binding affinity of PDZ2 to the HPV16 E6L/V peptide.

Please cite this article in press as: Luck, K., et al. The emerging contribution of sequence context to the specificity of protein interactions mediated by PDZ domains. FEBS Lett. (2012), http://dx.doi.org/10.1016/j.febslet.2012.03.056

Table 2
Mutagenesis performed to study sequence context of PDZ domains. The table summarizes the mutational data obtained from studies that analysed the influence of extended PDZ domains or multido-main constructs on peptide binding. If available, measured binding affinities are indicated. SS-bridge = disulfide bridge.

PDZ	peptide	peptide sequence	PDZ modification	affinity in μM	ref
DLG4-3/3	CRIPT	TKNYKQTSV	extended	1	[78]
			core	26	
		NYKQTSV	extended	3.6	[80]
			core	81	
			α3:phosphY397	14	
NHERF1-2/2	CFTR	...TEEEVQDTRL	core	5	[76]
			extended	0.26	
			β1:R153Q	0.93	
INADL-5/10	Kon/Perd	LLRRNQYWV	reduced	1.2	[7]
			oxidized	20.4	
INADL-45/10			reduced	2.5	
INADL-45/10	TRP	TGRMISGWL	reduced	140	
INADL-345/10	TRP	...TGRMISGWL	reduced	0.1	
INADL-5/10	PLCβ	KTQGKTEFYA	reduced	/	
INADL-45/10			reduced	30	
			oxidized	/	
Par6-1/1	Rhodamine	VKESLV	core	72	[88]
			CRIB-PDZ	54	
			CRIB-PDZ SS-bridge	13	
			CRIB-PDZ+Cdc42	6	
			CRIB-PDZ SS-bridge+Cdc42	6	
ZO1-PSG	JAM-A	EGEFKQTSSFLV	extended	35	[107]
ZO1-3/3				/	
ZO1-PSG			SH3:β2-β3:L549R	/	
ZO1-PSG	Connexin45	SGDGKTSVWI	extended	10	[68]
ZO1-3/3				75	
ZO1-PSG			SH3:β2-β3:L549R	63	
DLG-3/3	CRIPT	DTKNYKQTSV	-	14	[108]
DLG-PSG			-	0.8	
DLG-3/3	neuroligin	KRVHIQEISV	-	900	
DLG-PSG			-	71	

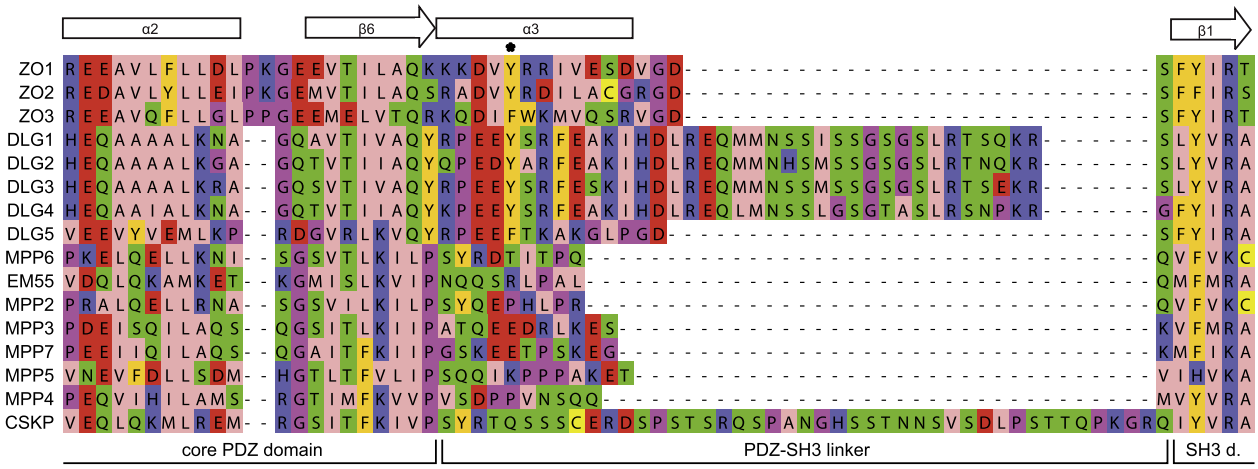


Fig. 3. Sequence alignment of the PDZ-SH3 linker of the human MAGUK family of proteins. Secondary structure elements on top of the alignment are indicated based on the structure of the PDZ3-SH3-GK module from ZO1 (PDB ID: 3SHW [68]). An initial alignment was built with Mafft [116] and corrected by hand using Jalview [117]. The asterisk indicates the phosphorylation site described for ZO1, which seems to be conserved for some members of the MAGUK family.

4.3.2. PDZ extensions as regulatory elements

DLG4 has a phosphorylation site (Y397) located in the helical C-terminal extension (hereafter called α3 helix) of PDZ3 [79]. Zhang et al. [80] studied the effect of this phosphorylation site on the structure of the extended PDZ3 domain and on ligand binding. Phosphorylated Y397 led to an equilibrium between a locally

unfolded and folded state of the α3 helix reflected by a fourfold decrease in affinity of phosphorylated PDZ3 in comparison to the unphosphorylated extended PDZ3 domain (Table 2).

DLG4 is a member of the membrane associated guanylate kinase (MAGUK) family of proteins together with four other DLG proteins, 3 ZO proteins, CASK (calcium/calmodulin-dependent serine protein kinase), and 7 MPP (membrane protein, palmitoylated) proteins [81]. All of these proteins share a common domain arrangement consisting of a PDZ domain followed by an SH3 (Src homology 3) and GK (guanylate kinase) domain (hereafter called PSG module). In addition to DLG4, the PDZ domain of the PSG module of ZO1 and DLG1 was also shown to possess an α3 helix. The α3 helix as well as the phosphorylation site seem to be conserved in the ZO and DLG subfamily of proteins (except for DLG5) (Fig. 3). Remarkably, the linker sequence between PDZ3 and SH3 of DLG1-4 is much longer than for the other members of the MAGUK-family. Three phosphorylated serines are reported in the linker region of DLG1-4 [82,83]. Furthermore, by using the ELM resource [84], we predicted in this linker a very likely actin binding site (LIG_Actin_WH2_2). To our knowledge, DLG proteins have not yet been demonstrated to bind to actin, yet they are already known to be involved in the regulation of actin filaments [85]. Therefore, it might be interesting to investigate the functional role of this actin binding site and its potential influence on peptide binding by the neighbouring PDZ3 and SH3 domains. The linker sequence of the MPP family is shorter than for the ZO and DLG proteins and does not display any conserved features. Furthermore, the region in MPP proteins that corresponds to the α3 helix observed in ZO and DLG proteins, tends to be proline-rich (Fig. 3) and is therefore unlikely to adopt a helical conformation.

In summary, this data suggests that a helical extension of the PDZ of the PSG module is a property shared by DLG1-4 and all three ZO proteins and that it might be used as a regulatory element to control peptide binding to PDZ3 via phosphorylation and actin binding.

4.3.3. PDZ extensions that modulate the conformation of the binding pocket

The PDZ domain of Par6 (partitioning defective 6 homolog) has in its unstructured N-terminal extension a CRIB (Cdc42/Rac-interactive binding) domain that adopts a β strand when bound by Cdc42 (Cell division control protein 42 homolog) [86]. The association of Cdc42 with the extended PDZ of Par6 leads to a 10-fold increase in affinity for C-terminal peptides bound by the PDZ [87] (Table 2). Several studies have been published investigating the mechanism and implications of the Cdc42-Par6 interaction that are in detail reviewed in Wang et al. [33]. Very recently, Whitney et al. [88] published structural details about how the signal resulting from Cdc42 binding to the CRIB domain is propagated to the binding pocket of the PDZ, thereby altering PDZ–peptide binding. By introducing a disulfide bridge between the otherwise very flexible CRIB domain and the core PDZ, they obtained a construct that appeared to mimick the structure of the CRIB-PDZ module when bound to Cdc42 and that was amenable for NMR studies. By comparing the structure of the disulfide-bridged mutant with a structure of the single PDZ and a structure from the Cdc42–Par6 complex, they revealed a switch in conformation of two residues (Lys and Leu) in the β1–β2 loop that reshapes the binding pocket upon Cdc42 binding and thereby facilitates peptide binding [88].

PDZ5 of INADL, a core component of the phototransduction pathway, has a pair of cysteines located in the peptide binding pocket, that is in reduced form in absence of light and forms a disulfide bridge upon light exposure [89]. C-terminal peptides were shown to bind significantly weaker to the oxidized form of PDZ5 in comparison to its reduced form [7] (Table 2). Interestingly, whereas the isolated PDZ5 was stable in its oxidized form, the reduced form of PDZ5 was prevailing within a PDZ4-PDZ5 tandem construct [7]. The crystal structure of PDZ4 and PDZ5 shows that they form a tight module that is stabilised by a C-terminal extension of PDZ5 that binds to the surface of PDZ4, and a N-terminal extension of PDZ4 that folds back onto PDZ4 [7]. Liu et al. [7] further provide interesting data that suggests that the C-terminal extension of PDZ5 is involved in a regulatory switch that changes the redox state of PDZ5 in the cell. Rapid light-induced acidification happens during signal transduction in the microvilli of eyes. This can lead to a local significant decrease of the pH that would be enough to protonate a histidine residue of PDZ4 leading to the disruption of its hydrogen bond with a threonine residue of the C-terminal extension. Consequently, the interaction between the C-terminal extension and PDZ4 is disturbed leading to a destabilisation of the whole module. This may result

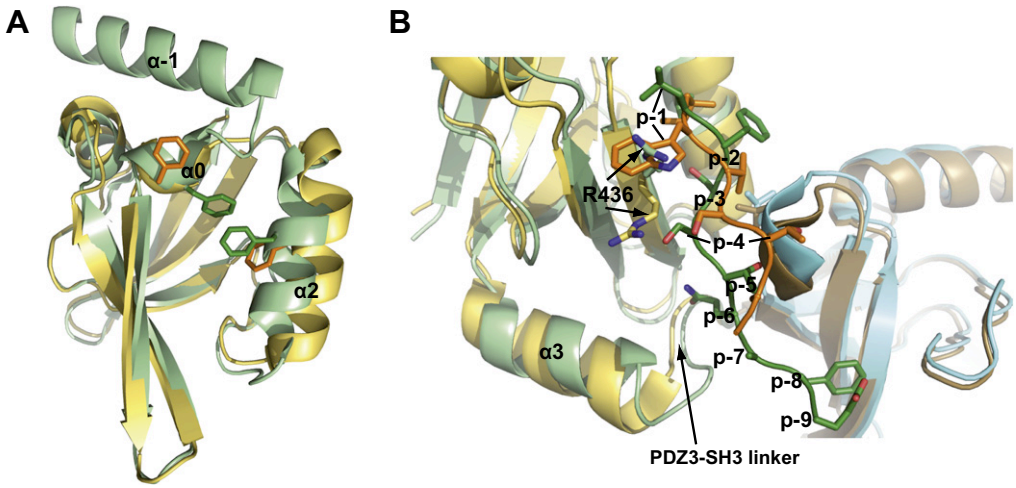


Fig. 4. Sequence context of PDZ domains. (A) Short (yellow, sPDZ, PDB ID: 2I6V) and long (green, lPDZ, PDB ID: 2I4S) version of the PDZ domain of the bacterial EpsC protein [75]. lPDZ has an additional N-terminal α helix (α−1) that leads to different conformations of α0 and α2. Two Phe are differently positioned (indicated in sticks) that lead in lPDZ to a closed peptide binding pocket. (B) Comparison of two complexes between the PDZ3-SH3-GK module of ZO1 and C-terminal peptides derived from JAM-A (forest-green) [107] and Cx45 (orange) [68]. The PDZ3 domains bound to JAM-A and Cx45 are shown in light-green and yellow, respectively, the corresponding SH3 domains in cyan and brown. The additional C-terminal α helix of PDZ3 is labelled α3. R436 possesses different conformations in the two complexes due to different residues at peptide position p−1 (W in Cx45 and L in JAM-A). The figures were created with Pymol [113] (see suppl. data for the pymol session files).

Please cite this article in press as: Luck, K., et al. The emerging contribution of sequence context to the specificity of protein interactions mediated by PDZ domains. FEBS Lett. (2012), http://dx.doi.org/10.1016/j.febslet.2012.03.056

in a change of the local environment of the cysteine pair allowing for the formation of the disulfide bond. Possibly, the redox switch is used to initiate the dissociation, from INADL, of proteins such as the Ca^2+-permeable channel protein TRP and the phospholipase PLCβ. These dissociation events would serve to mediate signalling in photoreceptors upon light exposure.

Several bacterial proteins such as EpsC, a component of the type 2 secretion system of *V. cholerae,* contain PDZ-like domains that display a circularly permuted topology as compared to canonical PDZ folds [10,75]. In addition to this structural particularity, the carboxylate binding loop of the PDZ-like domain of EpsC is replaced by a small helical structure (hereafter called α0) [75] (see Fig. 4A). Both, the isolated PDZ-like core domain (sPDZ) and a longer construct (lPDZ) were crystallized. The lPDZ construct revealed an additional N-terminal α helix, which establishes many hydrophobic interactions with residues of the PDZ-like core structure [75]. The structures of sPDZ and lPDZ vary in the positioning of the α 2 helix (average difference of Cα atoms is about 3), which is further apart from β2 in sPDZ, thereby opening a deep and narrow hydrophobic groove formed between β2, α0, and α2 [75] (the numbering of secondary structure elements corresponds to the canonical topology of PDZs). The two structures also display significant differences in the positioning of α0. These two restructuring events cause a change in conformation of two Phe residues (one from α0, one from α2) that close the hydrophobic groove in lPDZ (Fig. 4A). It will be very interesting to investigate the binding properties of this very special PDZ-like domain that might have evolved to bind different molecules as has been suggested by the authors [75]. These three latter examples illustrate very different mechanisms by which peptide binding to PDZ or PDZ-like domains can be directly modulated via extensions that influence the conformation of the peptide binding pocket. Other mechanisms of peptide binding pocket modulation via extensions may exist for other PDZ domains, which would be interesting to find and investigate.

4.3.4. PDZ extensions that influence the folding of the PDZ domain

The assembly of PDZ4 and PDZ5 of INADL into a tight module, which was described in the previous section, also represents an example where the N- and C-terminal extensions were essential for obtaining soluble and folded constructs [7]. Very similar observations were obtained for PDZ4 and PDZ5 of GRIP1 (Glutamate receptor-interacting protein 1). Removal of the N-terminal extension of PDZ4 of GRIP1 led to spontaneous unfolding of the PDZ45 tandem [90] and prevented peptide binding [91] (more in detail reviewed elsewhere [5,33]). Together with the N-terminal extension of PDZ1 of Harmonin [66], these are also three examples where extensions of PDZ domains contribute to the construction of multidomain arrangements that ultimately alter the peptide binding behaviours of the PDZ domains being involved [33].

4.3.5. Non-exhaustive collection of structured and unstructured extensions that would deserve further investigation

- PDZ1 of INADL (PDB ID: 2DB5, NMR, [92]) has an additional N-terminal α helix that folds into a hole formed between the β2–β3 sheet and the β1–β5 sheet, and that is parallel to the β4 strand.
- PDZ1 of MPDZ (PDB ID: 2O2T, Xray, [93]) has an additional N-terminal and C-terminal α helix that both, contact the PDZ. The N-terminal helix is very similar to the one mentioned for PDZ1 of INADL.
- PDZ3 of Harmonin (PDB ID: 1V6B, NMR, [94]) has an additional N- terminal α helix that does not fold back on the PDZ core but establishes hydrophobic contacts with V at the beginning of the β1 strand and F of the partially structured N-terminal extension that precedes the N-terminal < helix.
- The PDZ of MPP5 (PDB ID: 1VA8, NMR, [95]) has a small β hairpin that is formed at the N-terminus of the PDZ.
- PDZ7 of INADL (PDB ID: 2DAZ, NMR, [96]) has an N-terminal and C-terminal < helix that both, fold back on the PDZ core. The N-terminal helix is in slightly different orientation in comparison to PDZ1 of INAD and MPDZ. Here, the helix rather aligns to the β1- β5 sheet.
- PDZ2 of Harmonin (see section PBM extensions, PDB ID: 2KBS, NMR, [67]).
- PDZ2 (PDZ1) of MAGI2 (PDB ID: 1UEQ, NMR, [97]) has structured N-and C-terminal extensions that fold back onto the PDZ core and that are similar to those observed for PDZ2 of MAGI1 (see sections PBM extensions and PDZ extensions).
- The PDZ of PLCO (PDB ID: 1UJD, NMR, [98]) has a structured C-terminal extension that folds back onto the PDZ core.

4.3.6. General remarks

We have seen that PDZ extensions can influence the binding *affinity* of peptides to PDZ domains via a wide range of possibilities. Yet, do extensions also affect the binding *specificity* of peptides to PDZs? PDZ extensions that either directly interact with peptide residues (see section dealing with PBM extensions), as well as extensions that directly or indirectly alter the conformation of the peptide binding pocket, are very likely to increase the binding specificities of the corresponding PDZs. By contrast, PDZ extensions affecting the general fold, stability, dynamics of the PDZ, or participating in its general regulation, may more often impact indifferently the general binding behaviour of the PDZ to any of its targets, thereby not increasing binding specificity. Overall, these uncertainties strongly call for studies that will further focus on the impact of PDZ extensions on peptide binding specificities.

4.4. Influence of neighbouring domains on peptide binding to a PDZ

4.4.1. PDZ tandems

It is a well known property of PDZ domains that they often occur in multiple copies within one protein sequence. More and more evidence accumulates that neighbouring PDZ domains influence each other's structure and binding behaviour, especially those that are connected by very short linkers (reviewed in [5]). As structural data of such tandem PDZ domains indicate, they can be tightly packed and form one unit. The term supramodule was introduced to account for this property [5]. Two examples of such supramodules (PDZ45 of INADL and of GRIP1) were already mentioned in the previous section.

The PDZ12 tandem of DLG4 (PSD95) has been the subject of several publications. Based on the observation that PDZ1 and PDZ2 of DLG4 can bind to the same set of C-terminal peptides, it has been suggested that the two PDZs can bind simultaneously to the C-termini of homo- or heteromeric channel proteins of the postsynaptic density [99]. It has been shown that such synergistic binding leads to an increase in affinity and specificity as those ligands occurring in dimeric form are favoured as compared to monomeric ligands [99] (see review of Feng et al. [5] for more details). This property was successfully used to develop biomimetic divalent ligands that were much more efficient than monovalent ligands in disrupting the binding of DLG4 to its interaction partners [100,101]. Such a binding model would be favoured by an arrangement of PDZ1 and PDZ2 where the peptide binding pockets point to the same direction. Five very recent studies concentrated on the domain orientation of PDZ12 of DLG4 and revealed interesting findings. Using various techniques (NMR, crystallography, molecular dynamics simulations, single molecule FRET), very different conformations of the PDZ12 tandem were suggested ranging from a parallel alignment of the binding pockets to an antiparallel arrangement [99,100,102]. In addition, Wang et al. [103] provided data that

Please cite this article in press as: Luck, K., et al. The emerging contribution of sequence context to the specificity of protein interactions mediated by PDZ domains. FEBS Lett. (2012), http://dx.doi.org/10.1016/j.febslet.2012.03.056

suggests that the tandem PDZ being restricted in its interdomain orientations in the ligand-free state, encounters a dramatic increase in flexibility when bound to a C-terminal peptide. Equivalent observations have been obtained for PDZ12 of DLG4 when bound to a divalent ligand [101]. Remarkably, McCann et al. [102] could further demonstrate that domain orientation and flexibility of the single PDZ tandem was comparable to those obtained in full length DLG4. It seems that, depending on the cellular context, parallel alignment of the binding pockets of PDZ12, which would favour multivalent ligand binding, is as likely as antiparallel alignment, which would instead enable recruitment of cytosolic proteins to membrane receptors [102]. More studies are needed to investigate this model. The fact that very different conclusions can be drawn from studies using different biophysical methods may indicate that methodologies remain of limited accuracy for the study of large and dynamic systems such as tandem PDZs.

4.4.2. Autoinhibition of PDZ domains

NHERF1 and X11 (amyloid beta A4 precursor protein-binding family A member 1, reviewed in [5]) have PDZ domains whose peptide binding is regulated via an autoinhibitory mechanism that involves PBMs at their own C-termini [104,105]. In an interesting NMR study, Bhattacharya et al. [76] investigated the molecular details of this autoinhibition for PDZ2 of NHERF1. NHERF1 has two PDZ domains and a C-terminal EB (Ezrin binding) domain that binds to Ezrin (involved in linking cytoskeletal structures to the plasma membrane) and that overlaps with a PBM at the C-terminus of NHERF1. Binding of Ezrin to the EB domain increases affinity of PDZ2 to the C-terminal peptide of CFTR (Cystic fibrosis transmembrane conductance regulator) by 24-fold [106,76]. NMR data indicates that the PBM at the C-terminus of NHERF1 can very weakly and transiently bind into the peptide binding pocket of PDZ2. When Ezrin binds to the EB domain, the whole EB domain including its PBM adopts a helical conformation and dissociates from PDZ2, making the peptide binding pocket fully accessible for binding to PBMs of other proteins [76].

4.4.3. Influence on PDZ–peptide binding from neighbouring domains of different type

In 2011, two articles were published within less than a month, reporting the crystal structure of the PDZ3-SH3-GK (PSG) module of ZO1 com-plexed to C-terminal peptides derived from either the cell adhesion proteins JAM-A [107] or the gap junction protein Connexin45 (Cx45) [68]. Both studies agree on the overall extended shape of the PSG module (PDZ3 does not contact the GK domain). An additional C-terminal α helix (hereafter called α 3) of PDZ3 is located in the linker region between the PDZ3 and SH3 domain. Residues of the α 2 and α 3 helix as well as of the β2–β3 loop of PDZ3 interact with residues from the PDZ-SH3 linker and the SH3 domain. In both crystals, the peptides only partially insert into the binding pocket of PDZ3. The backbone of the peptides exhibit a shift towards the β2 strand, probably forced to this different conformation by the β2–β3 loop of the SH3 domain that is located in front of the peptide binding pocket (Fig. 4B) [107,68]. In both structures the PDZ3-SH3 linker inserts between the β2–β3 loop and the SH3 domain, making the loop inaccessible for interaction with peptide residues. Interestingly, no binding could be observed between JAM-A and the single PDZ3 domain [107] and Cx45 exhibited a 9-fold reduced binding affinity to the single PDZ3 in comparison to the affinity obtained for the PSG module [68]. Similar observations for CRIPT and Neuroligin peptides were independently obtained for the equivalent PSG module of DLG (Disks large) from *D. melanogaster* [108] (Table 2). This difference in affinity might mostly be due to a hydrophobic residue located at p-2 of the peptides (this residue is a Phe in JAM-A and a Val in Cx45) that inserts into a hydrophobic pocket formed by residues from both the α 2 helix of PDZ3 and the β2–β3 loop of the SH3 domain of ZO1 [107,68].

The side chains of PDZ3 exhibit almost identical conformations in the two complexes except of R436 from the β2 strand. The aliphatic part of the side chain of R436 establishes hydrophobic contacts with Leu and Trp at p-1 in JAM-A and Cx45, respectively. However, in the PSG-Cx45 complex, the side chain of W_{-1} of Cx45, being bulkier than L_{-1} of JAM-A, displaces the side chain of R436 of PSG, which consequently occupies space that is used by S_{-3} in the PSG-JAM-A complex (Fig. 4B). This in turn displaces S_{-3} and more upstream residues of the Cx45 peptide further away from the PDZ domain as compared to the equivalent residues from the JAM-A peptide. This may be the main reason for the differences in backbone conformation observed for the peptides in the two crystals (Fig. 4B). Based on these observations, one would expect that the JAM-A peptide binds stronger to the PSG module than the Cx45 peptide. Surprisingly, the contrary is the case (Table 2). In both peptides, residues at p−3 and p−4 are very similar being either serine or threonine, while residues further upstream were not observed to significantly contribute to the binding. Hence, the difference in affinity can only be explained from sequence differences at the last three peptide positions being either VWI for Cx45 or FLV for JAM-A. It seems likely that residues W_{-1} (based on previous observations [109]) and Ile at p0 (preference of Leu or Ile over Val at p0 [19]) are the main contributors to the higher affinity of Cx45 to PDZ3. Based on these observations, we speculate that despite of binding with less affinity, JAM-A seems to bind with higher specificity to the PSG module than Cx45.

In summary, the structures of these two complexes revealed very interesting findings. To our knowledge, they are the first to provide atomic details for the direct influence that non-PDZ domains can have on the peptide binding of their PDZ neighbours. They also show that, whenever possible, investigations of small protein fragments such as single PDZs should be complemented by and compared to investigations of larger protein constructs (comprising multiple domains) or even full length proteins. They also serve as a nice example that shows how peptide residues can influence each other's binding to the PDZ domain. Such "cooperative" effects are excluded from most current PDZ–peptide interaction prediction models, which for complexity reasons assume independence of the peptide residues.

4.5. Influence on PDZ–peptide binding from distal domains

A fourth type of sequence context has not been discussed in this review, namely regions in protein sequences that are not in the neighbourhood of PDZs but still influence their peptide binding behaviour. We are convinced that numerous such examples exist, but our current knowledge on this subject is very limited due to a current lack of biophysical methods that allow studying larger protein fragments or even full length proteins at the molecular level. The works of McCann et al. [102], who used single molecule FRET to study the tandem PDZ12 in full length DLG4 as well as of Pan et al. [68] and Nomme et al. [107] who published the structure of the PSG module of ZO1, are promising steps towards bridging this gap.

5. Concluding remarks

This review has been focussed on the influence of sequence context on PDZ–peptide interactions given the vast amount of data available for the PDZ domain family. Other types of globular domain-linear motif interactions are much less studied and our knowledge on sequence context in these systems is sparse. However, we think that it is very likely that sequence context will have

Please cite this article in press as: Luck, K., et al. The emerging contribution of sequence context to the specificity of protein interactions mediated by PDZ domains. FEBS Lett. (2012), http://dx.doi.org/10.1016/j.febslet.2012.03.056

similar importance on linear motif binding in other systems as has been demonstrated for the PDZ domain family. A few studies underpin this speculation (WW domain family [110], SH3 domain family [111]).

To which extent does sequence context influence specificity of PDZ–PBM interactions? An interaction between two proteins is specific when their mutual binding affinity is significantly higher than the affinities of their interactions with most other proteins. Therefore, in principle the specificity of a given PDZ–PBM interaction can only be assayed by comparing its binding affinity to the binding affinities of this particular PDZ domain towards a variety of other PBMs, and/or to the binding affinities of this particular PBM towards a variety of other PDZ domains. Most of the studies discussed here or elsewhere [5,33] show that sequence context influences binding affinity, yet they did not perform the comparative studies to address interaction specificity. There is a need for more studies including different interaction partners and protein sequences of various lengths [26] to better understand how sequence context influences specificity of domain–SLiM interactions in cell signalling processes.

Acknowledgements

We thank Bruno Kieffer and Yves Nomine for fruitful discussions on PDZ domain extensions. We especially thank Robert Weatheritt and Toby Gibson for very constructive comments on the manuscript. This work was supported by CNRS, University of Strasbourg and Association de Recherche contre le Cancer (ARC, grant 3171). KL was supported by the "Région Alsace", ARC, and the Collége Doctoral Européen de Strasbourg. SC was supported by grants from the Agence Nationale de la Recherche (ANR-MIME-2007, project EPI-HPV-3D) and the National Institute of Health (NIH, grant R01CA134737).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.febslet.2012.03.056.

References

[1] Diella, F., Haslam, N., Chica, C., Budd, A., Michael, S., Brown, N.P., Trave, G. and Gibson, T.J. (2008) Understanding eukaryotic linear motifs and their role in cell signaling and regulation. Front Biosci 13, 6580–6603.
[2] Davey, N.E., Roey, K.V., Weatheritt, R.J., Toedt, G., Uyar, B., Altenberg, B., Budd, A., Diella, F., Dinkel, H. and Gibson, T.J. (2012) Attributes of short linear motifs. Mol Biosyst 8 (1), 268–281.
[3] Ponting, C.P. (1997) Evidence for PDZ domains in bacteria, yeast, and plants. Protein Sci 6 (2), 464–468.
[4] Bilder, D. (2001) PDZ proteins and polarity: functions from the fly. Trends Genet 17 (9), 511–519.
[5] Feng, W. and Zhang, M. (2009) Organization and dynamics of PDZ-domain-related supramod-ules in the postsynaptic density. Nat Rev Neurosci 10 (2), 87–99.
[6] Reiners, J., Nagel-Wolfrum, K., Jürgens, K., Märker, T. and Wolfrum, U. (2006) Molecular basis of human Usher syndrome: deciphering the meshes of the Usher protein network provides insights into the pathomechanisms of the Usher disease. Exp Eye Res 83 (1), 97–119.
[7] Liu, W., Wen, W., Wei, Z., Yu, J., Ye, F., Liu, C.-H., Hardie, R.C. and Zhang, M. (2011) The INAD scaffold is a dynamic, redox-regulated modulator of signaling in the Drosophila eye. Cell 145 (7), 1088–1101.
[8] Iden, S. and Collard, J.G. (2008) Crosstalk between small GTPases and polarity proteins in cell polarization. Nat Rev Mol Cell Biol 9 (11), 846–859.
[9] Roh, M.H. and Margolis, B. (2003) Composition and function of PDZ protein complexes during cell polarization. Am J Physiol Renal Physiol 285 (3), F377–F387.
[10] Harris, B.Z. and Lim, W.A. (2001) Mechanism and role of PDZ domains in signaling complex assembly. J Cell Sci 114 (Pt 18), 3219–3231.
[11] C. Nourry, S.G.N. Grant, J.-P. Borg, PDZ domain proteins: plug and play!, Sci STKE 2003 (179) (2003) RE7. doi:10.1126/stke.2003.179.re7.
[12] Cabral, J.H.M., Petosa, C., Sutcliffe, M.J., Raza, S., Byron, O., Poy, F., Marfatia, S.M., Chishti, A.H. and Liddington, R.C. (1996) Crystal structure of a PDZ domain. Nature 382 (6592), 649–652.
[13] Hillier, B.J., Christopherson, K.S., Prehoda, K.E., Bredt, D.S. and Lim, W.A. (1999) Unexpected modes of PDZ domain scaffolding revealed by structure of nNOS-syntrophin complex. Science 284 (5415), 812–815.
[14] Lenfant, N., Polanowska, J., Bamps, S., Omi, S., Borg, J.-P. and Reboul, J. (2010) A genomewide study of PDZ-domain interactions in C. elegans reveals a high frequency of non-canonical binding. BMC Genomics 11, 671.
[15] Gallardo, R., Ivarsson, Y., Schymkowitz, J., Rousseau, F. and Zimmermann, P. (2010) Structural diversity of PDZ-lipid interactions. Chembiochem 11 (4), 456–467.
[16] Songyang, Z., Fanning, A.S., Fu, C., Xu, J., Marfatia, S.M., Chishti, A.H., Crompton, A., Chan, A.C., Anderson, J.M. and Cantley, L.C. (1997) Recognition of unique carboxyl-terminal motifs by distinct PDZ domains. Science 275 (5296), 73–77.
[17] Stricker, N.L., Christopherson, K.S., Yi, B.A., Schatz, P.J., Raab, R.W., Dawes, G., Bassett, D.E., Bredt, D.S. and Li, M. (1997) PDZ domain of neuronal nitric oxide synthase recognizes novel C-terminal peptide sequences. Nat Biotechnol 15 (4), 336342.
[18] Schleinkofer, K., Wiedemann, U., Otte, L., Wang, T., Krause, G., Oschkinat, H. and Wade, R.C. (2004) Comparative structural and energetic analysis of WW domain-peptide interactions. J Mol Biol 344 (3), 865–881.
[19] Tonikian, R., Zhang, Y., Sazinsky, S.L., Currell, B., Yeh, J.-H., Reva, B., Held, H.A., Appleton, A., Evangelista, M., Wu, Y., Xin, X., Chan, A.C., Seshagiri, S., Lasky, L.A., Sander, C., Boone, C., Bader, G.D. and Sidhu, S.S. (2008) A specificity map for the PDZ domain family. PLoS Biol 6 (9), e239.
[20] Linding, R., Jensen, L.J., Pasculescu, A., Olhovsky, M., Colwill, K., Bork, P., Yaffe, M.B. and Pawson, T. (2008) NetworKIN: a resource for exploring cellular phospho-rylation networks. Nucleic Acids Res 36 (Database issue), D695–D699.
[21] Tonikian, R., Xin, X., Toret, C.P., Gfeller, D., Landgraf, C., Panni, S., Paoluzi, S., Castagnoli, L., Currell, B., Seshagiri, S., Yu, H., Winsor, B., Vidal, M., Ger- stein, M.B., Bader, G.D., Volkmer, R., Cesareni, G., Drubin, D.G., Kim, P.M. and Sidhu, S.S. (2009) Boone, Bayesian modeling of the yeast SH3 domain interactome predicts spa-tiotemporal dynamics of endocytosis proteins. PLoS Biol 7 (10), e1000218.
[22] Panni, S., Montecchi-Palazzi, L., Kiemer, L., Cabibbo, A., Paoluzi, S., Santonico, E., Landgraf, C., Volkmer-Engert, R., Bachi, A., Castagnoli, L. and Cesareni, G. (2011) Combining peptide recognition specificity and context information for the prediction of the 14-3-3-mediated interactome in S. cerevisiae and H. sapiens. Proteomics 11 (1), 128–143.
[23] Stein, A. and Aloy, P. (2008) Contextual specificity in peptide-mediated protein interactions. PLoS One 3 (7), e2524.
[24] Chica, C., Diella, F. and Gibson, T.J. (2009) Evidence for the concerted evolution between short linear protein motifs and their flanking regions. PLoS One 4 (7), e6052.
[25] Gibson, T.J. (2009) Cell regulation: determined to signal discrete cooperation. Trends Biochem Sci 34 (10), 471–482.
[26] Luck, K., Fournane, S., Kieffer, B., Masson, M., Nominé, Y. and Travé, G. (2011) Putting into practice domain-linear motif interaction predictions for exploration of protein networks. PLoS One 6 (11), e25376.
[27] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The protein data bank. Nucleic Acids Res 28 (1), 235–242.
[28] Letunic, I., Doerks, T. and Bork, P. (2012) SMART 7: recent updates to the protein domain annotation resource. Nucleic Acids Res 40 (Database issue), D302–D305.
[29] Spaller, M.R. and globally, Act (2006) think locally: systems biology addresses the PDZ domain. ACS Chem Biol 1 (4), 207–210.
[30] Bhattacharyya, R.P., Remnyi, A., Yeh, B.J. and Lim, W.A. (2006) Domains, motifs, and scaffolds: the role of modular interactions in the evolution and wiring of cell signaling circuits. Annu Rev Biochem 75, 655–680.
[31] Javier, R.T. and Rice, A.P. (2011) Emerging theme: cellular PDZ proteins as common targets of pathogenic viruses. J Virol 85 (22), 11544–11556.
[32] Giallourakis, C., Cao, Z., Green, T., Wachtel, H., Xie, X., Lopez-Illasaca, M., Daly, M., Rioux, J. and Xavier, R. (2006) A molecular-properties-based approach to understanding PDZ domain proteins and PDZ ligands. Genome Res 16 (8), 1056–1072.
[33] Wang, C.K., Pan, L., Chen, J. and Zhang, M. (2010) Extensions of PDZ domains as important structural and functional elements. Protein Cell 1 (8), 737–751.
[34] Velthuis, A.J.W.T., Sakalis, P.A., Fowler, D.A. and Bagowski, C.P. (2011) Genome-wide analysis of PDZ domain binding reveals inherent functional overlap within the PDZ interaction network. PLoS One 6 (1), e16047.
[35] Chen, J.R., Chang, B.H., Allen, J.E., Stiffler, M.A. and MacBeath, G. (2008) Predicting PDZ domain-peptide interactions from primary sequences. Nat Biotechnol 26 (9), 1041–1045.
[36] Hui, S. and Bader, G.D. (2010) Proteome scanning to predict PDZ domain interactions using support vector machines. BMC Bioinformatics 11 (1), 507.
[37] Smith, C.A. and Kortemme, T. (2010) Structure-based prediction of the peptide sequence space recognized by natural and synthetic PDZ domains. J Mol Biol 402 (2), 460–474.
[38] Gerek, Z.N. and Ozkan, S.B. (2010) A flexible docking scheme to explore the binding selectivity of PDZ domains. Protein Sci 19 (5), 914–928.

[39] Karthikeyan, S., Leung, T. and Ladias, J.A. (2001) Structural basis of the Na+/H+ exchanger regulatory factor PDZ1 interaction with the carboxyl-terminal region of the cystic fibrosis transmembrane conductance regulator. J Biol Chem 276 (23), 19683–19686.
[40] Sheng, M. and Sala, C. (2001) PDZ domains and the organization of supramolecular complexes. Annu Rev Neurosci 24, 1–29.
[41] Lim, I.A., Hall, D.D. and Hell, J.W. (2002) Selectivity and promiscuity of the first and second PDZ domains of PSD-95 and synapse-associated protein 102. J Biol Chem 277 (24), 21697–21711.
[42] Préhaud, C., Wolff, N., Terrien, E., Lafage, M., Mégret, F., Babault, N., Cordier, F., Tan, G.S., Maitrepierre, E., Ménager, P., Chopy, D., Hoos, S., England, P., Delepierre, M., Schnell, M.J., Buc, H. and Lafon, M. (2010) Attenuation of rabies virulence: takeover by the cytoplasmic domain of its envelope protein. Sci Signal 3 (105), ra5.
[43] Babault, N., Cordier, F., Lafage, M., Cockburn, J., Haouz, A., Préhaud, C., Rey, F.A., Delepierre, M., Buc, H., Lafon, M. and Wolff, N. (2011) Peptides targeting the PDZ domain of PTPN4 are efficient inducers of glioblastoma cell death. Structure 19 (10), 1518–1524.
[44] Zhang, J., Yan, X., Shi, C., Yang, X., Guo, Y., Tian, C., Long, J. and Shen, Y. (2008) Structural basis of beta-catenin recognition by Tax-interacting protein-1. J Mol Biol 384 (1), 255–263.
[45] Banerjee, M., Huang, C., Marquez, J. and Mohanty, S. (2008) Probing the structure and function of human glutaminase-interacting protein: a possible target for drug design. Biochemistry 47 (35), 9208–9219.
[46] Yan, X., Zhou, H., Zhang, J., Shi, C., Xie, X., Wu, Y., Tian, C., Shen, Y. and Long, J. (2009) Molecular mechanism of inward rectifier potassium channel 2.3 regulation by tax-interacting protein-1. J Mol Biol 392 (4), 967–976.
[47] Im, Y.J., Kang, G.B., Lee, J.H., Park, K.R., Song, H.E., Kim, E., Song, W.K., Park, D. and Eom, S.H. (2010) Structural basis for asymmetric association of the betaPIX coiled coil and shank PDZ. J Mol Biol 397 (2), 457–466.
[48] Balana, B., Maslennikov, I., Kwiatkowski, W., Stern, K.M., Bahima, L., Choe, S. and Slesinger, P.A. (2011) Mechanism underlying selective regulation of G protein-gated inwardly rectifying potassium channels by the psychostimulant-sensitive sorting nexin 27. Proc Natl Acad Sci U S A 108 (14), 5831–5836.
[49] Mostarda, S., Gfeller, D. and Rao, F. (2012) Beyond the Binding Site: The Role of the ?2 - ?3 Loop and Extra-Domain Structures in PDZ Domains. PLoS Comput Biol 8 (3), e1002429.
[50] Zhang, Y., Dasgupta, J., Ma, R.Z., Banks, L., Thomas, M. and Chen, X.S. (2007) Structures of a human papillomavirus (HPV) E6 polypeptide bound to MAGUK proteins: mechanisms of targeting tumor suppressors by a high-risk HPV oncoprotein. J Virol 81 (7), 3618–3626.
[51] Fournane, S., Charbonnier, S., Chapelle, A., Kieffer, B., Orfanoudakis, G., Trav, G., Masson, M. and Nomin, Y. (2011) Surface plasmon resonance analysis of the binding of high-risk mucosal HPV E6 oncoproteins to the PDZ1 domain of the tight junction protein MAGI-1. J Mol Recognit 24 (4), 511–523.
[52] Charbonnier, S., Nominé, Y., Ramirez, J., Luck, K., Chapelle, A., Stote, R.H., Travé, G., Kieffer, B. and Atkinson, R.A. (2011) The structural and dynamic response of MAGI-1 PDZ1 with noncanonical domain boundaries to the binding of human pa-pillomavirus E6. J Mol Biol 406 (5), 745–763.
[53] Zhang, Z., Li, H., Chen, L., Lu, X., Zhang, J., Xu, P., Lin, K. and Wu, G. (2011) Molecular basis for the recognition of adenomatous polyposis coli by the Discs Large 1 protein. PLoS One 6 (8), e23507.
[54] Liu, Y., Henry, G.D., Hegde, R.S. and Baleja, J.D. (2007) Solution structure of the hDlg/SAP97 PDZ2 domain and its mechanism of interaction with HPV- 18 papillomavirus E6 protein. Biochemistry 46 (38), 10864–10874.
[55] Wang, L., Piserchio, A. and Mierke, D.F. (2005) Structural characterization of the inter-molecular interactions of synapse-associated protein-97 with the NR2B subunit of N-methyl-D-aspartate receptors. J Biol Chem 280 (29), 26992–26996.
[56] Madsen, K.L., Beuming, T., Niv, M.Y., Chang, C.-W., Dev, K.K., We- instein, H. and Gether, U. (2005) Molecular determinants for the complex binding specificity of the PDZ domain in PICK1. J Biol Chem 280 (21), 20539–20548.
[57] Tyler, R.C., Peterson, F.C. and Volkman, B.F. (2010) Distal interactions within the par3-VE- cadherin complex. Biochemistry 49 (5), 951–957.
[58] Feng, W., Wu, H., Chan, L.-N. and Zhang, M. (2008) Par-3-mediated junctional localization of the lipid phosphatase PTEN is required for cell polarity establishment. J Biol Chem 283 (34), 23440–23449.
[59] Skelton, N.J., Koehler, M.F.T., Zobel, K., Wong, W.L., Yeh, S., Pisabarro, M.T., Yin, J.P., Lasky, L.A. and Sidhu, S.S. (2003) Origins of PDZ domain ligand specificity. Structure determination and mutagenesis of the Erbin PDZ domain. J Biol Chem 278 (9), 7645–7654.
[60] Appleton, B.A., Zhang, Y., Wu, P., Yin, J.P., Hunziker, W., Skelton, N.J., Sidhu, S.S. and Wiesmann, C. (2006) Comparative structural analysis of the Erbin PDZ domain and the first PDZ domain of ZO-1. Insights into determinants of PDZ domain specificity. J Biol Chem 281 (31), 22312–22320.
[61] Birrane, G., Chung, J. and Ladias, J.A.A. (2003) Novel mode of ligand recognition by the Erbin PDZ domain. J Biol Chem 278 (3), 1399–1402.
[62] Kozlov, G., Gehring, K. and Ekiel, I. (2000) Solution structure of the PDZ2 domain from human phosphatase hPTP1E and its interactions with C-terminal peptides from the Fas receptor. Biochemistry 39 (10), 2572–2580.
[63] Kachel, N., Erdmann, K.S., Kremer, W., Wolff, P., Gronwald, W., Heumann, R. and Kalbitzer, H.R. (2003) Structure determination and ligand interactions of the PDZ2b domain of PTP-Bas (hPTP1E): splicing-induced modulation of ligand specificity. J Mol Biol 334 (1), 143–155.
[64] Walma, T., Aelen, J., Nabuurs, S.B., Oostendorp, M., van den Berk, L., Hendriks, W. and Vuister, G.W. (2004) A closed binding pocket and global destabilization modify the binding properties of an alternatively spliced form of the second PDZ domain of PTP-BL. Structure 12 (1), 11–20.
[65] Sierralta, J. and Mendoza, C. (2004) PDZ-containing proteins: alternative splicing as a source of functional diversity. Brain Res Brain Res Rev 47 (1-3), 105–115.
[66] Yan, J., Pan, L., Chen, X., Wu, L. and Zhang, M. (2010) The structure of the harmonin/sans complex reveals an unexpected interaction mode of the two Usher syndrome proteins. Proc Natl Acad Sci U S A 107 (9), 4040–4045.
[67] Pan, L., Yan, J., Wu, L. and Zhang, M. (2009) Assembling stable hair cell tip link complex via multidentate interactions between harmonin and cadherin 23. Proc Natl Acad Sci U S A 106 (14), 5575–5580.
[68] Pan, L., Chen, J., Yu, J., Yu, H. and Zhang, M. (2011) The structure of the PDZ3-SH3-GuK tandem of ZO-1 protein suggests a supramodular organization of the membrane-associated guanylate kinase (MAGUK) family scaffold protein core. J Biol Chem 286 (46), 40069–40074.
[69] Chen, J., Pan, L., Wei, Z., Zhao, Y. and Zhang, M. (2008) Domain-swapped dimerization of ZO-1 PDZ2 generates specific and regulatory connexin43-binding sites. EMBO J 27 (15), 2113–2123.
[70] Wu, J., Yang, Y., Zhang, J., Ji, P., Du, W., Jiang, P., Xie, D., Huang, H., Wu, M., Zhang, G., Wu, J. and Shi, Y. (2007) Domain-swapped dimerization of the second PDZ domain of ZO2 may provide a structural basis for the polymerization of claudins. J Biol Chem 282 (49), 35988–35999.
[71] Niraula, T.N., Yoneyama, M., Koshiba, S., Inoue, M., Kigawa, T. and Yokoyama, S. (2008) Riken Structural Genomics/Proteomics Initiative. Solution structure of PDZ domain of Rho GTPase Activating Protein 21.
[72] Akiva, E., Friedlander, G., Itzhaki, Z. and Margalit, H. (2012) A dynamic view of domain-motif interactions. PLoS Comput Biol 8 (1), e1002341.
[73] Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E.L.L., Eddy, S.R., Bateman, A. and Finn, R.D. (2012) The Pfam protein families database. Nucleic Acids Res 40 (Database issue), D290–D301.
[74] Zhang, Y., Appleton, B.A., Wu, P., Wiesmann, C. and Sidhu, S.S. (2007) Structural and functional analysis of the ligand specificity of the HtrA2/Omi PDZ domain. Protein Sci 16 (8), 1738–1750.
[75] Korotkov, K.V., Krumm, B., Bagdasarian, M. and Hol, W.G.J. (2006) Structural and functional studies of EpsC, a crucial component of the type 2 secretion system from Vibrio cholerae. J Mol Biol 363 (2), 311–321.
[76] Bhattacharya, S., Dai, Z., Li, J., Baxter, S., Callaway, D.J.E., Cowburn, D. and Bu, Z. (2010) A conformational switch in the scaffolding protein NHERF1 controls au- toinhibition and complex formation. J Biol Chem 285 (13), 9981–9994.
[77] Charbonnier, S., Stier, G., Orfanoudakis, G., Kieffer, B., Atkinson, R.A. and Trav, G. (2008) Defining the minimal interacting regions of the tight junction protein MAGI-1 and HPV16 E6 oncoprotein for solution structure studies. Protein Expr Purif 60 (1), 64–73.
[78] Petit, C.M., Zhang, J., Sapienza, P.J., Fuentes, E.J. and Lee, A.L. (2009) Hidden dynamic allostery in a PDZ domain. Proc Natl Acad Sci U S A 106 (43), 18249–18254.
[79] Ballif, B.A., Carey, G.R., Sunyaev, S.R. and Gygi, S.P. (2008) Large-scale identification and evolution indexing of tyrosine phosphorylation sites from murine brain. J Proteome Res 7 (1), 311–318.
[80] Zhang, J., Petit, C.M., King, D.S. and Lee, A.L. (2011) Phosphorylation of a PDZ domain extension modulates binding affinity and interdomain interactions in postsynaptic density-95 (PSD-95) protein, a membrane-associated guanylate kinase (MAGUK). J Biol Chem 286 (48), 41776–41785.
[81] Consortium, U. (2010) The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res 38 (Database issue), D142–D148.
[82] Dinkel, H., Chica, C., Via, A., Gould, C.M., Jensen, L.J., Gibson, T.J. and Diella, F. (2011) Phospho.ELM: a database of phosphorylation sites-update 2011. Nucleic Acids Res 39 (Database issue), D261–D267.
[83] Dephoure, N., Zhou, C., Villn, J., Beausoleil, S.A., Bakalarski, C.E., Elledge, S.J. and Gygi, S.P. (2008) A quantitative atlas of mitotic phosphorylation. Proc Natl Acad Sci U S A 105 (31), 10762–10767.
[84] Dinkel, H., Michael, S., Weatheritt, R.J., Davey, N.E., Roey, K.V., Altenberg, B., Toedt, G., Uyar, B., Seiler, M., Budd, A., Jdicke, L., Dammert, M.A., Schroeter, C., Hammer, M., Schmidt, T., Jehl, P., McGuigan, C., Dymecka, M., Chica, C., Luck, K., Via, A., Chatr-Aryamontri, A., Haslam, N., Grebnev, G., Edwards, R.J., Steinmetz, M.O., Meiselbach, H., Diella, F. and Gibson, T.J. (2012) ELM-the database of eu-karyotic linear motifs. Nucleic Acids Res 40 (Database issue), D242–D251.
[85] Round, J.L., Tomassian, T., Zhang, M., Patel, V., Schoenberger, S.P. and Miceli, M.C. (2005) Dlgh1 coordinates actin polymerization, synaptic T cell receptor and lipid raft aggregation, and effector function in T cells. J Exp Med 201 (3), 419–430.
[86] Garrard, S.M., Capaldo, C.T., Gao, L., Rosen, M.K., Macara, I.G. and Tomchick, D.R. (2003) Structure of Cdc42 in a complex with the GTPase-binding domain of the cell polarity protein, Par6. EMBO J 22 (5), 1125–1133.
[87] Peterson, F.C., Penkert, R.R., Volkman, B.F. and Prehoda, K.E. (2004) Cdc42 regulates the Par-6 PDZ domain through an allosteric CRIB-PDZ transition. Mol Cell 13 (5), 665–676.
[88] Whitney, D.S., Peterson, F.C. and Volkman, B.F. (2011) A conformational switch in the CRIB-PDZ module of Par-6. Structure 19 (11), 1711–1722.
[89] Mishra, P., Socolich, M., Wall, M.A., Graves, J., Wang, Z. and Ranganathan, R. (2007) Dynamic scaffolding in a G protein-coupled signaling system. Cell 131 (1), 80–92.

Please cite this article in press as: Luck, K., et al. The emerging contribution of sequence context to the specificity of protein interactions mediated by PDZ domains. FEBS Lett. (2012), http://dx.doi.org/10.1016/j.febslet.2012.03.056

[90] Feng, W., Shi, Y., Li, M. and Zhang, M. (2003) Tandem PDZ repeats in glutamate receptor-interacting proteins have a novel mode of PDZ domain-mediated target binding. Nat Struct Biol 10 (11), 972–978.
[91] Dong, H., O'Brien, R.J., Fung, E.T., Lanahan, A.A., Worley, P.F. and Huganir, R.L. (1997) GRIP: a synaptic PDZ domain-containing protein that interacts with AMPA receptors. Nature 386 (6622), 279–284.
[92] K. Inoue, C. Kurosaki, F. Hayashi, S. Yokoyama, Riken Structural Ge-nomics/Proteomics Initiative. Solution structure of the first PDZ domain of InaD-like protein (2006). .
[93] E. Papagrigoriou, C. Gileadi, C. Phillips, J. Elkins, C. Johansson, E. Salah, P. Savitsky, G. Berridge, D. Doyle, Structural Genomics Consortium. The crystal structure of the 1st PDZ domain of MPDZ. (2006).
[94] K. Yamada, N. Nameki, K. Saito, S. Koshiba, M. Inoue, T. Kigawa, S. Yokoyama, Riken Structural Genomics/Proteomics Initiative. Solution structure of the third PDZ domain of mouse harmonin (2004). .
[95] X.-R. Qin, F. Hayashi, S. Yokoyama, Riken Structural Genomics/Proteomics Initiative. Solution structure of the PDZ domain of Pals1 protein. (2005). .
[96] K. Inoue, T. Nagashima, K. Izumi, F. Hayashi, S. Yokoyama, Riken Structural Ge-nomics/Proteomics Initiative. Solution structure of the 7th PDZ domain of InaD-like protein. (2006).
[97] C. Zhao, T. Kigawa, N. Tochio, S. Koshiba, M. Inoue, S. Yokoyama, Riken Structural Genomics/Proteomics Initiative. Solution structure of the first PDZ domain of Human Atrophin-1 Interacting Protein 1 (KIAA0705 protein). (2003).
[98] W. Ohashi, H. Hirota, T. Yamazaki, Y. Muto, S. Yokoyama, Riken Structural Genomics/Proteomics Initiative. Solution structure of RSGI RUH-003, a PDZ domain of hypothetical KIAA0559 protein from human cDNA. (2004).
[99] Long, J.-F., Tochio, H., Wang, P., Fan, J.-S., Sala, C., Niethammer, M., Sheng, M. and Zhang, M. (2003) Supramodular structure and synergistic target binding of the N-terminal tandem PDZ domains of PSD-95. J Mol Biol 327 (1), 203–214.
[100] Sainlos, M., Tigaret, C., Poujol, C., Olivier, N.B., Bard, L., Breillat, C., Thi- olon, K., Choquet, D. and Imperiali, B. (2011) Biomimetic divalent ligands for the acute disruption of synaptic AMPAR stabilization. Nat Chem Biol 7 (2), 81–91.
[101] Bach, A., Clausen, B.H., Møller, M., Vestergaard, B., Chi, C.N., Round, A., Sørensen, P.L., Nissen, K.B., Kastrup, J.S., Gajhede, M., Jemth, P., Kris-tensen, A.S., Lundström, P., Lambertsen, K.L. and Strømgaard, K. (2012) A high-affinity, dimeric inhibitor of PSD-95 bivalently interacts with PDZ1-2 and protects against is- chemic brain damage. Proc Natl Acad Sci U S A 109 (9), 3317–3322.
[102] McCann, J.J., Zheng, L., Chiantia, S. and Bowen, M.E. (2011) Domain orientation in the N-Terminal PDZ tandem from PSD-95 is maintained in the full-length protein. Structure 19 (6), 810–820.
[103] Wang, W., Weng, J., Zhang, X., Liu, M. and Zhang, M. (2009) Creating conformational entropy by increasing interdomain mobility in ligand binding regulation: a revisit to N- terminal tandem PDZ domains of PSD-95. J Am Chem Soc 131 (2), 787–796.
[104] Li, J., Poulikakos, P.I., Dai, Z., Testa, J.R., Callaway, D.J.E. and Bu, Z. (2007) Protein kinase C phosphorylation disrupts Na+/H+ exchanger regulatory factor 1 autoinhibition and promotes cystic fibrosis transmembrane conductance regulator macromolecular assembly. J Biol Chem 282 (37), 27086–27099.
[105] Long, J.-F., Feng, W., Wang, R., Chan, L.-N., Ip, F.C.F., Xia, J., Ip, N.Y. and Zhang, M. (2005) Autoinhibition of X11/Mint scaffold proteins revealed by the closed conformation of the PDZ tandem. Nat Struct Mol Biol 12 (8), 722–728.
[106] Li, J., Dai, Z., Jana, D., Callaway, D.J.E. and Bu, Z. (2005) Ezrin controls the macromolecu-lar complexes formed between an adapter protein Na+/H+ exchanger regulatory factor and the cystic fibrosis transmembrane conductance regulator. J Biol Chem 280 (45), 37634–37643.
[107] Nomme, J., Fanning, A.S., Caffrey, M., Lye, M.F., Anderson, J.M. and Lavie, A. (2011) The Src Homology 3 domain is required for junctional adhesion molecule binding to the third PDZ domain of the scaffolding protein ZO-1. J Biol Chem 286 (50), 43352–43360.
[108] Qian, Y. and Prehoda, K.E. (2006) Interdomain interactions in the tumor suppressor discs large regulate binding to the synaptic protein gukholder. J Biol Chem 281 (47), 35757–35763.
[109] Luck, K. and Travé, G. (2011) Phage display can select over-hydrophobic sequences that may impair prediction of natural domain-peptide interactions. Bioinformatics 27 (7), 899–902.
[110] Fidan, Z., Younis, A., Schmieder, P. and Volkmer, R. (2011) Chemical synthesis of the third WW domain of TCERG 1 by native chemical ligation. J Pept Sci 17 (9), 644–649.
[111] Bauer, F., Schweimer, K., Meiselbach, H., Hoffmann, S., Rosch, P. and Sticht, H. (2005) Structural characterization of Lyn-SH3 domain in complex with a herpesviral protein reveals an extended recognition motif that enhances binding affinity. Protein Sci 14 (10), 2487–2498.
[112] Chen, Q., Niu, X., Xu, Y., Wu, J. and Shi, Y. (2007) Solution structure and backbone dynamics of the AF-6 PDZ domain/Bcr peptide complex. Protein Sci 16 (6), 10531062.
[113] W.L. DeLano, The PyMOL molecular graphics system (2002). URL http://www.pymol.org.
[114] Durney, M.A., Birrane, G., Anklin, C., Soni, A. and Ladias, J.A.A. (2009) Solution structure of the human Tax-interacting protein-1. J Biomol NMR 45 (3), 329–334.
[115] Im, Y.J., Lee, J.H., Park, S.H., Park, S.J., Rho, S.-H., Kang, G.B., Kim, E. and Eom, S.H. (2003) Crystal structure of the Shank PDZ-ligand complex reveals a class I PDZ interaction and a novel PDZ-PDZ dimerization. J Biol Chem 278 (48), 48099–48104.
[116] Katoh, K. and Toh, H. (2007) PartTree: an algorithm to build an approximate tree from a large number of unaligned sequences. Bioinformatics 23 (3), 372–374.
[117] Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M. and Barton, G.J. (2009) Jalview version 2-a multiple sequence alignment editor and analysis workbench. Bioinformatics 25 (9), 1189–1191.

Please cite this article in press as: Luck, K., et al. The emerging contribution of sequence context to the specificity of protein interactions mediated by PDZ domains. FEBS Lett. (2012), http://dx.doi.org/10.1016/j.febslet.2012.03.056

Part III.

Conclusion and perspectives

Studying cell signalling pathways implies the detection of their underlying protein interactions as well as the description of their temporal, spacial, and functional interplay. Many protein interactions that are involved in cell signalling are mediated by SLiMs that bind to globular domains. The importance of SLiMs for biological processes has been reflected by a significant increase in attention for SLiM-mediated protein interactions in various research fields including molecular, structural, and computational biology, as well as medical and drug research. Some SLiM–domain interactions have attracted particular attention, including those that involve SH2, SH3, and PDZ domains.

This thesis has focussed on studying the characteristics of protein interactions that are mediated by PDZ domains as well as the tools to computationally and experimentally detect them. Our findings contributed to a better understanding of the mechanisms that define specificity in PDZ interactions. To address this subject we combined experimental and computational methods being aware of the great potential of integrative approaches.

Specificity of PDZ-mediated protein interactions can and should be investigated under various aspects. Most studies, including bio-computational and experimental approaches, concentrate on minimal interacting protein fragments, such as core PDZs and core PBMs, to investigate the specificity of protein interactions. However, in this thesis, we have provided evidence that sequence context has the potential to change the binding affinities and specificities initially observed between minimal interacting protein fragments.

Altered binding affinity and specificity of PDZ–peptide interactions due to sequence context

In this thesis, the importance of sequence context for PDZ–peptide interactions has been demonstrated by various examples. In our structural study, an extended C-terminal peptide sequence derived from HPV16 E6 has been observed to interact with residues of the β2-β3 loop of PDZ2 of MAGI1 (see chapter 7). Similar interactions have also been observed in various other published studies as we have discussed in our review (see chapter 10). Our medium throughput SPR analysis demonstrated that interactions between residues of N-terminally extended peptides and of the β2-β3 loop can change the binding affinity and specificity in comparison to interactions involving core PBMs. On the PDZ side, our structural analysis provided insights into the mechanisms, by which the structured N-terminal and unstructured C-terminal extensions of PDZ2 of MAGI1 influence the stability, monomeric behaviour, and peptide binding properties of PDZ2 (see chapter 7). Our SPR data suggests intermolecular interactions between PDZ3 and PDZ4 of SCRIB that influence their peptide binding behaviour, maybe by forming a supramodule (see chapter 9). However, these initial data do not allow to draw any more detailed conclusions on the modes of interaction between PDZ3 and PDZ4 and their effects for peptide binding and thus, strongly call for further studies. Other published cases of extensions and neighbouring domains

of PDZs that influence their peptide binding behaviour have been reviewed by others [55, 158] and ourselves (see chapter 10).

From these studies several conclusions can be obtained. The β2-β3 loop of PDZs displays the highest sequence variability within all structural elements of the PDZ domain family and thus, might be an important player for modulation of interaction affinities and specificities. Available structures on PDZ-peptide complexes show a multitude of modes, by which this loop can contribute to peptide binding (see chapter 10) making it extremely difficult to be incorporated into any interaction prediction models. On the experimental side, it might be worth while to systematically use peptides of at least ten residues in length for PDZ interaction assessments. This would allow in most of the cases for an interaction between the residues of the extended peptide and the β2-β3 loop to take place.

The impact on interaction affinity and specificity of sequence context might depend on the mechanisms by which sequence context influences PDZ–peptide interactions. PDZ extensions that either directly interact with peptide residues, as well as extensions that directly or indirectly alter the conformation of the peptide binding pocket, have the potential to influence binding specificities of the corresponding PDZs. By contrast, PDZ extensions affecting the general fold, stability, dynamics of the PDZ, or participating in its general regulation, may more often impact indifferently the general binding behaviour of the PDZ to any of its targets, thereby not increasing binding specificity. Not only these speculations but also a general lack of studies on sequence context of PDZs that assess not only influences on binding affinity but also address interaction specificity, call for more comparative studies including different interaction partners and protein sequences of various lengths (such as ours, see chapter 9) to better understand how sequence context influences specificity of domain–SLiM interactions in cell signalling processes.

Our experimental results in conjunction with results published by others [81] point to the possibility that some (many?) PDZ-peptide interactions at the minimal interaction fragment level might display promiscuous rather than specific binding behaviour accompanied by very low binding affinities. These binding affinities are likely to (specifically) increase when extending the protein fragments. Interestingly, the changes in binding affinity that we observed due to fragment extensions were never at a scale where protein fragments that did not bind to each other in their short version started to bind to each other when being extended or *vice versa*. Thus, we conclude that quantitative binding affinity data obtained for minimal interacting fragments is not necessarily valid for extended fragments or the corresponding full length proteins. However, the qualitative results (i.e. binding or not binding) obtained for minimal interacting fragments should in general be transferable to full length proteins (see chapter 9).

Our findings on sequence context also point to the importance of the design of experimental protein constructs on the outcome of quantitative experimental protein interaction studies. Theoretical domain boundaries being defined by the beginning and end of the first and last secondary structure element, respectively, often do not result in soluble and stable domain constructs upon expression. Using available structural information, we have proposed manually curated constructs for the full human PDZ domain family (see chapter 10). This data in addition to our findings about the importance of fragment extensions will hopefully positively influence future design of experiments involving PDZ domains.

Prediction of PDZ–peptide interaction specificities

In this thesis, state-of-the-art prediction tools for PDZ-peptide interaction specificities have been assessed resulting in the identification of significant limitations in their application (see chapters 8 and 9). These findings are in contrast to the enormous amount of computational work that has been published in this field. The poor performance of some (many?) PDZ interaction predictors is likely to originate from over-simplification and miss-interpretation of biological data. Albert Einstein once said that *"everything should be made as simple as possible, but not simpler"*. It may be one of the major challenges in computational biology (and actually most other sciences) to find the right balance between simplification and appropriate model definition.

It might be worthwhile to think about more restrictive peer reviewing processes for interaction predictors. This can imply the need for experimental validation of predictions, which is probably the best way to assess the usefulness of a predictor in question. Asking for experimental validations might have an additional positive effect in encouraging computational biologists to engage in collaborations with biologists and to find putative applications for their predictors thereby directly addressing biological questions. Potential ways to improve PDZ interaction predictors might consist of taking one step back and to focus on particular sub-systems involving only one or a few PDZ domains and to make better use of the immense structural repertoire of PDZ domains that is currently available and ever growing.

Our knowledge on bio-physical properties of SLiMs is steadily increasing. Yet, is this knowledge adequately applied in experimental and computational approaches for SLiM-mediated protein interaction detection? Our findings on sequence bias in PDZ-related phage display data illustrate the risk of artificial selection methods in producing peptides with non-SLiM-like sequence properties (see chapter 8). There is a general need for careful analysis for any sort of bias in experimental data before its application for protein interaction prediction. SLiMs have defined properties (e.g. weak binding affinities, disorder propensity) that should be more in the focus during the design of experimental and computational studies on SLiM-mediated protein interactions. The ELM resource and iELM method (see sections A.1 and A.2) can serve as examples of such approaches where biological information on SLiMs obtained from careful analysis of experimental studies has been successfully incorporated into prediction tools.

Prediction of PDZ-mediated protein interaction networks

PDZ–peptide interaction predictors are developed with the prospect to be applied to protein–protein interaction (PPI) network predictions. In line with our experimental data (see chapter 9), predictions on affinity and specificity based on protein fragments might at best only qualitatively be transferred to full length proteins. Given the identified weaknesses in PDZ-mediated interaction prediction, fully automatic derivation of PPI networks will be highly error-prone. However, predictions of PDZ–peptide interactions combined with manual analysis and experimental validation can result in the identification of new potential interactions that are worth further experimental investigation (see chapter 9). The new potential interactors that we have identified for MAGI1 and SCRIB highlight the implication of PDZ proteins in G protein related signalling pathways where they most probably function as scaffolding proteins. Particular striking has been the finding that the Rho specific GTPase activating protein (GAP) ARHGAP6 might bind to PDZ3 of MAGI1. Together with previous findings that the RhoA specific GEF Net1 binds to PDZ2 of MAGI1, this suggests the possibility that MAGI1 brings closely together RhoA activating and inactivating enzymes, thereby becoming a player for G protein signal termination. However, this speculation has to be validated by testing *in vitro* and *in vivo* the existence of this ternary complex.

High-risk HPV E6 proteins have been shown to inactivate the human proteins MAGI1 and SCRIB thereby perturbing the numerous cellular protein interactions they mediate. Inactivation of MAGI1 and SCRIB has been suggested to be involved in tumour development in general [200, 201] and in the context of HPV infection [202]. Thus, it will be interesting to see if any of the new potential interactors identified for MAGI1 and SCRIB might play role in this process. But prior to such investigation is again the validation of the potential interactions in a full length and *in vivo* context.

Our phage display data analysis did not identify any sequence bias within the peptides that were selected for 17 out of the 54 human PDZ domains that have been used by Tonikian *et al.* [9] (see chapter 8). It may be very promising to use this phage display data for PDZ-mediated PPI network predictions (see our ongoing study described in section C.1).

From working with PPI networks, I obtained the impression that if we want to extract biological relevant information out of the ever growing number of protein interactions that are published, we have to map additional information onto PPI networks. This includes information about which interaction partners of a given protein will function together in protein complexes, which ones are mutually exclusive, under which cellular conditions they are active and finally, how they are regulated. Such knowledge can partially be derived from identifying the binding interfaces of interacting proteins. Thus, the iELM method (see section A.2) provides an important step into this direction. It will also be essential to add temporal (e.g. gene expression) and spacial (e.g. localisation) information (the cellular context) to these networks to be able to order the highly complex network of protein interactions that is currently

established. However, such efforts will strongly depend on the potential of current experimental methods to provide this kind of data.

Outlook

Published findings demonstrate that high affinity artificial peptides often cannot be found in natural protein sequences and that mutations of natural SLiMs can lead to significant increases in their binding affinity towards a domain. These observations suggest that evolution of SLiM-domain interactions may only partially be affinity driven, resulting in SLiMs with sequences that are sub-optimal for binding to a given domain in terms of binding affinity. Maybe SLiMs evolve to provide a balanced mix of advantageous and disadvantageous interface contacts with domain residues leading to weak yet specific interactions. It will be highly interesting to experimentally investigate this hypothesis.

Overall, I have the impression that although many publications address the specificity of PDZ–peptide interactions, only very few of them do so by determining and comparing binding affinities of various peptides to a PDZ domain or vice versa. Very interesting yet unperformed studies would consist of assessing the binding of one or a few peptides towards the human PDZome. As presented in chapter B.2, we have started such a project. An even more fascinating future project would be to assess in an HTP approach the (relative) binding strengths of most of the potential C-terminal PBMs in the human proteome versus the human PDZome. Similar studies have already been successfully performed for other types of SLiM–domain interactions [32]. I think that we need more of such comparative experimental studies in the PDZ field to advance in our understanding of the mechanisms that confer specificity to PDZ-mediated protein interactions and of the various levels of specificity they might have. Such studies may reveal that it is not possible to generalise findings on the level of specificity found for a subset of PDZ-peptide interactions to the whole PDZ domain family. There is probably space for both, instances of high specificity defined by molecular constraints in the PDZ binding pocket and instances of promiscuous binding behaviour. Which level of specificity will be required for a PDZ-peptide interaction is likely to depend on its biological function.

From diving into the PDZ literature during my PhD, I have been impressed by the amount of studies that were published in this field and that are weekly going to be published. For scientists, especially those that are newcomers to the field, it is impossible to keep track of all published PDZ-related work. Thus, many findings, especially those from single case studies, carry the risk to be overlooked, potentially leading to repetitive findings. Future efforts in data integration should focus on tool development that allow to gather all the published work relevant to one specific field into a single resource that can be updated (maybe in a wiki-like style) and queried by any interested scientist. I am convinced that such efforts would be paid off by allowing scientists to perform research more efficiently and of higher quality.

Part IV.

Appendix

A. Published articles under supervision of Toby Gibson

A.1. Prediction of instances of known linear motifs – the ELM resource

Summary

The ELM resource is dedicated to the annotation and prediction of SLiMs in protein sequences. Published information on SLiMs are gathered, manually annotated and classified in the ELM database. A pipeline has been implemented that aims at predicting instances of annotated SLiM classes in protein sequences. Both the database and pipeline are publicly available via a web interface. This article presents updates that have been made to the ELM resource since its last publication two years ago.
Updates: 24 new SLiM classes and more than 500 new SLiM instances have been annotated, leading in total to 170 SLiM classes and 1,800 SLiM instances that are currently stored in the ELM database. Particular attention has been put on the annotation of SLiM instances that have structural support and SLiM instances in viral proteins. Information on interacting domains has been added to each SLiM class. A candidate list has been made visible containing information about potential SLiM classes that await their annotation. The web interface has been changed to improve querying the database and running the prediction pipeline. A disorder prediction filter (see section 6.4.2) has been added to the SLiM instance prediction pipeline to improve identification of potential false positive hits. The graphical output of the prediction results has been enriched with annotations about known phosphorylation sites in the query protein that are annotated in Phospho.ELM.
Discussion/Conclusions: The ELM resource has been proven useful for the functional annotation of proteins and the guidance of experiments dedicated to the identification of protein functions. In addition, the information stored in the ELM database contributed to HTP-screenings and served as benchmark data set for numerous interaction predictors. Users of the prediction pipeline have to be aware of the fact that SLiM instance predictions suffer from a high FPR.
Contribution: I have assisted in the update of the SLiM class describing PDZ-binding motifs. I have contributed to the annotation of the interacting domains to SLiM classes.

D242–D251 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 21 November 2011*
doi:10.1093/nar/gkr1064

ELM—the database of eukaryotic linear motifs

Holger Dinkel[1], Sushama Michael[1], Robert J. Weatheritt[1], Norman E. Davey[1], Kim Van Roey[1], Brigitte Altenberg[1], Grischa Toedt[1], Bora Uyar[1], Markus Seiler[1], Aidan Budd[1], Lisa Jödicke[1], Marcel A. Dammert[1], Christian Schroeter[1], Maria Hammer[1], Tobias Schmidt[1], Peter Jehl[1], Caroline McGuigan[1], Magdalena Dymecka[2], Claudia Chica[3], Katja Luck[4], Allegra Via[5], Andrew Chatr-aryamontri[6], Niall Haslam[7], Gleb Grebnev[7], Richard J. Edwards[8], Michel O. Steinmetz[9], Heike Meiselbach[10], Francesca Diella[1,11] and Toby J. Gibson[1,*]

[1]Structural and Computational Biology, European Molecular Biology Laboratory, Heidelberg, Germany, [2]Laboratory of Bioinformatics and Systems Biology, M. Sklodowska-Curie Cancer Center and Institute of Oncology, WK Roentgena 5, 02-781 Warsaw, Poland, [3]Genoscope (CEA – Institut de Génomique), 2 rue Gaston Cremieux CP5706, 91057 Evry, [4]Group Oncoproteins, Unité CNRS-UDS UMR 7242, Institut de Recherche de l'Ecole de Biotechnologie de Strasbourg, 1, Bd Sébastien Brant, BP 10413, 67412 Illkirch – Cedex, France, [5]Biocomputing Group, Department of Physics, Sapienza University of Rome, P.le Aldo Moro 5, Rome, Italy, [6]School of Biological Sciences, University of Edinburgh, Mayfield Road, Edinburgh EH9 3JR, UK, [7]School of Medicine and Medical Science, University College, Dublin, Ireland, [8]Centre for Biological Sciences, Institute for Life Sciences, University of Southampton, UK, [9]Biomolecular Research, Paul Scherrer Institut, CH-5232 Villigen PSI, Switzerland, [10]Bioinformatik, Institut für Biochemie, Friedrich-Alexander-Universität, Fahrstraße 17, 91054 Erlangen-Nürnberg and [11]Molecular Health GmbH Belfortstr. 2, 69115 Heidelberg, Germany

Received September 13, 2011; Revised and Accepted October 27, 2011

ABSTRACT

Linear motifs are short, evolutionarily plastic components of regulatory proteins and provide low-affinity interaction interfaces. These compact modules play central roles in mediating every aspect of the regulatory functionality of the cell. They are particularly prominent in mediating cell signaling, controlling protein turnover and directing protein localization. Given their importance, our understanding of motifs is surprisingly limited, largely as a result of the difficulty of discovery, both experimentally and computationally. The Eukaryotic Linear Motif (ELM) resource at http://elm.eu.org provides the biological community with a comprehensive database of known experimentally validated motifs, and an exploratory tool to discover putative linear motifs in user-submitted protein sequences. The current update of the ELM database comprises 1800 annotated motif instances representing 170 distinct functional classes, including approximately 500 novel instances and 24 novel classes. Several older motif class entries have been also revisited, improving annotation and adding novel instances. Furthermore, addition of full-text search capabilities, an enhanced interface and simplified batch download has improved the overall accessibility of the ELM data. The motif discovery portion of the ELM resource has added conservation, and structural attributes have been incorporated to aid users to discriminate biologically relevant motifs from stochastically occurring non-functional instances.

INTRODUCTION

Short linear motifs (SLiMs, LMs or MiniMotifs) are regulatory protein modules characterized by their compact interaction interfaces (the affinity and specificity determining residues are usually encoded between 3 and 11 contiguous amino acids (1)) and their enrichment in natively unstructured, or disordered, regions of proteins (2). As a result of limited intermolecular contacts with their interaction partners, SLiMs bind with relatively

*To whom correspondence should be addressed. Tel: +49 (0) 6221 3878398; Fax: +49 (0) 6221 387517; Email: gibson@embl-heidelberg.de

low affinity (in the low-micromolar range), an advantageous attribute for use as transient, conditional and tunable interactions necessary for many regulatory processes. Due to the limited number of mutations necessary for the genesis of a novel motif, SLiMs are amenable to convergent evolution, functioning as a driver of network evolution by adding novel interaction interfaces, and thereby new functionality, to proteins. This evolutionary plasticity facilitates the rapid proliferation within a proteome, and as a result, motif use is ubiquitous in higher eukaryotes.

SLiMs play an important role for many regulatory processes such as signal transduction, protein trafficking and post-translational modification (3,4). Their importance to the correct functionality of the cell is also reflected by the outcome of motif deregulation. For example, point mutations in SLiMs have been shown to lead severe pathologies such as 'Noonan-like syndrome' (5), 'Liddle's syndrome' (6) or 'Retinitis pigmentosa' (7). Furthermore, mimicry of linear motifs by viruses to hijack their hosts' existing cellular machinery plays an important role in many viral life cycles (8). However, despite their obvious importance to eukaryotic cell regulation, our understanding of SLiM biology is relatively limited, and it has been suggested that, to date, we have only discovered a small portion of the human motifs (9).

Several resources are devoted to the annotation and/or detection of SLiMs [Prosite (10), MiniMotifMiner (11) and Scansite (12)]. Here, we report on the 2012 status of the Eukaryotic Linear Motif database.

THE ELM RESOURCE

The ELM initiative (http://elm.eu.org) has focused on gathering, storing and providing information about short linear motifs since 2003. It was established as the first manually annotated collection of SLiM classes and as a tool for discovering linear motif instances in proteins (13). As it was mainly focused on the eukaryotic sequences, it was termed the Eukaryotic Linear Motif resource, usually shortened to ELM. The ELM resource consists of two applications: the ELM database of curated motif classes and instances, and the motif detection pipeline to detect putative SLiM instances in query sequences. In the ELM database, SLiMs are annotated as 'ELM classes', divided into four 'types': cleavage sites (CLV), ligand binding sites (LIG), sites of post-translational modification (MOD) and subcellular targeting sites (TRG) (Table 1). Currently, the ELM database contains 170 linear motif classes with more than 1800 motif instances linked to more than 1500 literature references (Table 1). Each class is described by a regular expression capturing the key specificity and affinity determining amino acid residues. A regular expression is a computer-readable term for sequence annotation and is used by the ELM motif detection pipeline to scan proteins for putative instances of annotated ELM classes. The search form for sequence input is shown in Figure 1, while the results page showing the putative and annotated instances is illustrated in Figure 2.

The ELM resource is powered by a PostgreSQL relational database for data storage and a PYTHON web framework for data retrieval/visualization. The main tables within the database contain information about ELM classes, ELM instances, sequences, references, taxonomy and links to other databases [the database structure is described in greater detail in (14)].

New ELM classes

Since the last release (14), 24 new ELM classes have been added to the ELM database (Table 1) and several more have been updated. One of the newly annotated motif classes is the AGC kinase docking motif (LIG_AGCK_PIF), consisting of three distinct classes. It is present in the non-catalytic C-terminal tail of AGC kinases that constitute a family of serine/threonine kinases consisting of 60 members that regulate critical processes, including cell growth and survival. Deregulation of these enzymes is a causative factor in different diseases such as cancer and diabetes. The motif interacts with the PDK1 Interacting Fragment (PIF) pocket in the kinase domain of AGC kinases. It mediates intramolecular binding to the PIF pocket, serving as a *cis*-activating module together with other regulatory sequences in the C-tail. Interestingly, in some kinases the motif also acts as a PDK1 docking site that *trans*-activates PDK1, which itself lacks the regulatory C-tail, by interacting with the PDK1 PIF pocket. PDK1 in turn will phosphorylate and activate the docked kinase. Other novel classes (Table 2) include phosphodegrons, which are important mediators of phosphorylation-dependent protein destruction, and the LYPxL motif, which is involved in endosomal

Table 1. Summary of data stored in the ELM database[a]

Number of functional site entries		ELM motif classes		ELM motif instances		Links to PDB structures	GO terms		Pubmed links	
Totals	115		170		1840	195		340		1561
By category		LIG	111	Human	1004					
		MOD	30	Mouse	160		Biological process	173	From ELM motif	787
		TRG	21	Rat	102					
		CLV	8	Fly	67		Cell compartment	74	From instance	1071
				Yeast	90					
				Other	417		Molecular function	93		

[a]As of October 2011.

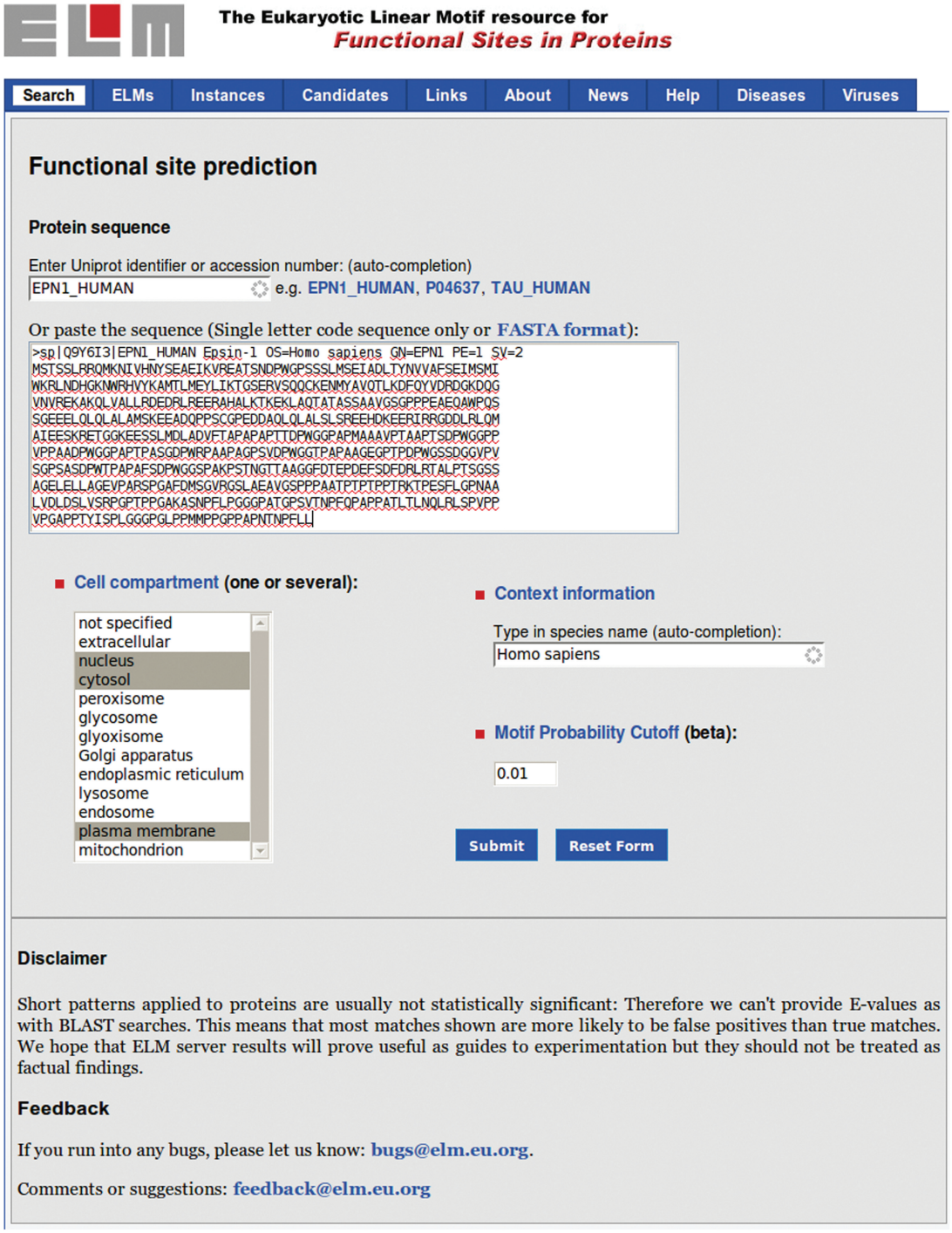


Figure 1. ELM start page. The user can submit a query sequence to the motif detection pipeline either as UniProt accession number or in FASTA format. Filtering criteria such as taxonomic range or cellular compartment should be activated to limit the resulting list of SLiM instances.

sorting of membrane proteins but is also implicated in retrovirus budding.

New ELM instances

Annotated ELM instances serve as representative examples of the respective ELM class. They are also invaluable for the computational analysis and classification of motifs (15). Therefore, special emphasis has been put on the curation of more than 500 novel ELM instances (in 40 different classes) by scanning and annotating more than 400 articles. The number of protein databank (PDB) entries annotated have been increased to 195 (Table 1), meaning that for ~10% of all instances there is a 3D

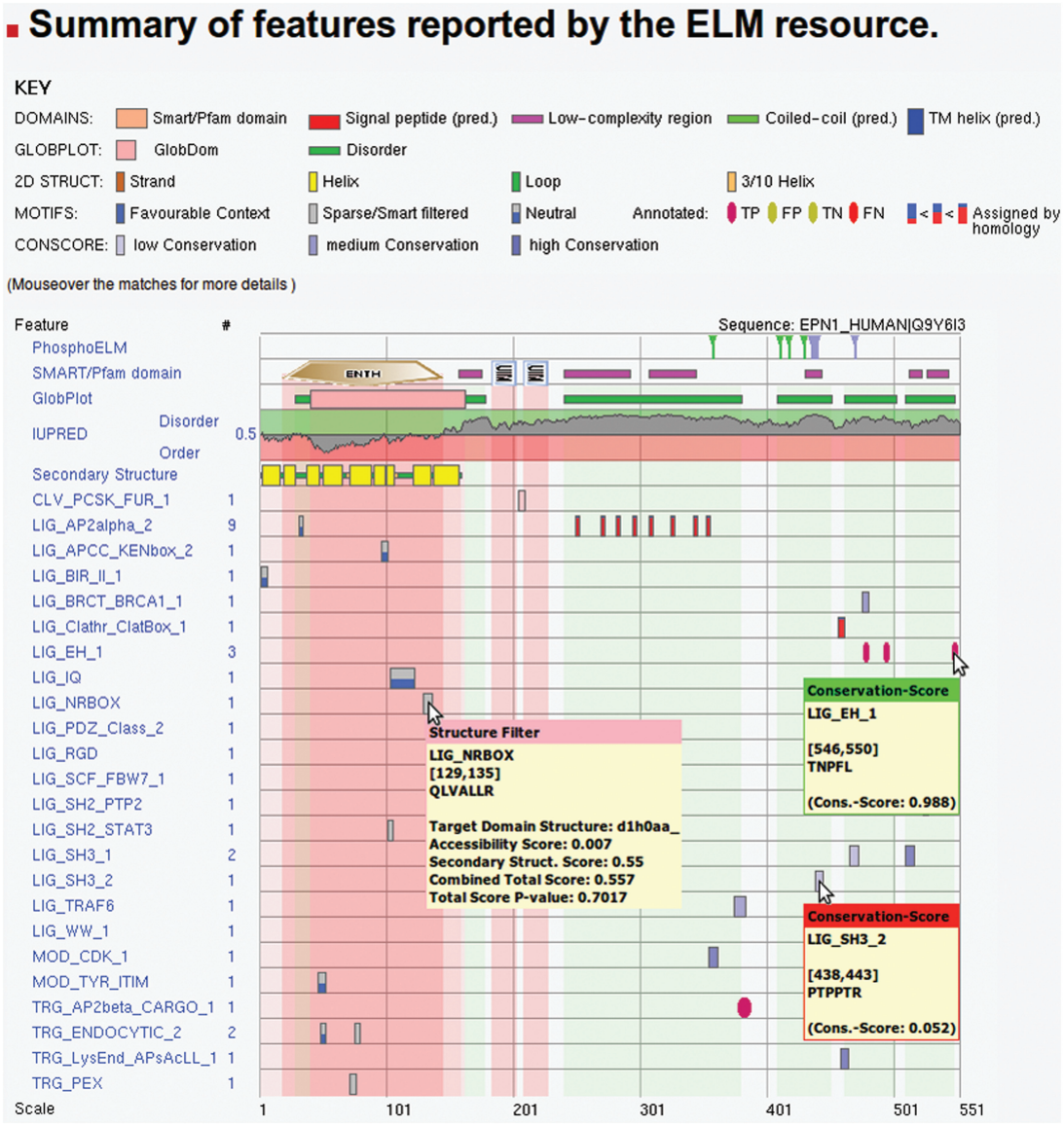


Figure 2. ELM motif detection pipeline output page. The top legend explains the different colors/symbols used. The graphical output of ELM concentrates the output of multiple sequence classification algorithms; phosphorylation sites from Phospho.ELM, protein domains detected by SMART/Pfam, disorder predictions by GlobPlot and IUPred and secondary structure (18). The lower part contains the annotated and putative ELM instances for the given protein sequence (Epsin1, UniProt accession Q9Y6I3). The background is colored according to the structural information available. Each box represents one ELM instance, the color of which indicates the likelihood that this instance is functional: grey instances are buried within structured regions, while shades of blue represent instances outside of structured regions and hint on sequence conservation, with pale blue representing weak sequence conservation and dark blue indicating strong sequence conservation. Red ellipses or boxes mark instances that are annotated in the query sequence or a homologous sequence, respectively.

protein structure annotated, giving more detailed information about the biological context of the respective motif.

NEW FEATURES

The ELM website at http://elm.eu.org can be used in two ways: first, as a front-end to explore the ELM database of curated ELM classes and instances, and second, to run the motif detection pipeline to detect putative SLiM instances in query sequences. Both interfaces have been improved with the most notable changes listed below.

User interface

The database user interface, having been stable for many years, has been overhauled and replaced by a novel interface introducing several new features (Figure 1). Up-to-date web technologies have been used to improve the general user experience: the PYTHON framework DJANGO (http://www.djangoproject.com) dynamically creates and serves all HTML pages, while JavaScript was used to make the whole site more interactive and thus improve the user experience. In particular, the ELM detail pages (Figure 3), which hold the most

Table 2. List of novel ELM classes[a]

Identifier	Description
LIG_Actin_WH2_1 LIG_Actin_WH2_2 LIG_Actin_RPEL_3	Motifs, present in proteins in several repeats, which mediate binding to the hydrophobic cleft created by subdomains 1 and 3 of G-actin
LIG_AGCK_PIF_1 LIG_AGCK_PIF_2 LIG_AGCK_PIF_3	The AGCK docking motif mediates intramolecular interactions to the PDK1 Interacting Fragment (PIF) pocket, serving as a *cis*-activating module
LIG_BIR_II_1 LIG_BIR_III_1 LIG_BIR_III_2 LIG_BIR_III_3 LIG_BIR_III_4	IAP-binding motifs are found in pro-apoptotic proteins and function in the abrogation of caspase inhibition by inhibitor of apoptosis proteins in apoptotic cells
LIG_eIF4E_1 LIG_eIF4E_2	Motif binding to the dorsal surface of eIF4E
LIG_EVH1_3	A proline-rich motif binding to EVH1/WH1 domains of WASP and N-WASP proteins
LIG_HCF-1_HBM_1	The DHxY Host Cell Factor-1 binding motif interacts with the N-terminal kelch propeller domain of the cell cycle regulator HCF-1
LIG_Integrin_isoDGR_1	Present in proteins of extracellular matrix which upon deamidation forms biologically active isoDGR motif which binds to various members of integrin family
LIG_LYPXL_L_2 LIG_LYPXL_S_1	The LYPxL motif binds the V-domain of Alix, a protein involved in endosomal sorting
LIG_PAM2_1	Peptide ligand motif that directly interacts with the MLLE/PABC domain found in poly(A) binding proteins and HYD E3 ubiquitin ligases
LIG_PIKK_1	Motif located in the C terminus of Nbs1 and its homologous interacting with PIKK family members
LIG_Rb_pABgroove_1	The LxxLFD motif binds in a deep groove between pocket A and pocket B of the Retinoblastoma protein
LIG_SCF_FBW7_1 LIG_SCF_FBW7_2	The TPxxS phospho-dependent degron binds the FBW7 F box proteins of the SCF (Skp1-Cullin-Fbox) complex
LIG_SPAK-OSR1_1	SPAK/OSR1 kinase binding motif acts as a docking site which aids the interaction with their binding partners including the upstream activators and the phosphorylated substrates

[a]As of October 2011.

important information about each ELM class including references, regular expression, taxonomic distribution and gene ontology terms (Table 3), have been updated by annotating the protein domain interacting with the respective motif. Where available, a 3D model of representative protein databank structures of linear motif interactions was added to the ELM detail page (Figure 3, top right).

To cope with the increasing amount of annotated classes as well as instances, a novel query interface was introduced to assist the user in finding information of interest. The ELM browser (Figure 4) now features a search interface for free text search. In addition, the search results can also be filtered and reordered using buttons (Figure 4, left side) and table headers, respectively, and be downloaded as tab-separated values (TSV).

Further, improvements to the ELM database include revising the experimental methods used for annotation by using a standardized methods vocabulary [in sync with PSI-MI ontology (16,17)].

A candidate page has been introduced to display novel ELM classes that have not yet been annotated in detail or are currently undergoing annotation. We invite researchers to send us their feedback and expert opinion on these classes and to contribute novel motif classes that will be added to the candidate page and ultimately be turned into full ELM classes (Figure 5). Minimum requirements are at least one literature reference as well as a short description. In addition, a draft regular expression or a 3D structure showing the relevant interaction would also be helpful. Currently, the number of possible ELM classes on this candidate list (awaiting further annotation) exceeds the number of completely annotated classes, indicating the great demand for further annotation.

Graphical representation of sequence search

The ELM motif detection pipeline scans protein sequences for matches to the regular expressions of annotated ELM classes (Figure 2). The query output combines these putative instances with information from the database (annotated ELM instances) as well as predictions from different algorithms/filters. The ELM resource employs a structural filter (18) to highlight and mask secondary structure elements, as well as SMART (19) to detect protein domains. Furthermore, an additional disorder prediction algorithm (IUPred) (20) has been included to predict ordered/disordered regions within the protein. IUPred uses a cutoff of 0.5 to classify a sequence region as either structured or disordered, with values above this threshold corresponding to disorder, highlighted in green background and lower values indicating structured regions, displayed in red background in the output graph. Disorder and domain information is combined by

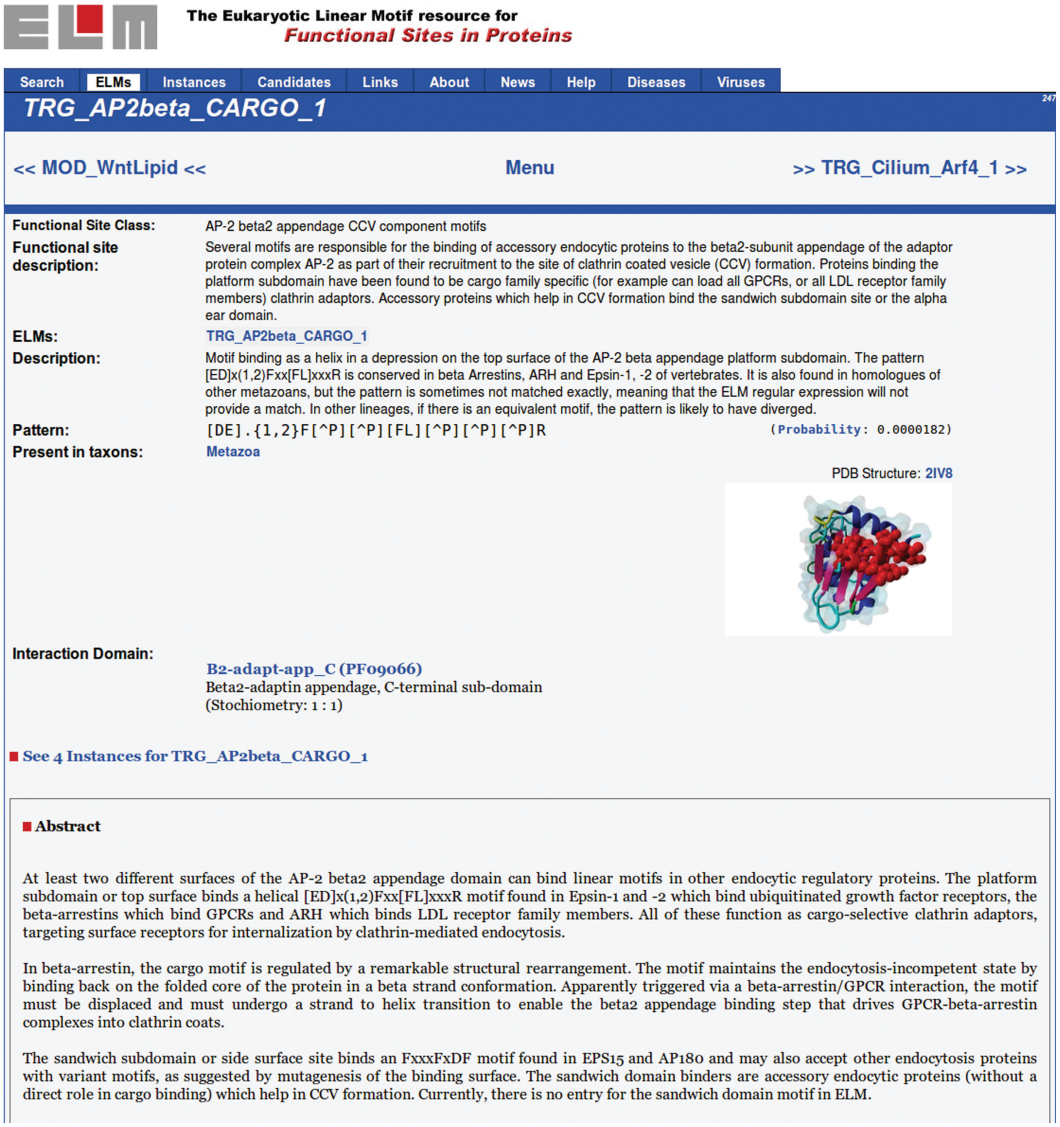


Figure 3. ELM detail page showing information about the ELM class TRG_AP2beta_CARGO_1.

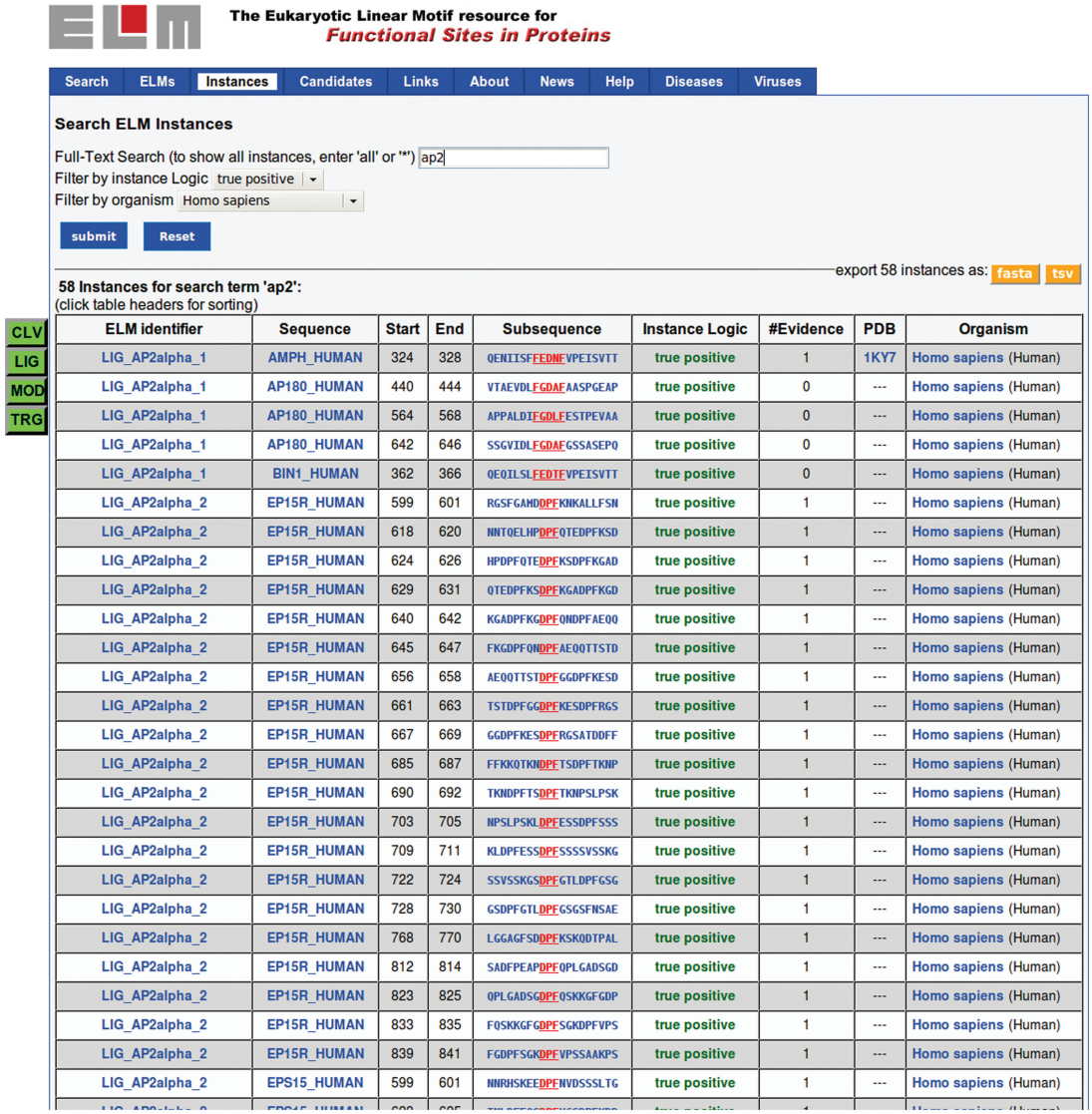
Table 3. Main cellular compartments used in ELM annotation

Count	GO Id	GO term
98	GO:0005829	Cytosol
69	GO:0005634	Nucleus
17	GO:0005576	Extracellular
12	GO:0005794	Golgi apparatus
10	GO:0005886	Plasma membrane
9	GO:0009898	Internal side of plasma membrane
9	GO:0005783	Endoplasmic reticulum
6	GO:0005739	Mitochondrion
5	GO:0005643	Nuclear pore
5	GO:0045334	Clathrin-coated endocytic vesicle

background coloring to highlight structured regions within the protein, which allows inspection of SLiMs that reside at domain boundaries and emphasizes motifs in disordered regions.

The conservation of linear motifs can help in assessing the functional relevance of putative instances, with functional instances showing higher overall sequence conservation than non-functional ones (21). Therefore, sequence conservation of the query protein is calculated using a tree-based conservation scoring method (22) and highlighted in the graphical output. Here, lighter shades of blue represent low conservation while dark blue shading corresponds to high-sequence conservation. The actual conservation score can be inspected by moving the mouse over the respective ELM instance (Figure 2).

The functionality of linear motifs can be modulated by modifications such as phosphorylation (23,24). To enable the user to investigate phosphorylation data in the context of putative linear motif instances, phosphorylation annotations from the Phospho.ELM resource (25) have been added to the graphical output (Figure 2, top row).



ELM identifier	Sequence	Start	End	Subsequence	Instance Logic	#Evidence	PDB	Organism
LIG_AP2alpha_1	AMPH_HUMAN	324	328	QENIISFFEDNFVPEISVTT	true positive	1	1KY7	Homo sapiens (Human)
LIG_AP2alpha_1	AP180_HUMAN	440	444	VTAEVDLFGDAFAASPGEAP	true positive	0	---	Homo sapiens (Human)
LIG_AP2alpha_1	AP180_HUMAN	564	568	APPALDIFGDLFESTPEVAA	true positive	0	---	Homo sapiens (Human)
LIG_AP2alpha_1	AP180_HUMAN	642	646	SSGVIDLFGDAFGSSASEPQ	true positive	0	---	Homo sapiens (Human)
LIG_AP2alpha_1	BIN1_HUMAN	362	366	QEQILSLFEDTFVPEISVTT	true positive	0	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	599	601	RGSFGAMDDPFKNKALLFSN	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	618	620	NNTQELHPDPFQTEDPFKSD	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	624	626	HPDPFQTEDPFKSDPFKGAD	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	629	631	QTEDPFKSDPFKGADPFKGD	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	640	642	KGADPFKGDPFQNDPFAEQQ	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	645	647	FKGDPFQNDPFAEQQTTSTD	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	656	658	AEQQTTSTDPFGGDPFKESD	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	661	663	TSTDPFGGDPFKESDPFRGS	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	667	669	GGDPFKESDPFRGSATDDFF	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	685	687	FFKKQTKNDPFTSDPFTKNP	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	690	692	TKNDPFTSDPFTKNPSLPSK	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	703	705	NPSLPSKLDPFESSDPFSSS	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	709	711	KLDPFESSDPFSSSSVSSKG	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	722	724	SSVSSKGSDPFGTLDPFGSG	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	728	730	GSDPFGTLDPFGSGSFNSAE	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	768	770	LGGAGFSDDPFKSKQDTPAL	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	812	814	SADFPEAPDPFQPLGADSGD	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	823	825	QPLGADSGDPFQSKKGFGDP	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	833	835	FQSKKGFGDPFSGKDPFVPS	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	839	841	FGDPFSGKDPFVPSSAAKPS	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EPS15_HUMAN	599	601	NNRHSKEEDPFNVDSSSLTG	true positive	1	---	Homo sapiens (Human)

Figure 4. ELM instances browse page. A full-text search (here, search term used was 'AP2', filtering for 'true positive' instances in taxon '*Homo sapiens*', yielding 58 instances) assists in finding annotated instances. A search can be restricted to a particular taxonomy or instance logic (top) or ELM class type (buttons on the left). The list can also be exported to TSV or FASTA format for further processing.

The phosphorylated residues are highlighted in different colors (serine: green, threonine: blue, tyrosine: red); each phosphorylation site is linked to a page showing detailed information about the respective modification site from the manually curated data set of the Phospho.ELM resource.

VIRAL INSTANCES

The importance of the short linear motifs in virus–host interactions makes the ELM resource an important tool for the viral research community. For example, Cruz *et al.* (26) analyzed a protein phosphatase 1 (PP1) docking motif in 'protein 7' of transmissible gastroenteritis virus using the ELM class LIG_PP1. This conserved sequence motif mediates binding to the PP1 catalytic subunit, a key regulator of the cellular antiviral defense mechanisms, and is also found in other viral proteomes, suggesting that it might be a recurring strategy to counteract the hosts' defense against RNA viruses by dephosphorylating eukaryotic translation initiation factor 2α and ultimately ribonuclease L.

To reflect our increasing awareness of viral motifs (8), special focus has been attributed to the annotation of viral instances in the ELM database: in the latest release, more than 200 novel ELM instances found in 84 different viral taxons have been added. The notion of viruses abusing existing SLiMs in their hosts is demonstrated by viral instances being annotated alongside instances in their hosts' proteins. For example, the ELM class LIG_PDZ_Class_1 contains 12 instances in human proteins but has recently been expanded with 5 instances from 5 different human pathogenic virus proteins.

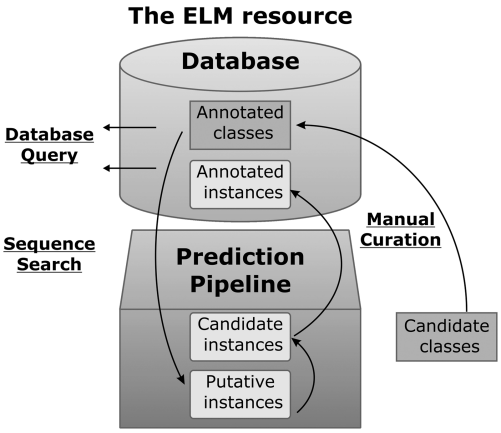


Figure 5. Schema of the ELM resource and data life cycle. Annotated ELM classes, and instances thereof, can be searched by database query. Via sequence search by the motif detection pipeline, annotated ELM classes yield putative instances in query sequences. By adding experimental evidence and references, these putative instances become candidate instances for annotation, and, with further curation, ultimately become fully annotated instances.

LINEAR MOTIFS AND DISEASES

The importance of SLiMs is further corroborated by the occurrence of pathologies that are caused by mutations that either mutate existing linear motifs or create novel linear motifs (of undesired function) (27). Examples include 'Usher's syndrome' (28), 'Liddle's Syndrome' (6) or 'Golabi-Ito-Hall Syndrome' (29). The developmental disorder 'Noonan Syndrome' can be caused by mutations in Raf-1 that abrogate the interaction with 14-3-3 proteins mediated by corresponding SLiMs and thereby deregulate the Raf-1 kinase activity (30) (the Raf-1 protein sequence features two LIG_14-3-3_1 binding sites that are annotated at 256-261 and 618-623 in the ELM resource). A related disease, 'Noonan-like Syndrome', is caused by an S to G mutation at position 2 of the SHOC2 protein, creating a novel myristoylation site (annotated as ELM class MOD_NMyristoyl). This irreversible modification results in aberrant targeting of SHOC2 to the plasma membrane and impaired translocation to the nucleus upon growth factor stimulation (5). More information about the implication of short linear motifs on diseases is collected at http://elm.eu.org/infos/diseases.html.

APPLICATION OF THE ELM RESOURCE

By providing a high-quality, manually curated data set of linear motif classes with experimentally validated SLiM instances, the ELM database has proven to be invaluable to the community: small-scale (single protein) analyzes benefit from the detailed annotation of each ELM class in attributing novel features to proteins of interest. By using *in vitro* and *in vivo* studies, von Nandelstadh *et al.* (31) could validate a PDZ class III motif, detected by ELM at the carboxy terminus of myotilin and the FATZ (calsarcin/myozenin) families. This evolutionarily conserved carboxy-terminal motif mediates binding to PDZ domains of ZASP/Cypher and other Enigma family members (ALP, CLP-36 and RIL) and disruption of these interactions results in myofibrillar myopathies (32). Additionally, ELM annotations can contribute to high-throughput screenings (33) as well as development of novel algorithms (34–36), methods (37) and databases (38). Furthermore, the highly curated data of the ELM resource are used as a benchmarking data set to evaluate the accuracy of prediction algorithms (21,39,40).

For any such analysis, the user should be aware that many matches to ELM regular expressions are false positives. Before conducting experiments based on ELM results, it is strongly advisable to check if a motif match is conserved, exposed in a cell compartment in which the motif is known to be functional. The ELM resource applies several filters to provide the user with such information that should ideally also be supported by the experimental evidence.

SUMMARY

The importance of SLiMs is highlighted by the growing number of instances with relevance to diseases or viruses. Yet, despite their importance and abundance, our understanding of linear motifs is still limited. This is mainly owing to the fact that they are still quite difficult to predict computationally and to investigate experimentally (3,41,42). By better understanding the biology of linear motifs, we hope to increase our insight into diseases and viruses (and vice versa). The ELM resource tries to aid the researcher in the search for putative SLiM instances by providing a feature-rich toolset for sequence analysis. Consequently, with the aforementioned additions and changes, we hope that the ELM resource continues to be a valuable asset to the community.

ACKNOWLEDGEMENTS

The authors would like to thank the users of the ELM resource as well as all colleagues, contributors and annotators of the ELM resource.

FUNDING

EMBL international PhD program (to R.J.W.); EMBL Interdisciplinary PostDoc fellowship (EIPOD to N.E.D.); NGFN framework by the Federal Government Department of Education and Science [FKZ01GS0862 (DiGtoP) to M.S. and M.H.]; European Community's Seventh Framework Programme FP7/2009 (SysCilia) (241955 to G.T.) and (SyBoSS) (242129 to K.V.R.); Polish Ministry of Science and Higher Education within Iuventus Plus project (IP2010-0483-70 to M.D.); Biotechnology and Biological Sciences Research Council (BB/F010486/1 to A.C.); Région Alsace and Collège Doctoral Européen (to K.L.); Science Foundation Ireland (08/IN.1/B1864 to G.G.); BBSRC New Investigator Award (BB/I006230/1 to R.J.E.); German Research Foundation (SFB796 Project A2 to H.M.); grants from the Swiss National Science Foundation (to M.O.S.). Funding for open access charge: EMBL.

Conflict of interest statement. None declared.

REFERENCES

1. Davey,N.E., Van Roey,K., Weatheritt,R.J., Toedt,G., Uyar,B., Altenberg,B., Budd,A., Diella,F., Dinkel,H. and Gibson,T.J. (2011) Attributes of short linear motifs. *Mol Biosyst.*, September 12 (doi:10.1039/c1mb05231d; epub ahead of print).
2. Fuxreiter,M., Tompa,P. and Simon,I. (2007) Local structural disorder imparts plasticity on linear motifs. *Bioinformatics*, **8**, 950–956.
3. Diella,F., Haslam,N., Chica,C., Budd,A., Michael,S., Brown,N.P., Trave,G. and Gibson,T.J. (2008) Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Front. Biosci.*, 6580–6603.
4. Gibson,T.J. (2009) Cell regulation: determined to signal discrete cooperation. *Trends Biochem. Sci.*, **34**, 471–482.
5. Cordeddu,V., Di Schiavi,E., Pennacchio,L.A., Ma'ayan,A., Sarkozy,A., Fodale,V., Cecchetti,S., Cardinale,A., Martin,J., Schackwitz,W. *et al.* (2009) Mutation of SHOC2 promotes aberrant protein N-myristoylation and causes Noonan-like syndrome with loose anagen hair. *Nat. Genet.*, **9**, 1022–1026.
6. Furuhashi,M., Kitamura,K., Adachi,M., Miyoshi,T., Wakida,N., Ura,N., Shikano,Y., Shinshi,Y., Sakamoto,K., Hayashi,M. *et al.* (2005) Liddle's syndrome caused by a novel mutation in the proline-rich PY motif of the epithelial sodium channel beta-subunit. *J. Clin. Endocrinol. Metab.*, **1**, 340–344.
7. Deretic,D., Schmerl,S., Hargrave,P.A., Arendt,A. and McDowell,J.H. (1998) Regulation of sorting and post-Golgi trafficking of rhodopsin by its C-terminal sequence QVS(A)PA. *Proc. Natl Acad. Sci. USA.*, **18**, 10620–10625.
8. Davey,N.E., Trave,G. and Gibson,T.J. (2011) How viruses hijack cell regulation. *Trends Biochem. Sci.*, **3**, 159–169.
9. Neduva,V., Linding,R., Su-Angrand,I., Stark,A., de Masi,F., Gibson,T.J., Lewis,J., Serrano,L. and Russell,R.B. (2005) Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol.*, **12**, e405.
10. Hulo,N., Bairoch,A., Bulliard,V., Cerutti,L., Cuche,B.A., de Castro,E., Lachaize,C., Langendijk-Genevaux,P.S. and Sigrist,C.J. (2008) The 20 years of PROSITE. *Nucleic Acids Res.*, **36**, D245–D249.
11. Rajasekaran,S., Balla,S., Gradie,P., Gryk,M.R., Kadaveru,K., Kundeti,V., Maciejewski,M.W., Mi,T., Rubino,N., Vyas,J. *et al.* (2009) Minimotif miner 2nd release: a database and web system for motif search. *Nucleic Acids Res.*, **37**, D185–D190.
12. Obenauer,J.C., Cantley,L.C. and Yaffe,M.B. (2003) Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.*, **13**, 3635–3641.
13. Puntervoll,P., Linding,R., Gemund,C., Chabanis-Davidson,S., Mattingsdal,M., Cameron,S., Martin,D.M., Ausiello,G., Brannetti,B., Costantini,A. *et al.* (2003) ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res.*, **13**, 3625–3630.
14. Gould,C.M., Diella,F., Via,A., Puntervoll,P., Gemund,C., Chabanis-Davidson,S., Michael,S., Sayadi,A., Bryne,J.C. *et al.* (2010) ELM: the status of the 2010 eukaryotic linear motif resource. *Nucleic Acids Res.*, **38**, D167–D180.
15. Davey,N.E., Edwards,R.J. and Shields,D.C. (2010) Computational identification and analysis of protein short linear motifs. *Front. Biosci.*, **15**, 801–825.
16. Hermjakob,H., Montecchi-Palazzi,L., Bader,G., Wojcik,J., Salwinski,L., Ceol,A., Moore,S., Orchard,S., Sarkans,U., von Mering,C. *et al.* (2004) The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. *Nat. Biotechnol.*, **2**, 177–183.
17. Cote,R.G., Jones,P., Apweiler,R. and Hermjakob,H. (2006) The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics.*, **7**.
18. Via,A., Gould,C.M., Gemund,C., Gibson,T.J. and Helmer-Citterich,M. (2009) A structure filter for the Eukaryotic Linear Motif Resource. *BMC Bioinformatics.*, **10**.
19. Letunic,I., Doerks,T. and Bork,P. (2009) SMART 6: recent updates and new developments. *Nucleic Acids Res.*, **37**, D229–D232.
20. Dosztanyi,Z., Csizmok,V., Tompa,P. and Simon,I. (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **16**, 3433–3434.
21. Dinkel,H. and Sticht,H. (2007) A computational strategy for the prediction of functional linear peptide motifs in proteins. *Bioinformatics*, **24**, 3297–3303.
22. Chica,C., Labarga,A., Gould,C.M., Lopez,R. and Gibson,T.J. (2008) A tree-based conservation scoring method for short linear motifs in multiple alignments of protein sequences. *BMC Bioinformatics*, **9**.
23. Balagopalan,L., Coussens,N.P., Sherman,E., Samelson,L.E. and Sommers,C.L. (2010) The LAT story: a tale of cooperativity, coordination, and choreography. *Cold Spring Harb. Perspect. Biol.*, **8**, a005512.
24. Pawson,T. and Scott,J.D. (2005) Protein phosphorylation in signaling–50 years and counting. *Trends Biochem. Sci.*, **6**, 286–290.
25. Dinkel,H., Chica,C., Via,A., Gould,C.M., Jensen,L.J., Gibson,T.J. and Diella,F. (2011) Phospho.ELM: a database of phosphorylation sites–update 2011. *Nucleic Acids Res.*, **39**, D261–D267.
26. Cruz,J.L., Sola,I., Becares,M., Alberca,B., Plana,J., Enjuanes,L. and Zuniga,S. (2011) Coronavirus gene 7 counteracts host defenses and modulates virus virulence. *PLoS Pathog.*, **6**, e1002090.
27. Kadaveru,K., Vyas,J. and Schiller,M.R. (2008) Viral infection and human disease–insights from minimotifs. *Front. Biosci.*, **13**, 6455–6471.
28. Weil,D., El-Amraoui,A., Masmoudi,S., Mustapha,M., Kikkawa,Y., Laine,S., Delmaghani,S., Adato,A., Nadifi,S., Zina,Z.B. *et al.* (2003) Usher syndrome type I G (USH1G) is caused by mutations in the gene encoding SANS, a protein that associates with the USH1C protein, harmonin. *Hum. Mol. Genet.*, **5**, 463–471.
29. Tapia,V.E., Nicolaescu,E., McDonald,C.B., Musi,V., Oka,T., Inayoshi,Y., Satteson,A.C., Mazack,V., Humbert,J., Gaffney,C.J. *et al.* (2010) Y65C missense mutation in the WW domain of the Golabi-Ito-Hall syndrome protein PQBP1 affects its binding activity and deregulates pre-mRNA splicing. *J. Biol. Chem.*, **25**, 19391–19401.
30. Pandit,B., Sarkozy,A., Pennacchio,L.A., Carta,C., Oishi,K., Martinelli,S., Pogna,E.A., Schackwitz,W., Ustaszewska,A., Landstrom,A. *et al.* (2007) Gain-of-function RAF1 mutations cause Noonan and LEOPARD syndromes with hypertrophic cardiomyopathy. *Nat. Genet.*, **8**, 1007–1012.
31. von Nandelstadh,P., Ismail,M., Gardin,C., Suila,H., Zara,I., Belgrano,A., Valle,G., Carpen,O. and Faulkner,G. (2009) A class III PDZ binding motif in the myotilin and FATZ families binds enigma family proteins: a common link for Z-disc myopathies. *Mol. Cell Biol.*, **3**, 822–834.
32. Selcen,D. and Engel,A.G. (2004) Mutations in myotilin cause myofibrillar myopathy. *Neurology*, **8**, 1363–1371.
33. Gfeller,D., Butty,F., Wierzbicka,M., Verschueren,E., Vanhee,P., Huang,H., Ernst,A., Dar,N., Stagljar,I., Serrano,L. *et al.* (2011) The multiple-specificity landscape of modular peptide recognition domains. *Mol. Syst. Biol.*, **7**.
34. Bauer,D.C., Willadsen,K., Buske,F.A., Le Cao,K.A., Bailey,T.L., Dellaire,G. and Boden,M. (2011) Sorting the nuclear proteome. *Bioinformatics*, **13**, i7–i14.
35. Walsh,I., Martin,A.J., Di Domenico,T., Vullo,A., Pollastri,G. and Tosatto,S.C. (2011) CSpritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs. *Nucleic Acids Res.*, **39**, W190–W196.
36. Lieber,D.S., Elemento,O. and Tavazoie,S. (2010) Large-scale discovery and characterization of protein regulatory motifs in eukaryotes. *PLoS One*, **12**, e14444.
37. Pless,O., Kowenz-Leutz,E., Dittmar,G. and Leutz,A. (2011) A differential proteome screening system for post-translational modification-dependent transcription factor interactions. *Nat. Protoc.*, **3**, 359–364.

38. Goel,R., Muthusamy,B., Pandey,A. and Prasad,T.S. (2011) Human protein reference database and human proteinpedia as discovery resources for molecular biotechnology. *Mol. Biotechnol.*, **1**, 87–95.
39. Edwards,R.J., Davey,N.E. and Shields,D.C. (2007) SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. *PLoS One*, **10**, e967.
40. Edwards,R.J., Davey,N.E. and Shields,D.C. (2008) CompariMotif: quick and easy comparisons of sequence motifs. *Bioinformatics*, **10**, 1307–1309.
41. Perkins,J.R., Diboun,I., Dessailly,B.H., Lees,J.G. and Orengo,C. (2010) Transient protein-protein interactions: structural, functional, and network properties. *Structure*, **10**, 1233–1243.
42. Edwards,R.J., Davey,N.E., Brien,K.O. and Shields,D.C. (2011) Interactome-wide prediction of short, disordered protein interaction motifs in humans. *Mol. Biosyst.*, August 30 (doi:10.1039/c1mb05212h; epub ahead of print).

A.2. Prediction of instances of known linear motifs using protein interaction data – iELM

Summary

A rapidly growing number of protein-protein interactions is being published, often resulting from HTP proteomic studies. For most of these interactions, we miss information about the molecular details of their interaction interfaces. Such knowledge would be important to better understand the function of an interaction in its cellular context and to develop potential inhibitors. In this article, we describe the method iELM that aims at predicting potential SLiM-mediated binding inferfaces of protein interactions.
Approach: iELM is based on the information that a particular class of SLiMs will bind to a particular class of domains. Interacting domain instances have been manually annotated for all SLiM classes that are currently stored in the ELM resource. Using these domain instances and their orthologues, HMMs have been created with the aim to capture domain properties that divide domains into (sub)classes based on recognition of subsets of SLiMs. These HMMs have been used to search protein sequences for instances of domains. If a domain instance has been found, interaction partners of this protein have been consequently searched for instances of the corresponding SLiM class. SLiM instances have been searched using the SLiMSearch program. SLiMSearch uses regular expressions to identify potential SLiM instances and scores them based on their local conservation and disorder propensity. In addition, potential SLiM instances are assessed for overlap with known Pfam domains. The scores from these different sub-methods have been combined into a SVM. Resulting domain-SLiM interfaces have been assessed by PepSite for their ability to biophysically interact with each other on condition that a known 3D structure of at least 30% similarity to the domain instance was available.
Findings: HMMs of domains stored in Pfam have often been found to be too general for being used in SLiM-domain interface prediction necessitating the definition of own HMMs. iELM performed well on benchmark data sets with sensitivities around 75% and specificities around 80%. iELM has been applied to all human protein interactions stored in the STRING database (306,211 in total). Of those, more than 12,000 interactions were predicted to be likely to be mediated by a SLiM-domain interface. The predicted SLiM-domain interfaces comprised published and unknown ones.
Discussion/Conclusions: iELM currently only works for SLiM classes that are annotated in the ELM resource but can easily be extended to other SLiM classes. The number of protein interactions for which SLiM–domain interfaces can be predicted is likely to increase with increasing numbers of annotated SLiM classes. Detection of SLiM–domain interfaces in protein interactions can provide information about the directionality of interactions, e.g. by providing information about which protein will posttranslationally modify which other protein. In addition, prediction of overlapping binding interfaces within one protein chain can be used to suggest interactions that are mutually exclusive. This is very valuable information when trying to decipher the

different functions that one protein can be engaged with.
Contribution: The development of iELM has been resulted from my work in Toby Gibson's lab at EMBL during my Master's thesis that focussed on using protein interaction information for improving linear motif prediction. During my PhD thesis, I have finalised my Master's project and assisted the first author of this article in further method development. I have helped structuring and improving the manuscript.

BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 7 2012, pages 976–982
doi:10.1093/bioinformatics/bts072

Systems biology

Advance Access publication February 10, 2012

The identification of short linear motif-mediated interfaces within the human interactome

R. J. Weatheritt[1], K. Luck[2], E. Petsalaki[3,4], N. E. Davey[1,5] and T. J. Gibson[1,*]

[1]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany, [2]Group Oncoproteins, Unité CNRS-UDS UMR 7242, Institut de Recherche de l'Ecole de Biotechnologie de Strasbourg, 1, Bd Sébastien Brant, BP 10413, 67412 Illkirch - Cedex, France, [3]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, ON M5G 1X5, [4]Department of Genetics, University of Toronto, Toronto, ON M5S 3E1, Canada and [5]Chemical Biology Core Facility, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Eukaryotic proteins are highly modular, containing multiple interaction interfaces that mediate binding to a network of regulators and effectors. Recent advances in high-throughput proteomics have rapidly expanded the number of known protein–protein interactions (PPIs); however, the molecular basis for the majority of these interactions remains to be elucidated. There has been a growing appreciation of the importance of a subset of these PPIs, namely those mediated by short linear motifs (SLiMs), particularly the canonical and ubiquitous SH2, SH3 and PDZ domain-binding motifs. However, these motif classes represent only a small fraction of known SLiMs and outside these examples little effort has been made, either bioinformatically or experimentally, to discover the full complement of motif instances.

Results: In this article, interaction data are analysed to identify and characterize an important subset of PPIs, those involving SLiMs binding to globular domains. To do this, we introduce iELM, a method to identify interactions mediated by SLiMs and add molecular details of the interaction interfaces to both interacting proteins. The method identifies SLiM-mediated interfaces from PPI data by searching for known SLiM–domain pairs. This approach was applied to the human interactome to identify a set of high-confidence putative SLiM-mediated PPIs.

Availability: iELM is freely available at http://elmint.embl.de

Contact: toby.gibson@embl.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 26, 2011; revised on January 9, 2012; accepted on January 28, 2012

1 INTRODUCTION

Short linear motifs (SLiMs) are compact domain binding interfaces ubiquitous in eukaryotic proteomes. They mediate a range of important cellular processes including protein scaffolding [e.g. SOS1 SH3 motifs (Kaneko *et al.*, 2008)], cell signalling [e.g. PDZ motifs (Lee and Zheng, 2010)], subcellular compartment targeting (e.g. nuclear localization signals (Fontes *et al.*, 2003)], post-translational modification [e.g. sumoylation (Yang and Gregoire, 2006)] and cleavage [e.g. caspase 3 cleavage sites (Pop and Salvesen, 2009)]. SLiMs consist of ~3–10 amino acids though usually only 2–4 residues are strictly required for binding. As a result of the limited number of residues contacting their binding partner, SLiMs bind with low affinity [usually between 1.0 and 150 micromolar (Diella *et al.*, 2008)] distinguishing them from domain–domain interactions that often have an affinity in the nanomolar range (Neduva *et al.*, 2005). This attribute of a weak-binding affinity renders SLiM-mediated interactions difficult to detect experimentally (Diella *et al.*, 2008). A number of resource- and time-intensive experiments are therefore required to properly validate a SLiM, ranging from mutational analysis to structural studies (Davey *et al.*, 2012). The use of bioinformatics is therefore an important technique to direct or augment the experimental elucidation of SLiMs.

A number of databases have been developed to facilitate our understanding of SLiMs. The Eukaryotic Linear Motif (ELM) resource (Dinkel *et al.*, 2012) contains over 1600 experimentally validated SLiM instances while the Minimotif Miner (Mi *et al.*, 2012) database has collected over 880 consensus sequences. These datasets generate insights into the attributes of SLiMs, such as their conservation among homologues and enrichment in disorder. This enables the development of prediction servers within both the ELM and Minimotif Miner resources to filter novel instances based on the attributes of the curated regular expressions. However, both servers have issues with over-prediction. The SLiMSearch resource (Davey *et al.*, 2011) expands this methodology to whole proteome searches. This method scores a SLiM instance by assessing the sequence conservation of the motif in its orthologous proteins, however, disordered regions are often poorly aligned and this can lead to an artificially low score for some motifs (Perrodou *et al.*, 2008). The Anchor (Meszaros *et al.*, 2009) predictors rely on the propensity for SLiMs to undergo a disorder-to-order transition upon binding and α-MORF-Pred (Mohan *et al.*, 2006) identifies patterns in a disorder prediction output. Other resources have focused on a subset of SLiMs (Hui and Bader, 2010; Li *et al.*, 2008), for example,

*To whom correspondence should be addressed.

ScanSite (Obenauer *et al.*, 2003) was established to identify short protein sequence motifs based on peptide library and phage display experiments.

The growth in the number of protein complexes with a determined 3D structure has facilitated the development of structural tools to predict SLiM specificities (Betel *et al.*, 2007; Encinar *et al.*, 2009; King and Bradley, 2010; Petsalaki *et al.*, 2009; Stein and Aloy, 2010). The ADAN database (Encinar *et al.*, 2009) utilizes the FoldX algorithm (Schymkowitz *et al.*, 2005) to perform an assessment of the stability and affinity of peptide–domain complexes under *in silico* mutagenesis analysis. However, the requirement for extensive knowledge of these interfaces has generally curtailed this type of method to well-studied and ubiquitous domains, such as the SH3, SH2 and PDZ domains (Encinar *et al.*, 2009; Stein and Aloy, 2010). The exception is PepSite (Petsalaki *et al.*, 2009), which provides a generic method to predict peptide binding by using a position-specific scoring matrix to predict peptide binding though this all-encompassing approach lead to a decrease in accuracy when compared with domain-specific methods. SLiM prediction has also taken advantage of the recent advances in high-throughput proteomics (Beltrao and Serrano, 2005; Edwards *et al.*, 2007; Linding *et al.*, 2007; Neduva *et al.*, 2005), for example, Dilimot (Neduva *et al.*, 2005) and SLiMFinder (Edwards *et al.*, 2007) identify novel SLiM classes by searching for enriched motifs within interaction data while NetworKIN (Linding *et al.*, 2007) uses protein–protein interaction (PPI) data to elucidate the kinase associated with a particular phosphorylation site. However, the inherent noise within PPI networks hinders these methods. Despite these advances in the area of SLiM discovery tools, outside the intensively experimentally studied SH3, SH2 and PDZ domains, the expected deluge of new SLiM instances and classes has not occurred. Nevertheless, there is clearly signal in each of the methods described as demonstrated by the positive results produced in the analyses of Translin (Neduva *et al.*, 2005), EH-1 (Copley, 2005) and KENBox (Michael *et al.*, 2008) SLiM classes, as well as, the identification of kinases associated with particular phosphorylation sites by NetworKin (Linding *et al.*, 2007).

In this study, we produce a high-confidence list of human SLiM-mediated interfaces by creating a method (iELM) that identifies SLiM–domain partners from interaction data. A dataset of SLiM-binding domains and SLiM-mediated interactions was manually curated from the literature. These annotated domains were used to train Hidden Markov Models (HMMs) to specifically recognize SLiM-binding domains associated with a particular ELM class. To identify true SLiM instances a combination of methods, relying on known SLiM attributes, were incorporated allowing the assessment of a binary interaction for a complimentary SLiM-domain partnership. This association is also assessed for structural feasibility by the structural bioinformatics tool, PepSite. The iELM method enables the analysis of the human interactome for SLiM-mediated interfaces and interactions. A list of high-confidence SLiM-mediated interfaces for the human interactome is produced and can be accessed at http://elmint.embl.de.

2 METHODS

iELM assesses a binary interaction for a SLiM–domain interface and, if present, outputs the SLiM sequence and the globular domain putatively responsible for binding.

2.1 Datasets

The SLiM functional classes used in iELM were extracted, in the form of a regular expression, from the ELM database (2011-03). The ELM resource annotation did not include information about the binding partners and binding domain for each ELM class. To identify this information, the 3DID resource (Stein *et al.*, 2011) was parsed for the SLiM-binding domains in complex with a peptide from an ELM class; however, this search only identified 28% (44) of the binding domains for the ELM classes. To identify the remaining 72% (112) of SLiM-binding domains a literature search was undertaken. The annotation process recorded the UniProt ID, the binding domain and the domain's position within the sequence as well as, when possible, the affinity of the binding (see Supplementary Table S3).

2.1.1 Annotation of true positive SLiM-mediated interface dataset The true positive dataset is the experimentally annotated dataset of SLiM–domain interaction interfaces (SLiMDoM dataset) based on the aforementioned literature survey and the crystal structures retrieved from the 3DID database. The SLiMDom test dataset consists of 1080 SLiM–domain-mediated interactions and the training set comprises of 434 SLiM–domain-mediated interactions. This dataset was divided for each ELM class in a 3:1 divide with respect to testing and training.

A second true positive dataset based on the annotation from the Domino (Ceol *et al.*, 2007) resource (version 2009-10) was also assembled. The Domino database annotates the sequences of peptides experimentally shown to bind to a particular globular domain. With our *a priori* knowledge of the Pfam domain (Finn *et al.*, 2010) that binds an ELM class, the appropriate ELM regular expression (Dinkel *et al.*, 2012) was used to search within the binding peptides. The results were recorded and are referred to as the Domino dataset (Supplementary Table S4) consisting of 1684 interactions.

2.1.2 False positive or control SLiM-mediated interface datasets Experimentally validated negative instances are too rare to be used as a control group. Instead a false positive dataset of SLiM-mediated interfaces unlikely to be true was constructed. The majority of these interfaces are likely to be true negatives, however, since our knowledge of SLiMs and PPIs is incomplete, this set will undoubtedly contain functional instances and true interactions.

Two false positive datasets (SLiMDoM- and Domino-False Positive Datasets) were created to be specific controls for each of the aforementioned true positive datasets and the same procedure was applied to each. First, all proteins in these datasets were collected along with their associated ELM class(es). These proteins were combined in all possible combinations such that in a dataset of 10 proteins, each protein would have nine interactions. This list was then filtered for proteins associated with the same ELM class as well as for known interactions [using STRING resource v9.0 (Szklarczyk *et al.*, 2011)]. After these filtering steps, 211 600 protein pairs were present for the false positive SLiMDoM dataset and 111 156 pairs were present within the false positive Domino dataset. These datasets were pruned to produce two datasets each containing 30 000 interactions. The datasets used to train the support vector machine (SVM) algorithm are described in the Supplementary Material.

A final test dataset was constructed to assess the performance of the iELM method on 'real-world' PPI data from the BioGrid (Stark *et al.*, 2011) database (version 3.1.70). This PPI network was randomized by node degree conservation using the Neat web server (Brohee *et al.*, 2008) to ensure the underlying structure of the network remained intact.

2.2 HMM production

The HMMs were trained on a multiple sequence alignment consisting of the experimentally annotated SLiM-binding domain instance and its orthologous proteins. The underlying assumption of this being that the orthologous domains of the annotated domain would also bind the motif. The orthologous sequences of the annotated protein were identified using the Gopher programme (Davey *et al.*, 2007) to search the UniProt database

(UniProt release 2011-05) (UniProt Consortium, 2010) by BLAST reciprocal best hit for each species (Altschul *et al.*, 1990). These orthologous proteins were aligned using the multiple sequence alignment programme Muscle (Edgar, 2004) and the position of the SLiM-binding domain identified within the alignment. To remove poorly sequenced and/or incorrectly identified orthologues, aligned domains with indels covering >10% of the reference domain sequence were removed. The sequences were then iteratively realigned and poorly aligned sequences removed until a set of orthologues were identified with <10% indel coverage compared with the curated reference SLiM-binding domain. The HMMs were trained on this alignment using the HMMer programme's (Eddy, 1998) HMMBuild. The HMMs produced by this process are the 'domain identifier' HMMs. For the benchmarking, only the 434 HMMs made from the SLiMDom training set were used.

2.3 Modelling domains for PepSite

PepSite requires a Protein Data Bank (PDB) structure in order to predict the binding position of a peptide. The sequences of all the 3D structures from the PDBe database (Velankar and Kleywegt, 2011) were blasted against the human UniProt (UniProt Consortium, 2010) sequences for matches with a sequence identity of >30%. For all the non-identical matches detected, structural models of the domain were produced using the MODELLER programme (Eswar *et al.*, 2006) (see Supplementary Fig. S4 for receiver-operating characteristic curve (ROC) for PepSite benchmarking on models).

2.4 Training SVM kernel

The score for the iELM resource is calculated using a SVM learning algorithm (Joachims, 2002). The SVM algorithm was trained on the SVM true positive and SVM false positive datasets (see Supplementary Material). The iELM method was run with 75% of the data used as a training dataset and 25% as a test dataset and a SVM trained model produced.

2.5 Method outline

2.5.1 Domain identifier The HMMer package's HMMSearch programme was used to search a sequence using the domain identifier HMMs. The domain identifier uses an *E*-value cut-off of 0.01 (Finn *et al.*, 2010) and, in order to remove fragment hits, all hits with a length of <80% of the annotated SLiM-binding domain's length were also rejected; if a result is returned, the *E*-value score(s) is converted into a domain score. The domain score is a similarity score to the optimal score of an annotated SLiM-binding domain of similar length. This calculation was based on the equation of the regression line calculated from the optimal *E*-value hit for each domain against the length of the annotated HMM (Pearson's correlation value 0.96). The HMM_length is the length of the HMM used to make the prediction and the *E*-value is the estimated likelihood calculated by the HMMSearch programme:

$$X = \frac{-1.93E - \text{value}}{\text{HMM\_length} - 1.076}$$

2.5.2 iELM method iELM predicts the SLiM-mediated interfaces of a single binary interaction by combining the domain identifier with the motif discovery programme SLiMSearch (Davey *et al.*, 2011), the disorder predictor IUPred (Dosztanyi *et al.*, 2005) and the structural analysis programme PepSite (Petsalaki *et al.*, 2009) (see workflow in Fig. 1).

2.5.3 Interface-pair identification A binary interaction is first queried for interacting domains as annotated in the 3DID resource (Stein *et al.*, 2011). The identification of a putative domain–domain interaction between the binary partners leads to the search being discontinued and the domain–domain interaction being returned. Otherwise, the two proteins in the binary interaction are searched using the following two procedures. The Domain identifier searches a sequence using the domain identifier HMMs in order to identify putative SLiM-binding domains. If a putative SLiM-binding domain is present, a search is undertaken for the corresponding SLiM of the same ELM class in the interacting protein. The SLiMSearch programme uses a regular expression, annotated within the ELM resource, to identify potential SLiMs and assigns a Relative Local Conservation (RLC) score of the residues based on a multiple alignment of the sequence and its orthologues [see Davey *et al.* (2011) for details]. The SLiM and its surrounding residues are then assessed for their propensity to be in a region of intrinsic disorder using IUPred. The SLiMSearch programme also outputs a score for the Conservation Score (Chica *et al.*, 2008) and a RLC variance score indicating the differences in conservation between the individual amino acids of the SLiM instance. Contextual information such as overlapping Pfam Domains and PDB structures (Velankar and Kleywegt, 2011) is also included.

2.5.4 Interface-pair scoring If a complimentary SLiM–domain association is found then the score from the domain identifier and the SLiM detection methods are assessed using a SVM trained model, otherwise the search discontinues. The following scores are considered using SVM_{light} classify programme (Joachims, 2002) for assessment: Domain score, RLC score, RLC variance, IUPred disorder score, the Conservation score and HMM length. Finally, the SLiM–domain interface is assessed using PepSite, to test whether or not the binding is biophysically feasible. This requires a PDB structure (or a model) of the putative SLiM-binding domain. If such a 3-dimensional structure is available, PepSite analyses the SLiM-binding domain for the likely binding position of the peptide, producing a putative binary complex and a score for the likelihood of the interaction. This score is not included in the iELM score calculated by the SVM, because identified SLiM-binding domains often do not have known 3D structures with >30% sequence identity and therefore cannot be assessed using PepSite.

2.6 Method assessment

2.6.1 Dataset assessment The datasets were split into training and test datasets and assessed for sensitivity and specificity:

$$\text{Sensititvity} = \frac{\text{Number of true positives}}{\text{Number of true positives} + \text{Number of false negatives}}$$

$$\text{Specificity} = \frac{\text{Number of true negatives}}{\text{Number of true negatives} + \text{Number of false positives}}$$

The true hits are considered correct if the annotated SLiM and SLiM-binding domain positions were predicted to bind with a score above the set threshold.

3 RESULTS

3.1 Features of SLiM–domain interfaces

The annotated SLiMDom dataset reveals that many globular domain classes bind to multiple ELM classes (156 ELM classes annotated to 85 globular domain functional classes). Those globular domain Pfam families that bind multiple ELM classes can be broadly divided into two categories. The first type, have an over-arching canonical SLiM with subgroups, in general, defined by slight differences in flanking residues of the motif. These classes partially overlap with changes in binding affinity distinguishing closely related subgroups [e.g. Huang *et al.* (2008); Kay *et al.* (2000)]. For example, the core constituent of the canonical SH3-binding SLiM is PxxP (x = any amino acid) with the specificity of the subgroups of this domain class arising from the flanking residues (e.g. YxxPxxP as compared to PxxPxR) (Li, 2005). The second category can also be divided into subgroups, however in contrast to the first type, no over-arching canonical SLiM can be defined, as the SLiMs associated with

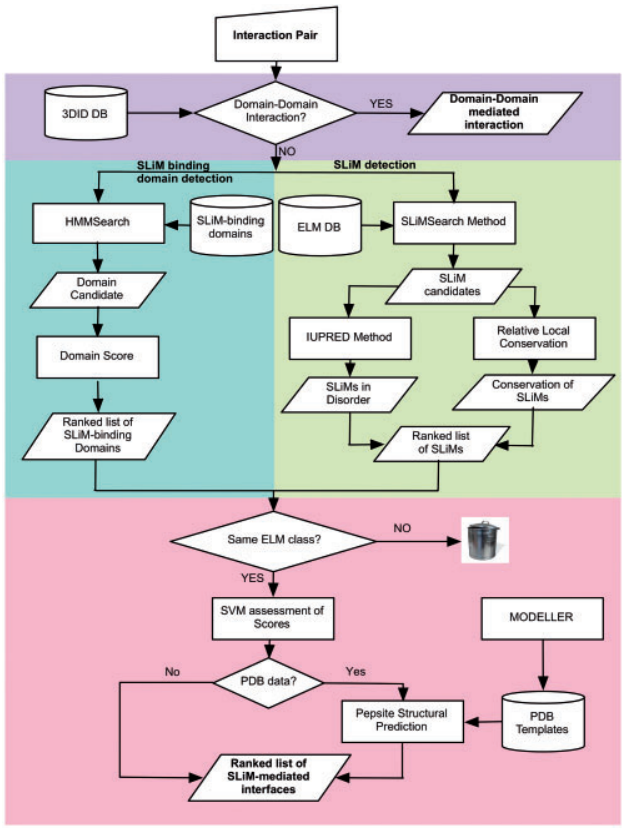


Fig. 1. A workflow for the iELM method. The pipeline proceeds through four major stages utilizing the 3DID resource (purple), the SLiM-binding domain identification method (green), the SLiMSearch methods (yellow) and the PepSite structural bioinformatic methods (red). After this step, the bioinformatic pipeline ends and laboratory verification is required.

this type of domain family are too diverse. These subgroups often contain only paralogous proteins and have SLiM specificities that are very definitive and often exclusive to each subgroup. For example, the WD40 domains of beta-TrCP (uniprot: Q9Y297) bind to a phospho-dependent degron SLiM (LIG_SCF_TrCP_1 - DSGxxS) while the WD40 repeats of PEX7 (uniprot: O00628) binds to a seemingly unrelated SLiM (TRG_PTS2 – Rxxx[LIV]xx[HQ][LIF]) (Stirnimann *et al.*, 2010).

A method for identifying SLiM-domains must be able to distinguish between the aforementioned subgroups. The use of HMMs to identify globular domains and transmembrane regions is well established (Eddy, 1998; Finn *et al.*, 2010) and incorporated into resources such as Pfam. The HMMs trained by Pfam recognize functional domain groups and could therefore be used to identify SLiM-binding domains. However, these HMMs are not able to distinguish the aforementioned intra-domain binding specificities, since the training of Pfam HMMs does not take into account the subcategorization of a domain family by SLiM specificities. We therefore used the annotated and experimentally validated SLiM-binding domains (and their orthologues) to train HMMs. By incorporating known binding specificities, those HMMs trained to recognize SLiM-binding domain should distinguish the subgroups of those functional globular domains that bind multiple ELM classes (see Supplementary Material for details).

3.2 Benchmarking the domain identifier

Two types of HMMs were used: those extracted from Pfam (version 25.0) and those that we generated based on the experimentally validated SLiM-binding domains (domain identifier HMMs) (from the training set—see Section 2 for details). For each of the SLiM–domain interactions from the SLiMDoM dataset, the benchmarking assessed whether either the Pfam- or domain identifier HMMs identified the known binding domain. The domain identifier HMMs achieved a sensitivity of 84.0% (907/1080) and a specificity of 90.1% [false positive rate (FPR): 2696/30 000]. Pfam HMMs accomplished a sensitivity and specificity of 65.1% (703/1080) and 72.1% (FPR: 8370/30 000), respectively (see ROC curves in Fig. 2a) suggesting that the use of HMMs trained on SLiM-binding domains is a more effective way of identifying putative SLiM-binding domains. The domain identifier HMMs were also assessed for intra-domain specificities using the annotated SH2 and SH3 domains. The domain identifier HMMs achieved a specificity of 83.9% and a sensitivity of 80.3% (see Supplementary Fig. S2).

3.3 iELM benchmark

The iELM method was benchmarked using two separate datasets. The first consists of experimentally validated SLiM-mediated interaction data (SLiMDoM dataset) and the second is based on the Domino dataset, which is curated from the Domino database's experimentally annotated peptide–domain interactions (for full results see Supplementary Table S5). The performance of iELM on the SLiMDoM dataset using the domain identifier HMMs (cut-off = −1.0) was a sensitivity of 84.8% (916/1080) and a specificity of 86.5% (FPR: 4050/30 000) while using the Pfam HMMs decreased both the sensitivity and specificity scores to 76.1% (822/1080) and 80.4% (FPR: 5880/30 000), respectively (Fig. 2b). Using iELM (cut-off = −1.0) with the domain identifier HMMs on the Domino benchmark dataset achieved a sensitivity of 75.5% (1272/1684) and a specificity of 83.4% (FPR: 4980/30 000). In comparison, the use of Pfam HMMs managed a sensitivity and specificity of 60.9% (1025/1684) and 79.4% (FPR: 6180/30 000), respectively (Fig. 2c). The application of the SVM was contrasted to using a cut-off system, based on the recommendations in the respective papers. The cut-off version of iELM (IUPred: 0.4; Motif score: 0.5; Domain score: 0.4) on the SLiMDoM dataset achieved a slightly better specificity 89.3% (FPR: 3111/30 000) but a much lower sensitivity of 70.4% (760/1080) than the SVM-based method.

The iELM method was also benchmarked on 'real world' data whose interactions were collected independently of whether or not they were SLiM mediated. The BioGrid interaction dataset and a randomized version of this dataset (both containing 46 676 interactions) were assessed using iELM (cut-off = −1.0) with the domain identifier HMMs. Within the BioGrid interaction dataset, 11 153 SLiM-mediated interactions were identified compared to 1112 in the randomized network suggesting a FPR of 9.97%.

3.4 Human interactome analysis

The interfaces for the majority of PPIs are still unknown and it is therefore of interest to detect novel motif-mediated interfaces on a proteome-wide scale. A human PPI network comprising 306 211

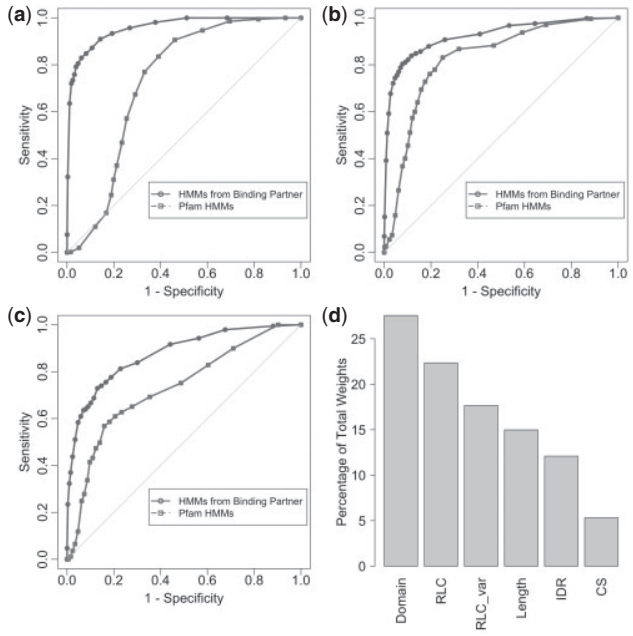
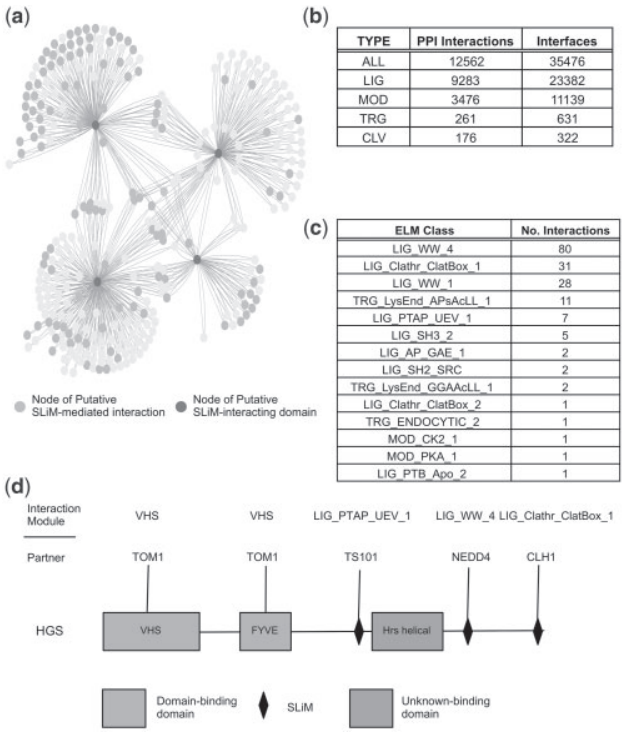


Fig. 2. ROC curves and SVM kernel weights. Plots describing the properties of the domain identifier and iELM methods. (a), (b) and (c) are ROC curves. These curves are a graphical plot of sensitivity versus 1 - specificity as compared with random (the grey line). The ROC curves demonstrate that for detecting SLiM-binding domains and SLiM-mediated interfaces, respectively, the two methods are a considerable improvement over random. Furthermore, they illustrate the advantages of training HMMs on annotated SLiM-binding domains. (**a**) Benchmark dataset results for domain identifier method. (**b**) The iELM method as benchmarked against SLiMDom dataset. (**c**) The iELM method as benchmarked against Domino data. (**d**) A bar plot of the percentage of the total weight as assigned by the SVM kernel. (Domain = Domain Score, RLC_var = RLC variance, length = domain-length, IDR = intrinsic disordered regions, CS = conservation score). The ratio of weights was consistent during multiple testing with a standard deviation of 0.0087, 0.019, 0.012, 0.0068, 0.016 and 0.017 for the domain score, RLC, RLC_var, length, IDR and CS, respectively.



TYPE	PPI Interactions	Interfaces
ALL	12562	35476
LIG	9283	23382
MOD	3476	11139
TRG	261	631
CLV	176	322

ELM Class	No. Interactions
LIG_WW_4	80
LIG_Clathr_ClatBox_1	31
LIG_WW_1	28
TRG_LysEnd_APsAcLL_1	11
LIG_PTAP_UEV_1	7
LIG_SH3_2	5
LIG_AP_GAE_1	2
LIG_SH2_SRC	2
TRG_LysEnd_GGAAcLL_1	2
LIG_Clathr_ClatBox_2	1
TRG_ENDOCYTIC_2	1
MOD_CK2_1	1
MOD_PKA_1	1
LIG_PTB_Apo_2	1

Fig. 3. SLiM-mediated Human Interface Interactome. A summary of the iELM results for the human interactome. (**a**) A cytoscape image (Cline *et al.*, 2007) of a subset of the interactions found to be motif-mediated within the human interactome. The heavily-shaded and highly connected nodes (in dark purple) are the SLiM-binding-domain-containing proteins (in a clockwise order from the top left are): NEDD4, TS101, GGA3 and CLH1. In a slightly lighter shading are highlighted those nodes, identified by iELM as, containing SLiMs binding to the aforementioned SLiM-binding domains. (**b**) Statistics for the number of interactions and interfaces for all the SLiM-mediated interactions and then divided by type using ELM resource distinctions (LIG = ligand, MOD = modification, TRG = targeting, CLV = cleavage). (**c**) A table derived from the interactome shown in (a) depicting those ELM classes found with the number of times they occur. (**d**) The modular interactions of HGS found from the previous network. Also mapped on are interactions found from the 3DID resource (in orange or lighter shading).

interactions [extracted from STRING (Szklarczyk *et al.*, 2011) v9.0; PPIs; cut-off = 0.6] was assessed using the iELM method (cut-off = −1.0). In total, 12 562 PPIs and 35 476 interfaces were predicted as SLiM-mediated by iELM, including 7251 predicted structures (PepSite score < 0.25) (Fig. 3b and Supplementary Table S2). A large number of these PPIs are mediated by multiple SLiM classes or SLiM instances, for example, in the interaction between GRB2 (uniprot:P62993) and SOS1 (uniprot:Q07889); SOS1 has seven putative SH3 motifs and GRB2 has two SH3 domains, potentially this can equate to 14 binding interfaces for a single PPI. The putative motif interface map of the human interactome, produced by the iELM method, identified a large number of potentially novel SLiM-mediated interfaces as well as demonstrating the ability of iELM to automatically annotate the edges of interactions within a PPI network. To explore the interactome produced by iELM, the putative SLiM-mediated-interaction interfaces associated with the cell division cycle protein 20 (CDC20; uniprot: Q12834) were studied. CDC20 is a regulatory subunit of the anaphase-promoting complex (APC/C) that targets proteins for ubiquitination and subsequent degradation by the 26S proteasome (Peters, 2006). In early mitosis, CDC20 joins the APC/C complex and targets substrates for ubiquitination containing either a destruction box SLiM (Glotzer *et al.*, 1991) (D-box – RxxLxxφ – φ = hydrophobic amino acid) or a KEN-box (Pfleger and Kirschner, 2000) (xKENx). The iELM method identified 34 PPIs (from 246 binary interactions) with 41 putative SLiM-mediated interfaces that bind to CDC20 via a D-box motif. All the experimentally annotated (seven instances) ELM instances of D-box SLiMs (including human orthologues of non-human instances) were identified as well as five additional experimentally validated SLiMs (Peters, 2006). iELM identified a number of interesting candidate interfaces binding to CDC20 including the sperm-associated antigen 5 (SPAG5), a protein necessary for spindle formation during mitosis, a process whose completion synchronizes with the formation of the APC/C complex (Song and Rape, 2010) (see Supplementary Fig. S3).

In addition, we investigated a subnetwork of the human SLIM-mediated PPI network associated with four SLiM binding proteins: Clathrin heavy chain 1 (CLH1) (uniprot: Q00610), ADP-ribosylation factor-binding protein GGA3 (GGA3) (uniprot: Q9NZ52), E3 ubiquitin-protein ligase NEDD4 (NEDD4) (uniprot: P46934) and tumour susceptibility gene 101 protein (TSG101) (uniprot: Q99816), and their interactions (Fig. 3a). This subnetwork contains 810 interactions, 173 of which are predicted by iELM as SLiM-mediated interactions. This number includes SLiM interfaces from three different categories of ELM (LIG or ligand, MOD or modification, and TRG or targeting) and 14 different classes (Fig. 3c). Of these 173 putative interactions, approximately half are predicted to bind to NEDD4 via a WW-binding motif associated with ubiquitinating substrates. The remainder of the putative protein interfaces function within endocytic-related pathways; for example, the Clathrin-Box motif-mediated interactions are associated with clathrin-mediated vesicular trafficking. The protein hepatocyte growth factor-regulated tyrosine kinase substrate (HGS) (uniprot: O14964) was extracted from this network and its module architecture investigated (Fig. 3d). This protein contains putative SLiMs for targeting HGS for ubiquitination (via NEDD4), clathrin-mediated endocytosis (via CLH1), signalling via Grb2 and P85A (uniprot: P27986) as well as an annotated PTAP SLiM, involved in the ESCRT signalling. Furthermore, 3DID data predict a domain–domain interaction with Tom1 (uniprot: O60784). This subnetwork highlights the information about functionality and directionality that can be garnered by mapping SLiM-predictions onto PPI networks.

4 DISCUSSION

SLiM-mediated binding interfaces are key components of the human proteome (Jorgensen and Linding, 2008) and are abundant within the signalling pathways of the cell (Pawson, 2007). In this article, we manually annotated domain-binding partners for 156 ELM classes and curated 1514 SLiM-mediated interfaces, thus generating a high-quality dataset for studying the interfaces between specific ELM classes and their interacting domains. This dataset enabled us to train HMMs for identifying SLiM-binding domains. These models were then incorporated into a novel method called iELM with the aim of detecting SLiM-mediated interfaces. iELM was able to distinguish specificities within SLiM-binding domains (see Supplementary Fig. S2), as well as identify SLiM-mediated interactions from a background of PPIs (Fig. 3). The iELM method uses an SVM algorithm in preference to a simple cut-off system due to our wish to develop a method with the best ratio between sensitivity and specificity. A comparison of these two techniques identifies the SVM model as having a higher sensitivity but a lower specificity, with the ratio weighted in favour of the SVM model. This suggests that using the SVM will identify a greater number of true positive interactions with only a slight increase in the FPR. iELM, so far, covers only linear motifs as they are annotated in the ELM resource, but is easily extendible to any SLiM, in the form of a regular expression, for which the interacting SLiM-binding domain is known.

The importance of a number of canonical and ubiquitous domains (e.g. SH2, SH3, PDZ and Pkinase) in signalling and regulatory networks has lead to a great deal of work focusing on their SLiM-binding properties (Beltrao and Serrano, 2005; Encinar *et al.*, 2009; Gfeller *et al.*, 2011; Huang *et al.*, 2008; Hui and Bader, 2010; Li, 2005; Linding *et al.*, 2007; Stein and Aloy, 2010). These domains are abundant in higher eukaryotes with small differences in amino acid composition leading to subtle shifts in specificities (Encinar *et al.*, 2009; Gfeller *et al.*, 2011; Huang *et al.*, 2008). In the ELM resource, and by association in iELM, however, these subtle shifts in specificity are not necessarily fully explored. This is because for a particular SLiM functional class the ELM resource's annotation process aims to curate the full spectrum of variation within eukaryotes; potentially this can allow too broad a specificity for a SLiM and lead to false positive results. Despite these potential problems, the iELM method performed strongly on benchmarking datasets and was able to distinguish specificities for these ubiquitous domains. More importantly, iELM incorporates the less well-known SLiM classes (over two-thirds of those annotated in ELM) that do not have this overlapping intra-domain specificity enabling a more extensive array of SLiM-mediated interfaces to be predicted for the human interactome. This is illustrated by those interactions associated with targeting proteins for destruction, using D-box motifs, as well as by a subnetwork of interconnected SLiM-mediated interactions linked to endocytosis.

The automatic annotation of the molecular detail of a protein–protein interface is an important step in understanding the function of many of the interactions identified by proteomic experiments. In this study, we developed a novel method enabling for the first time, to our knowledge, the fast and automatic annotation of SLiM-mediated interactions on large-scale datasets. The development of iELM permitted us to produce an edge-based interactome of 12 562 interactions with 35 476 interfaces representing ~4% of the known human interactome. This number is likely to represent only a small fraction of the SLiM-mediated interactions within the interactome, as it is only based on 156 ELM classes and SLiM-mediated interactions are known to be under-represented in mass spectrometry-derived proteomic data (Gavin *et al.*, 2006). The final percentage is difficult to estimate as the total number of SLiM classes is unknown but taking into consideration that there are over 13 000 globular domain classes annotated in Pfam, the potential influence of SLiM-mediated interactions is prodigious.

The annotation of the edges of PPI networks allows a more biologically realistic edge-based analysis of PPI networks to be implemented. This is important, as proteins are modular entities whose function can vary depending on their interaction partners. Furthermore, as proteins have a finite number of binding sites, an appreciation of the location of their interaction surface will facilitate models to consider mutually exclusive binding. The use of a node-based view generalizes these properties and therefore loses the subtleties of a protein's behaviour, while an edge-based view would distinguish this difference enabling a more accurate portrayal of cellular networks.

ACKNOWLEDGEMENTS

We would like to thanks members of the Gibson Team for their support and advice; in particular, Kim van Roey for his critical reading of the manuscript and Holger Dinkel for assistance with the website.

Funding: R.J.W. was supported by the EMBL international PhD programme. K.L. was supported by the Région Alsace and

Collége Doctoral Européen. N.E.D. was supported by an EMBL Interdisciplinary Postdoctoral (EIPOD) fellowship.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
Beltrao,P. and Serrano,L. (2005) Comparative genomics and disorder prediction identify biologically relevant SH3 protein interactions. *PLoS Comput. Biol.*, **1**, e26.
Betel,D. *et al.* (2007) Structure-templated predictions of novel protein interactions from sequence information. *PLoS Comput. Biol.*, **3**, 1783–1789.
Brohee,S. *et al.* (2008) NeAT: a toolbox for the analysis of biological networks, clusters, classes and pathways. *Nucleic Acids Res.*, **36**, W444–W451.
Ceol,A. *et al.* (2007) DOMINO: a database of domain-peptide interactions. *Nucleic Acids Res.*, **35**, D557–D560.
Chica,C. *et al.* (2008) A tree-based conservation scoring method for short linear motifs in multiple alignments of protein sequences. *BMC Bioinformatics*, **9**, 229.
Cline,M.S. *et al.* (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.*, **2**, 2366–2382.
Copley,R.R. (2005) The EH1 motif in metazoan transcription factors. *BMC Genomics*, **6**, 169.
Davey,N.E. *et al.* (2007) The SLiMDisc server: short, linear motif discovery in proteins. *Nucleic Acids Res.*, **35**, W455–W459.
Davey,N.E. *et al.* (2011) SLiMSearch 2.0: biological context for short linear motifs in proteins. *Nucleic Acids Res.*, **39** (Suppl. 2), W56–W60.
Davey,N.E. *et al.* (2012) Attributes of short linear motifs. *Mol. Biosyst.*, **8**, 268–281.
Diella,F. *et al.* (2008) Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Front Biosci.*, **13**, 6580–6603.
Dinkel,H. *et al.* (2012) ELM–the database of eukaryotic linear motifs. *Nucleic Acids Res.*, **40**, D242–D251.
Dosztanyi,Z. *et al.* (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**, 3433–3434.
Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
Edwards,R.J. *et al.* (2007) SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. *PLoS One*, **2**, e967.
Encinar,J.A. *et al.* (2009) ADAN: a database for prediction of protein-protein interaction of modular domains mediated by linear motifs. *Bioinformatics*, **25**, 2418–2424.
Eswar,N. *et al.* (2006) Comparative protein structure modeling using Modeller. *Curr. Protoc. Bioinformatics*, **Chapter 5**, Unit 5.6.
Finn,R.D. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
Fontes,M.R. *et al.* (2003) Structural basis for the specificity of bipartite nuclear localization sequence binding by importin-alpha. *J. Biol. Chem.*, **278**, 27981–27987.
Gavin,A.C. *et al.* (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**, 631–636.
Gfeller,D. *et al.* (2011) The multiple-specificity landscape of modular peptide recognition domains. *Mol. Syst. Biol.*, **7**, 484.
Glotzer,M. *et al.* (1991) Cyclin is degraded by the ubiquitin pathway. *Nature*, **349**, 132–138.
Huang,H. *et al.* (2008) Defining the specificity space of the human SRC homology 2 domain. *Mol. Cell. Proteomics*, **7**, 768–784.
Hui,S. and Bader,G.D. (2010) Proteome scanning to predict PDZ domain interactions using support vector machines. *BMC Bioinformatics*, **11**, 507.
Joachims,T. (2002) *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. Springer, Germany.
Jorgensen,C. and Linding,R. (2008) Directional and quantitative phosphorylation networks. *Brief. Funct. Genomic Proteomic*, **7**, 17–26.
Kaneko,T. *et al.* (2008) The SH3 domain–a family of versatile peptide- and protein-recognition module. *Front Biosci.*, **13**, 4938–4952.
Kay,B.K. *et al.* (2000) The importance of being proline: the interaction of proline-rich motifs in signaling proteins with their cognate domains. *FASEB J.*, **14**, 231–241.
King,C.A. and Bradley,P. (2010) Structure-based prediction of protein-peptide specificity in Rosetta. *Proteins*, **78**, 3437–3449.
Lee,H.J. and Zheng,J.J. (2010) PDZ domains and their binding partners: structure, specificity, and modification. *Cell Commun. Signal.*, **8**, 8.
Li,L. *et al.* (2008) Prediction of phosphotyrosine signaling networks using a scoring matrix-assisted ligand identification approach. *Nucleic Acids Res.*, **36**, 3263–3273.
Li,S.S. (2005) Specificity and versatility of SH3 and other proline-recognition domains: structural basis and implications for cellular signal transduction. *Biochem. J.*, **390**, 641–653.
Linding,R. *et al.* (2007) Systematic discovery of in vivo phosphorylation networks. *Cell*, **129**, 1415–1426.
Meszaros,B. *et al.* (2009) Prediction of protein binding regions in disordered proteins. *PLoS Comput. Biol.*, **5**, e1000376.
Mi,T. *et al*. (2012) Minimotif Miner 3.0: database expansion and significantly improved reduction of false-positive predictions from consensus sequences. *Nucleic Acids Res.*, **40**, D252—D260.
Michael,S. *et al.* (2008) Discovery of candidate KEN-box motifs using cell cycle keyword enrichment combined with native disorder prediction and motif conservation. *Bioinformatics*, **24**, 453–457.
Mohan,A. *et al.* (2006) Analysis of molecular recognition features (MoRFs). *J. Mol. Biol.*, **362**, 1043–1059.
Neduva,V. *et al.* (2005) Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol.*, **3**, e405.
Obenauer,J.C. *et al.* (2003) Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.*, **31**, 3635–3641.
Pawson,T. (2007) Dynamic control of signaling by modular adaptor proteins. *Curr. Opin. Cell. Biol.*, **19**, 112–116.
Perrodou,E. *et al.* (2008) A new protein linear motif benchmark for multiple sequence alignment software. *BMC Bioinformatics*, **9**, 213.
Peters,J.M. (2006) The anaphase promoting complex/cyclosome: a machine designed to destroy. *Nat. Rev. Mol. Cell Biol.*, **7**, 644–656.
Petsalaki,E. *et al.* (2009) Accurate prediction of peptide binding sites on protein surfaces. *PLoS Comput. Biol.*, **5**, e1000335.
Pfleger,C.M. and Kirschner,M.W. (2000) The KEN box: an APC recognition signal distinct from the D box targeted by Cdh1. *Genes Dev.*, **14**, 655–665.
Pop,C. and Salvesen,G.S. (2009) Human caspases: activation, specificity, and regulation. *J. Biol. Chem.*, **284**, 21777–21781.
Schymkowitz,J. *et al.* (2005) The FoldX web server: an online force field. *Nucleic Acids Res.*, **33**, W382–W388.
Song,L. and Rape,M. (2010) Regulated degradation of spindle assembly factors by the anaphase-promoting complex. *Mol. Cell*, **38**, 369–382.
Stark,C. *et al.* (2011) The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res.*, **39**, D698–D704.
Stein,A. and Aloy,P. (2010) Novel peptide-mediated interactions derived from high-resolution 3-dimensional structures. *PLoS Comput. Biol.*, **6**, e1000789.
Stein,A. *et al.* (2011) 3DID: identification and classification of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.*, **39**, D718–D723.
Stirnimann,C.U. *et al.* (2010) WD40 proteins propel cellular networks. *Trends Biochem. Sci.*, **35**, 565–574.
Szklarczyk,D. *et al.* (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.*, **39**, D561–D568.
UniProt Consortium (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.
Velankar,S. and Kleywegt,G.J. (2011) The Protein Data Bank in Europe (PDBe): bringing structure to biology. *Acta Crystallogr. D Biol. Crystallogr.*, **67**, 324–330.
Yang,X.J. and Gregoire,S. (2006) A recurrent phospho-sumoyl switch in transcriptional repression and beyond. *Mol. Cell*, **23**, 779–786.

B. Manuscripts in preparation

B.1. Structural basis for hijacking of cellular LxxLL motifs by a papillomavirus E6 oncoprotein

The protein E6 from papillomavirus (PV) possesses two Zn-binding domains that form one globular domain. This domain has been shown to recognize leucine-rich linear motifs of the form LxxLL where x represents any amino acid. Important cellular interaction partners of E6, such as the ubiquitin ligase E6-associated protein (E6AP) and the cellular adhesion protein paxillin, bear such LxxLL motifs that are preferentially bound by HPV16 E6 and bovine papillomavirus 1 (BPV1) E6, respectively. It has been unknown whether these LxxLL motifs constitute a particular type of eukaryotic SLiM or whether they are overlapping with already known types of SLiMs such as LD motifs. In this article, we describe the structure of full length BPV1 E6 bound to an LxxLL motif derived from paxillin. In addition, we used our experimental data to predict new potential cellular proteins with LxxLL motifs that were successfully validated for binding to E6. This summary strongly focusses on the computational part of this article.
Approach: The structure of the fusion construct MBP - LxxLL motif (MDDLDALLAD) - BPV1 E6 has been solved by crystallography. Phage display and mutagenesis have been performed to identify preferences for residues at positions surrounding the conserved leucines. These data in combination with observations from the complex structure have been used to define a regular expression and PSSM. The human and bovine proteomes have been screened using the regular expression. Resulting matches were filtered based on their disorder propensity (rejected if predicted to be in ordered regions), scored and ranked using the PSSM. Functional annotation terms of the proteins bearing identified matches have been searched for significant enrichments.
Results: The structure reveals a helical LxxLL motif that binds into a pocket that is formed by both Zn-binding domains of E6. Proteins predicted to contain LxxLL motifs have equally been identified in independent co-precipitation experiments that were combined with mass spectrometry. The best matches of LxxLL motifs have also been validated *in vitro* to bind to E6. Proteins carrying identified LxxLL motifs were found to be frequently annotated with cell adhesion, cytoskeletal dynamics and organisation as well as transcription regulation, cell proliferation and cell death terms.
Discussion/Conclusions: The LD motif might represent a subclass of LxxLL motifs. Identified potential interaction partners of E6 are worth further investigation to identify their eventual implications in the viral life cycle and in malignant transformation via interaction with E6. The structure of the E6-paxillin complex has shown how E6

proteins specifically capture acidic LxxLL motifs to hijack multiple cellular functions.
Contribution: I have contributed to the definition of the regular expression and the PSSM. I have implemented and performed the proteome-wide screens and filtering strategies as well as the annotation term enrichment analysis. I have been involved in the data analysis and figure preparation.
This manuscript had been once submitted. A modified manuscript is in preparation comprising in addition to the BPV1 E6 structure the HPV16 E6 structure. The computational part and related experimental data will be published separately.
Author list: Sebastian Charbonnier, Nicole Brimer, Abdellahi Ould M'hamed Ould Sidi, Katja Luck, Katia Zanier, Khaled Ould Babah, Tina Ansari, Isabelle Muller, Pierre Poussin, Vincent Cura, Charles Lyons, Jean Cavarelli, Scott Vande Pol, Gilles Travé.

Extraction of supplementary methods: Proteome-wide prediction of protein binders to E6
The human and bovine proteome together with annotations were downloaded from Ensembl v58 [203]. The proteomes were pre-screened with a regular expression representing a raw definition of the leucine-rich E6-binding motif. The regular expression was defined as follows: ...LD.L[LFM].. — ..[DE]L[ˆE].L[LFM].. A dot represents any amino acid, amino acids in brackets are allowed at this position, ˆ represents not and — represents the logical or. Matches of the regular expression were analysed for their propensities of being in disordered regions of the proteins using IUPred [140]. The mean IUPred scores of the motifs and, if possible of the 10 amino acids upstream and downstream were calculated. Motifs with at least one of these 3 mean scores ¡ 0.4 were rejected. The above procedure retained 387 potential BPV-1 E6-binding motifs, which were subsequently ranked using a score calculated with a Position Specific Scoring Matrix (PSSM) based on the work of Mount *et al.* [204]. The PSSM contains values for the likeliness of a residue to be part of the motif for each amino acid at each position of the motif. These values were determined by taking the logarithm of the frequency of occurrence of an amino acid at a particular position divided by the background frequency of this amino acid in the proteome. Frequencies of amino acids at motif positions were based on the mutagenesis experiments, the phage display data and observations based on the BPV-1 E6-LxxLL structure presented in this manuscript. Amino acids that do not appear at specific motif positions were given a very small pseudo count of 10E-10. The scored motifs were further annotated for functional, structural and cellular location information using the Database for Annotation, Visualisation and Integrated Discovery (DAVID) v6.7 [205] (see supplemental Table 3). A motif was marked for its occurrence in coiled-coil regions if at least one of its residues was predicted to be part of coiled-coil [206]. Note that several proteins, such as PXN, AP1G1 or TGFB1I1, were found to contain several instances of potential BPV1 E6-binding motifs (see Figure 3A).

Abstract

Papillomavirus E6 oncoproteins are key players in epithelial tumours induced by papillomaviruses in vertebrates, including cervical cancer in humans. Despite their small size (~150 residues) E6 proteins recognize large numbers of host proteins. Here, the crystal structure of bovine papillomavirus E6 bound to cellular focal adhesion protein paxillin reveals that E6 proteins possess a basic-hydrophobic pocket that specifically recognises acidic LxxLL helical motifs. We identified various E6-binding host proteins related to transformation and immortalisation that contain this motif, and inactivation of the LxxLL binding site disrupted the transforming phenotype of E6. Thus, the structural basis of E6 oncogenic activity resides in its peptide binding pocket, which allows E6 to hijack a large family of sequence motifs mediating protein-protein interaction networks related to oncogenesis.

Papillomaviruses (PV) infect the cutaneous or mucosal epithelia of vertebrates, with more than 200 PV types so far identified and sequenced (*1*). Whereas most PVs produce epithelial hyperplasias, infections by a subset of types known as "high-risk" PVs may eventually lead to cancer. In particular, cervical cancers are caused by High-risk mucosal Human PVs (Hrm-HPVs) (*2*) and some skin cancers have been associated with high risk cutaneous HPVs (*3*). Bovine Papillomavirus Type 1 (BPV-1) is a model system for papillomavirus transcription, transformation and replication (*4*) and induces tumours in its natural host (cattle) and in a heterologous host (equids).

PV carcinogenesis is primarily linked to two PV oncoproteins, E6 and E7. Hrm-HPV E6 recruits the ubiquitin ligase E6AP and the tumour suppressor p53, leading to ubiquitin–mediated degradation of p53 (*5*). Hrm-HPV E6 also interacts with many other cellular proteins, sometimes resulting in their proteasome-dependent degradation (*6*). Hrm-HPV E6 recognises several of its target proteins (including E6AP (*5*) and IRF-3 (*7*)) *via* acidic Leucine-rich motifs containing an LxxLL consensus sequence (*8*, *9*). Low-risk mucosal HPV-11 also interacts with the Leucine-rich motif of E6AP (*10*). Finally, BPV-1 E6 recognises within the focal adhesion protein paxillin several acidic sequences, containing the LxxLL consensus plus additional conserved features, known as "LD motifs" (*11*). E6 binding to LD motifs on paxillin is required for cellular transformation by E6 (*12*, *9*, *13*). Within the host cell, LD motifs regulate cell motility, cell adhesion and gene expression by mediating the interaction of paxillin and related proteins with partner proteins including Focal Adhesion Kinase (FAK), Vinculin, GIT1 (*11*) and alpha Parvin (*14*).

Most mammalian PV E6 proteins are small cysteine-rich proteins consisting of two zinc-binding domains, E6N and E6C. Whereas the solution structure of a soluble mutant of

HPV-16 E6C domain has been determined (*15*), full-length mammalian E6 proteins, including BPV-1 E6, undergo self-oligomerisation processes (*16*), which have precluded their structural analysis for more than twenty years.

To circumvent this problem, we fused together a highly soluble, crystallisation-prone mutant of the bacterial Maltose Binding Protein (MBP), the E6-binding LxxLL sequence present in the LD1 motif of paxillin, and the BPV-1 E6 protein (Supplemental Fig. 1). The resulting MBP-LxxLL-E6 triple fusion construct was produced as a soluble monomer, which readily crystallised in the space group $C222_1$, yielding diffraction data at a resolution better than 2.3 Å using synchrotron radiation. The structure was solved by molecular replacement using the known structure of MBP as a template (Supplemental Fig. 2) (*17*).

The structure of E6 bound to paxillin comprises two zinc-binding domains connected by a linker helix (Fig. 1A, Supplemental Fig. 4). The C-terminal domain (E6C, residues 58-137) adopts a zinc-binding fold similar to that of the isolated HPV16 E6C domain in solution (*15*). The resolved region of the N-terminal domain (E6N, residues 11-57) shares common structural features with the corresponding region of E6C, onto which it can be well superimposed (Supplemental Fig. 5).

The E6-bound motif (sequence $M_1D_2D_3L_4D_5A_6L_7L_8A_9D_{10}$) adopts a helical conformation from residue D_2 to residue L_8 (Fig. 1A, B, D) and inserts inside a groove composed by the two zinc-binding domains and the linker helix of E6 (Fig. 1A, C, D). The three leucine residues L_4, L_7 and L_8 defining the LxxLL motif are plugged into a pocket of hydrophobic residues (W19, F37, V40, A49, L54, C57, L58) exclusively contributed by the E6N domain (Fig. 1, C, E). The E6N domain also contributes an electrostatic component to the complex, via R42 whose side chain forms a salt bridge with D_3 of the peptide and is

also proximal to D_2 (Fig. 1B, bottom right). However, most E6-peptide charge interactions are contributed by E6C. The surface of E6C oriented towards the peptide (Fig. 1B, D) bears seven basic residues (H79, K81, R85, R89, K96, R116 and R121), resulting in a large positively charged surface, which provides a favourable electrostatic environment likely to attract negatively charged peptides into the binding pocket. The side-chain of R116 establishes a salt bridge with the side chain of residue D_5 of the peptide. Other positive residues of the E6C surface participate to a network of interactions mediated by water molecules connecting the peptide and the binding pocket (Fig. 1B).

The viral oncoprotein E6 displays a novel mode of LxxLL motif recognition as compared to cellular proteins that bind LxxLL motifs. The cellular FAT and CH domains are helical domains that recognise the LD motif with rather weak equilibrium affinity constants (K_D ~ 100 μM to 1 mM) via surface interactions, with the bound motif occupying a peripheral position (*18*, *19*) (Fig. 1G). Helix-shaped LxxLL motifs are also found within the "NR boxes" that mediate the interaction of transcription coactivators with the LBD domains of nuclear receptors (*20*). Like the FAT and CH domains, the LBD domains are fully helical and bind rather weakly (K_D ~ 1 μM (*21*)) to the LxxLL motif placed in a rather peripheral position (Fig. 1G). Comparatively, the LD peptide of paxillin is significantly more buried when clamped into the hydrophobic pocket of viral E6 (Fig. 1B, D). Accordingly, E6 binds tightly to the paxillin LD motif, with an equilibrium dissociation constant K_D ~ 50 nM (Ould Sidi M'hamed et al., submitted to publication). Thus, E6 should be able to compete very efficiently with host proteins for binding to paxillin and other proteins displaying suitable LD or LxxLL motifs.

The sequences of E6 proteins of various papillomaviruses are well aligned with that of BPV-1 E6 (Fig. 2A), indicating the conservation of the overall structure composed of two zinc-binding domains connected via a linker helix. Most hydrophobic and polar positions critical for peptide binding are well conserved in nature, yet not always identical. This suggests that HPV E6 proteins will recognise similar acidic / hydrophobic helical motifs, albeit with fine variations in their sequence recognition specificities.

The E6-peptide structure reveals 21 BPV-1 E6 residues (indicated on top of the alignment in Fig. 2A) that are involved in atomic contacts with the peptide motif. 12 single point mutations altering residues of the motif-binding site of E6 were generated, among which 8 mutants were defective for LxxLL motif recognition (Fig. 2B). Remarkably, E6 mutants disrupted for LxxLL-motif binding systematically lost the transformation phenotype in living cells (Fig. 2B). Therefore, an intact LxxLL motif-binding pocket is essential to the tumourigenic phenotype of E6.

Two ion bridges involving basic residues of E6 and acidic residues of the LxxLL motif (R42-D3 and R116-D5) constitute electrostatic clamps (Fig. 1, B and E). Charge swapping mutagenesis experiments targetting these clamps (Fig. 2C, 2D and Supplemental Fig. 6) altered E6 peptide recognition specificities. In particular, E6 R116D became more selective than E6 wt towards a panel of mutants of the LxxLL motif, and displayed a strong preference for an LxxLL motif bearing the inverse charge mutation D5R. This shows that R116 plays a critical role in recognition and indicates that a swapped R5/D116 ion bridge has successfully replaced the original D5/R116 ion bridge in the mutated complex.

To further define the critical determinants of LxxLL-E6 interaction, 12-meric peptides binding selectively to BPV-1 E6 were isolated out of a random peptide library

using phage display (Supplemental Fig. 7). Selected sequences were strikingly similar to the sequences of the E6-binding paxillin repeats LD1, LD2 and LD4, allowing us to propose a more refined 10-residue long regular expression for BPV-1 E6-binding LxxLL sequences: $\Phi_1X_2D_3L_4D_5[-]_6L_7$ $(F/L)_8X_9[-]_{10}$ (Fig. 1F and Supplemental Fig. 7). The net charge of the E6-binding peptides was always negative, in agreement with the structural data, which showed that the E6C domain displayed a strongly positive surface prone to attract negatively charged ligands (Fig. 1 D).

By combining the information gained from structural analysis, phage display and mutagenesis data (*12*), we defined precisely the LxxLL sequence motif recognised by BPV-1 E6 (Material and Methods) and built a position specific scoring matrix (PSSM) profile (*22*) (supplemental table 1) that were used to perform a sequence similarity search on the human proteome. Since short protein interaction motifs are mainly found in non-folded areas (*23*), we restricted our search to regions of the proteome predicted to be non-folded. This procedure allowed us to generate and rank a short list of host cell proteins potentially targeted by BPV-1 E6 (Fig. 3A and supplemental Table 2). Remarkably, 9 human proteins (Fig. 3) predicted by our approach to be highly likely binders of E6, were independently identified by co-precipitation experiments coupled with mass spectroscopy analysis (Fig. 3). Furthermore, all the short acidic LxxLL motifs detected within these proteins by our *in silico* search were found to bind to BPV-1 E6 *in vitro* (Fig. 3). Therefore, a thorough structural analysis can foster accurate proteome-wide bioinformatic predictions of the cellular targets of a viral protein, which may be used to increase the significance of high-throughput experimental interactomics data and/or decrease the number of putative targets deserving to be explored for further biological validation.

The list of potential targets of E6 is highly enriched in proteins, which, like paxillin, participate in cell adhesion and cytoskeletal dynamics and organisation, as well as numerous regulators of transcription, cell proliferation or cell death (Fig. 3, supplemental Table 2, supplemental Table 3). While all sequences in the list present acidic leucine-rich patterns compatible with the E6 peptide binding pocket, only a small number of them display the consensus sequence LDxLLxxL representative of the full LD motif (*11*). Therefore, the LD motif represents only a sub-class of a larger family of acidic leucine-rich motifs potentially targeted by the E6 protein.

Selective interactions between short linear interaction motifs and their cognate target domains are known to mediate protein-protein interaction networks involved in particular cellular functions (*24*), which actually constitute an Achilles' heel for viral attack (*25*). Our data show that E6 has evolved a fold specialised in capturing a family of acidic hydrophobic interaction motifs, including LD motifs, which participate in biological functions related to transformation and immortalisation (Fig. 4). Such a motif hijacking strategy is in principle extremely efficient, because a single binding pocket recognising the key conserved residues of a target motif is able to capture a large number of instances of the motif present among numerous cellular proteins and therefore strongly disrupt the entire functional pathway mediated by the motif. This strategy may be further potentiated when, as shown here for E6, the viral protein pocket binds tighter to the motif than its natural cellular partners. Indeed, it was previously proposed that BPV-1 E6 might transform cells through such competitive interactions at the LD motifs of paxillin (*26*).

In the absence of target peptide, E6 is likely to adopt a different overall structure. The few interactions observed between E6N and E6C in the complex (Supplemental Fig. 8)

should be insufficient to maintain the two domains and the linker helix in their relative positions. In addition, solvent exposure of the large hydrophobic pocket hidden by the peptide (see Fig. 1C) should be energetically unfavourable. Whether the free structure would resemble the model previously proposed for unbound HPV16 E6 (*15*) remains to be investigated. Indeed, the propensity to aggregation of unbound wild-type E6 proteins (*16*) and their strong affinity for target motifs suggest that most E6 molecules preferentially exist as target-bound complexes in infected cells.

The structure of the E6-paxillin complex has shown how PV E6 proteins specifically capture acidic LxxLL motifs to hijack multiple cellular functions. Structure-guided inactivation of this LxxLL motif-binding site efficiently disrupted the transforming properties of E6. This work, together with comparable structural studies of high-risk human PV E6 oncoproteins bound to their cognate target motifs, should help us to design, screen and rationally improve small molecule inhibitors of papillomavirus mediated oncogenesis.

Acknowledgments

This work was supported by institutional support from CNRS, Université de Strasbourg, INSERM, the European Commission SPINE2-Complexes project (contract n° LSHG-CT-2006-031220) and grants from ARC (n° 3171) and ANR (ANR-MIME-2007 EPI-HPV-3D). S.C. was supported by ANR, A.o.M.o.S. by ARC and K.L by Région Alsace and College Doctoral Européen. S.V.P, N.B, and T.A. were supported by NIH grants (CA120352 and CA08093) to S.V.P, and institutional support from the University of Virginia and the Department of Pathology Mass Spectrometry Facility. The authors thank all members of the ESRF-EMBL joint structural biology groups, all members of SOLEIL

for the use of their synchrotron beamline facilities and for help during data collection, all members of the Structural Genomics Platform of IGBMC for setting up automated procedures, members of the IGBMC's common services for assistance, and all members of the Oncoprotein group for helpful discussions and advice.

Author Contributions

S.C., N.B., A.o.M.o.S., K.Z., T.A., K.o.B., I.M., P.P., V.C., C.L. and S.V.P. performed experiments; K.L. performed bioinformatics predictions; J.C. performed structure determination. S.C., K.L., J.C., S.V.P. and G.T. analysed the data. G.T., S.V.P., S.C. and J.C. prepared the manuscript. G.T., S.V.P. and J.C. supervised the work.

corresponding authors: G.T., S.V.P. and J.C.

gilles.trave@unistra.fr vandepol@virginia.edu cava@igbmc.fr

The authors declare no competing financial interests.

References and notes

1. H. U. Bernard *et al*., *Virology* 401, 70 (May 25, 2010).
2. H. zur Hausen, *Semin Cancer Biol* 9, 405 (Dec, 1999).
3. H. Pfister *et al*., *Arch Dermatol Res* 295, 273 (Dec, 2003).
4. L. Nasir, M. S. Campo, *Vet Dermatol* 19, 243 (Oct, 2008).
5. J. M. Huibregtse, M. Scheffner, P. M. Howley, *Mol Cell Biol* 13, 4918 (Aug, 1993).
6. S. S. Tungteakkhun, P. J. Duerksen-Hughes, *Arch Virol*. 153, 397 (2008).
7. L. V. Ronco, A. Y. Karpova, M. Vidal, P. M. Howley, *Genes Dev* 12, 2061 (Jul 1, 1998).
8. X. Be *et al*., *Biochemistry* 40, 1293 (Feb 6, 2001).
9. S. B. Vande Pol, M. C. Brown, C. E. Turner, *Oncogene* 16, 43 (Jan 8, 1998).
10. N. Brimer, C. Lyons, S. B. Vande Pol, *Virology* 358, 303 (Feb 20, 2007).
11. D. A. Tumbarello, M. C. Brown, C. E. Turner, *FEBS Lett* 513, 114 (Feb 20, 2002).
12. J. Bohl, K. Das, B. Dasgupta, S. B. Vande Pol, *Virology* 271, 163 (May 25, 2000).
13. R. Wade, N. Brimer, S. Vande Pol, *J Virol*. 82, 5962 (June, 2008).
14. X. Wang *et al*., *J Biol Chem* 283, 21113 (Jul 25, 2008).
15. Y. Nomine *et al*., *Mol Cell* 21, 665 (Mar 3, 2006).
16. K. Zanier *et al*., *J Mol Biol* 396, 90 (Feb 12, 2010).
17. The 10 first N-terminal and the 7 last C-terminal residues of E6 are not visible in the electron density map. Both regions are distal from the peptide recognition pocket and do not seem essential for target motif recognition (Supplemental Fig. 3).
18. M. K. Hoellerer *et al*., *Structure* 11, 1207 (Oct, 2003).
19. S. Lorenz *et al*., *Structure* 16, 1521 (Oct 8, 2008).
20. R. S. Savkur, T. P. Burris, *J Pept Res* 63, 207 (Mar, 2004).
21. E. Hur *et al*., *PLoS Biol* 2, E274 (Sep, 2004).
22. D. W. Mount, *Sequence and Genome Analysis* Bioinformatics: (Cold Spring Harbor Laboratory Press, New York City, ed. 2nd Edition, 2004), pp.
23. F. Diella *et al*., *Bioinformatics* 25, 1 (Jan 1, 2009).
24. F. Diella *et al*., *Front Biosci* 13, 6580 (2008).

25. N. E. Davey, G. Trave, T. J. Gibson, *Trends Biochem Sci*. 36, 159 (2011).

26. X. Tong, R. Salgia, J. L. Li, J. D. Griffin, P. M. Howley, *J Biol Chem* 272, 33373 (Dec 26, 1997).

27. K. Das, J. Bohl, S. B. Vande Pol, *J Virol* 74, 812 (Jan, 2000).

28. R. Ned, S. Allen, S. Vande Pol, *J Virol* 71, 4866 (Jun, 1997).

29. S. Charbonnier, K. Zanier, M. Masson, G. Trave, *Protein Expr Purif* 50, 89 (Nov, 2006).

30. This work was supported by institutional support from CNRS, Université de Strasbourg, INSERM, the European Commission SPINE2-Complexes project (contract n° LSHG-CT-2006-031220) and grants from ARC (n° 3171) and ANR (ANR-MIME-2007 EPI-HPV-3D). S.C. was supported by ANR, A.o.M.o.S. by ARC and K.L by Région Alsace and College Doctoral Européen. S.V.P, N.B, and T.A. were supported by NIH grants (CA120352 and CA08093) to S.V.P, and institutional support from the University of Virginia and the Department of Pathology Mass Spectrometry Facility. The authors thank all members of the ESRF-EMBL joint structural biology groups, all members of SOLEIL for the use of their synchrotron beamline facilities and for help during data collection, all members of the Structural Genomics Platform of IGBMC for setting up automated procedures, members of the IGBMC's common services for assistance, and all members of the Oncoprotein group for helpful discussions and advice. The refined model and the structure factor amplitudes have been deposited in the Protein Data Bank under the PDB code 3PY7.

Figure legends

Figure 1.

A. Structure of E6 bound to paxillin LD1 motif. Blue: E6N, grey: linker helix, gold: E6C, green: LxxLL peptide corresponding to residues 1-10 of paxillin.

B. Charged, polar and water-mediated interactions between E6 and peptide. Note the electrostatic charge clamps R42 - D3 and R116 - D5.

C. The hydrophobic pocket (coloured pink) responsible for LxxLL motif recognition.

D. E6 surface charge potential (spectrum range –12 kT/e to +12 kT/e). Blue: positive charge, red: negative charge. Acidic residues of the helical peptide are coloured red.

E. Key contacts between E6 residues and bound peptide. Magenta: hydrophobic residues; blue: basic residues; pink dashed lines: hydrophobic van der Waals contacts; black lines: polar contacts involving the side chain (plain black lines) or the main chain (interrupted black lines) of an E6 residue.

F. The consensus motif recognised by BPV-1 E6, derived from aligned sequences of paxillin motifs LD1, LD2 and LD4 and of E6-binding phage peptides (Supplemental Fig. 5).

G. Whereas LxxLL motifs bind tightly to viral E6, they bind superficially to cellular domains. From left to right, the structures of LxxLL motifs bound to a FAT domain (PDB: 1OW7), a CH domain (PDB: 2VZI) and a Nuclear Receptor LBD (PDB: 3ERD).

Figure 2.

A. Alignment of BPV-1 E6 with a subset of BPV and HPV sequences. Residues whose solvent accessibility is higher when calculated in absence (black bars) than in presence (white bars) of the bound peptide are labelled as peptide-binding positions. Residues whose mutagenesis disrupted transforming activity (see panel B) are boxed.

B. Upper part: yeast two-hybrid interactions (dark spots) between peptide binding site mutants of E6 and full-length paxillin (PXN) or paxillin LD1 motif (M_1DDLDALLADL$_{11}$). Lower left part: transformation phenotype of E6 mutants, quantified using numbers of anchorage independent colonies (*9*) normalised to results from E6 wt. '*' indicate previously published results (*27*, *28*, *9*). Boxed labels indicate E6 mutants significantly impaired for transformation ($P<0.05$). Right part: instances of E6 constructs proficient (E6-wt and E6-V40A) or impaired (vector and E6-R89L) for transformation.

C. Mutagenic analysis of the charge clamps R42 - D3 and R116 - D5 (see Fig. 1 B). Interactions between charge inversion mutants of E6 (R42D and R116D) and of the peptide (D3R, D5R and D3R-D5R) were analysed by yeast two-hybrid.

D. The charge clamp mutants were also assayed by GST-pull-down (Supplementary Fig. 6). Binding intensities were quantified by autoradiograph scanning, normalised to 100% for the strongest signal, and plotted as a heat map.

Figure 3.

A. Proteome-wide prediction of putative cellular BPV-1 E6 binding motifs, ranked according to computed likelihood (see supplemental Table 2 for the full list, containing 387 instances). "yes" denotes coincidence of a predicted motif with a LD motif. 'x' denotes occurence of a keyword highly enriched in annotation terms of the predicted binders of E6

(supplemental Table 3). Coloured protein names denote BPV-1 E6 binders that we independently identified using TAP-MS experiments (see panel B). Red and blue names denote the best scored and the lower scored motifs found in each E6-coprecipitated protein, respectively. Six among the eight best ranked predicted motifs belong to experimentally determined E6 binders. The rectangular distribution plot above the list shows that most other experimentally validated predictions are also well ranked.

B. Representative SDS-PAGE of potential cellular binders of FLAG-E6 (lane 1) or FLAG-LxxLL-E6 (lane 2). Note that internally fused LxxLL motif inhibits recruitment of most E6 partners. MS-identified proteins interacting with E6 and not with LxxLL-E6 are labeled left of lane 1.

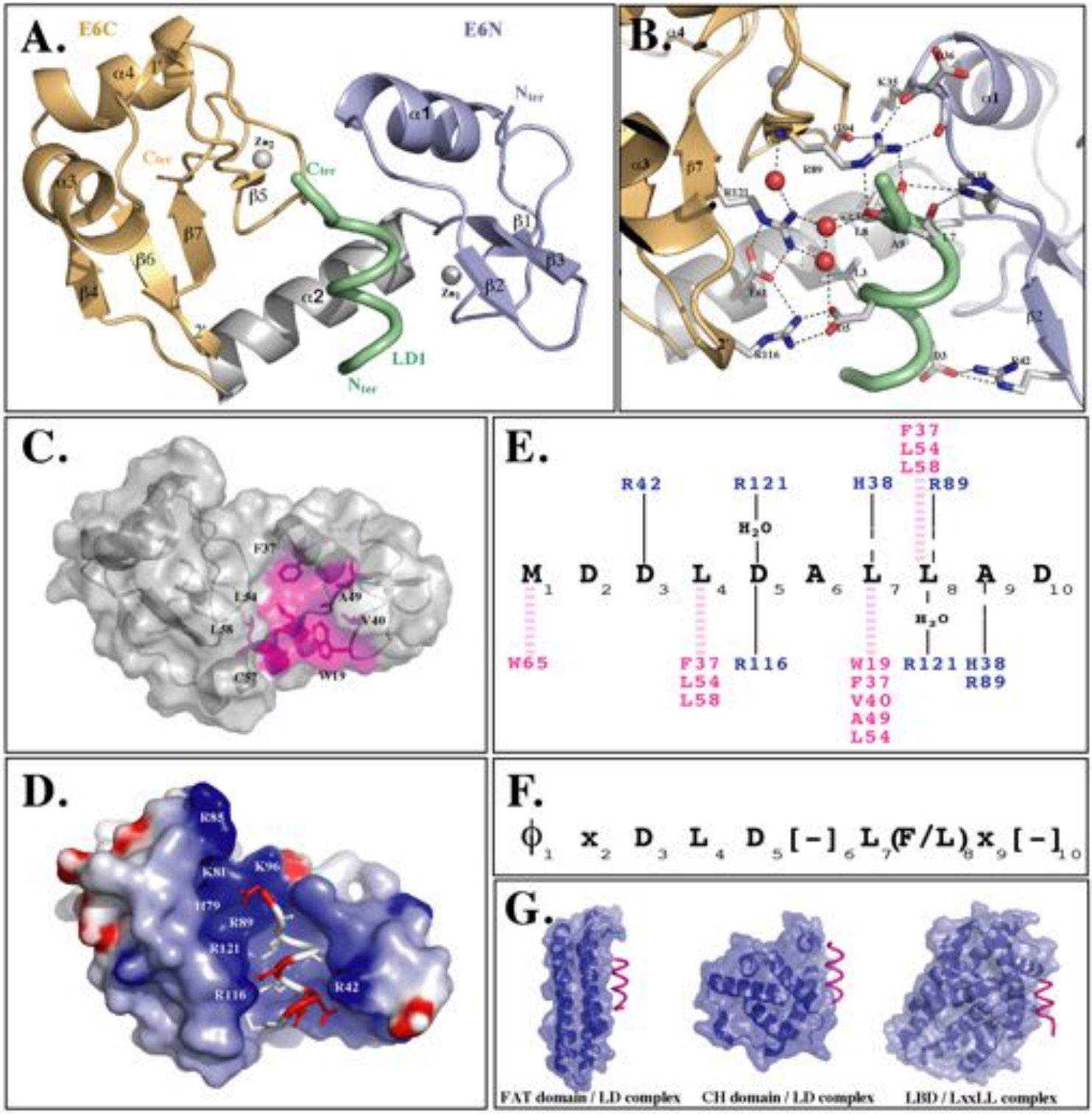
C. All predicted LxxLL motifs belonging to proteins identified by TAP-MS bind significantly to MBP-fused E6 as compared to the two negative control peptides, CTLR-1 and CTLR-2. We performed "holdup" assays (*29*) in which the amount of MBP-E6 bound to each peptide is proportional to the difference in intensity between flow-throughs from biotin-saturated resin (left band, white bar) and peptide-saturated resin (right band, and grey bar). Intensities of peptide flow-throughs (normalized against biotine flow-throughs) were obtained from densitometric analysis of three independent experiments.

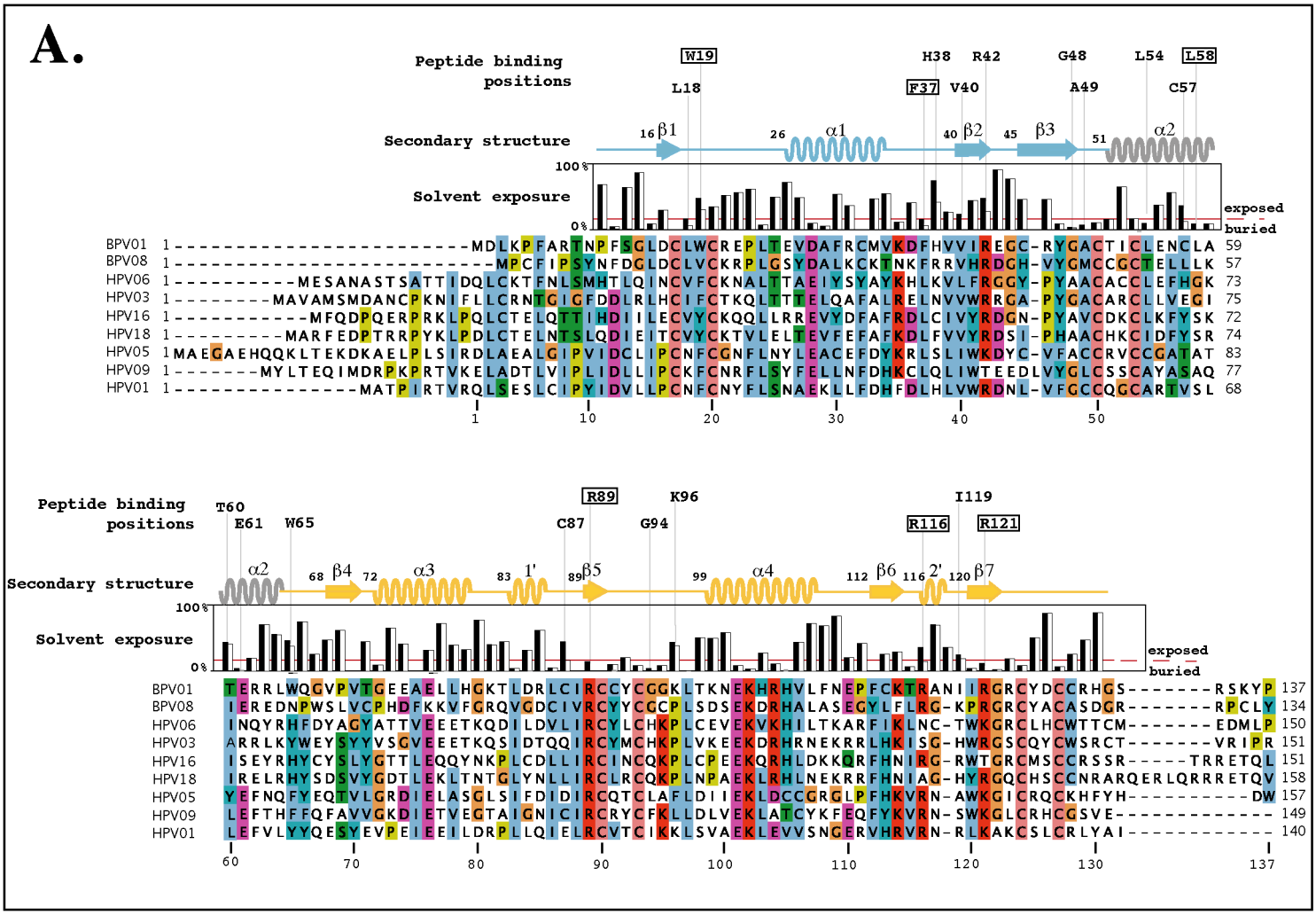
Figure 4. E6-LxxLL motif structure reveals an highly efficient strategy for viral hijacking of cellular functions.

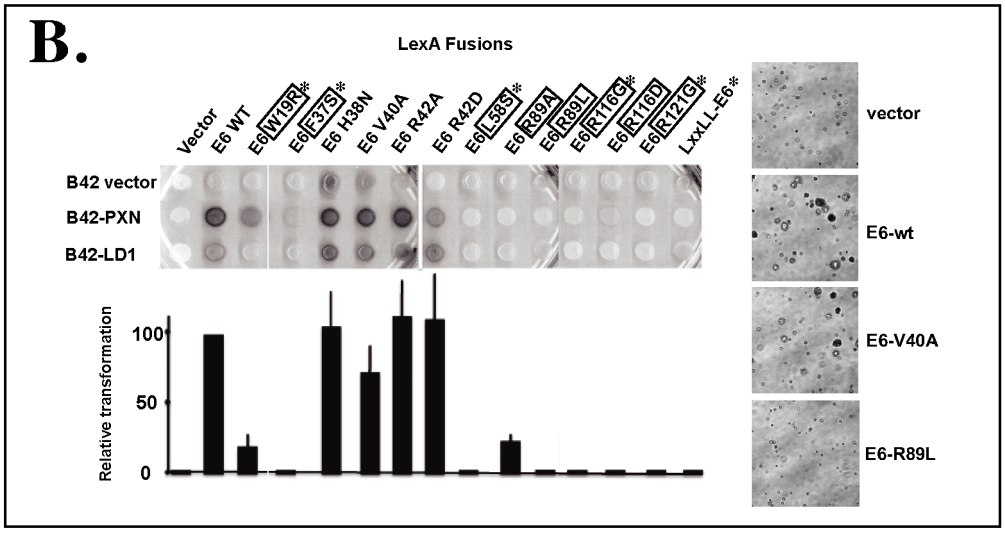
A. Within the host cells, numerous functions rely on protein-protein interaction networks mediated by small sequence motifs preferentially located in natively unfolded regions of proteins, which bind selectively yet with moderate affinity to folded domains.

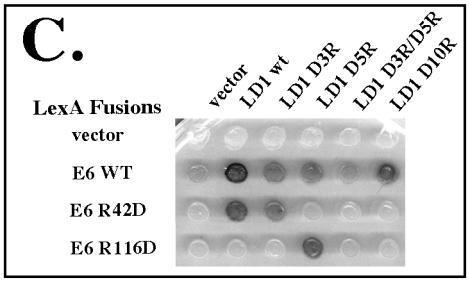
The acidic LxxLL sequences recognized by E6 appears to represent a family of sequence motifs, including LD motifs, that are involved in the regulation of various key functions generally preventing transformation and immortalisation.

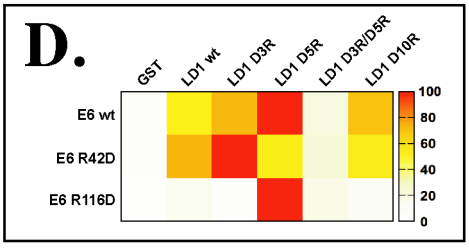
B. Within a papillomavirus-infected cell, the E6 oncoproteins capture with high efficiency, as demonstrated by the structural data, numerous members of the LxxLL motif family, thereby perturbing the biological processes mediated by these motifs. The perturbation may result from a simple competition mechanism (as shown on the figure) that prevents the cellular domains from interacting with their target LxxLL motifs. More sophisticated mechanisms may involve alteration of the activity of the LxxLL motif-containg protein, as well as formation of multiple complexes involving E6, the LxxLL motif-containing protein and other cellular proteins. These cumulated perturbations release the control upon transformation and immortalisation, turning the cell into a proliferative and immortalised state proficient for papillomavirus replication, which may eventually lead to oncogenesis.


A.
E6C
E6N
α4
β4
Cter
Zn2
β5
Cter
α1
Nter
α3
β7
β6
β1
β3
α2
Zn1
β2
LD1
Nter
B.
α4
α1
α3
β7
β2
C.
F37
L54
L58
A49
V40
C57
W19
D.
R85
K81
K96
I179
R89
R121
R116
R42
E.
F37
L54
L58
R42
R121
H38
R89
H2O
M1 D2 D3 L4 D5 A6 L7 L8 A9 D10
H2O
W65
F37
L54
L58
R116
W19
F37
V40
A49
L54
R121
H38
R89
F.
ϕ1 x2 D3 L4 D5 [-]6 L7 (F/L)8 x9 [-]10
G.
FAT domain / LD complex
CH domain / LD complex
LBD / LxxLL complex

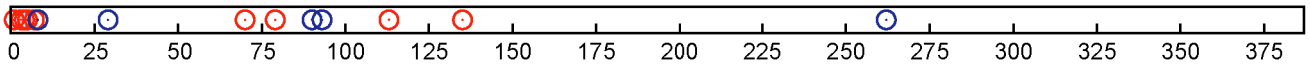

A.
Peptide binding positions
W19
L18
H38
R42
F37
V40
G48
A49
L54
L58
C57
Secondary structure
β1
α1
β2
β3
α2
Solvent exposure
exposed
buried
BPV01
BPV08
HPV06
HPV03
HPV16
HPV18
HPV05
HPV09
HPV01
Peptide binding positions
T60
E61
W65
R89
C87
K96
G94
I119
R116
R121
Secondary structure
α2
β4
α3
β5
α4
β6
β7
Solvent exposure
exposed
buried
BPV01
BPV08
HPV06
HPV03
HPV16
HPV18
HPV05
HPV09
HPV01


B.
LexA Fusions
Vector
E6 WT
E6 W19R*
E6 F37S*
E6 H38N
E6 V40A
E6 R42A
E6 R42D
E6 L56S*
E6 R89A
E6 R89L
E6 R116G*
E6 R116D
E6 R121G*
LxxLL-E6*
B42 vector
B42-PXN
B42-LD1
Relative transformation
vector
E6-wt
E6-V40A
E6-R89L


C.
LexA Fusions
vector
E6 WT
E6 R42D
E6 R116D
vector
LD1 wt
LD1 D3R
LD1 D5R
LD1 D3R/D5R
LD1 D10R

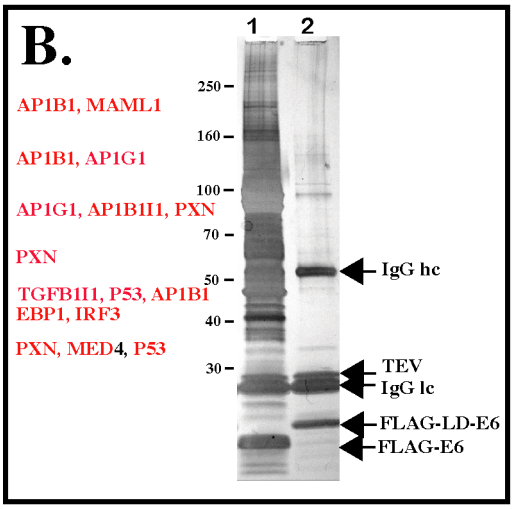

D.
GST
LD1 wt
LD1 D3R
LD1 D5R
LD1 D3R/D5R
LD1 D10R
E6 wt
E6 R42D
E6 R116D

A. Distribution of hits from proteomic screen in list of predicted binders

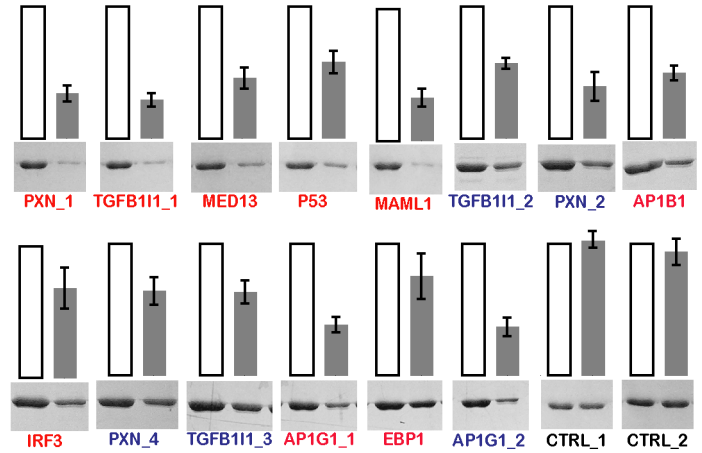


n° prediction (out of 387)	target protein	motif-containing sequence	localization in sequence	net charge	LD-motif	regulation of transcription	Epithelium	cytoskeleton
1	paxillin (PXN_1)	MDDLDALLAD LESTT	1-11	-4	yes		x	x
2	ninein	QKRLS WDKLDHLMNE EQQLL	1838-1848	-1	no			x
3	TGFβ1 induced transcript 1 (TGFB1I1_1)	MEDLDALLSD LETTT	1-11	-4	yes	x		x
4	mediator complex subunit 13 (MED13)	DLAVS YTDLDNLFNS DEDEL	778-788	-2	no	x	x	
5	tumor protein p53 (P53)	LSQET FSDLWKLLPE NNVLS	19-29	-1	no	x		
7	mastermind-like 1 (MAML1)	TSEEW MSDLDDLLGS Q	1006-1016	-3	no	x		
8	TGFβ1 induced transcript 1 (TGFB1I1_2)	VLGTG LCELDRLLQE LNATQ	73-83	-2	yes	x		x
9	amyloid beta (A4) precursor protein-binding. family B. member 1 interacting protein	LNALE DQDLDALMAD LVADI	61-71	-4	yes			x
14	meningioma 1	KSAMS TIDLDSLMAE HSAAW	1213-1223	-3	no		x	
19	Decidual protein induced by progesterone	PMADT VDPLDWLFGE SQEKQ	102-112	-3	no			
20	golgi-associated gamma adaptin ear containing ARF binding protein 3	SALHH LDALDQLLEE AKVTS	516-526	-4	no			
29	paxillin (PXN_2)	SLGSN LSELDRLLLE LNAVQ	142-152	-2	yes		x	x
43	kinesin light chain 2	KGDVP KDTLDDLFPN EDEQS	161-171	-2	no		x	x
44	spindlin family. member 2B	ITQWK GTVLDQLLDD YKEGD	70-80	-3	no			
58	consortin. connexin sorting protein	CGNNQ ISDLGILLPE VCMAP	516-526	-2	no			
64	protocadherin 10	HSTLE RKELDGLLTN TRAPY	1015-1025	0	no			
70	adaptor-related protein complex 1. beta 1 subunit (AP1B1)	AVDLL GGGLDSLMGD EPEGI	660-670	-2	no			
79	interferon regulatory factor 3 (IRF3)	TSDTQ EDILDELLGN MVLAP	137-147	-4	no	x		
90	paxillin (PXN_4)	SASSA TRELDELMAS LSDFK	263-273	-2	yes		x	x
93	TGFβ1 induced transcript 1 (TGFB1I1_3)	SATSA TLELDRLMAS LSDFR	136-146	-1	yes	x		x
95	clathrin interactor 1	PAASN SSDLFDLMGS SQATM	402-412	-2	no		x	
97	eukaryotic translation initiation factor 4E nuclear import factor 1	QKA KVDLKPLLSS LSANK	4-14	1	no			
105	centrobin. centrosomal BRCA2 interacting protein	QQVAE DYELRLLLLD PPAPG	520-530	-2	no			x
113	adaptor-related protein complex 1. gamma 1 subunit (AP1G1_1)	KPSSA GGELLDLLGD INLTG	651-661	-3	no			
135	proliferation-associated 2G4 (EBP1)	EMEVQ DAELKALLQS SASRK	349-359	-1	no	x	x	
262	adaptor-related protein complex 1. gamma 1 subunit (AP1G1_2)	QPTSQ ANDLLDLLGG NDITP	626-636	-2	no			

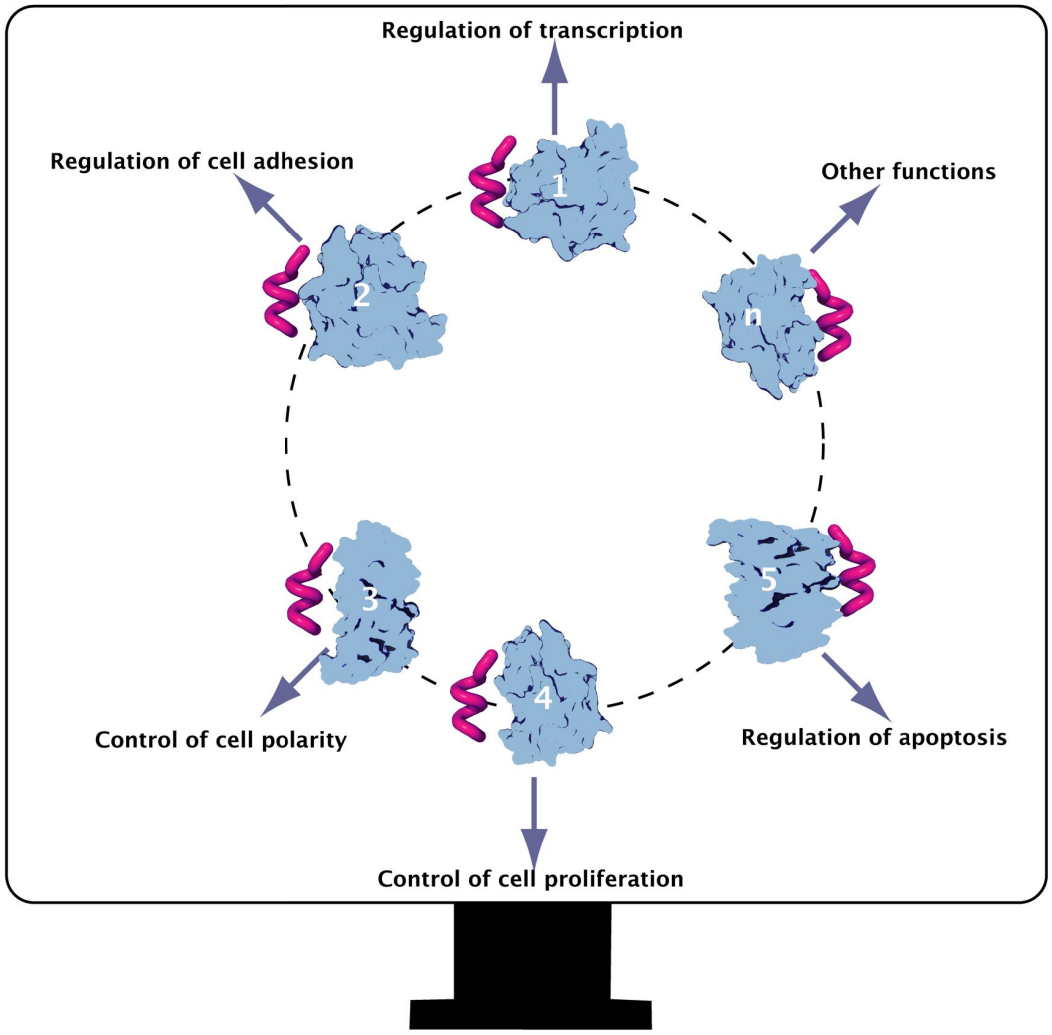
B.



C.

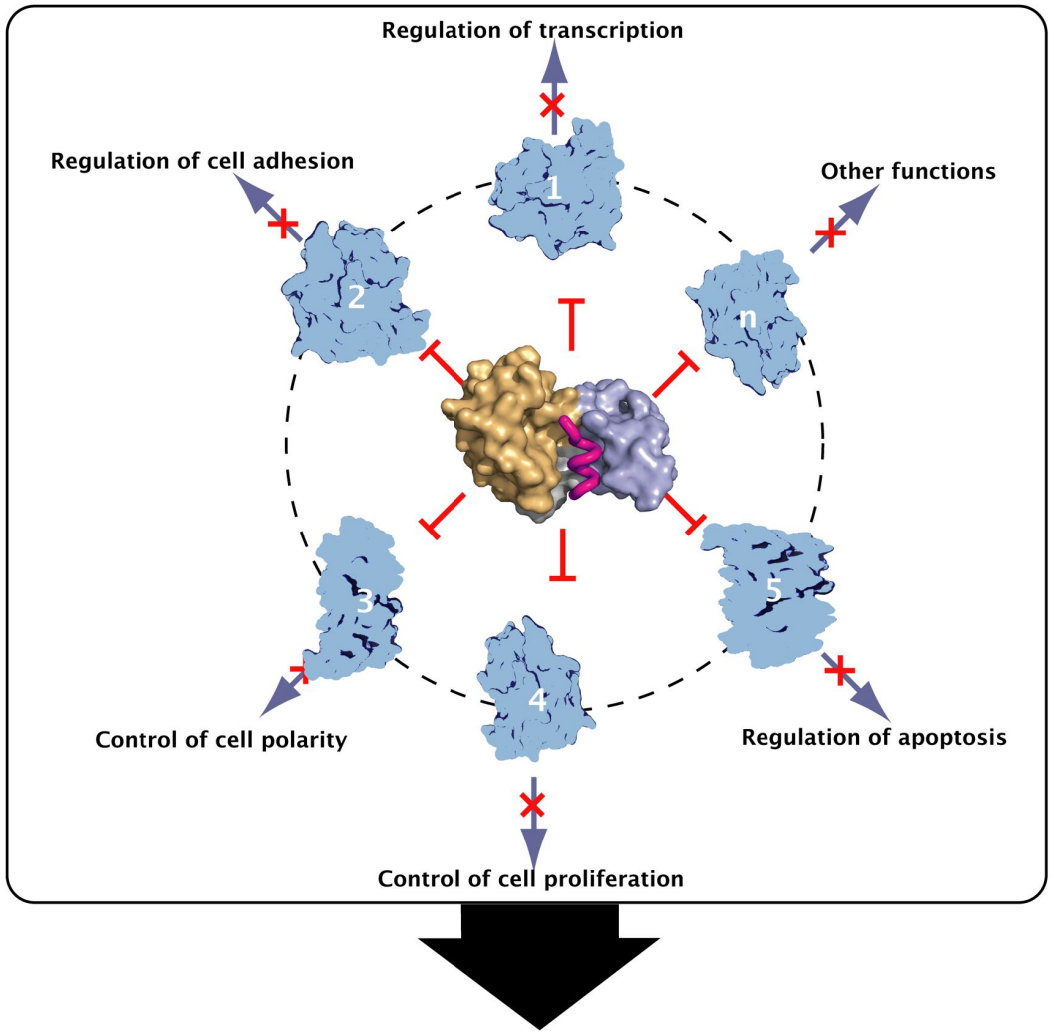


"Normal physiological state"



Transformation and Immortalization : OFF

"Papillomavirus infected state"



Transformation and Immortalization : ON

Supplemental Table 1.

Position	1	2	3	4	5	6	7	8	9	10
Residue										
A	-2,4880	-0,4880	-1,4880	-35,7073	-1,4880	-0,4880	-35,7073	-35,7073	-0,4880	-35,7073
C	-0,8969	1,1031	0,1031	-34,1162	0,1031	1,1031	-34,1162	-34,1162	1,1031	-34,1162
D	-1,9097	0,0903	3,4122	-35,1290	3,5497	0,0903	-35,1290	-35,1290	0,0903	3,0903
E	-2,4926	-0,4926	-0,4926	-35,7118	-35,7118	-0,4926	-35,7118	-35,7118	-0,4926	2,5074
F	2,3550	0,4482	-0,5518	-34,7711	-0,5518	0,4482	-34,7711	3,6181	0,4482	-34,7711
G	-2,4070	-0,4070	-1,4070	-35,6262	-1,4070	-0,4070	-35,6262	-35,6262	-0,4070	-35,6262
H	-1,0756	0,9244	-0,0756	-34,2949	-0,0756	0,9244	-34,2949	-34,2949	0,9244	-34,2949
I	-1,7908	0,2092	-0,7908	-35,0101	-0,7908	0,2092	-35,0101	-35,0101	0,2092	-35,0101
K	-2,1811	-0,1811	-1,1811	-35,4004	-1,1811	-0,1811	-35,4004	-35,4004	-0,1811	-35,4004
L	0,9051	-1,0018	-2,0018	3,3201	-2,0018	-1,0018	3,3201	2,1681	-1,0018	-36,2211
M	1,5386	1,2166	0,2166	-34,0026	0,2166	1,2166	-34,0026	2,2166	1,2166	-34,0026
N	-1,5089	0,4911	-0,5089	-34,7282	-0,5089	0,4911	-34,7282	-34,7282	0,4911	1,4911
P	1,5657	-0,3412	-1,3412	-35,5605	-1,3412	-0,3412	-35,5605	-35,5605	-0,3412	-35,5605
Q	-1,9277	0,0723	-0,9277	-35,1470	-0,9277	0,0723	-35,1470	-35,1470	0,0723	-35,1470
R	-2,1950	-0,1950	-1,1950	-35,4143	-1,1950	-0,1950	-35,4143	-35,4143	-0,1950	-35,4143
S	-2,7298	-0,7298	-1,7298	-35,9491	-1,7298	-0,7298	-35,9491	-35,9491	-0,7298	0,2702
T	-2,0825	-0,0825	-1,0825	-35,3017	-1,0825	-0,0825	-35,3017	-35,3017	-0,0825	-35,3017
V	-2,2600	-0,2600	-1,2600	-35,4793	-1,2600	-0,2600	-35,4793	-35,4793	-0,2600	-35,4793
W	3,9009	1,9940	0,9940	-33,2252	0,9940	1,9940	-33,2252	-33,2252	1,9940	-33,2252
Y	-1,0794	0,9206	-0,0794	-34,2986	-0,0794	0,9206	-34,2986	-34,2986	0,9206	-34,2986

Position Specific Scoring Matrix (PSSM) defining binding preferences of BPV-1 E6 for LxxLL motifs. Position Specific Scoring Matrix (PSSM) used for scoring a selection of identified BPV-1 E6 binding LxxLL motifs out of non-folded regions of proteins from the human proteome. The PSSM contains values for the likeliness of a residue to be part of the motif for each amino acid at each position of the motif. These values were determined by taking the logarithm of the frequency of occurrence of an amino acid at a particular position divided by the background frequency of this amino acid in the proteome. Frequencies of amino acids at motif positions were based on mutagenesis experiments, phage display data and observations based on the BPV-1 E6-LxxLL structure presented in this manuscript (also refer to materials and methods section). Amino acids that do not appear at specific motif positions were given a very small pseudo count of 10^{-10}.

Supplemental Table 3.: Gene onthology terms-based enrichment (N.B.: Figure legends provided at the bottom)

nr	Category	Term	Count	%	p-value
1	SP_PIR_KEYWORDS	coiled coil	101	32	4,50E-26
2	SP_PIR_KEYWORDS	phosphoprotein	200	63,3	7,90E-22
3	GOTERM_BP_FAT	transcription	76	24,1	5,10E-13
4	UP_SEQ_FEATURE	compositionally biased region:Pro-rich	50	15,8	7,50E-13
5	GOTERM_BP_FAT	regulation of transcription	86	27,2	8,90E-13
6	UP_TISSUE	Epithelium	90	28,5	3,40E-12
7	GOTERM_CC_FAT	intracellular non-membrane-bounded organelle	82	25,9	4,50E-12
8	GOTERM_CC_FAT	non-membrane-bounded organelle	82	25,9	4,50E-12
9	SP_PIR_KEYWORDS	transcription regulation	75	23,7	7,90E-12
10	SP_PIR_KEYWORDS	Transcription	75	23,7	2,30E-11
11	GOTERM_CC_FAT	cytoskeleton	54	17,1	7,60E-11
12	SP_PIR_KEYWORDS	activator	33	10,4	8,60E-11
13	SP_PIR_KEYWORDS	nucleus	119	37,7	1,40E-10
14	GOTERM_CC_FAT	nuclear lumen	55	17,4	1,50E-10
15	SP_PIR_KEYWORDS	alternative splicing	174	55,1	7,30E-10
16	SP_PIR_KEYWORDS	cytoplasm	97	30,7	1,90E-09
17	GOTERM_MF_FAT	transcription regulator activity	58	18,4	3,10E-09
18	UP_SEQ_FEATURE	splice variant	172	54,4	3,60E-09
19	GOTERM_CC_FAT	microtubule cytoskeleton	29	9,2	1,50E-08
20	GOTERM_CC_FAT	nucleoplasm	37	11,7	3,50E-08
21	GOTERM_MF_FAT	transcription factor binding	29	9,2	4,40E-08
22	GOTERM_MF_FAT	transcription cofactor activity	24	7,6	5,40E-08
23	GOTERM_MF_FAT	transcription activator activity	25	7,9	1,20E-07
24	GOTERM_CC_FAT	organelle lumen	56	17,7	1,80E-07
25	GOTERM_CC_FAT	intracellular organelle lumen	55	17,4	2,10E-07
26	UP_SEQ_FEATURE	compositionally biased region:Poly-Ser	26	8,2	2,50E-07
27	GOTERM_CC_FAT	nucleoplasm part	27	8,5	2,70E-07
28	GOTERM_CC_FAT	membrane-enclosed lumen	56	17,7	3,40E-07
29	UP_SEQ_FEATURE	compositionally biased region:Ser-rich	24	7,6	4,80E-07
30	SP_PIR_KEYWORDS	cytoskeleton	30	9,5	4,90E-07
31	GOTERM_BP_FAT	regulation of RNA metabolic process	56	17,7	5,70E-07

32	GOTERM_CC_FAT	cytoskeletal part	36	11,4	7,00E-07
33	UP_SEQ_FEATURE	compositionally biased region:Glu-rich	19	6	8,90E-07
34	SP_PIR_KEYWORDS	ubl conjugation	28	8,9	1,10E-06
35	GOTERM_BP_FAT	positive regulation of transcription	26	8,2	2,40E-06
36	GOTERM_BP_FAT	regulation of transcription, DNA-dependent	53	16,8	3,60E-06
37	UP_SEQ_FEATURE	sequence variant	232	73,4	4,10E-06
38	GOTERM_BP_FAT	positive regulation of gene expression	26	8,2	4,10E-06
39	GOTERM_BP_FAT	positive regulation of transcription, DNA-dependent	23	7,3	5,50E-06
40	GOTERM_BP_FAT	positive regulation of RNA metabolic process	23	7,3	6,30E-06
41	UP_SEQ_FEATURE	compositionally biased region:Gln-rich	13	4,1	6,50E-06
42	GOTERM_BP_FAT	positive regulation of nitrogen compound metabolic process	27	8,5	8,30E-06
43	GOTERM_BP_FAT	positive regulation of cellular biosynthetic process	28	8,9	8,40E-06
44	SP_PIR_KEYWORDS	polymorphism	223	70,6	8,60E-06
45	GOTERM_BP_FAT	positive regulation of macromolecule biosynthetic process	27	8,5	1,10E-05
46	GOTERM_BP_FAT	positive regulation of biosynthetic process	28	8,9	1,10E-05
47	GOTERM_BP_FAT	positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	26	8,2	1,40E-05
48	GOTERM_MF_FAT	transcription coactivator activity	15	4,7	1,70E-05
49	GOTERM_CC_FAT	nuclear speck	10	3,2	2,50E-05
50	GOTERM_CC_FAT	nuclear body	12	3,8	4,80E-05
51	SP_PIR_KEYWORDS	isopeptide bond	17	5,4	6,30E-05
52	GOTERM_BP_FAT	in utero embryonic development	12	3,8	9,90E-05
53	GOTERM_MF_FAT	transcription factor activity	34	10,8	1,00E-04
54	GOTERM_MF_FAT	DNA binding	63	19,9	1,10E-04
55	GOTERM_CC_FAT	kinesin complex	5	1,6	1,30E-04
56	GOTERM_MF_FAT	cytoskeletal protein binding	22	7	1,30E-04
57	UP_SEQ_FEATURE	compositionally biased region:Poly-Pro	19	6	1,50E-04
58	GOTERM_MF_FAT	motor activity	11	3,5	1,50E-04
59	GOTERM_BP_FAT	positive regulation of macromolecule metabolic process	29	9,2	1,60E-04
60	SP_PIR_KEYWORDS	chromosomal rearrangement	15	4,7	1,80E-04
61	GOTERM_BP_FAT	positive regulation of transcription from RNA polymerase II promoter	17	5,4	2,30E-04
62	INTERPRO	Clathrin adaptor, alpha/beta/gamma-adaptin, appendage, Ig-like subdomain	4	1,3	2,70E-04
63	UP_SEQ_FEATURE	short sequence motif:Nuclear localization signal	16	5,1	2,80E-04
64	UP_TISSUE	T-cell	16	5,1	2,80E-04

65	SP_PIR_KEYWORDS	motor protein	10	3,2	2,90E-04
66	SP_PIR_KEYWORDS	microtubule	13	4,1	3,80E-04
67	UP_TISSUE	Cervix carcinoma	17	5,4	4,10E-04
68	GOTERM_BP_FAT	regulation of transcription from RNA polymerase II promoter	25	7,9	4,20E-04
69	GOTERM_MF_FAT	actin binding	16	5,1	4,50E-04
70	UP_SEQ_FEATURE	compositionally biased region:Poly-Glu	20	6,3	4,60E-04
71	UP_SEQ_FEATURE	short sequence motif:Nuclear export signal	6	1,9	5,60E-04
72	UP_SEQ_FEATURE	compositionally biased region:Poly-Arg	11	3,5	5,80E-04
73	SP_PIR_KEYWORDS	transcription factor	7	2,2	6,10E-04
74	GOTERM_BP_FAT	microtubule-based process	13	4,1	6,20E-04
75	INTERPRO	Leupaxin	3	0,9	6,80E-04
76	INTERPRO	Kinesin light chain repeat	3	0,9	6,80E-04
77	INTERPRO	Neurogenic mastermind-like, N-terminal	3	0,9	6,80E-04
78	GOTERM_BP_FAT	chordate embryonic development	15	4,7	7,00E-04
79	GOTERM_BP_FAT	embryonic development ending in birth or egg hatching	15	4,7	7,70E-04
80	UP_SEQ_FEATURE	short sequence motif:LD motif 1	3	0,9	7,70E-04
81	UP_SEQ_FEATURE	short sequence motif:LD motif 2	3	0,9	7,70E-04
82	UP_SEQ_FEATURE	short sequence motif:LD motif 3	3	0,9	7,70E-04
83	GOTERM_CC_FAT	microtubule associated complex	8	2,5	8,80E-04
84	UP_TISSUE	Testis	95	30,1	9,00E-04
85	GOTERM_CC_FAT	microtubule	13	4,1	9,40E-04
86	UP_SEQ_FEATURE	repeat:ANK 5	9	2,8	1,30E-03
87	GOTERM_CC_FAT	transcription factor complex	11	3,5	1,30E-03
88	GOTERM_MF_FAT	enzyme binding	20	6,3	1,50E-03
89	UP_SEQ_FEATURE	compositionally biased region:Arg-rich	9	2,8	1,60E-03
90	GOTERM_CC_FAT	microtubule organizing center	12	3,8	1,60E-03
91	INTERPRO	Zinc finger, PHD-finger	7	2,2	1,90E-03
92	UP_SEQ_FEATURE	mutagenesis site	51	16,1	1,90E-03
93	GOTERM_MF_FAT	microtubule motor activity	7	2,2	1,90E-03
94	GOTERM_CC_FAT	nucleolus	22	7	2,00E-03
95	GOTERM_CC_FAT	centrosome	11	3,5	2,10E-03
96	UP_TISSUE	Foreskin	6	1,9	2,20E-03
97	INTERPRO	Clathrin adaptor, gamma-adaptin, appendage	3	0,9	2,20E-03
98	INTERPRO	Kinesin light chain	3	0,9	2,20E-03

99	SP_PIR_KEYWORDS	actin-binding	12	3,8	2,30E-03
100	SP_PIR_KEYWORDS	dna-binding	47	14,9	2,30E-03
101	INTERPRO	Pleckstrin homology-type	13	4,1	2,40E-03
102	INTERPRO	Zinc finger, PHD-type	7	2,2	2,50E-03
103	UP_SEQ_FEATURE	domain:GAE	3	0,9	2,50E-03
104	INTERPRO	Phosphotyrosine interaction region	5	1,6	2,60E-03
105	GOTERM_CC_FAT	clathrin coat	5	1,6	2,70E-03
106	GOTERM_CC_FAT	cytosol	34	10,8	2,80E-03
107	UP_SEQ_FEATURE	cross-link:Glycyl lysine isopeptide (Lys-Gly) (interchain with G-Cter in SUMO)	7	2,2	3,20E-03
108	INTERPRO	Rab11-binding domain , FIP domain, C-terminal	3	0,9	3,30E-03
109	INTERPRO	Rabaptin, GTPase-Rab5 binding	3	0,9	3,30E-03
110	UP_SEQ_FEATURE	compositionally biased region:Lys-rich	8	2,5	3,40E-03
111	GOTERM_BP_FAT	cytoskeleton organization	16	5,1	3,50E-03
112	GOTERM_MF_FAT	protein transporter activity	7	2,2	3,60E-03
113	INTERPRO	Ubiquitin-associated/translation elongation factor EF1B, N-terminal, eukaryote	5	1,6	3,70E-03
114	UP_SEQ_FEATURE	domain:RBD-FIP	3	0,9	3,70E-03
115	UP_SEQ_FEATURE	domain:UBA	5	1,6	3,90E-03
116	UP_TISSUE	Brain	154	48,7	3,90E-03
117	UP_SEQ_FEATURE	repeat:ANK 4	9	2,8	4,20E-03
118	UP_SEQ_FEATURE	DNA-binding region:Basic motif	9	2,8	4,70E-03
119	INTERPRO	DNA-binding SAP	4	1,3	4,90E-03
120	UP_SEQ_FEATURE	cross-link:Glycyl lysine isopeptide (Lys-Gly) (interchain with G-Cter in ubiquitin)	10	3,2	5,20E-03
121	INTERPRO	ENTH/VHS	4	1,3	5,50E-03
122	GOTERM_BP_FAT	epithelial cell differentiation	8	2,5	5,70E-03
123	UP_TISSUE	Eye	29	9,2	5,70E-03
124	UP_SEQ_FEATURE	domain:SAP	4	1,3	5,80E-03
125	INTERPRO	Tensin phosphotyrosine-binding domain	3	0,9	6,10E-03
126	INTERPRO	Phosphatase tensin type	3	0,9	6,10E-03
127	INTERPRO	Tensin phosphatase, C2 domain	3	0,9	6,10E-03
128	GOTERM_BP_FAT	transcription, DNA-dependent	12	3,8	6,20E-03
129	SP_PIR_KEYWORDS	host-virus interaction	12	3,8	6,60E-03
130	GOTERM_BP_FAT	RNA biosynthetic process	12	3,8	6,80E-03
131	UP_SEQ_FEATURE	domain:C2 tensin-type	3	0,9	6,80E-03
132	GOTERM_MF_FAT	small GTPase binding	7	2,2	7,40E-03

133	GOTERM_CC_FAT	actin cytoskeleton	11	3,5	7,80E-03
134	INTERPRO	Basic helix-loop-helix dimerisation region bHLH	7	2,2	7,90E-03
135	GOTERM_BP_FAT	microtubule-based movement	7	2,2	8,50E-03
136	UP_SEQ_FEATURE	domain:Phosphatase tensin-type	3	0,9	8,70E-03
137	GOTERM_BP_FAT	establishment of protein localization	22	7	8,90E-03
138	GOTERM_BP_FAT	chromosome organization	16	5,1	9,10E-03
139	GOTERM_CC_FAT	adherens junction	8	2,5	9,10E-03
140	GOTERM_CC_FAT	cell projection	20	6,3	9,20E-03
141	GOTERM_CC_FAT	clathrin adaptor complex	4	1,3	9,30E-03
142	SP_PIR_KEYWORDS	DNA binding	13	4,1	9,40E-03
143	GOTERM_CC_FAT	microtubule organizing center part	5	1,6	1,00E-02
144	GOTERM_CC_FAT	AP-type membrane coat adaptor complex	4	1,3	1,00E-02
145	GOTERM_MF_FAT	GTPase binding	7	2,2	1,10E-02
146	UP_SEQ_FEATURE	domain:Helix-loop-helix motif	7	2,2	1,10E-02
147	INTERPRO	Ankyrin	10	3,2	1,10E-02
148	SP_PIR_KEYWORDS	zinc-finger	41	13	1,10E-02
149	GOTERM_BP_FAT	transcription from RNA polymerase II promoter	10	3,2	1,10E-02
150	UP_SEQ_FEATURE	compositionally biased region:Poly-Gln	8	2,5	1,20E-02
151	SP_PIR_KEYWORDS	protein transport	16	5,1	1,20E-02
152	UP_SEQ_FEATURE	zinc finger region:C3H1-type 1	4	1,3	1,20E-02
153	UP_SEQ_FEATURE	zinc finger region:C3H1-type 2	4	1,3	1,20E-02
154	UP_SEQ_FEATURE	zinc finger region:PHD-type 2	4	1,3	1,20E-02
155	UP_SEQ_FEATURE	domain:LIM zinc-binding 4	3	0,9	1,30E-02
156	UP_SEQ_FEATURE	repeat:ANK 1	10	3,2	1,30E-02
157	UP_SEQ_FEATURE	repeat:ANK 2	10	3,2	1,30E-02
158	GOTERM_CC_FAT	cilium	7	2,2	1,30E-02
159	UP_SEQ_FEATURE	repeat:ANK 3	9	2,8	1,40E-02
160	GOTERM_BP_FAT	cell morphogenesis involved in differentiation	10	3,2	1,40E-02
161	UP_SEQ_FEATURE	repeat:ANK 7	5	1,6	1,40E-02
162	UP_TISSUE	Hepatoma	10	3,2	1,40E-02
163	GOTERM_BP_FAT	intracellular transport	19	6	1,50E-02
164	UP_SEQ_FEATURE	compositionally biased region:Asp-rich	4	1,3	1,50E-02
165	GOTERM_CC_FAT	coated membrane	5	1,6	1,50E-02
166	GOTERM_CC_FAT	membrane coat	5	1,6	1,50E-02

167	GOTERM_MF_FAT	sequence-specific DNA binding	19	6	1,50E-02
168	UP_SEQ_FEATURE	domain:Chromo 1	3	0,9	1,50E-02
169	GOTERM_CC_FAT	anchoring junction	8	2,5	1,50E-02
170	GOTERM_BP_FAT	cell cycle process	17	5,4	1,60E-02
171	SP_PIR_KEYWORDS	zinc	49	15,5	1,60E-02
172	GOTERM_BP_FAT	protein transport	21	6,6	1,60E-02
173	SP_PIR_KEYWORDS	ank repeat	10	3,2	1,60E-02
174	UP_SEQ_FEATURE	zinc finger region:PHD-type 1	4	1,3	1,60E-02
175	GOTERM_BP_FAT	regulation of cell morphogenesis	7	2,2	1,70E-02
176	UP_SEQ_FEATURE	repeat:ANK 6	6	1,9	1,70E-02
177	SP_PIR_KEYWORDS	cell cycle	15	4,7	1,70E-02
178	GOTERM_BP_FAT	regulation of growth	12	3,8	1,80E-02
179	GOTERM_BP_FAT	embryonic organ development	8	2,5	1,80E-02
180	INTERPRO	Clathrin/coatomer adaptor, adaptin-like, N-terminal	3	0,9	1,90E-02
181	GOTERM_BP_FAT	cell cycle	21	6,6	1,90E-02
182	GOTERM_CC_FAT	focal adhesion	6	1,9	1,90E-02
183	GOTERM_MF_FAT	Ras GTPase binding	6	1,9	1,90E-02
184	GOTERM_BP_FAT	protein localization	23	7,3	2,00E-02
185	GOTERM_BP_FAT	negative regulation of cell differentiation	9	2,8	2,00E-02
186	GOTERM_BP_FAT	hemopoietic or lymphoid organ development	10	3,2	2,10E-02
187	UP_SEQ_FEATURE	compositionally biased region:Poly-Leu	7	2,2	2,20E-02
188	GOTERM_BP_FAT	cellular component morphogenesis	13	4,1	2,20E-02
189	GOTERM_CC_FAT	cell-substrate adherens junction	6	1,9	2,20E-02
190	GOTERM_BP_FAT	mitosis	9	2,8	2,20E-02
191	GOTERM_BP_FAT	nuclear division	9	2,8	2,20E-02
192	GOTERM_MF_FAT	protein dimerization activity	17	5,4	2,20E-02
193	GOTERM_MF_FAT	ligand-dependent nuclear receptor transcription coactivator activity	4	1,3	2,30E-02
194	GOTERM_BP_FAT	negative regulation of cellular component organization	7	2,2	2,40E-02
195	SP_PIR_KEYWORDS	repressor	14	4,4	2,40E-02
196	GOTERM_BP_FAT	cell morphogenesis	12	3,8	2,40E-02
197	GOTERM_CC_FAT	spindle	7	2,2	2,40E-02
198	GOTERM_BP_FAT	M phase of mitotic cell cycle	9	2,8	2,40E-02
199	INTERPRO	Zinc finger, LIM-type	5	1,6	2,50E-02
200	INTERPRO	Kinesin, motor region, conserved site	4	1,3	2,60E-02

201	INTERPRO	Kinesin, motor region	4	1,3	2,60E-02
202	GOTERM_BP_FAT	cell adhesion	19	6	2,60E-02
203	GOTERM_BP_FAT	positive regulation of NF-kappaB transcription factor activity	4	1,3	2,60E-02
204	GOTERM_BP_FAT	biological adhesion	19	6	2,60E-02
205	GOTERM_BP_FAT	epithelium development	9	2,8	2,60E-02
206	UP_TISSUE	Lymph	19	6	2,60E-02
207	GOTERM_BP_FAT	negative regulation of gene expression	15	4,7	2,70E-02
208	GOTERM_BP_FAT	organelle fission	9	2,8	2,70E-02
209	GOTERM_CC_FAT	cell-substrate junction	6	1,9	2,70E-02
210	SP_PIR_KEYWORDS	cell division	10	3,2	2,70E-02
211	SP_PIR_KEYWORDS	LIM domain	5	1,6	2,70E-02
212	GOTERM_BP_FAT	microtubule cytoskeleton organization	7	2,2	2,80E-02
213	SP_PIR_KEYWORDS	notch signaling pathway	4	1,3	2,80E-02
214	GOTERM_BP_FAT	negative regulation of transcription	14	4,4	2,80E-02
215	GOTERM_BP_FAT	chromatin modification	10	3,2	2,80E-02
216	GOTERM_BP_FAT	cell cycle phase	13	4,1	2,90E-02
217	GOTERM_BP_FAT	immune system development	10	3,2	2,90E-02
218	INTERPRO	Protein of unknown function DM15	2	0,6	3,00E-02
219	INTERPRO	CAZ complex, RIM-binding protein	2	0,6	3,00E-02
220	INTERPRO	Adaptor protein complex AP-1, gamma subunit	2	0,6	3,00E-02
221	GOTERM_BP_FAT	mitotic cell cycle	12	3,8	3,10E-02
222	UP_TISSUE	Embryonal rhabdomyosarcoma	3	0,9	3,20E-02
223	UP_SEQ_FEATURE	zinc finger region:PHD-type 5	2	0,6	3,20E-02
224	UP_SEQ_FEATURE	zinc finger region:PHD-type 4	2	0,6	3,20E-02
225	UP_SEQ_FEATURE	region of interest:Class A specific domain	2	0,6	3,20E-02
226	UP_SEQ_FEATURE	short sequence motif:LD motif 4	2	0,6	3,20E-02
227	UP_SEQ_FEATURE	domain:Kinesin-motor	4	1,3	3,20E-02
228	GOTERM_BP_FAT	regulation of cell morphogenesis involved in differentiation	5	1,6	3,20E-02
229	GOTERM_CC_FAT	chromosomal part	12	3,8	3,20E-02
230	GOTERM_BP_FAT	M phase	11	3,5	3,40E-02
231	GOTERM_CC_FAT	basolateral plasma membrane	8	2,5	3,40E-02
232	UP_TISSUE	Teratocarcinoma	16	5,1	3,50E-02
233	GOTERM_BP_FAT	regulation of insulin receptor signaling pathway	3	0,9	3,50E-02
234	GOTERM_BP_FAT	chromatin organization	12	3,8	3,50E-02

235	GOTERM_BP_FAT	positive regulation of growth	5	1,6	3,60E-02
236	GOTERM_BP_FAT	ectoderm development	8	2,5	3,70E-02
237	GOTERM_MF_FAT	protein N-terminus binding	5	1,6	3,70E-02
238	GOTERM_MF_FAT	transcription corepressor activity	7	2,2	3,70E-02
239	INTERPRO	Zinc finger, RING-type	10	3,2	3,70E-02
240	GOTERM_BP_FAT	negative regulation of cell development	4	1,3	3,90E-02
241	UP_TISSUE	Human lung	4	1,3	3,90E-02
242	GOTERM_BP_FAT	regulation of cell proliferation	20	6,3	3,90E-02
243	INTERPRO	Helix-loop-helix DNA-binding	5	1,6	3,90E-02
244	UP_SEQ_FEATURE	compositionally biased region:Gly-rich	9	2,8	4,00E-02
245	UP_SEQ_FEATURE	zinc finger region:C3H1-type 3	3	0,9	4,10E-02
246	GOTERM_BP_FAT	negative regulation of BMP signaling pathway	3	0,9	4,20E-02
247	GOTERM_BP_FAT	regulation of cell development	8	2,5	4,20E-02
248	GOTERM_MF_FAT	SMAD binding	4	1,3	4,30E-02
249	INTERPRO	Zinc finger, CCCH-type	4	1,3	4,40E-02
250	UP_SEQ_FEATURE	short sequence motif:LXXLL motif 1	3	0,9	4,50E-02
251	UP_SEQ_FEATURE	short sequence motif:LXXLL motif 2	3	0,9	4,50E-02
252	GOTERM_CC_FAT	chromosome	13	4,1	4,60E-02
253	INTERPRO	High mobility group, HMG1/HMG2	4	1,3	4,70E-02
254	GOTERM_BP_FAT	Notch signaling pathway	4	1,3	4,70E-02
255	GOTERM_BP_FAT	cytoskeleton-dependent intracellular transport	4	1,3	4,70E-02
256	GOTERM_BP_FAT	muscle organ development	8	2,5	4,80E-02
257	UP_SEQ_FEATURE	region of interest:Acidic	2	0,6	4,80E-02
258	UP_SEQ_FEATURE	region of interest:Interaction with CTNNB1	2	0,6	4,80E-02
259	GOTERM_BP_FAT	liver development	4	1,3	5,00E-02
260	GOTERM_CC_FAT	cell junction	14	4,4	5,00E-02
261	GOTERM_MF_FAT	zinc ion binding	50	15,8	5,10E-02
262	UP_SEQ_FEATURE	compositionally biased region:Poly-Asp	5	1,6	5,60E-02
263	SP_PIR_KEYWORDS	chromatin regulator	8	2,5	5,70E-02
264	GOTERM_BP_FAT	embryonic organ morphogenesis	6	1,9	5,70E-02
265	GOTERM_BP_FAT	positive regulation of multicellular organism growth	3	0,9	5,70E-02
266	GOTERM_BP_FAT	negative regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	14	4,4	5,80E-02
267	UP_TISSUE	Embryo	11	3,5	5,80E-02

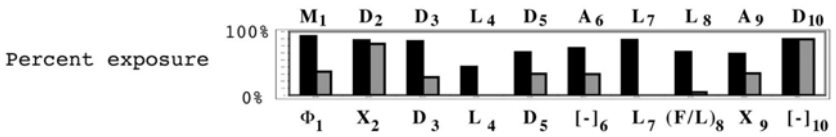
268	INTERPRO	FY-rich, C-terminal subgroup	2	0,6	5,90E-02
269	GOTERM_BP_FAT	gland development	6	1,9	6,00E-02
270	UP_SEQ_FEATURE	domain:LIM zinc-binding 3	3	0,9	6,10E-02
271	GOTERM_BP_FAT	steroid hormone receptor signaling pathway	4	1,3	6,20E-02
272	UP_SEQ_FEATURE	region of interest:Interaction with F-actin	2	0,6	6,30E-02
273	GOTERM_BP_FAT	negative regulation of nitrogen compound metabolic process	14	4,4	6,30E-02
274	GOTERM_BP_FAT	actin cytoskeleton organization	8	2,5	6,40E-02
275	GOTERM_BP_FAT	negative regulation of macromolecule metabolic process	18	5,7	6,70E-02
276	GOTERM_BP_FAT	positive regulation of transcription factor activity	4	1,3	6,70E-02
277	UP_TISSUE	Placenta	66	20,9	6,80E-02
278	GOTERM_BP_FAT	epidermis development	7	2,2	6,90E-02
279	GOTERM_BP_FAT	intracellular protein transport	11	3,5	6,90E-02
280	GOTERM_BP_FAT	regulation of multicellular organism growth	4	1,3	7,00E-02
281	UP_SEQ_FEATURE	domain:SH2	5	1,6	7,00E-02
282	GOTERM_CC_FAT	extrinsic to membrane	13	4,1	7,10E-02
283	UP_TISSUE	Uterus	37	11,7	7,20E-02
284	GOTERM_MF_FAT	promoter binding	4	1,3	7,20E-02
285	GOTERM_BP_FAT	developmental maturation	5	1,6	7,30E-02
286	INTERPRO	FY-rich, N-terminal subgroup	2	0,6	7,40E-02
287	INTERPRO	FY-rich, C-terminal	2	0,6	7,40E-02
288	INTERPRO	FY-rich, N-terminal	2	0,6	7,40E-02
289	INTERPRO	Spindlin/spermiogenesis-specific protein	2	0,6	7,40E-02
290	GOTERM_CC_FAT	myosin complex	4	1,3	7,50E-02
291	GOTERM_BP_FAT	mammary gland development	4	1,3	7,50E-02
292	SP_PIR_KEYWORDS	mitosis	7	2,2	7,50E-02
293	GOTERM_BP_FAT	B cell lineage commitment	2	0,6	7,60E-02
294	GOTERM_BP_FAT	hemopoiesis	8	2,5	7,70E-02
295	UP_SEQ_FEATURE	short sequence motif:LXXLL motif 6	2	0,6	7,80E-02
296	SP_PIR_KEYWORDS	Proto-oncogene	8	2,5	7,90E-02
297	UP_TISSUE	Clones donated by Kazusa DNA Research Inst.	3	0,9	7,90E-02
298	GOTERM_BP_FAT	regulation of BMP signaling pathway	3	0,9	7,90E-02
299	GOTERM_CC_FAT	centriole	3	0,9	7,90E-02
300	UP_SEQ_FEATURE	short sequence motif:Bipartite nuclear localization signal	3	0,9	7,90E-02
301	GOTERM_MF_FAT	insulin receptor binding	3	0,9	8,10E-02

302	GOTERM_MF_FAT	double-stranded DNA binding	5	1,6	8,30E-02
303	GOTERM_BP_FAT	keratinocyte differentiation	4	1,3	8,40E-02
304	GOTERM_BP_FAT	ear morphogenesis	4	1,3	8,40E-02
305	GOTERM_BP_FAT	actin filament-based process	8	2,5	8,40E-02
306	INTERPRO	Chromo domain	3	0,9	8,50E-02
307	GOTERM_BP_FAT	negative regulation of macromolecule biosynthetic process	14	4,4	8,70E-02
308	GOTERM_BP_FAT	enzyme linked receptor protein signaling pathway	10	3,2	8,70E-02
309	INTERPRO	MLL Transcription Factor	2	0,6	8,80E-02
310	GOTERM_BP_FAT	palate development	3	0,9	8,90E-02
311	SP_PIR_KEYWORDS	Apoptosis	11	3,5	8,90E-02
312	SP_PIR_KEYWORDS	acetylation	52	16,5	8,90E-02
313	INTERPRO	SH2 motif	5	1,6	8,90E-02
314	GOTERM_BP_FAT	blood vessel development	8	2,5	8,90E-02
315	GOTERM_CC_FAT	lamellipodium	4	1,3	8,90E-02
316	INTERPRO	Keratin, type I	3	0,9	8,90E-02
317	INTERPRO	PAS	3	0,9	8,90E-02
318	GOTERM_MF_FAT	tubulin binding	5	1,6	9,00E-02
319	SP_PIR_KEYWORDS	cytoplasmic vesicle	8	2,5	9,00E-02
320	GOTERM_BP_FAT	cell division	9	2,8	9,10E-02
321	GOTERM_MF_FAT	Rab GTPase binding	3	0,9	9,10E-02
322	UP_TISSUE	Cerebellum	17	5,4	9,30E-02
323	SP_PIR_KEYWORDS	steroid hormone receptor	2	0,6	9,30E-02
324	GOTERM_BP_FAT	positive regulation of cell development	4	1,3	9,30E-02
325	UP_TISSUE	Glial cell	2	0,6	9,60E-02
326	GOTERM_BP_FAT	positive regulation of DNA binding	4	1,3	9,60E-02

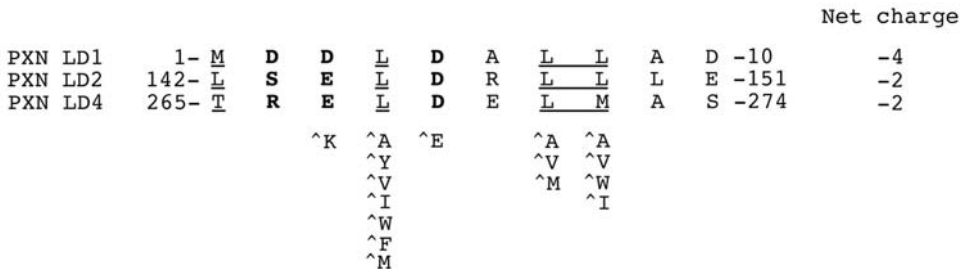
Supplemental Table 3.: Gene onthology terms-based enrichment of the predicted BPV-1 E6 binding proteins. We investigated the full *in silico* predicted BPV-1 E6 target list for functional annotation terms using the database DAVID (http://david.abcc.ncifcrf.gov/). The enrichment of functional terms according to the LxxLL regular expression was ranked by p-value. **Nr.:** ranking number; Category: origin and annotation of the enriched term; **Term:** description of the term; **Count:** number of genes that were identified by our bioinformatics screen and for which the enriched term was found in the DAVID database; **%:** percentage of genes annotated with the term, out of all genes idetified by our screen and that could be mapped to the DAVID database (316 genes in total); **p-value:** significance of enrichment of the term (the lower the number, the more significant the enrichment).

Supplemental Fig. 7

A.



B.



C.

Consensus				1- Φ	x	D	L	D	[-]	L	(F/L) -8			Net charge
HS	A	K	Y	P	D	*N	G	L	L	D	L F			GGG -1
HS	I	N	F	S	P	A	D	L	D	D	L L			GGG -3
HS	L	P	T	Y	W	A	P	M	M	E	L F			GGG -1
HS	N	I	T	N	L	Y	D	L	D	L	L F			GGG -2
HS		E	H	N	F	Q	W	L	D	H	L L			GGG -1
HS	N	I	T	N	L	Y	D	L	D	L	L F			GGG -2
HS	Y	S	D	N	F	G	N	L	D	A	L L			GGG -2
HS	R	Y	N	Q	M	D	Y	L	D	A	L L			GGG -1
HS	D	S	A	D	W	Q	S	F	D	E	L L			GGG -3
HS	N	D	N	G	L	L	W	L	D	T	L F			GGG -2
HS	T	Q	W	N	P	A	D	L	*N	E	L F			GGG -2
HS	T	S	F	R	P	F	D	L	D	D	L F			GGG -2
HS				S	W	E	D	L	S	A	L F	N	E	RGGG -2
HS			Y	A	P	Y	D	L	Y	M	L F	R	D	GGG -1
HS				D	H	S	P	L	D	E	L F	G	E	SGGG -2
HS			Q	D	F	A	M	L	D	L	L L	R	E	GGG -2

Identifying key positions of the LxxLL motif that are critical for recognition of BPV-1 E6.

A. Exposure plot of the LxxLL peptide : The percentage of solvent exposure of each peptide residue was calculated using the PyMol program in the presence (grey bar) or the absence (black bar) of the bound BPV-1 E6 protein, and plotted against the peptide sequence (above) and the consensus sequence derived from the phage display (bottom). The residues exhibiting a higher increase in exposure in the absence of bound E6 are the residues most involved in binding. **B.** Sequence alignment of the three BPV-1 E6 binding LD1 motifs of paxillin. The alignment reveals the preference for a hydrophobic residue at position 1, a negative charge at position 3 and 5, strongly conserved Leucine residues at position 4, 7 and 8 and a negative charge at position 10. Residues listed on the bottom of this figure labeled with a « ^ » are residues that can be excluded according to this alignment. **C.** Phage display results. 16 phage peptides selected after several rounds of phage display vs chitin resin-bound BPV-1 E6 are listed and aligned. The net charge of each peptide is plotted at the right and varies between -1 and -3. The resulting consensus sequence is plotted on top of the alignment. The coloured consensus plot shown above the consensus was created with WebLogo (http://weblogo.berkeley.edu/logo.cgi). The consensus features of the selected phage peptides are extremely similar to those of the E6-binding LD motifs of paxillin.

All the information presented in A, B and C was used to perform the in silico peptide prediction on the human proteome (Fig. 3 and supplemental Table 1 and 2).

B.2. Automated holdup assay for high-throughput determination of domain–linear motif affinities

HoldUp is a comparative chromatographic retention assay that has been developed to measure at equilibrium interactions between proteins or protein fragments (see Figure B.1) [207]. Here, we present the automation of this method, its benchmarking and application to large-scale interaction measurements between PDZ domains and C-terminal peptides. This summary will focus on the data analysis part of this study.
Results: The HoldUp method has been implemented on a multiplate liquid handling robot in 96 and 384 well plate format. Binding signals were quantified using a Caliper LabChip Gx/GX II machine, performing automated capillary electrophoresis. The Caliper software provides different data output formats. The fluorescence signals can be obtained in flat files, displayed as electropherograms or as artificial gels (see Figure B.2). Based on the raw data flat files provided by Caliper, a protocol for binding intensity calculation has been developed that takes into account variations in sample input quantities due to experimental imprecision (see Figure B.4 and Methods below). The combined experimental and data analysis protocol has been successfully benchmarked on 200 PDZ-peptide interactions, of which binding intensities were previously measured with SPR (see chapter 9). The benchmarking validated the usefulness of the protocol to obtain quantitative interaction data and the possibility of HoldUp to work with unpurified protein samples (see Figure B.3). We obtained clones of more than 240 human PDZ domains from collaborators that we tested for interaction to a C-terminal PBM derived from HPV16 E6. Results correlated very well with published literature on PDZ–E6 interactions (see Figure B.5).
Discussion/Conclusions: HoldUp has various strengths in comparison to other HTP methods for protein interaction measurements. Interaction data is recorded at equilibrium, thus it can in principle be used to estimate dissociation constants. HoldUp provides quantitative interaction data, which opens a wide range of diverse applications of this method. In the context of domain–linear motif interactions, it has great potential in being applied to binding affinity and specificity assessments of one peptide *versus* a Domainome (e.g. all instances of a domain type of one organism) or *vice versa*, of a domain versus a large set of peptides.
Contribution: I have been mainly involved in the data analysis including the development of an automated protocol for binding intensity calculation as well as data treatment. I mainly conceived and created most of the figures.

Author list:
Sebastian Charbonnier(*), Katja Luck(*), Jolanda Polanowska, Julie Abdat, Marilyne Blémont, François Iv, Yves Nominé, Jérôme Reboul, Gilles Travé(&), Renaud Vincentelli(&).
(*) Equal contributors to the work.
(&) Corresponding authors and equal supervisors of the work.

Extraction of methods:

Raw Caliper data processing

The electropherogram data that can be exported from Caliper contains the fluorescence signal versus the time and versus calculated protein sizes. For data analysis, we decided to work with calculated protein sizes instead of the time because according to the Caliper manual these were already corrected for signal shifts. The data series were uniformed for comparison by rounding the protein sizes to .5 kDa and taking the mean of the fluorescence signals recorded for the same rounded protein size. For each plate, a window containing the PDZ peaks and windows for input and background correction (see below) were defined by manual inspection. From caliper we extracted the fluorescence signals of PDZ domains obtained for runs launched either with biotinylated peptide or biotin. The more PDZ domain retained on the gel, the lower the fluorescence signal of the PDZ domain of the biotinylated peptide in comparison to the biotin control. This difference in signals is directly correlated to the binding strength of the PDZ biotinylated peptide interaction. This signal difference is normalised dividing it by the control signal to take into account variations in PDZ amounts loaded on the well plate. Therefore, we define the binding strength S as:

$$S = \frac{f_b - f_s}{f_b} = 1 - \frac{f_s}{f_b} \tag{B.1}$$

where f_s and f_b correspond to the fluorescence signal of the PDZ peak of the sample and the biotin run, respectively (see Figure B.4A). This way, binding strengths were calculated for measurements recorded with purified protein samples. For measurements carried out with unpurified (crude) protein samples, the fluorescence signals were subjected to two corrections before used for binding strength calculations: 1) input correction and 2) background correction.

Input correction

In theory, the signals of the sample and biotin runs should be identical except in the area of the PDZ peak. In practice, we observed signal variations (ask Gilles for reasons) that we corrected by introducing an α factor to rescale the sample to the biotin signals:

$$S = 1 - \frac{f_s \alpha}{f_b} \tag{B.2}$$

α was obtained by minimizing the difference $diff = d_b - (\alpha d_s)$ between the data series d_s of the sample and the biotin d_b (see Figure B.4B). The data series were obtained within the molecular weight window ranging from 24 to 50 kDa excluding the areas where PDZ peaks or system peaks were expected. We used the least square minimization method implemented in the "optimize" module of scipy to determine

α. Mostly, more than 80 % of α are comprised between 0.9 and 1.1. However, in a few cases, when sample or biotin runs displayed negative fluorescence signals the multiplication with α factor would lead to incorrect scaling. Therefore, the minimal fluorescence signals m_s and m_b determined within a window between 24 and 70 kDa were substracted from the sample and biotin runs, respectively:

$$S = 1 - \frac{f_s - m_s\alpha}{f_b - m_b} \tag{B.3}$$

Background correction

The background correction was designed to remove from a sample PDZ peak the signal from other proteins of the same size as the PDZ construct (referred to as background peaks) that otherwise could lead to an overestimation of binding strength. To estimate the height of the background peaks, all data series corrected for input variations from measurements with PDZ constructs of significant different size (e.g. single versus tandem PDZ constructs, see Figure B.4D) from the same plate were used to construct a mean curve. The sample and biotin runs were scaled to this mean curve by using again the least square optimization method to determine the α_r factor. Of the mean curve, the highest signal h in the window of the sample PDZ peak was determined and substracted from the sample and biotin PDZ peak to obtain the final binding strength (see Figure B.4E):

$$S = 1 - \frac{f_s - m_s\alpha\alpha_r - h}{f_b - m_b\alpha_r - h} \tag{B.4}$$

Background corrections were only performed for samples for which data series with similar peak signature but with PDZ constructs of different size were available. In rare cases, input or background correction led to binding intensities above 1 or below 0. In those cases, binding strengths were set to 1 or 0, respectively.

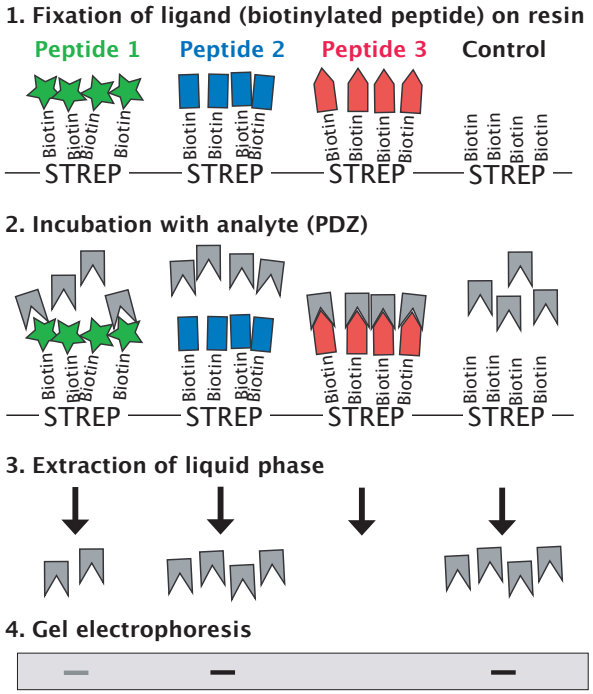


Figure B.1. Schematic representation of the HoldUp method. 1. The ligands (in our case C-terminal peptides) are fixed on a resin, e.g. via a biotin-streptavidin system. The negative control consists of an empty resin. 2. The fixed peptides are incubated with the analyte (in our case a PDZ domain) that will bind with different affinities to each of the proposed peptides. 3. The liquid phase is extracted and analysed via 4. gel electrophoresis. The amount of analyte that will be detected on the gel strongly depends on the affinity that it displayed towards a ligand (no analyte detectable if strong interaction with ligand, more amounts of analyte detected for weaker interactions). The amount of analyte detected on the gel will be normalised to the amount of analyte detected from the negative control to obtain relative binding intensities (figure adapted from S. Charbonnier).

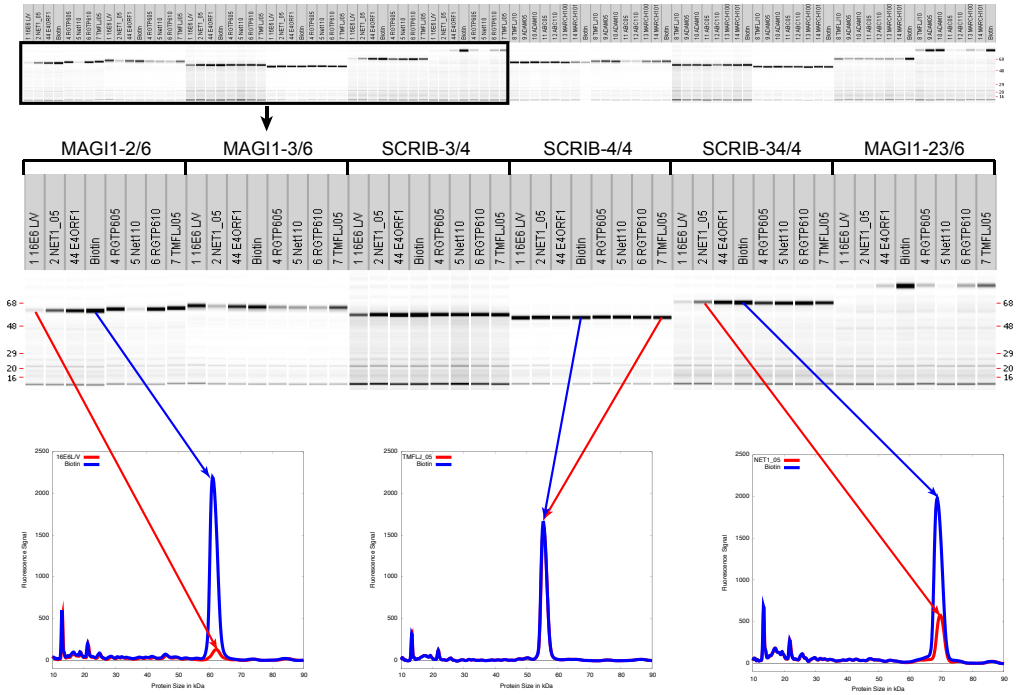


Figure B.2. Illustration of Caliper output in gel and electropherogram format. Three examples of interaction measurements were chosen to illustrate the electropherograms that are produced by Caliper and translated into artificial gels. The fluorescence signals obtained from the liquid phase (with the PDZ) after incubation with a fixed peptide on the resin are coloured in red, those of the negative control (no fixed peptide) in blue. From left to right: Strong interaction between PDZ2 of MAGI1 with a mutated C-terminal peptide derived from HPV16; no interaction between PDZ4 of SCRIB and a C-terminal peptide derived from transmembrane protein 215; weak interaction between the tandem construct comprising PDZ3 and PDZ4 of SCRIB and a C-terminal peptide derived from Net1 (Neuroepithelial cell-transforming gene 1).

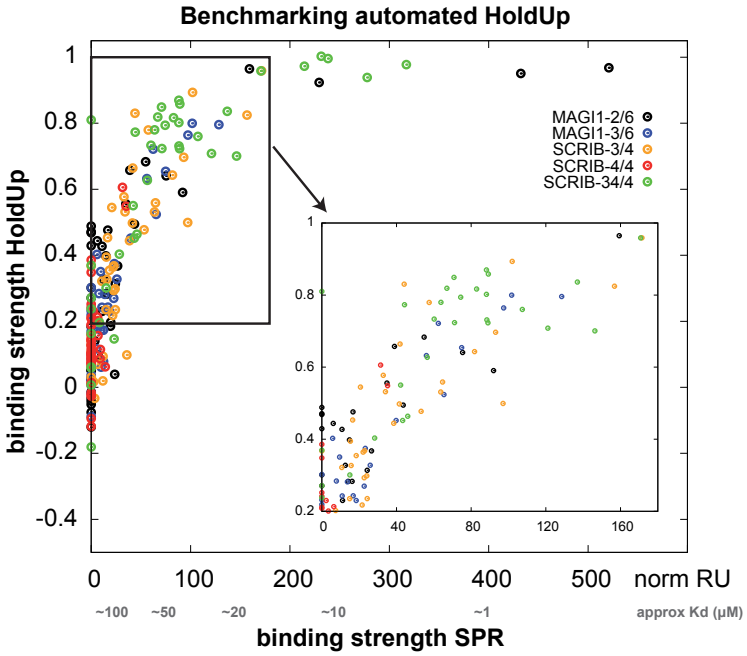


Figure B.3. Comparison of binding strengths obtained with HoldUp to binding strength measured with SPR. We benchmarked the automated HoldUp method on an interaction data set obtained with SPR. Relative binding strengths obtained from HoldUp are plotted versus the normalised Response Units (RUs) that were obtained with SPR (see chapter 9). There is a linear correlation between HoldUp and SPR binding intensities in a range of approximately 15 to 80 µM. Binding intensities determined with HoldUp for stronger interactions are saturated but can be dissolved when diluting analyte and/or ligand samples (data not shown). HoldUp seems to be more sensitive towards very weak interactions in comparison to SPR.

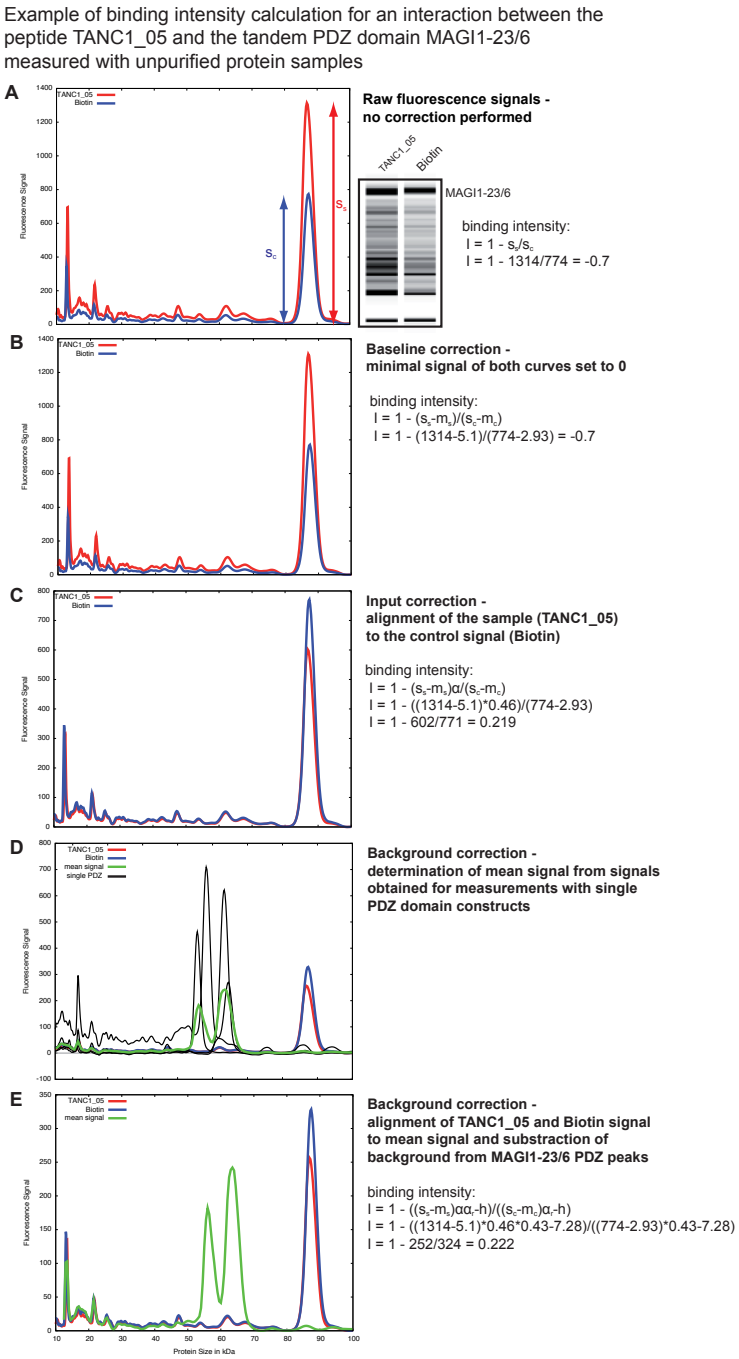


Figure B.4. Data processing of interaction measurements obtained with unpurified protein samples. More information on the individual electropherogram diagrams can be found in the methods text. The fluorescence signals obtained from the liquid phase (with the tandem PDZ construct MAGI1-23/6) after incubation with a C-terminal peptide derived from the human protein TANC1 fixed on the resin are coloured in red, those of the corresponding negative control in blue.

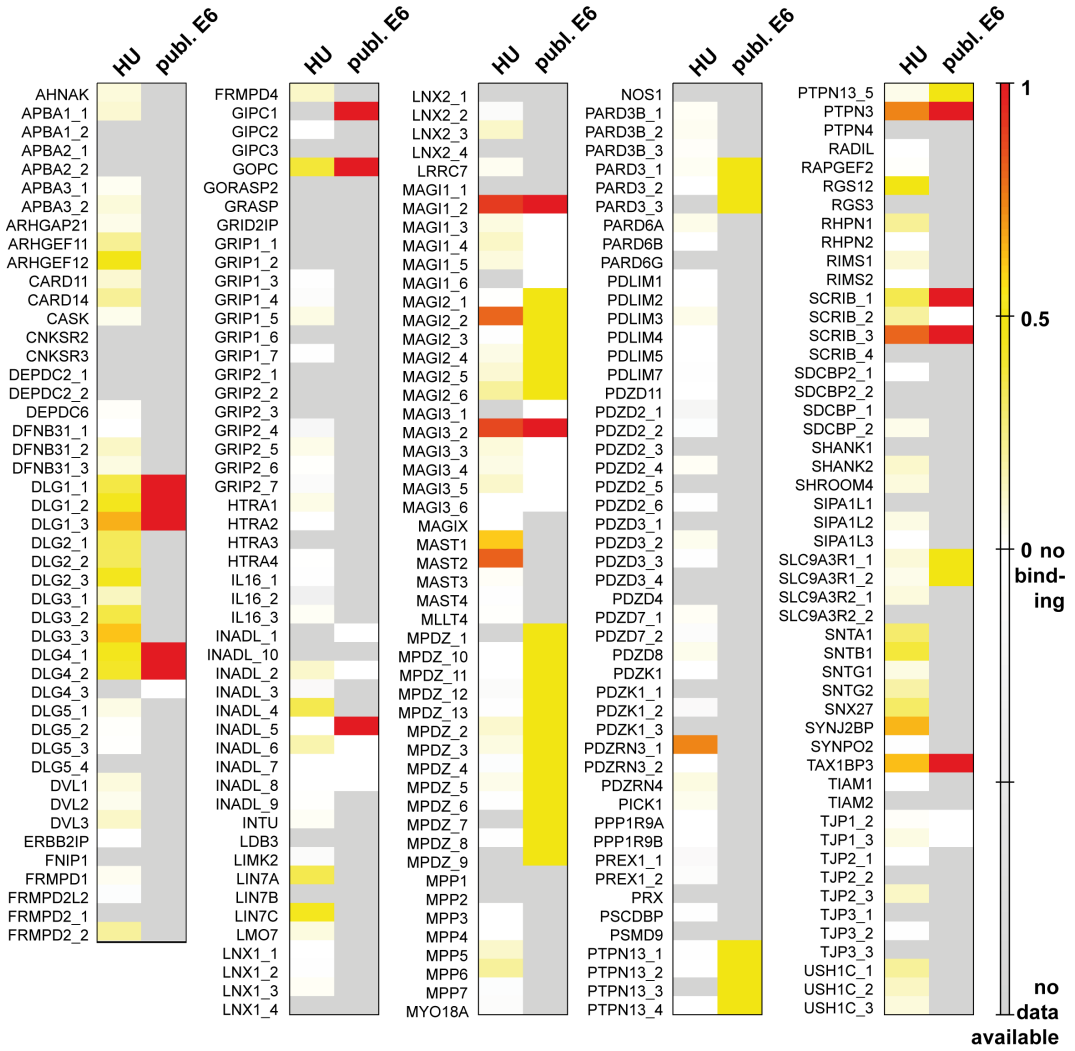


Figure B.5. PDZome vs. HPV16 E6 C-terminal peptide - comparison to published data. The binding intensities measured with automated HoldUp (HU) between the human PDZome and HPV16 E6 C-terminal peptide are illustrated with a colour code ranging from white for no binding over yellow to red for strong binding. Published interactions between HPV16 or HPV18 E6 and PDZ domains were grouped into non-interactions (white) and interactions (red). Published interactions between a PDZ protein and E6 without the detailed information, which PDZ domain mediated the binding, are marked with yellow. All pairs of PDZs and peptides for which neither published nor HoldUp information was available are coloured in grey. Almost all published PDZ binders of E6 were validated by HoldUp that in addition, identified some previously unknown potential PDZ binders of E6.

C. Advanced projects

C.1. Towards a comprehensive PDZ-mediated interaction network of the cell polarity regulator protein SCRIB

The human protein SCRIB is an important regulator of cell polarity. SCRIB has been termed tumour suppressor as its disregulation (e.g. by oncogenic viruses) can lead to tumorigenesis. Numerous proteins have been identified to be directly or indirectly associated with SCRIB, of which quite a few bind to one of the four PDZ domains of SCRIB. We aimed at contributing to a better understanding of the cellular functions of SCRIB, of which many are linked to cancer, by identifying new potential interactions of SCRIB and combining them with published data into a comprehensive network of protein interactions centred on SCRIB.
Approach: We built on our previous findings on sequence bias in published phage display data [9] (see chapter 8). Within the one third of phage display data that did not display any bias towards hydrophobic sequences were phage display peptides selected for PDZ1, PDZ2, and PDZ3 of SCRIB. We built PSSMs for these three PDZ domains and screened the human proteome for C-terminal peptides that were likely to bind to any of these three PDZs. Based on our previous findings and hypotheses on the particular peptide recognition properties of PDZ4 of SCRIB (see chapter 9), we designed regular expressions to additionally screen the human proteome for C-terminal peptides that might bind to PDZ4 of SCRIB. We selected 56 peptides including published binders of SCRIB, for experimental validation using automated HoldUp (see section B.2). We determined relative binding strengths between each of these 56 peptides and seven PDZ domain constructs comprising the four single PDZ domains of SCRIB as well as the tandem, triple, and quadruple construct PDZ34, PDZ234, and PDZ1234, respectively (see Figure C.1). In parallel, we have manually curated from published literature all proteins that were shown to be directly or indirectly associated with SCRIB. We used STRING [138] to identify interaction partners of published and new potential binding partners of SCRIB and to search for protein interactions between all those proteins. The resulting PPI network has been visualised and analysed with Cytoscape [208].
Results: 25 peptides bound at least to one of the four single PDZ domains of SCRIB with dissociation constants better than 100 μM. Those were merged with about 70 directly and indirectly associated published proteins of SCRIB into one network. Using STRING with a cutoff of 0.6 and a focus on data with experimental evidence, we identified more than 500 interaction partners of published and new binders of SCRIB establishing in total about 3,000 interactions with each other. Within this network we

searched the proteins for gene ontology terms that were statistically over-represented in comparison to the human proteome. This analysis revealed enrichments for terms like establishment and maintenance of cell polarity, cell adhesion, cell junction organisation, regulation of cell death, and regulation of localisation. These terms are all known functions that SCRIB is known to be associated with, validating the functional relevance of the established PPI network. From this unexpectedly big network we constructed a subnetwork comprising all those potential new interactors and published interactors that were found to be directly or indirectly (sharing a common interaction partner) linked with each other via protein interactions (see Figure C.2). An initial analysis revealed interesting functional links between the new potential interactors DOCK2, GUCSA2, YAP2 and SCRIB involving some of its known interaction partners.

To Dos: The analysis of the established PPI network has to be completed. Many of the new potential interactors remain to be analysed for their potential functional links to SCRIB. We have to find ways to graphically represent the accumulated knowledge on the tumour suppressor SCRIB that are straightforward for exploration by other researchers. Apart from investigating the biological roles of SCRIB, the experimental data obtained in this study also has great potential to provide insights into PDZ interaction specificities. This study is one of the very few (the only?) where numerous C-terminal PBMs were assayed for binding to all single and several multiple PDZ domain constructs of one protein. By analysing the determined binding strengths, we might be able to answer questions like: How do PDZ domains of one protein differ in peptide recognition? Do they have distinct recognition preferences or overlapping selectivity spaces? Do neighbouring PDZ domains influence peptide binding of each other? Answers to our questions might allow to better understand the biological function behind the four PDZ domains of SCRIB: did these PDZs evolve to bind different targets and act as scaffold for the assembly of signalling complexes or do they target the same PBMs thereby increasing binding affinity and maybe specificity towards interaction partners of SCRIB?

Contribution: I have performed the predictions and analysed the experimental data resulting from automated HoldUp experiments. I have carried out the manual curation of published associated proteins of SCRIB as well as the network constructions and their initial analysis.

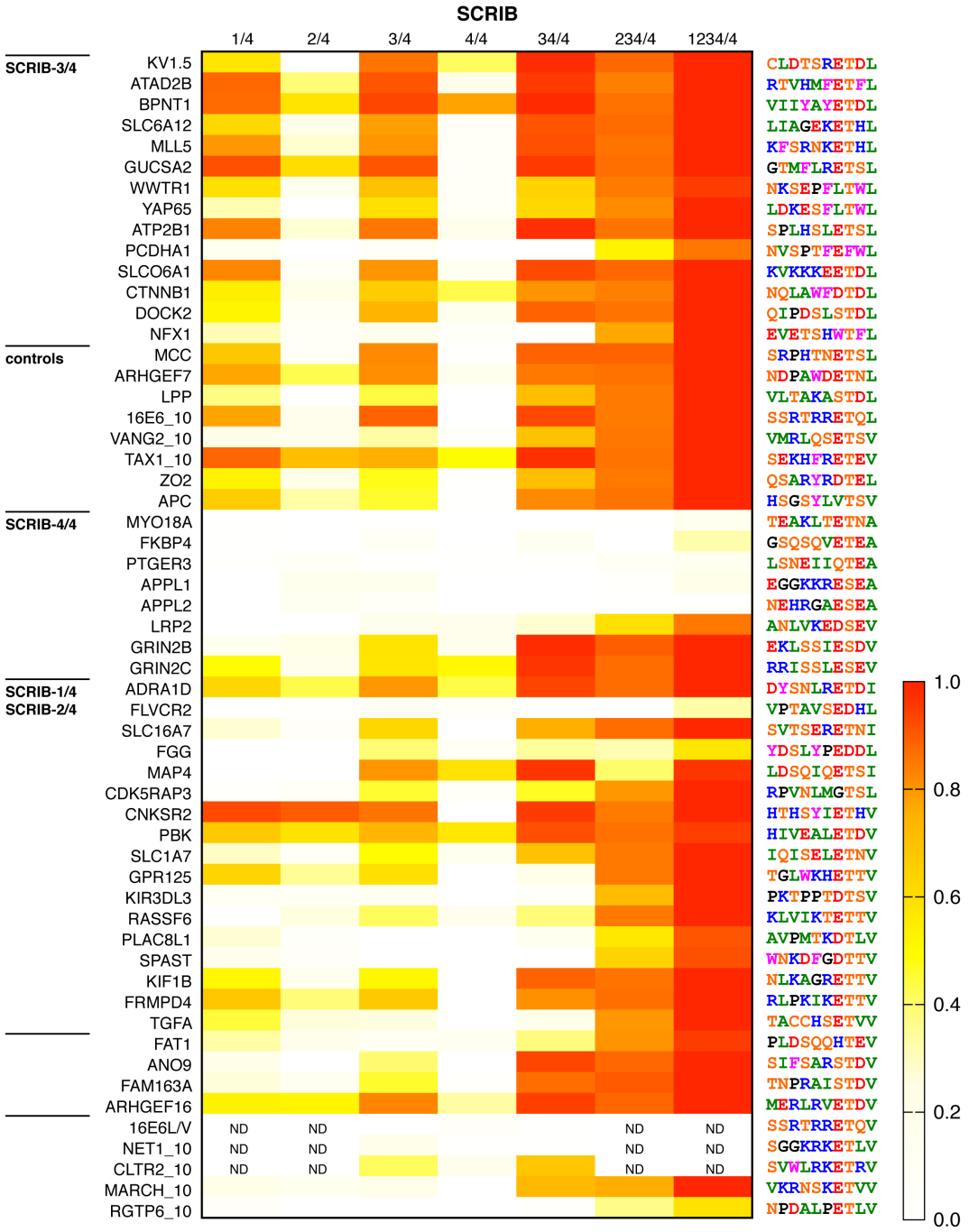


Figure C.1. Binding intensities obtained for 56 peptides and 7 PDZ constructs using automated HoldUp. A colour code has been used to indicate the different binding intensities obtained (white = no interaction detected, red = strong interactions (probably 5 μM or better)). Names of peptides are indicated to the left, their sequences to the right. Names of the PDZ constructs are indicated on top of the diagram (e.g. 1/4 is the first PDZ domain of SCRIB out of four.). The peptides were grouped together based on the PDZ domain to which they were predicted to bind or whether they served as control (indicated at the very left of the figure). ND = not determined.

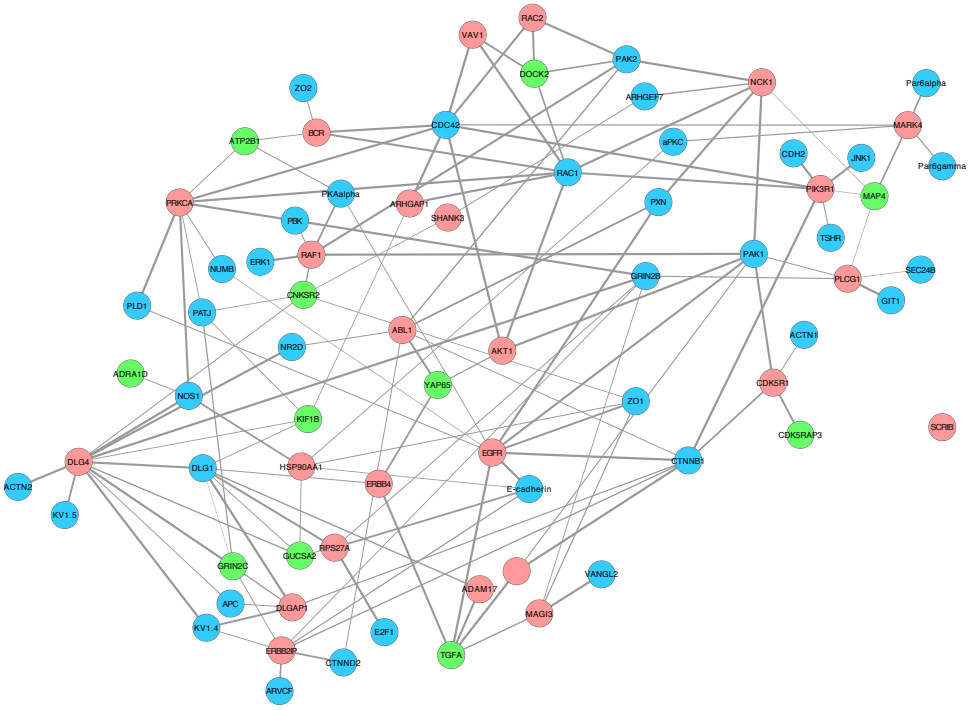


Figure C.2. Network of protein interactions around SCRIB combining new potential interactors of SCRIB with published ones. Nodes represent proteins, lines represent interactions between them as obtained from the STRING database [138]. Nodes coloured in green represent new potential interactors of SCRIB that we identified in automated HoldUp experiments. Nodes coloured in blue represent published directly or indirectly associated proteins of SCRIB. Interactions between SCRIB and its known and new binders have been hidden for clarity. Nodes coloured in brown represent proteins that bind to new and known binders of SCRIB. The line thickness of illustrated interactions correlates with their reliability and is based on the interaction score from STRING.

D. Selected supplemental material of published articles presented in this thesis

D.1. Supplemental material of the phage display article (see chapter 8)

Supplementary methods, Figure S1, Figure S2.

SUPPLEMENTAL METHODS

The programming and data analysis was done using python (www.python.org), scipy (www.scipy.org), R (cran.r-project.org), and gnuplot (www.gnuplot.info).

The human proteome

The human proteome was downloaded from Ensembl v50 (Hubbard *et al.*, 2009). From this proteome incomplete proteins (no asterisk at the end of the sequence) and isoforms of a gene that are identical in the last 5 residues were removed except one representative (Dataset S3).

Processing of phage display data

The phage display peptides of 54 human PDZ domains from Tonikian *et al.* (2008) were processed as follows. The peptides were cut to a length of 5 (from the C-terminus) and every peptide containing an X as amino acid or being identical to another peptide in the list was removed.

Calculation of mean hydrophobicity

The hydrophobicity of peptides was assessed using the hydrophobicity scale of Kidera *et al.* (1985) obtained from the AAindex database (Kawashima *et al.*, 2008). This scale assigns a value to each amino acid representing its hydrophobicity: positive values = hydrophilicity, 0 = neutrality, negative values = hydrophobicity. The sum over all these 20 values equals in zero. The mean hydrophobicity of a list of equally long peptides is the sum of the hydrophobicity values of all amino acids devided by the number of peptides.

Calculation of PSSMs

Position Specific Scoring Matrices (PSSMs) were used to search for similar human C-termini to phage display peptides. An entry in a PSSM was calculated as follows:

$$f_{a,i} = \frac{N_{a,i} + \frac{\sqrt{N_p}}{20}}{N_p + \sqrt{N_p}} \tag{1}$$

where a stands for an amino acid, i for a column (peptide position), $N_{a,i}$ means the number of occurrences of amino acid a at peptide position i and N_p represents the number of peptides. $f_{a,i}$ is the frequency of amino acid a at peptide position i for a given set of peptides. Pseudocounts were taken into account by the term $\frac{\sqrt{N_p}}{20}$, which is important for amino acids that were never observed at a particular peptide position. These matrix entries were then weighted with the amino acid frequencies extracted from the human proteome that was used throughout this work:

$$PSSM_{a,i} = \log_2(\frac{f_{a,i}}{F_a}) \tag{2}$$

where F_a stands for the frequency of amino acid a in the human proteome. This PSSM construction was done after Mount (2004).

A PSSM was constructed for each peptide list of the 54 human PDZ domains. Each PSSM was used to score every C-terminal peptide in the human proteome (score of peptide = sum of PSSM values). A first approach considered an individual score threshold for each domain based on the minimal score that a peptide of the phage display list obtained for the corresponding PSSM. This led to an extremely wide range of numbers of similar peptides that were returned (0 to 16205 matches, data not shown). A second simpler approach consisted in taking the best 25 hits per domain. This number was chosen because it corresponds to the mean number of peptides in the phage display peptide lists, which makes further comparison more robust. Additionally, this number of interactors is in a suitable range for experimental verification.

Calculation of distance between PSSMs

A PSSM was constructed based on the 25 best-matching human C-termini of each PDZ domain. For each PDZ domain, the distance D of the PSSM of the corresponding phage display peptides to the PSSM of the corresponding 25 best-matching human C-termini was calculated as follows (according to Tonikian *et al.* (2008)):

$$D = \sum_{a=1}^{20} \sum_{i=1}^{5} (PSSM_{a,i} - PSSM_{a,i})^2 \tag{3}$$

where a represents one of the 20 possible amino acids and i one of the 5 possible peptide positions.

REFERENCES

Hubbard, T. J. P. *et al.* (2009). Ensembl 2009. *Nucleic Acids Res*, **37**(Database issue), D690–D697.

Kawashima, S. *et al.* (2008). Aaindex: amino acid index database, progress report 2008. *Nucleic Acids Res*, **36**(Database issue), D202–D205.

Kidera, A. *et al.* (1985). Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *Journal of Protein Chemistry*, **4**(1), 23–55.

Mount, D. W. (2004). *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press.

Tonikian, R. *et al.* (2008). A specificity map for the pdz domain family. *PLoS Biol*, **6**(9), e239.

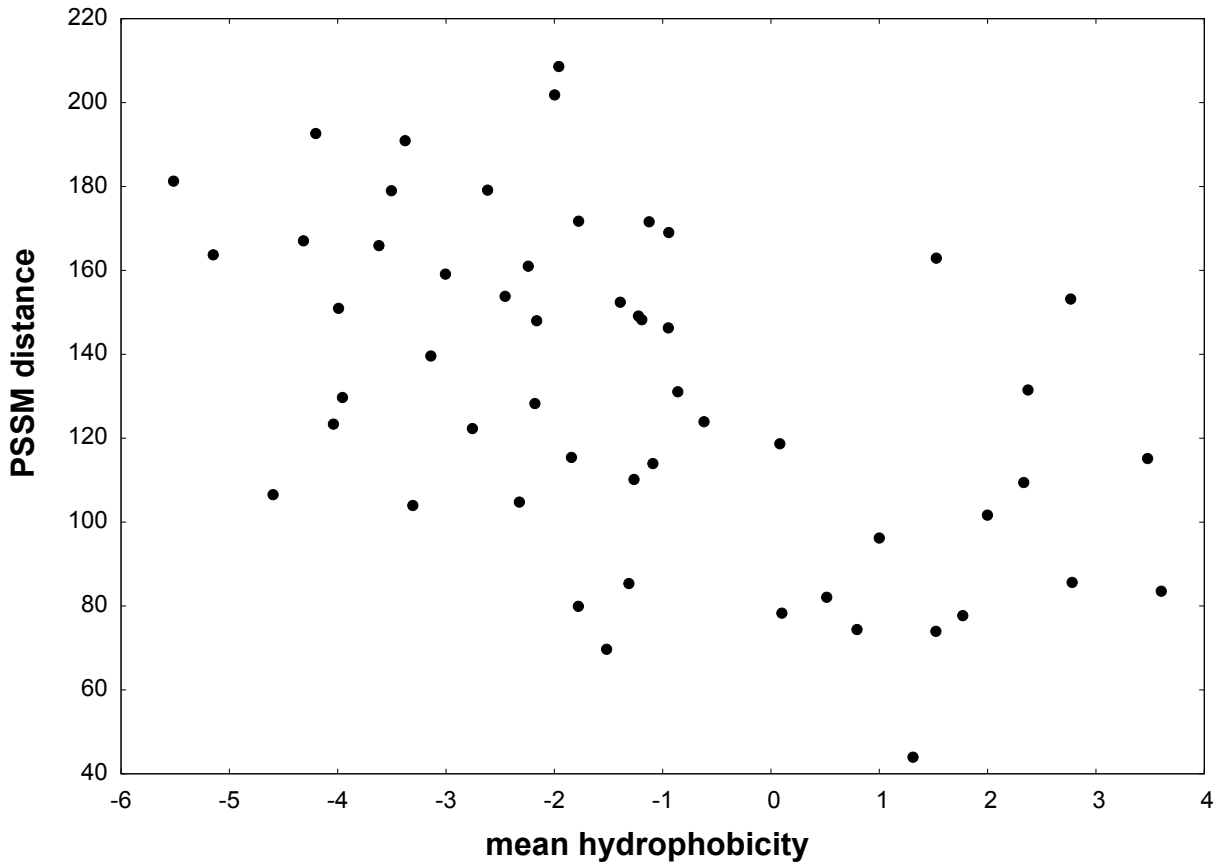


Figure S1: Correlation between PSSM distance and mean hydrophobicity of phage display peptide lists. For each of the 54 phage display peptide lists of Tonikian *et al*. a PSSM was constructed and used to determine the 25 best-matching human C-termini out of the human proteome. For each of these resulting 54 lists of best-matching human C-termini a PSSM was constructed and its distance determined to the PSSM of the corresponding phage display peptide list. This distance value is plotted against the mean hydrophobicity of the phage display peptide list. The more hydrophilic the phage display peptides, the smaller the distance between the PSSMs. This correlation indicates that the best-matching human C-termini for hydrophilic phage display peptides tend to be more similar to the sequence profile defined by the phage display peptides than it is the case for hydrophobic phage display peptides.

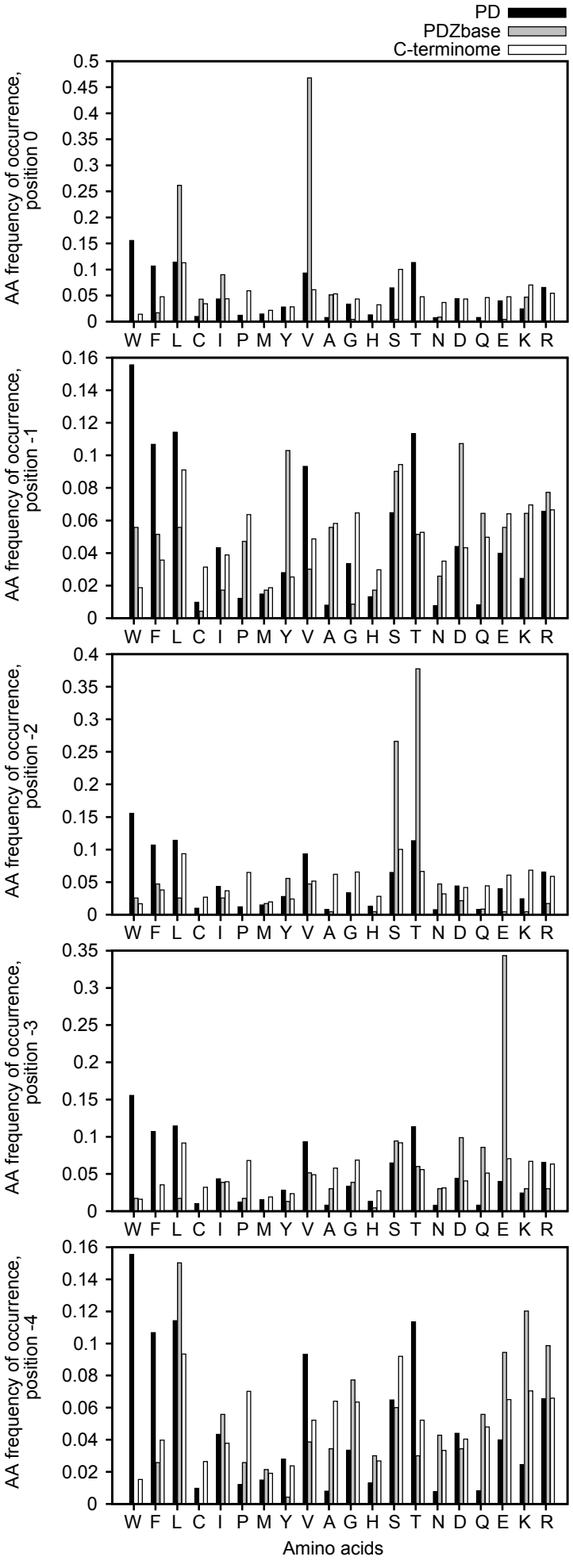


Figure S2: Amino acid composition of phage display peptides vs. the human C-terminome and PDZ-binding peptides from the PDZbase determined per peptide position. Amino acids are sorted from the most hydrophobic (left) to the most hydrophilic (right) according to the hydrophobicity scale of Kidera *et al.* (1985). All sequences were cut to a length of five residues. Peptides from the PDZbase seem to be enriched for class 1 PDZ-binding motifs because S and T are predominant at position -2. In contrast, phage display peptides seem to consist of equal amounts of class 2 (hydrophobic) and class 1 PDZ-binding motifs. Phage display peptides show much more sequence diversity at position 0 than peptides from the PDZbase.

D.2. Supplemental material of the SPR article (see chapter 9)

Table S1, Table S2, Table S3, Table S4, Text S1, sensorgrams of all measurements performed.

Table S1: Diversity of amino acids at five peptide positions in the training data of Chen et al.

ligand pos[a]	observed amino acids
0	A C F I L V
-1	A D E F G H I K L M N P Q R S T V W Y
-2	A D E F G H I N Q S T V W Y
-3	A D E F G H I K L N P Q R S T V W Y
-4	A C D E F G H I K L M N P Q R S T V Y

[a]Peptide positions are labelled from 0 to -4 going backwards from the very last amino acid to the fifth last.

Table S2: Filtered numbers of proteins predicted to bind to 1, 2, 3, ... or all PDZ domains of MAGI1 (6 PDZs) or Scribble (4 PDZs)

num. domains	MAGI1	Scribble
1	453	145
2	138	97
3[a]	103	29
4	39	68
5	9	/
6	0	/

[a]e.g. 103 and 29 human proteins were predicted to bind to 3 out of the 6 PDZ domains of MAGI1 and 3 out of the 4 PDZ domains of Scribble, respectively.

name	C-terminal sequence	UniprotID	long name	organism	function linked to PDZ	interactions with PDZ-containing proteins				this study
						protein	PDZ	Kd (μM)	source (PMID)	
16E6	SSRTRRETQL	VE6_HPV16	early protein E6	human papilloma virus 16	Binds and targets human PDZ-domain containing proteins to degradation.	MAGI1	2/6	2.5	Fournane et al., 11571640	3, 9 μM
						MAGI1			19285702	
						GOPC	1/1		16878151	
						PTN3	1/1		17166906, 17947517	
						TIP1	1/1		15492812	
						DLG1	2/3		9326658	
						DLG1	3/3		9326658	
						DLG4	1/3		17121805	
						DLG4	2/3		17121805	
						SCRIB			11027293, 19285702, 18160445	✓
TAX1	SEKHFRETEV	TAX_HTL1A	Trans-activating transcriptional regulatory protein of HTLV-1	HTLV-1A	Interaction with PDZ domain-containing proteins induces IL2-independent growth, which may be a factor in multi-step leukemogenesis. Inhibits the action of at least three cellular tumor suppressors TP53/p53, RB1 and DLG1.	DLG1, MAGI3, TIP1			Uniprot	
						MAGI1	2/6	3	Fournane et al.	53 μM
						Erbin			17633453, 19472191	
						SCRIB			18661220	✓
16E6L/V	SSRTRRETQV		early protein E6 **L158V**	human papilloma virus 16		MAGI1	2/6	0.8	Fournane et al., 11571640	1, 2, 18 μM
						DLG1			9326658, 11571640	
						DLG4			17121805	
						CAL	1/1		16878151	
						PTN3	1/1		17947517	

ABC1	QDEKVKESYV	ABCA1_HUMAN	ATP-binding cassette sub-family A member 1	human	cAMP-dependent and sulfonylurea sensitive anion transporter.	SNTA1			14722086	
						SNTB1, SNTB2			16192269, 12054535	
						ARGHEF11, ARGHEF12, DLG2, DLG3, LIN7A, LIN7B, LIN7C, MPDZ			16192269	
NET1	SGGKRKETLV	ARHG8_HUMAN	Neuroepithelial cell-transforming gene 1 protein	human	Acts as guanine nucleotide exchange factor (GEF) for RhoA GTPase.	DLG1, DLG3, DLG4, LIN7C			17938206	
						MAGI1	2/6	5	Fournane et al., 11350080	3 µM
						MAGI1	3/6	no binding	11350080	binding
PTEN	EDQHTQITKV	PTEN_HUMAN	Phosphatidylinositol-3,4,5-trisphosphate 3-phosphatase and dual-specificity protein phosphatase PTEN	human	Modulates cell cycle progression and cell survival, inhibits cell migration and integrin-mediated cell spreading and focal adhesion formation, synapse formation.	MAGI1	3/6		15629897	✓
						MAGI2	3/6		10760291	
						MAGI3	3/6		10748157	
						DLG1, MAST1, MAST2, MAST3			15951562	
VANG2	VMRLQSETSV	VANG2_HUMAN	Vang-like protein 2	human	Plays a role in the regulation of planar cell polarity	SCRIB	3/4		16687519	✓
						SCRIB	34/4		16687519, 16791850	✓
						SCRIB	2/4	weak	16687519	
						SCRIB	4/4	weak	16687519	
						SCRIB	23/4		16791850	
						DVL1, DVL2, DVL3			15456783	
						MAGI3	2/6		15195140	
ADAM17	NRVDSKETEC	ADA17_HUMAN	Disintegrin and metalloproteinase domain-containing protein 17	human	cleaves membrane-anchored and cell-surface proteins for their activation or degradation	DLG1			18930083	
						DLG1	3/3		12668732	
						PTPH1	1/1		12207026	
FZD4	KPGKGSETVV	FZD4_HUMAN	Frizzled-4	human	may be involved in transduction and intercellular transmission of polarity information during tissue morphogenesis and/or in differentiated tissues	MAGI3	2/6		15195140	

GLAST	EKPIDSETKM	EAA1_HUMAN	Excitatory amino acid transporter 1	human	Essential for terminating the postsynaptic action of glutamate by rapidly removing released glutamate from the synaptic cleft.	NHERF1	1/2	17048262
						NHERF2		20430067
DLL1	KDECVIATEV	DLL1_HUMAN	Delta-like protein 1	human				
ARHGAP6	NPDALPETLV	RHG06_HUMAN	Rho GTPase-activating protein 6	human	Could regulate the interactions of signaling molecules with the actin cytoskeleton			
TANC1	PKRSFIESNV	TANC1_HUMAN	Tetratricopeptide repeat, ankyrin repeat and coiled-coil domain-containing protein 1	human	may be a scaffold component in the postsynaptic density, interacts probably directly with DLG1 and DLG4			
GLUT7	TASPAKETSF	GTR7_HUMAN	Solute carrier family 2, facilitated glucose transporter member 7	human				
TMEM215	QGRWDHETIV	TM215_HUMAN	Transmembrane protein 215	human				
MARCH3	VKRNSKETVV	MARH3_HUMAN	E3 ubiquitin-protein ligase MARCH3	human	Mutational analyses revealed that the PDZ-binding motif and RING finger are essential for the subcellular localization of MARCH-III and the inhibitory effect on transferrin uptake (16428329), MARCH2 binds PDZs of DLG1 (17980554) but has different C-terminus than MARCH3: LKKVAEETPV			
MAS	CNTVTVETVV	MAS_HUMAN	Proto-oncogene Mas	human	Receptor for angiotensin 1-7, belongs to the G-protein coupled receptor 1 family.			
ATP1A1	GGWVEKETYY	AT1A1_HUMAN	Sodium/potassium-transporting ATPase subunit alpha-1	human				
CYSLTR2	SVWLRKETRV	CLTR2_HUMAN	Cysteinyl leukotriene receptor 2	human				
TTC24	PMESGICTIV	TTC24_HUMAN	Tetratricopeptide repeat protein 24	human				

BS = bad signal

NM = not measured

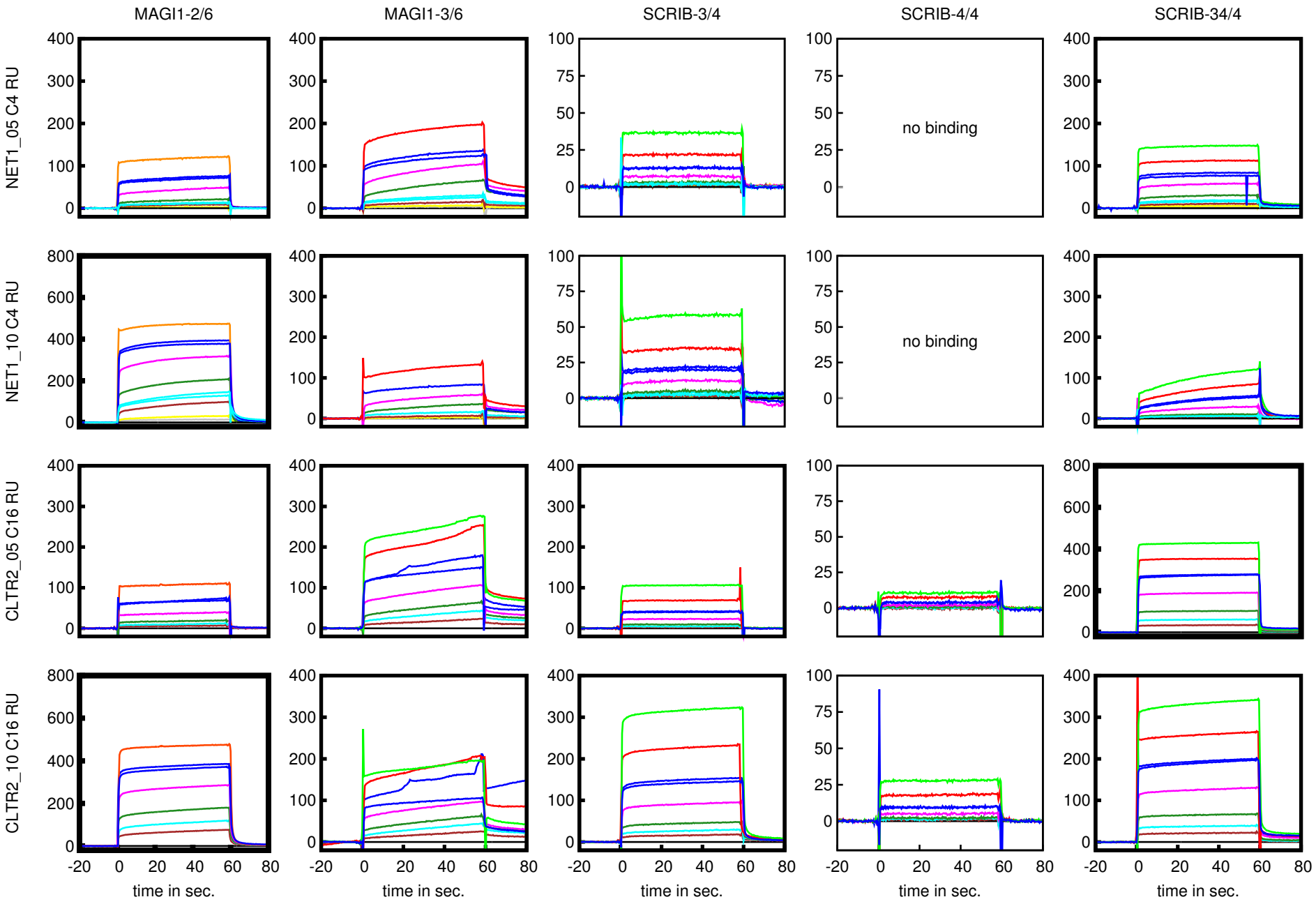
MBP-PDZ analyte	peptide	tentative K_D (µM)	RU at 10 µM, 1st exp.	RU at 10 µM, 2nd exp.	norm. RU at 10 µM
MAGI1-2/6	DLL1_05	32	6.4		3.7
MAGI1-3/6	DLL1_05	246	9.8		5.5
SCRIB-3/4	DLL1_05	200	24.8		15.3
SCRIB-34/4	DLL1_05	NM			
SCRIB-4/4	DLL1_05	893	15.1		9.3
MAGI1-2/6	DLL1_10		0.0	0.0	0.0
MAGI1-3/6	DLL1_10	238	13.0		10.0
SCRIB-3/4	DLL1_10	172	42.8		35.9
SCRIB-34/4	DLL1_10	NM			
SCRIB-4/4	DLL1_10	88	16.8		14.1
MAGI1-2/6	16E6_05	7	37.4		24.3
MAGI1-3/6	16E6_05		0.0	0.0	0.0
SCRIB-3/4	16E6_05	12	225.0		156.7
SCRIB-34/4	16E6_05	1	545.0		317.0
SCRIB-4/4	16E6_05		0.0	0.0	0.0
MAGI1-2/6	16E6L/V	2	801.0		520.6
MAGI1-3/6	16E6L/V	69	62.3		39.8
SCRIB-3/4	16E6L/V	18	133.5		93.0
SCRIB-34/4	16E6L/V	3	477.0		277.5
SCRIB-4/4	16E6L/V	47	12.8		8.9
MAGI1-2/6	ABC1_05	91	23.2	21.4	26.6
MAGI1-3/6	ABC1_05	26	88.9	77.1	97.4
SCRIB-3/4	ABC1_05	229	15.0	13.7	18.3
SCRIB-34/4	ABC1_05	10	63.5	55.9	63.7
SCRIB-4/4	ABC1_05		0.0	0.0	0.0
MAGI1-2/6	ARHGAP6_05	2000	10.5	10.3	12.6
MAGI1-3/6	ARHGAP6_05	58	47.0	46.6	55.9
SCRIB-3/4	ARHGAP6_05	216	12.1	12.0	15.7
SCRIB-34/4	ARHGAP6_05	31	40.3	37.2	42.1
SCRIB-4/4	ARHGAP6_05	104	5.0	5.5	6.8
MAGI1-2/6	ARHGAP6_10	18	20.6	18.4	23.7
MAGI1-3/6	ARHGAP6_10	41	54.5	49.9	62.4
SCRIB-3/4	ARHGAP6_10	541	8.9	7.6	10.7
SCRIB-34/4	ARHGAP6_10	28	8.5	6.8	8.3
SCRIB-4/4	ARHGAP6_10		0.0	0.0	0.0
MAGI1-2/6	ABC1_10	45	13.5	12.7	16.2
MAGI1-3/6	ABC1_10	20	92.9	74.0	101.6
SCRIB-3/4	ABC1_10	107	25.8	23.7	32.8
SCRIB-34/4	ABC1_10	8		54.3	60.2
SCRIB-4/4	ABC1_10		0.0	0.0	0.0
MAGI1-2/6	NET1_05	62		60.9	75.4
MAGI1-3/6	NET1_05	23	110.0	101.0	128.5
SCRIB-3/4	NET1_05	308	12.8	21.7	22.9
SCRIB-34/4	NET1_05	15	83.7	76.9	89.0
SCRIB-4/4	NET1_05		0.0	0.0	0.0
MAGI1-2/6	NET1_10	3	394.0	377.0	432.0
MAGI1-3/6	NET1_10	37	67.9		74.8
SCRIB-3/4	NET1_10	217	21.8	34.9	34.1
SCRIB-34/4	NET1_10	63	29.4	27.0	28.3

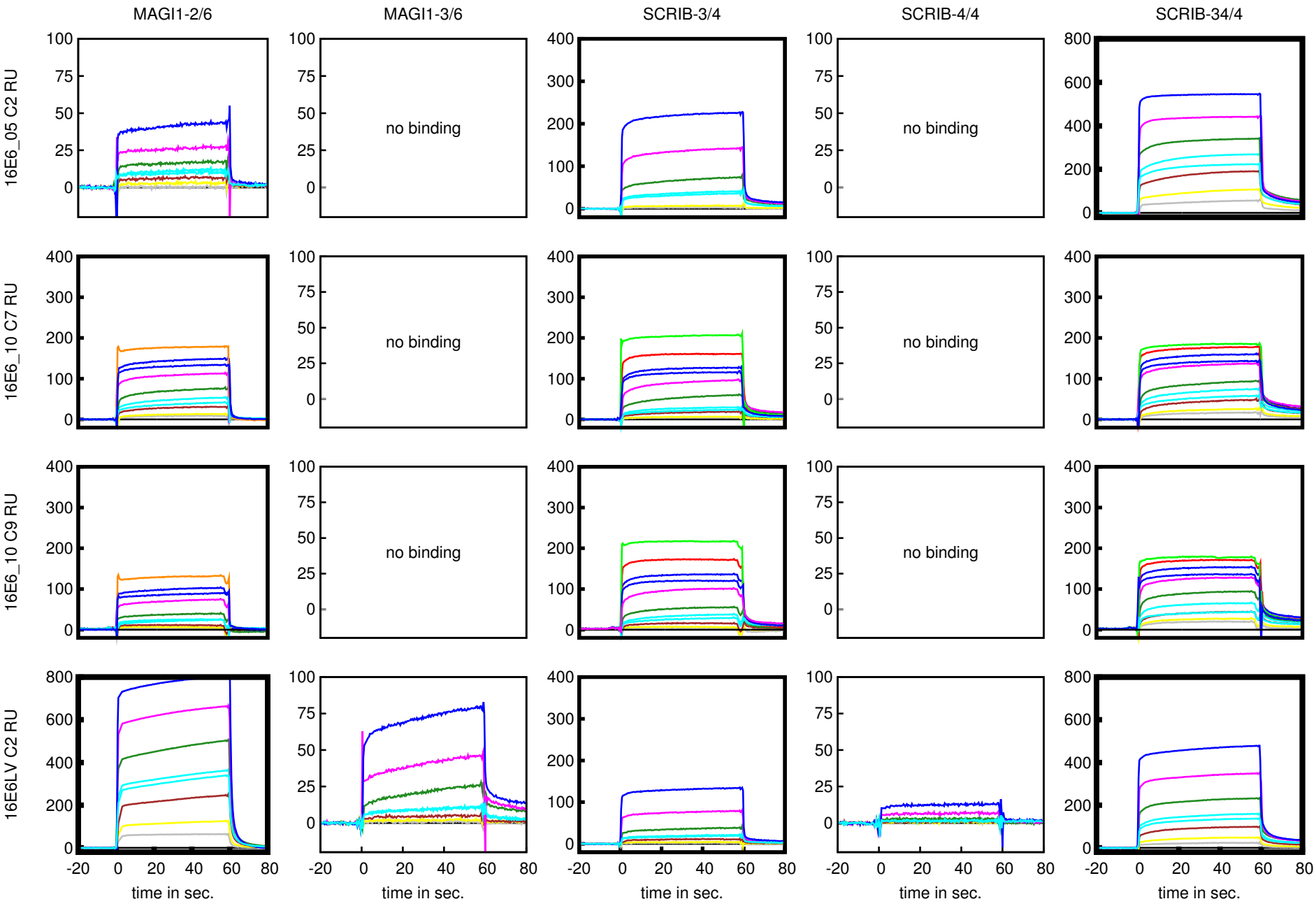
SCRIB-4/4	NET1_10		0.0	0.0	0.0
MAGI1-2/6	PTEN_05		0.0	0.0	0.0
MAGI1-3/6	PTEN_05	453	25.6	24.0	22.6
SCRIB-3/4	PTEN_05		0.0	0.0	0.0
SCRIB-34/4	PTEN_05		0.0	0.0	0.0
SCRIB-4/4	PTEN_05		0.0	0.0	0.0
MAGI1-2/6	PTEN_10		0.0	0.0	0.0
MAGI1-3/6	PTEN_10	144	25.0		23.3
SCRIB-3/4	PTEN_10		0.0	0.0	0.0
SCRIB-34/4	PTEN_10		0.0	0.0	0.0
SCRIB-4/4	PTEN_10		0.0	0.0	0.0
MAGI1-2/6	TANC1_05		0.0	0.0	0.0
MAGI1-3/6	TANC1_05	14	7.6	6.8	5.8
SCRIB-3/4	TANC1_05	498	24.0	23.0	20.7
SCRIB-34/4	TANC1_05	9	127.0	113.0	88.4
SCRIB-4/4	TANC1_05		0.0	0.0	0.0
MAGI1-2/6	TANC1_10	287	9.8	8.4	10.8
MAGI1-3/6	TANC1_10		0.0	0.0	0.0
SCRIB-3/4	TANC1_10	39	67.8	60.1	81.7
SCRIB-34/4	TANC1_10	2	201.0		214.5
SCRIB-4/4	TANC1_10		0.0	0.0	0.0
MAGI1-2/6	TAX1_05	33	10.7	11.8	14.1
MAGI1-3/6	TAX1_05	75	11.1	11.1	13.6
SCRIB-3/4	TAX1_05	96	32.2	30.3	41.9
SCRIB-34/4	TAX1_05	1	213.0		238.3
SCRIB-4/4	TAX1_05	50	26.9	25.6	35.2
MAGI1-2/6	TAX1_10	53	28.1	25.1	34.9
MAGI1-3/6	TAX1_10	1230	6.5	5.6	7.8
SCRIB-3/4	TAX1_10	29	77.4	67.5	102.0
SCRIB-34/4	TAX1_10		197.0		231.5
SCRIB-4/4	TAX1_10	48	24.2	20.5	31.5
MAGI1-2/6	16E6	3	148.0	134.0	179.7
MAGI1-3/6	16E6		0.0	0.0	0.0
SCRIB-3/4	16E6	18	112.0	103.0	146.8
SCRIB-34/4	16E6	3	143.0	129.0	155.1
SCRIB-4/4	16E6		0.0	0.0	0.0
MAGI1-2/6	GLUT7_05		0.0	0.0	0.0
MAGI1-3/6	GLUT7_05		0.0	0.0	0.0
SCRIB-3/4	GLUT7_05	56	33.8	30.8	44.1
SCRIB-34/4	GLUT7_05	5	89.2	81.7	97.5
SCRIB-4/4	GLUT7_05		0.0	0.0	0.0
MAGI1-2/6	GLUT7_10		0.0	0.0	0.0
MAGI1-3/6	GLUT7_10		0.0	0.0	0.0
SCRIB-3/4	GLUT7_10	45	45.4	40.4	57.5
SCRIB-34/4	GLUT7_10	5	66.5	60.0	70.8
SCRIB-4/4	GLUT7_10		0.0	0.0	0.0
MAGI1-2/6	16E6L/V	18	224.0		277.3
MAGI1-3/6	16E6L/V	76	17.9	19.4	22.7
SCRIB-3/4	16E6L/V	41	33.1	27.1	39.9
SCRIB-34/4	16E6L/V	4	105.0	95.9	111.3
SCRIB-4/4	16E6L/V		0.0	0.0	0.0
MAGI1-2/6	16E6	9	90.7		138.7
MAGI1-3/6	16E6		0.0	0.0	0.0

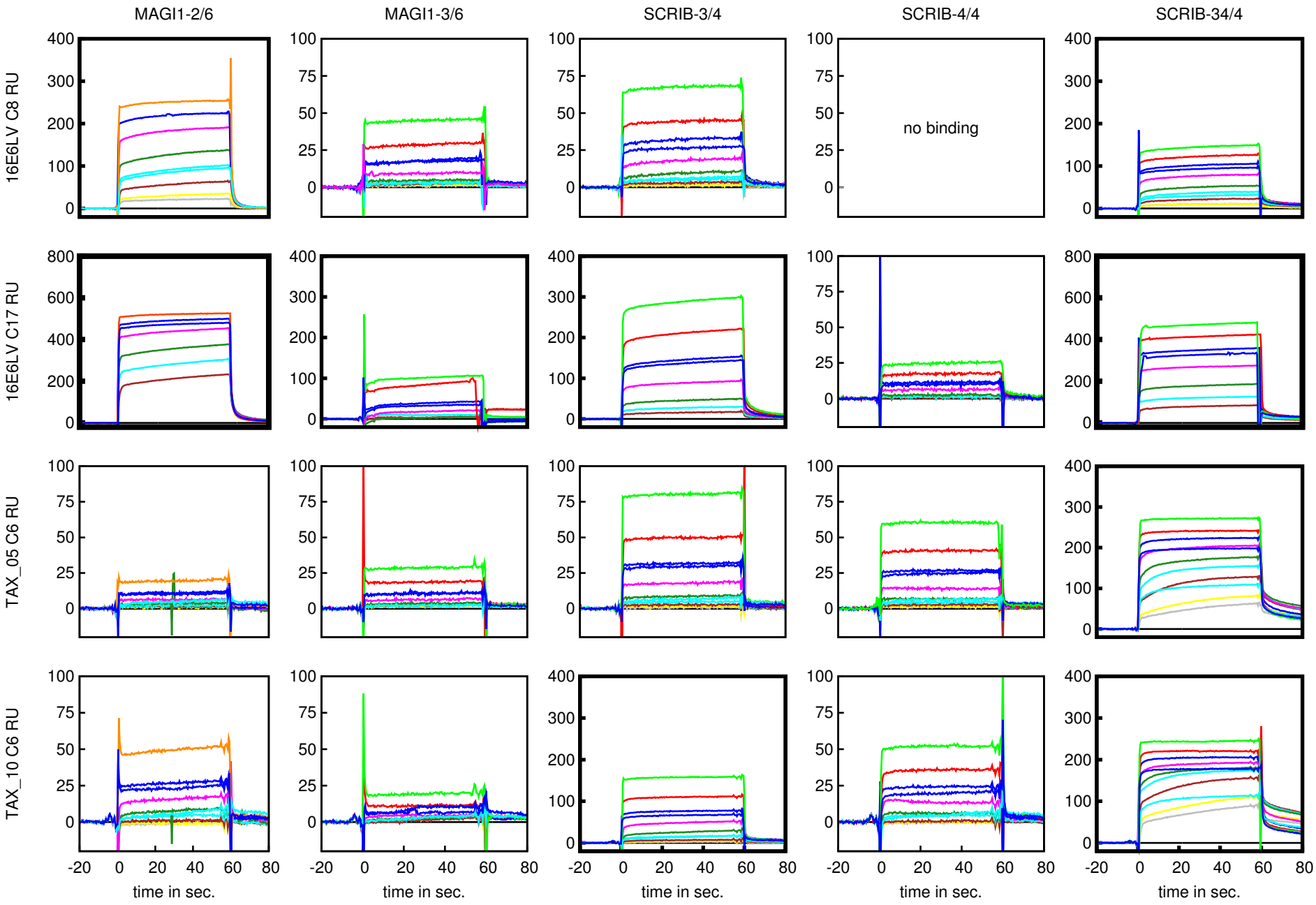
SCRIB-3/4	16E6	15	126.0	113.0	195.9
SCRIB-34/4	16E6	3	144.0	128.0	186.2
SCRIB-4/4	16E6		0.0	0.0	0.0
MAGI1-2/6	TMEM215_05		5.0		6.1
MAGI1-3/6	TMEM215_05	116	13.2	14.6	16.8
SCRIB-3/4	TMEM215_05	213	5.6		7.4
SCRIB-34/4	TMEM215_05	15	44.4	39.3	45.9
SCRIB-4/4	TMEM215_05		0.0	0.0	0.0
MAGI1-2/6	TMEM215_10		0.0	0.0	0.0
MAGI1-3/6	TMEM215_10	67	20.9	21.6	25.9
SCRIB-3/4	TMEM215_10	128	13.2	11.7	16.5
SCRIB-34/4	TMEM215_10	9	54.4	47.5	56.5
SCRIB-4/4	TMEM215_10		0.0	0.0	0.0
MAGI1-2/6	ADAM17_05		0.0	0.0	0.0
MAGI1-3/6	ADAM17_05		0.0	0.0	0.0
SCRIB-3/4	ADAM17_05		0.0	0.0	0.0
SCRIB-34/4	ADAM17_05		0.0	0.0	0.0
SCRIB-4/4	ADAM17_05		0.0	0.0	0.0
MAGI1-2/6	VANG2_05		0.0	0.0	0.0
MAGI1-3/6	VANG2_05		0.0	0.0	0.0
SCRIB-3/4	VANG2_05	24	31.4	30.4	53.1
SCRIB-34/4	VANG2_05	4	61.7	61.1	88.2
SCRIB-4/4	VANG2_05		0.0	0.0	0.0
MAGI1-2/6	VANG2_10		0.0	0.0	0.0
MAGI1-3/6	VANG2_10		0.0	0.0	0.0
SCRIB-3/4	VANG2_10	12	40.2	38.2	63.7
SCRIB-34/4	VANG2_10	5	50.3	48.4	66.9
SCRIB-4/4	VANG2_10		0.0	0.0	0.0
MAGI1-2/6	ADAM17_10		0.0	0.0	0.0
MAGI1-3/6	ADAM17_10		0.0	0.0	0.0
SCRIB-3/4	ADAM17_10	97		17.6	19.8
SCRIB-34/4	ADAM17_10	BS			
SCRIB-4/4	ADAM17_10		0.0	0.0	0.0
MAGI1-2/6	MARCH3_05	225	12.9	13.5	19.7
MAGI1-3/6	MARCH3_05	BS			
SCRIB-3/4	MARCH3_05	46	9.0	9.7	14.9
SCRIB-34/4	MARCH3_05	7	55.7		74.3
SCRIB-4/4	MARCH3_05		0.0	0.0	0.0
MAGI1-2/6	MARCH3_10	103	36.6	35.3	54.7
MAGI1-3/6	MARCH3_10	686	7.8	6.8	10.9
SCRIB-3/4	MARCH3_10	100	14.4	13.4	22.7
SCRIB-34/4	MARCH3_10	16	65.6	63.8	88.1
SCRIB-4/4	MARCH3_10		0.0	0.0	0.0
MAGI1-2/6	FZD4_05	146	12.3	11.5	19.4
MAGI1-3/6	FZD4_05	BS			
SCRIB-3/4	FZD4_05		0.0	0.0	0.0
SCRIB-34/4	FZD4_05		0.0	0.0	0.0
SCRIB-4/4	FZD4_05		0.0	0.0	0.0
MAGI1-2/6	MAS_05	2740	11.3	10.8	16.7
MAGI1-3/6	MAS_05	90	12.7	12.0	18.4
SCRIB-3/4	MAS_05	158	13.6	13.0	21.5
SCRIB-34/4	MAS_05	16	53.1	51.9	71.0
SCRIB-4/4	MAS_05	433	2.3	2.0	3.5

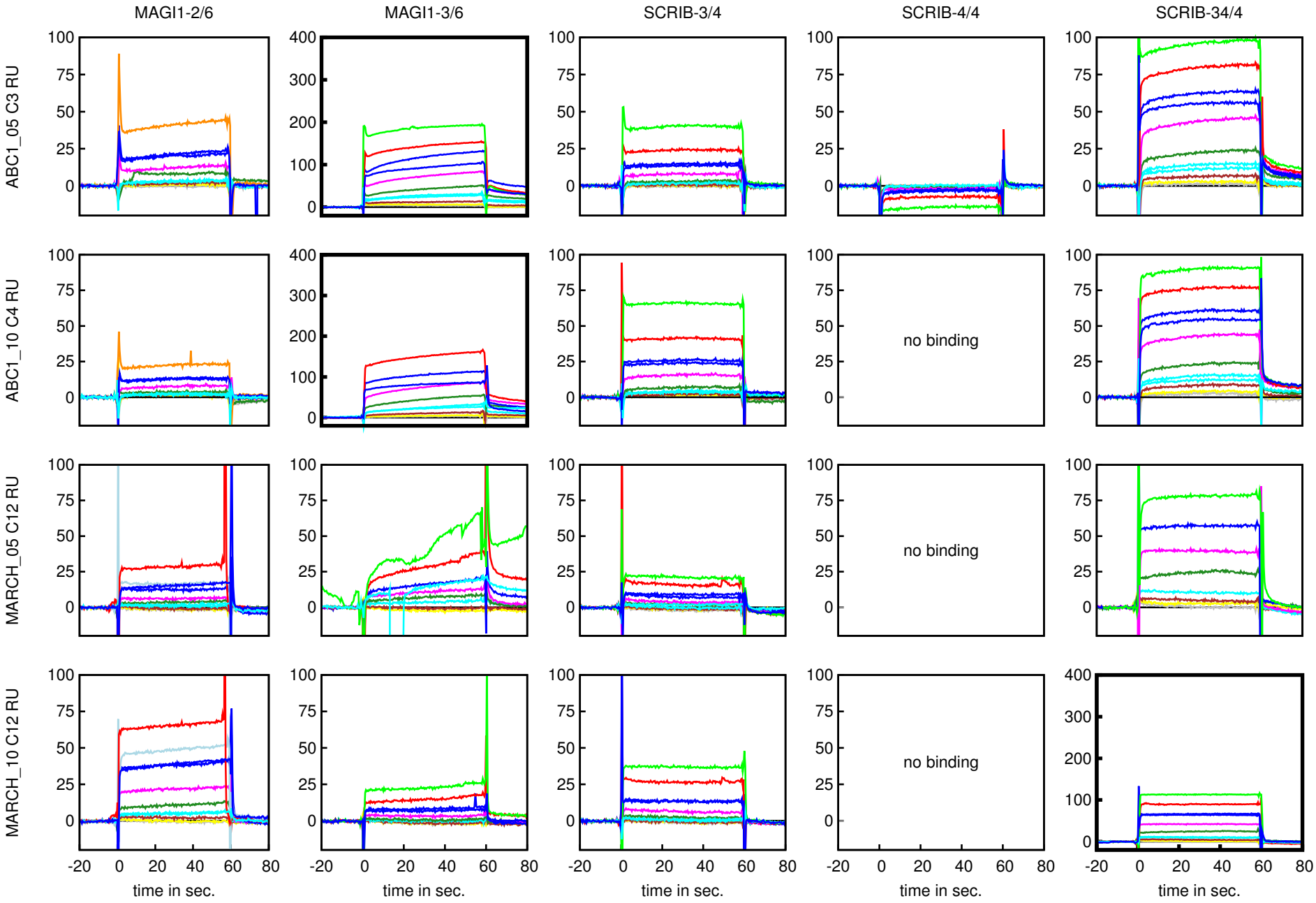
MAGI1-2/6	MAS_10		0.0	0.0	0.0
MAGI1-3/6	MAS_10	123	6.5	5.9	9.7
SCRIB-3/4	MAS_10	73	14.6	13.6	24.1
SCRIB-34/4	MAS_10	3	64.2	60.8	89.1
SCRIB-4/4	MAS_10		0.0	0.0	0.0
MAGI1-2/6	FZD4_10	23	61.2	56.8	92.0
MAGI1-3/6	FZD4_10		0.0	0.0	0.0
SCRIB-3/4	FZD4_10		0.0	0.0	0.0
SCRIB-34/4	FZD4_10		0.0	0.0	0.0
SCRIB-4/4	FZD4_10		0.0	0.0	0.0
MAGI1-2/6	GLAST_05		0.0	0.0	0.0
MAGI1-3/6	GLAST_05		0.0	0.0	0.0
SCRIB-3/4	GLAST_05	182	7.7	7.6	12.0
SCRIB-34/4	GLAST_05	179	11.6	11.4	15.0
SCRIB-4/4	GLAST_05		0.0	0.0	0.0
MAGI1-2/6	GLAST_10		0.0	0.0	0.0
MAGI1-3/6	GLAST_10		0.0	0.0	0.0
SCRIB-3/4	GLAST_10		0.0	0.0	0.0
SCRIB-34/4	GLAST_10		0.0	0.0	0.0
SCRIB-4/4	GLAST_10		0.0	0.0	0.0
MAGI1-2/6	ADAM17_10		0.0	0.0	0.0
MAGI1-3/6	ADAM17_10		0.0	0.0	0.0
SCRIB-3/4	ADAM17_10	79	33.4	30.7	22.0
SCRIB-34/4	ADAM17_10	8	78.0	76.3	44.3
SCRIB-4/4	ADAM17_10		0.0	0.0	0.0
MAGI1-2/6	ATP1A1_05		0.0	0.0	0.0
MAGI1-3/6	ATP1A1_05		0.0	0.0	0.0
SCRIB-3/4	ATP1A1_05	1280	3.1	3.4	2.5
SCRIB-34/4	ATP1A1_05		0.0	0.0	0.0
SCRIB-4/4	ATP1A1_05		0.0	0.0	0.0
MAGI1-2/6	ATP1A1_10		0.0	0.0	0.0
MAGI1-3/6	ATP1A1_10		0.0	0.0	0.0
SCRIB-3/4	ATP1A1_10	88	5.8	4.9	3.5
SCRIB-34/4	ATP1A1_10		0.0	0.0	0.0
SCRIB-4/4	ATP1A1_10		0.0	0.0	0.0
MAGI1-2/6	CYSLTR2_05	30	68.2	72.9	38.9
MAGI1-3/6	CYSLTR2_05	21	121.0	120.0	65.4
SCRIB-3/4	CYSLTR2_05	103	41.8	40.2	24.2
SCRIB-34/4	CYSLTR2_05	10	278.0	276.0	136.8
SCRIB-4/4	CYSLTR2_05	123	4.0	3.9	2.3
MAGI1-2/6	CYSLTR2_10	4	385.0	370.0	229.3
MAGI1-3/6	CYSLTR2_10	BS			
SCRIB-3/4	CYSLTR2_10	34	153.0	145.0	97.0
SCRIB-34/4	CYSLTR2_10	15	199.0	196.0	107.4
SCRIB-4/4	CYSLTR2_10	202	9.8	9.8	6.4
MAGI1-2/6	TTC24_05	65	11.5	12.9	6.7
MAGI1-3/6	TTC24_05	34	20.2	25.1	12.3
SCRIB-3/4	TTC24_05	239	11.3	11.1	6.6
SCRIB-34/4	TTC24_05	25	87.8	87.2	43.3
SCRIB-4/4	TTC24_05		0.0	0.0	0.0
MAGI1-2/6	16E6L/V	1	475.0	457.0	368.0
MAGI1-3/6	16E6L/V	202	42.3		32.9
SCRIB-3/4	16E6L/V	32	136.0	127.0	111.3

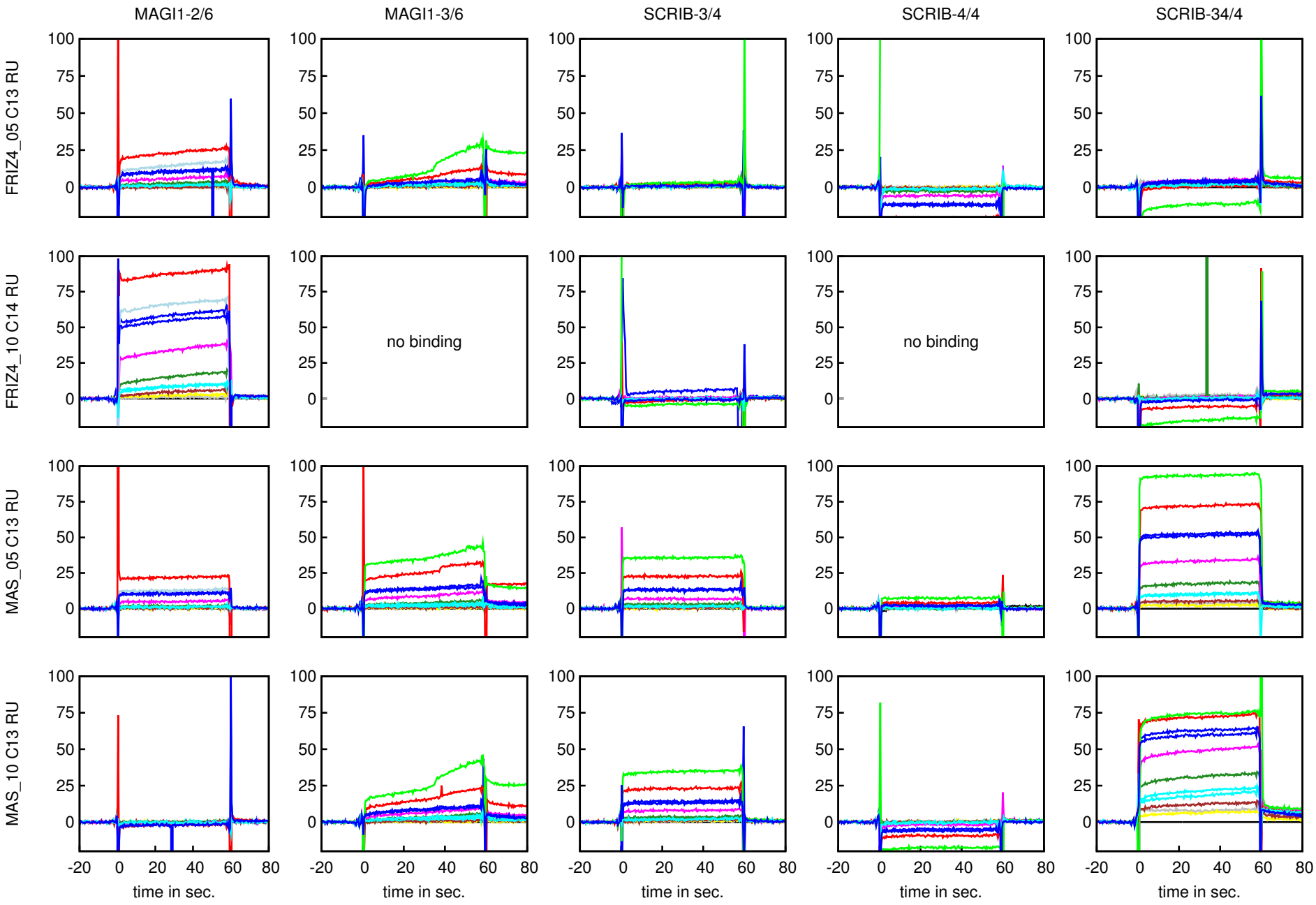
SCRIB-34/4	16E6L/V	4	339.0	317.0	231.8
SCRIB-4/4	16E6L/V	55	11.7	10.3	9.3
MAGI1-2/6	TTC24_10		0.0	0.0	0.0
MAGI1-3/6	TTC24_10	37	20.0	18.5	12.4
SCRIB-3/4	TTC24_10	117	17.1	16.2	11.7
SCRIB-34/4	TTC24_10	49	39.3	39.6	23.1
SCRIB-4/4	TTC24_10		0.0	0.0	0.0

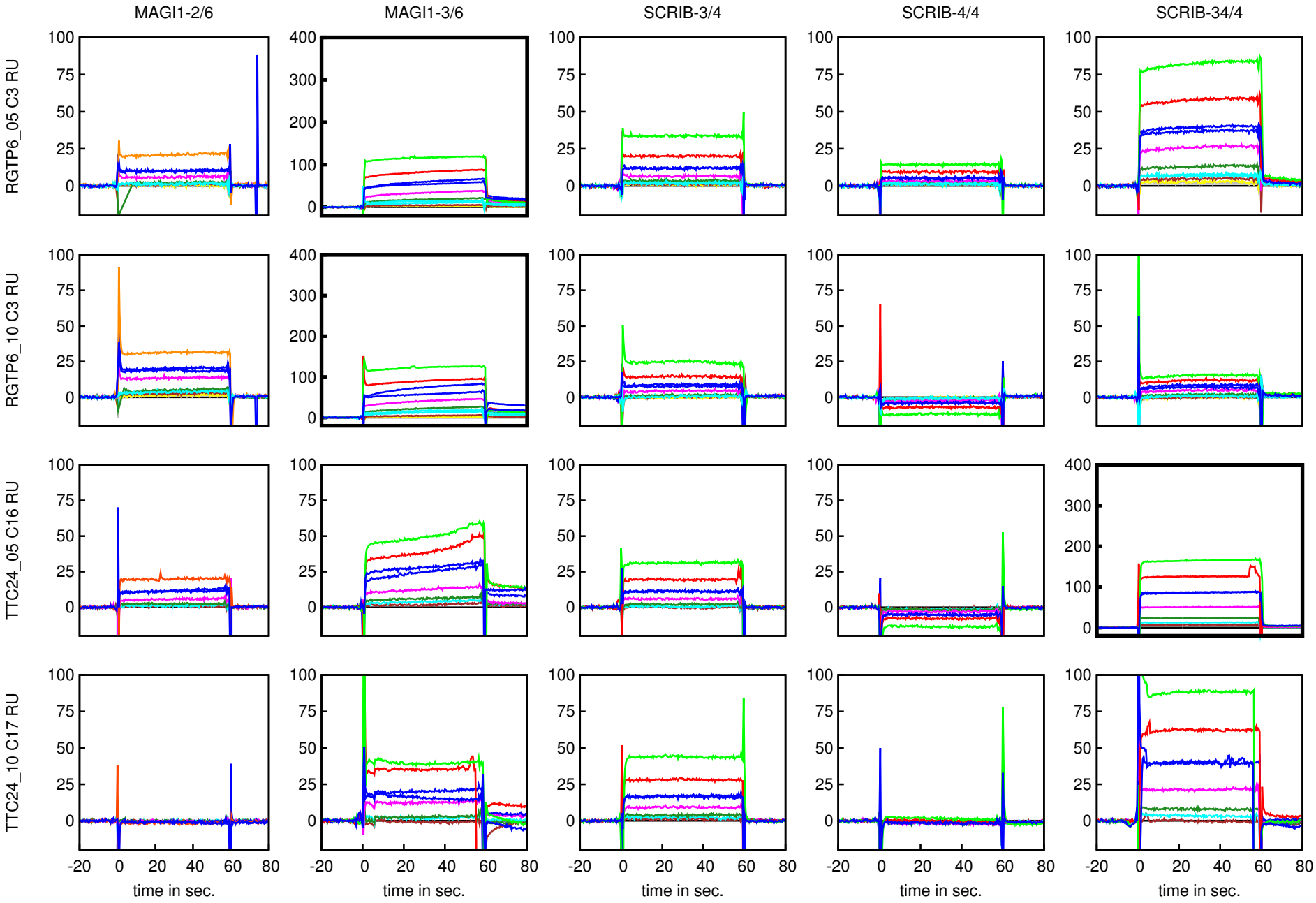

MAGI1-2/6
MAGI1-3/6
SCRIB-3/4
SCRIB-4/4
SCRIB-34/4
NET1_05 C4 RU
NET1_10 C4 RU
CLTR2_05 C16 RU
CLTR2_10 C16 RU
no binding
no binding
time in sec.
time in sec.
time in sec.
time in sec.
time in sec.

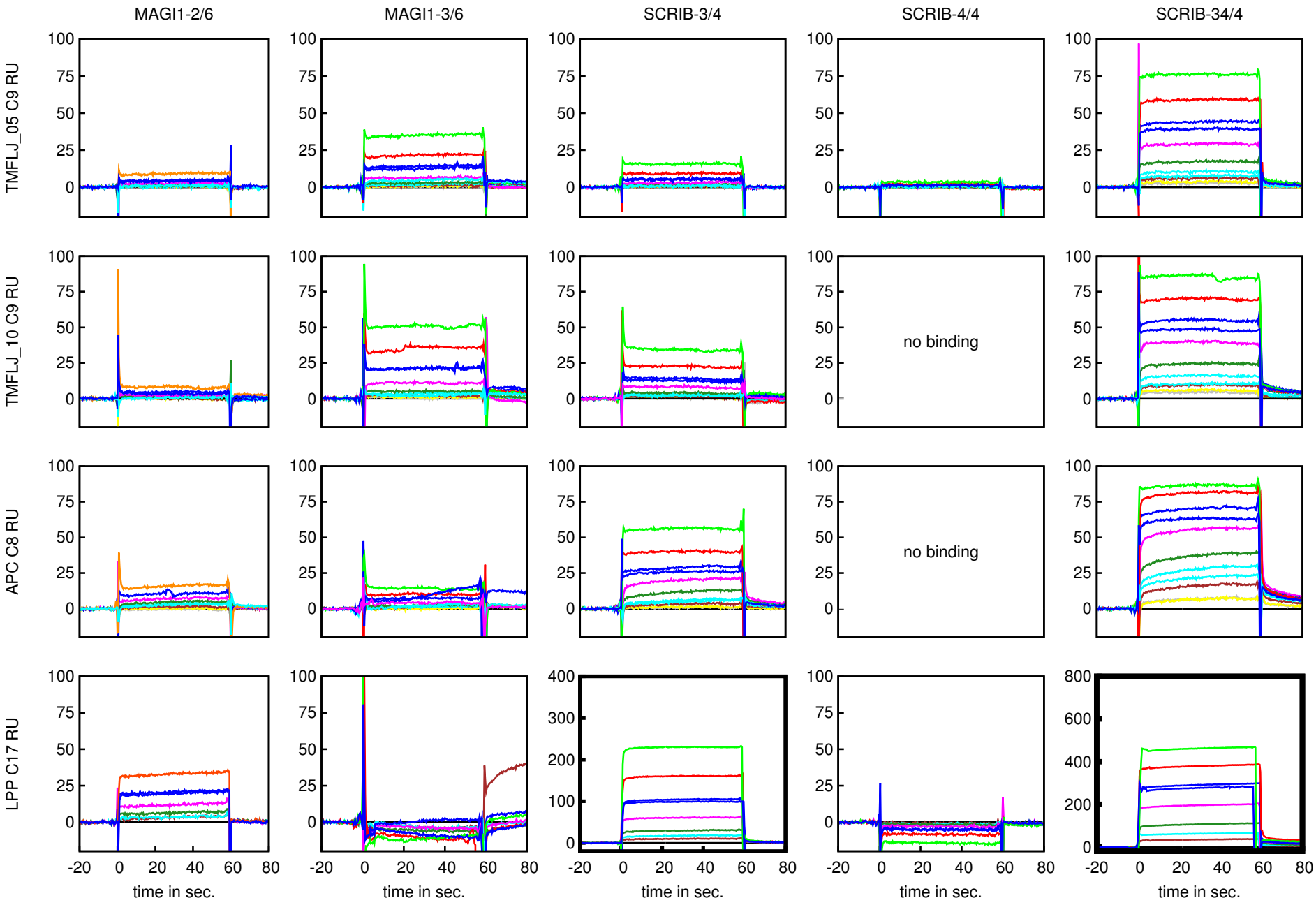

MAGI1-2/6
MAGI1-3/6
SCRIB-3/4
SCRIB-4/4
SCRIB-34/4
16E6_05 C2 RU
16E6_10 C7 RU
16E6_10 C9 RU
16E6LV C2 RU
no binding
time in sec.

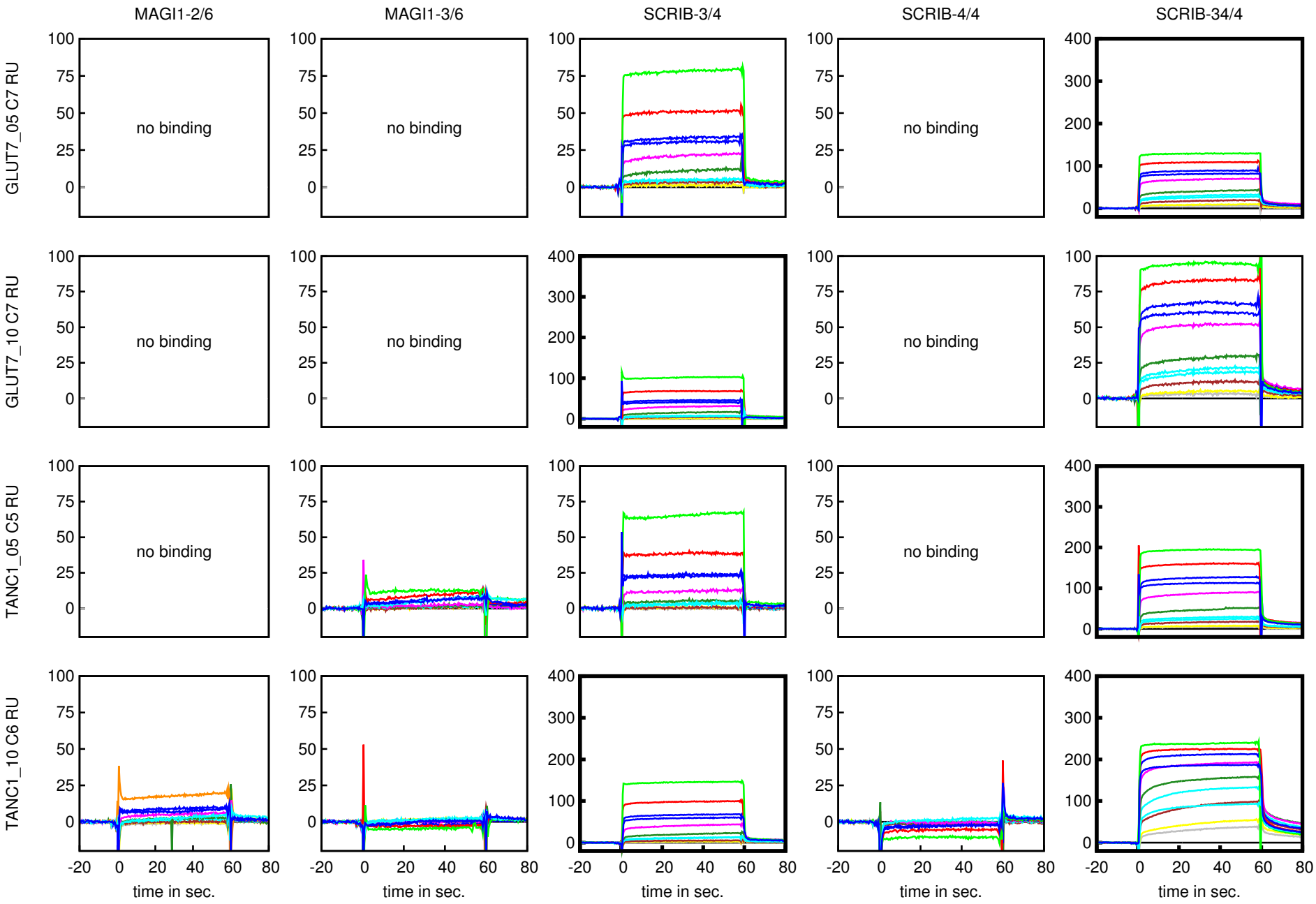

MAGI1-2/6
MAGI1-3/6
SCRIB-3/4
SCRIB-4/4
SCRIB-34/4
16E6LV C8 RU
16E6LV C17 RU
TAX_05 C6 RU
TAX_10 C6 RU
no binding
time in sec.

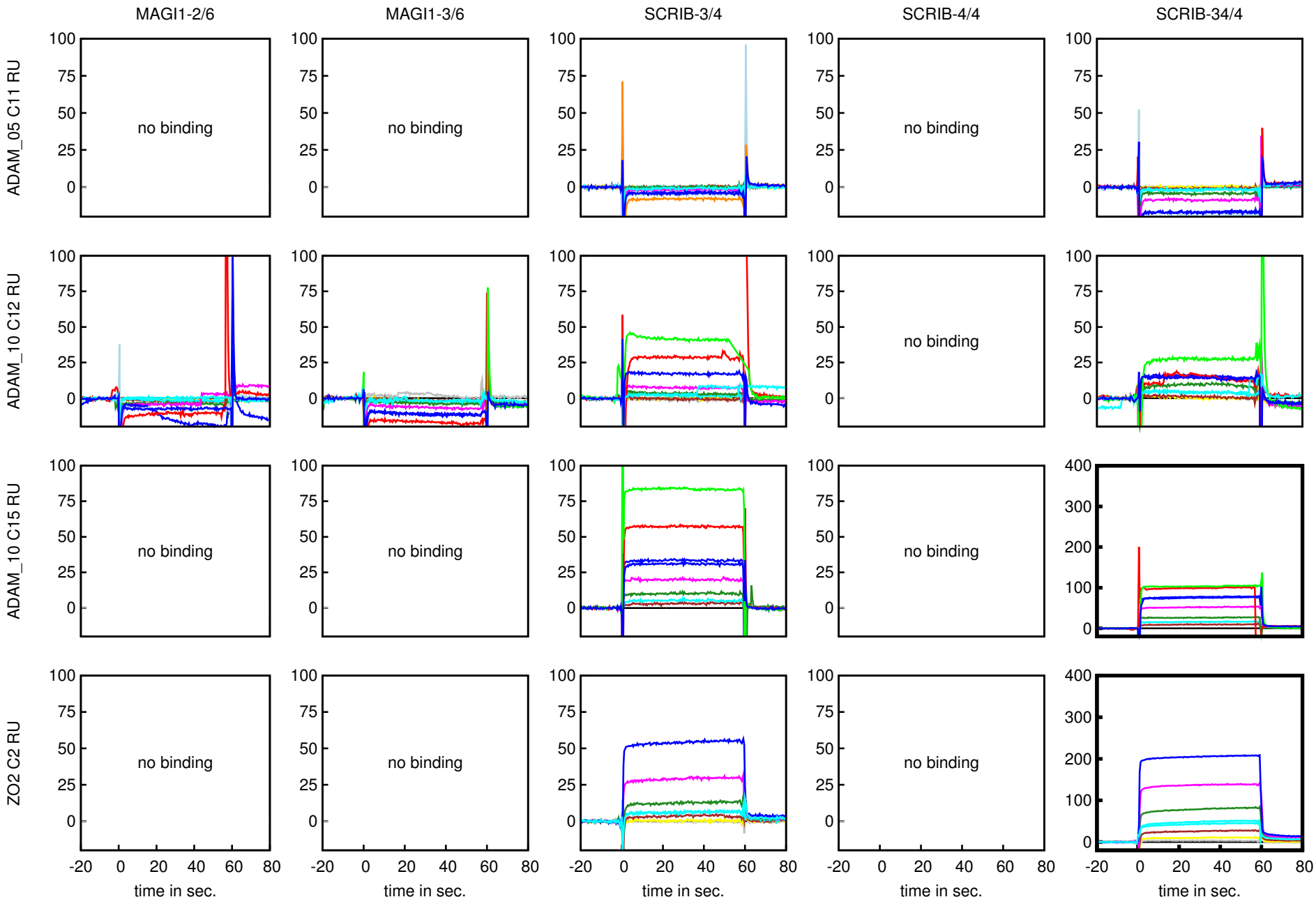

MAGI1-2/6
MAGI1-3/6
SCRIB-3/4
SCRIB-4/4
SCRIB-34/4
ABC1_05 C3 RU
ABC1_10 C4 RU
MARCH_05 C12 RU
MARCH_10 C12 RU
no binding
no binding
no binding
time in sec.
time in sec.
time in sec.
time in sec.
time in sec.

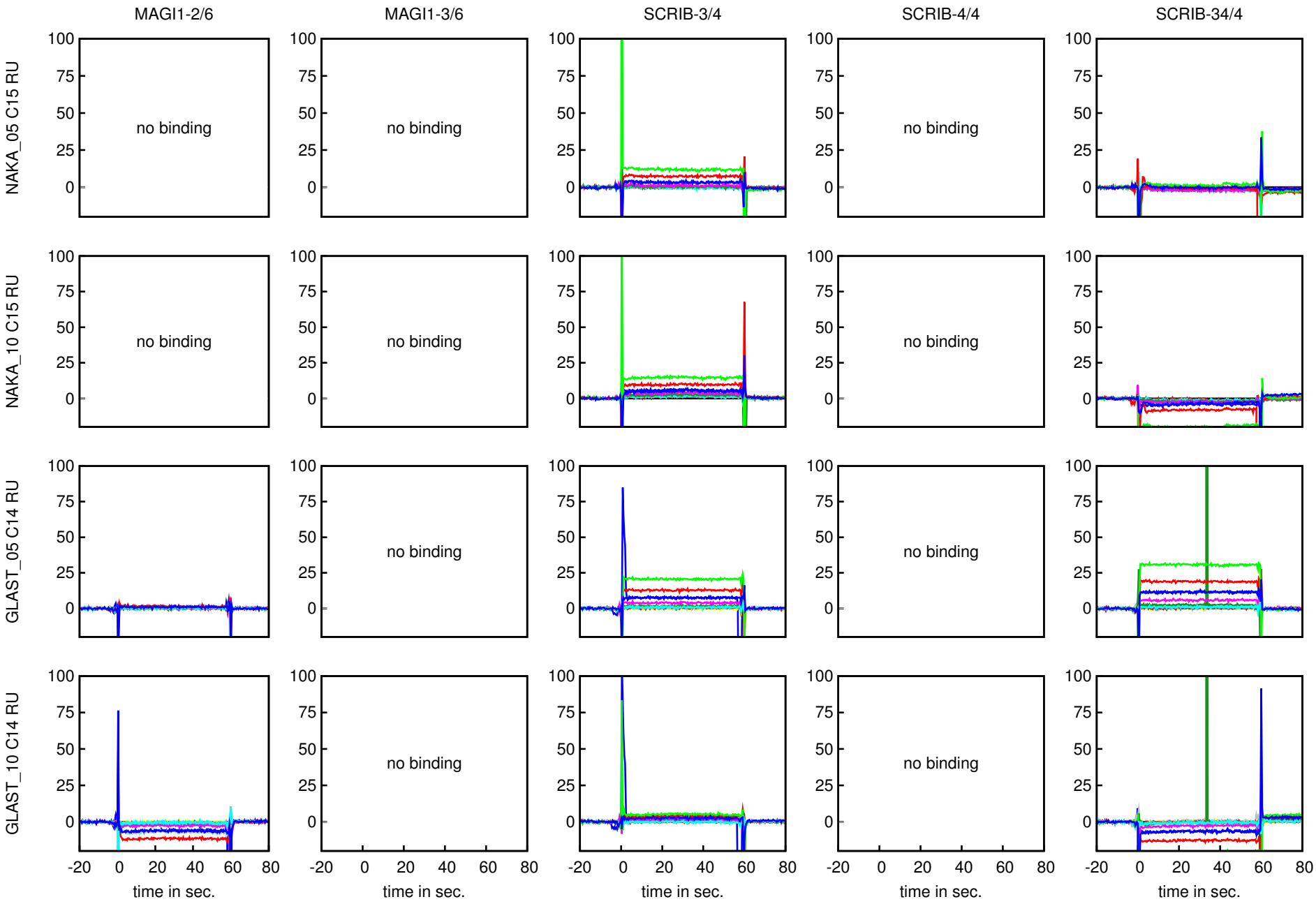

MAGI1-2/6
MAGI1-3/6
SCRIB-3/4
SCRIB-4/4
SCRIB-34/4
FRIZ4_05 C13 RU
FRIZ4_10 C14 RU
MAS_05 C13 RU
MAS_10 C13 RU
no binding
no binding
100
75
50
25
0
-20
0
20
40
60
80
time in sec.

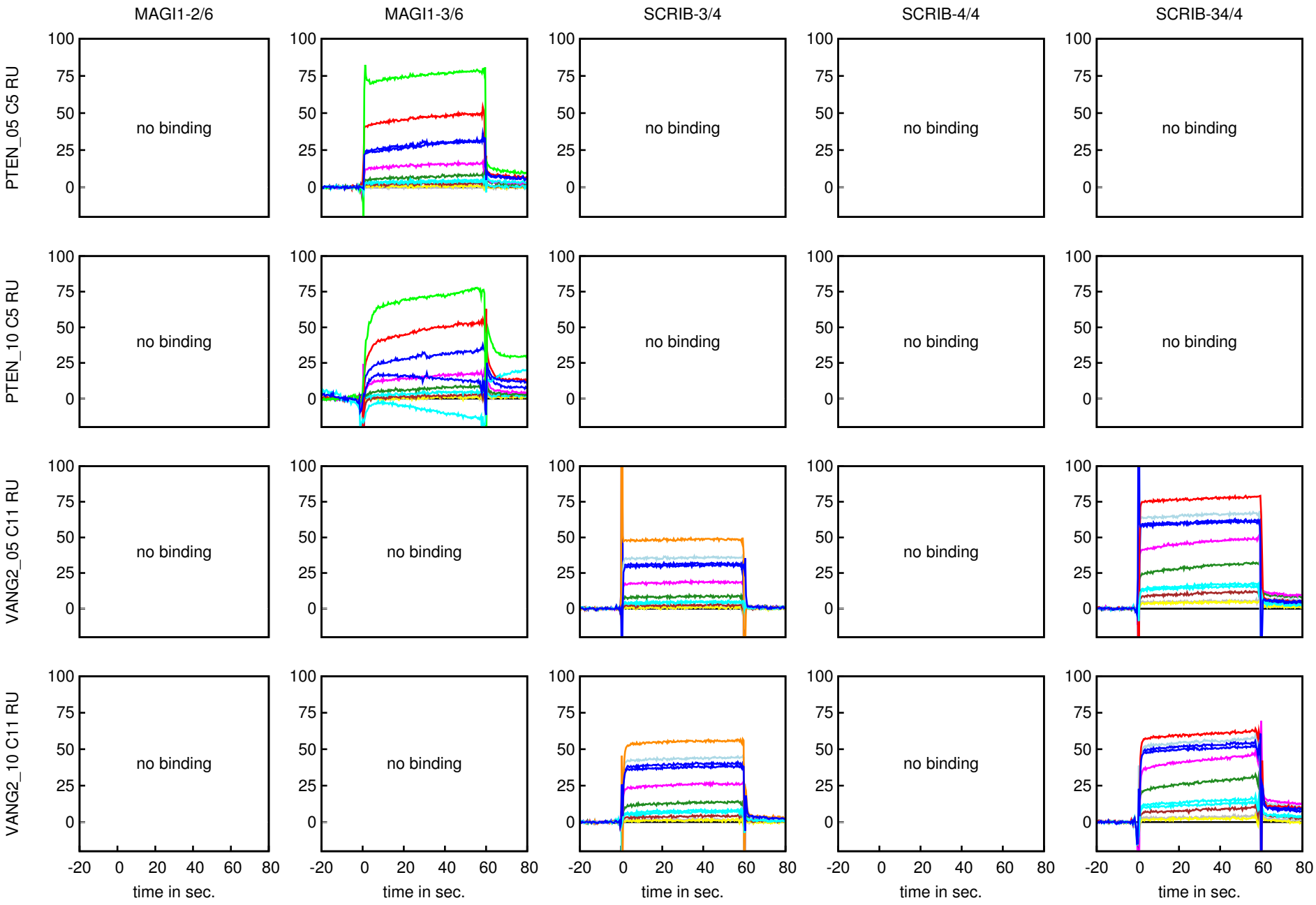

MAGI1-2/6
MAGI1-3/6
SCRIB-3/4
SCRIB-4/4
SCRIB-34/4
RGTP6_05 C3 RU
RGTP6_10 C3 RU
TTC24_05 C16 RU
TTC24_10 C17 RU
time in sec.

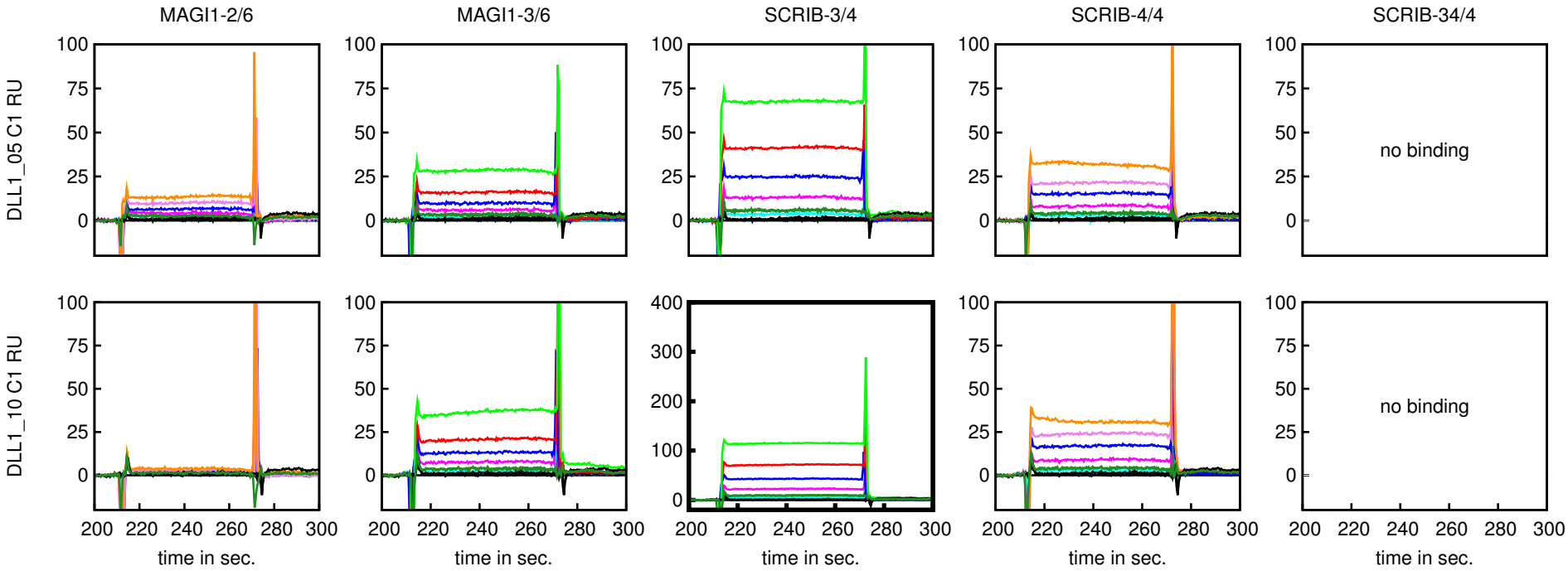

MAGI1-2/6
MAGI1-3/6
SCRIB-3/4
SCRIB-4/4
SCRIB-34/4
TMFLJ_05 C9 RU
TMFLJ_10 C9 RU
APC C8 RU
LPP C17 RU
no binding
no binding
time in sec.
time in sec.
time in sec.
time in sec.
time in sec.


MAGI1-2/6
MAGI1-3/6
SCRIB-3/4
SCRIB-4/4
SCRIB-34/4
GLUT7_05 C7 RU
GLUT7_10 C7 RU
TANC1_05 C5 RU
TANC1_10 C6 RU
no binding
time in sec.


MAGI1-2/6
MAGI1-3/6
SCRIB-3/4
SCRIB-4/4
SCRIB-34/4
ADAM_05 C11 RU
ADAM_10 C12 RU
ADAM_10 C15 RU
ZO2 C2 RU
no binding
time in sec.


MAGI1-2/6
MAGI1-3/6
SCRIB-3/4
SCRIB-4/4
SCRIB-34/4
NAKA_05 C15 RU
NAKA_10 C15 RU
GLAST_05 C14 RU
GLAST_10 C14 RU
no binding
100
75
50
25
0
-20
0
20
40
60
80
time in sec.


MAGI1-2/6
MAGI1-3/6
SCRIB-3/4
SCRIB-4/4
SCRIB-34/4
PTEN_05 C5 RU
PTEN_10 C5 RU
VANG2_05 C11 RU
VANG2_10 C11 RU
no binding
100
75
50
25
0
-20
0
20
40
60
80
time in sec.


MAGI1-2/6
MAGI1-3/6
SCRIB-3/4
SCRIB-4/4
SCRIB-34/4
DLL1_05 C1 RU
DLL1_10 C1 RU
no binding
no binding
time in sec.
time in sec.
time in sec.
time in sec.
time in sec.

Text S1: Recommendations for application of the predictor of Chen *et al.* followed by experimental validation

1. filter out predicted peptides with unlikely amino acids at last peptide position (use option -f in command line program provided)
2. check the number of domain positions of your query PDZ domain that have amino acids that did not occur in PDZ domains of the training data of the predictor (information provided in output file of command line program)
3. collect additional information from Uniprot (www.uniprot.org) and Pubmed (www.pubmed.com) for predicted binding protein with regard to biological functions linked to PDZ domains or known interactions with PDZ domains; check the Gene Ontology terms with which the binding protein is annotated (e.g. in Uniprot)
4. do not compare directly scores that a peptide obtained for different PDZ domains to guess the PDZ domain to which the peptide will bind strongest
5. expect that about half of the predicted peptides could be false positives; we suggest to choose at least 10 peptides for experimental validation to increase the likeliness that some binders can be confirmed
6. for experimental validation use peptides with a length of about 10 residues or full length proteins
7. (optionally) check the number of positions in your predicted peptide that have amino acids that did not occur in the training data of the predictor and check the number of position pairs that have amino acids that did not occur in the training data of the predictor (information provided in output file of command line program)

D.3. Supplemental material of the review article (see chapter 10)

Figure S1, construct design protocol, additional references, Table S1.

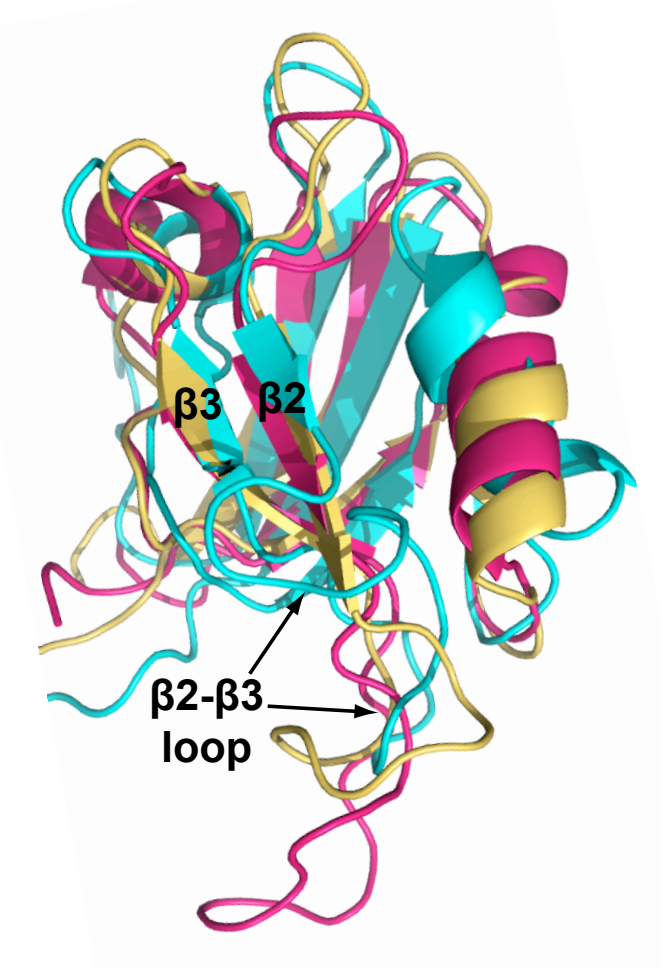


Figure 1. PDZ2 and its alternatively spliced isoform (PDZ2as) of hPTP1E. PDZ2as has an insertion of 5 residues (VLFDK) at the start of the β2-β3 loop. Two available NMR structures of PDZ2as considerably vary in their overall conformation and especially in conformation of the β2-β3 loop. yellow: PDZ2 (PDB ID: 1GM1 [1]); pink: PDZ2as (PDB ID: 1OZI [2]); cyan: PDZ2as (PDB ID: 1Q7X [3]). Figure was created with Pymol [4] (see suppl. data for the pymol session file).

Precision of the total number of PDZ domains in the human proteome - additional references.

Non-exhaustive list of articles that specify the total number of human PDZs of being about 250 but without providing the source for this information: [5, 6].
Non-exhaustive list of articles that specify the total number of human PDZs of being about 400 or more but without providing the source for this information: [7–11].
Non-exhaustive list of articles that specify the total number of human PDZs of being about 400 or more and indicate SMART as the source for this information: [12–16].

Design of experimental constructs for 266 human PDZ domains

Table S1:
The table contains for each of the 266 human PDZ domains an entry providing the sequence of the proposed construct and additional information. Column 1 and column 2: Commonly used names for the PDZ-domain containing protein. Column 3: Index of the PDZ domain in the protein. Column 4: PDB code of the structure that was used for construct design. Column 5: Sequence similarity between the PDZ domain in question and the PDZ domain of the structure (column 4) based on the local alignment produced by BLAST [17]. Column 6 and 7: Number of residues that the proposed construct is longer at the N- and C-terminus than the domain sequence predicted by SMART [18] (negative numbers indicate the number of residues that the proposed construct was shorter than the sequence returned by SMART). Column 8: Sequence of proposed construct.

Protocol of construct design:
Using the SMART web service [18], we searched the human proteome (from Ensembl [19] taking per gene the longest isoform) for hits of the Hidden Markov Models from SMART and Pfam [20] representing the PDZ domain. This revealed in total 265 PDZ domains. This list was compared to the recently published list of 267 and 269 human PDZ domains from Velthuis *et al.* [21] and Wang *et al.* [22], respectively. Redundancies, suspicious variants, and fragments of the three united data sets were removed, resulting in a final list of 266 human PDZ domains from 151 proteins.
For each of these PDZ sequences a BLAST search was performed against the sequences of the structures stored in the PDB [23] using the PDB web services (this search was performed in 2009). The three most similar structures reported together with the DSSP prediction of their secondary structure elements were considered for construct definition. If available, NMR structures were preferred as the reported structures were observed in solution. The PDZ sequence was aligned to the sequences of the structures and together with structural information and secondary structure predictions used to define the boundaries of the core PDZ domain in question. The previous steps were automatically performed using python (www.python.org) and biopython scripts [24]. The last step and

the following ones were manually carried out.
We compared the core domain boundaries to the boundaries of the constructs used to obtain the structures. If there was a structure with a highly similar sequence to the PDZ in question and if the construct boundaries of the structure comprised our defined core domain boundaries, we adopted the construct boundaries from the structure including eventual extensions. If there were only structures with less sequence similarity to the PDZ in question, eventual extensions present in the structure were not included, especially, if they seemed not to be conserved. In the latter case or if the boundaries observed from the structure were shorter than our predicted core domain boundaries, the final construct was designed by extending our defined core domain boundaries on both sides by a few residues, if they were not hydrophobic or if they seemed to extend the first and last β strand of the PDZ.

References

[1] Walma, T., Spronk, C. A. E. M., Tessari, M., Aelen, J., Schepens, J., Hendriks, W., and Vuister, G. W. Mar 2002 *J Mol Biol* **316(5)**, 1101–1110.

[2] Walma, T., Aelen, J., Nabuurs, S. B., Oostendorp, M., van denBerk, L., Hendriks, W., and Vuister, G. W. Jan 2004 *Structure* **12(1)**, 11–20.

[3] Kachel, N., Erdmann, K. S., Kremer, W., Wolff, P., Gronwald, W., Heumann, R., and Kalbitzer, H. R. Nov 2003 *J Mol Biol* **334(1)**, 143–155.

[4] DeLano, W. L. The PyMOL molecular graphics system (2002).

[5] Ernst, A., Sazinsky, S. L., Hui, S., Currell, B., Dharsee, M., Seshagiri, S., Bader, G. D., and Sidhu, S. S. (2009) *Sci Signal* **2(87)**, ra50.

[6] Ernst, A., Gfeller, D., Kan, Z., Seshagiri, S., Kim, P. M., Bader, G. D., and Sidhu, S. S. Oct 2010 *Mol Biosyst* **6(10)**, 1782–1790.

[7] Harris, B. Z. and Lim, W. A. Sep 2001 *J Cell Sci* **114(Pt 18)**, 3219–3231.

[8] Zhang, M. and Wang, W. Jul 2003 *Acc Chem Res* **36(7)**, 530–538.

[9] Kim, E. and Sheng, M. Oct 2004 *Nat Rev Neurosci* **5(10)**, 771–781.

[10] Kurakin, A., Swistowski, A., Wu, S. C., and Bredesen, D. E. (2007) *PLoS ONE* **2(9)**, e953.

[11] Joo, S. H. and Pei, D. Mar 2008 *Biochemistry* **47(9)**, 3061–3072.

[12] Jeleń, F., Oleksy, A., Smietana, K., and Otlewski, J. (2003) *Acta Biochim Pol* **50(4)**, 985–1017.

[13] vanHam, M. and Hendriks, W. Jun 2003 *Mol Biol Rep* **30(2)**, 69–82.

[14] Nourry, C., Grant, S. G. N., and Borg, J.-P. Apr 2003 *Sci STKE* **2003(179)**, RE7.

[15] Kay, B. K. and Kehoe, J. W. Apr 2004 *Chem Biol* **11(4)**, 423–425.

[16] Beuming, T., Farid, R., and Sherman, W. Aug 2009 *Protein Sci* **18(8)**, 1609–1619.

[17] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Oct 1990 *J Mol Biol* **215(3)**, 403–410.

[18] Letunic, I., Doerks, T., and Bork, P. Jan 2012 *Nucleic Acids Res* **40(Database issue)**, D302–D305.

[19] Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Khri, A. K., Keefe, D., Keenan, S., Kinsella, R., Komorowska, M., Koscielny, G., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Muffato, M., Overduin, B., Pignatelli, M., Pritchard, B., Riat, H. S., Ritchie, G. R. S., Ruffier, M., Schuster, M., Sobral, D., Tang, Y. A., Taylor, K., Trevanion, S., Vandrovcova, J., White, S., Wilson, M., Wilder, S. P., Aken, B. L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernndez-Suarez, X. M., Harrow, J., Herrero, J., Hubbard, T. J. P., Parker, A., Proctor, G., Spudich, G., Vogel, J., Yates, A., Zadissa, A., and Searle, S. M. J. Jan 2012 *Nucleic Acids Res* **40(Database issue)**, D84–D90.

[20] Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., Bateman, A., and Finn, R. D. Jan 2012 *Nucleic Acids Res* **40(Database issue)**, D290–D301.

[21] Velthuis, A. J. W. T., Sakalis, P. A., Fowler, D. A., and Bagowski, C. P. (2011) *PLoS One* **6(1)**, e16047.

[22] Wang, C. K., Pan, L., Chen, J., and Zhang, M. Aug 2010 *Protein Cell* **1(8)**, 737–751.

[23] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. Jan 2000 *Nucleic Acids Res* **28(1)**, 235–242.

[24] Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and deHoon, M. J. L. Jun 2009 *Bioinformatics* **25(11)**, 1422–1423.

gene name	gene name	nr of PDZ in protein	PDB ID	BLAST % similarity	proposed construct
AFAD	AF6	1	1T2M	100	RKEPEIITVTLKKQNGMGLSIVAAKGAGQDKLGIYVKSVVKGGAADVDGRLAAGDQLLSVDGRSLVGLSQERAAELMTRTSSVVTLEVAKQGAIYH
AHNAK	AHNAK	1	2QBW	35.14	EKEETTRELLLPNWQGSGSHGLTIAQRDDGVFVQEVTQNSPAARTGVVKEGDQIVGATIYFDNLQSGEVTQLLNTMGHHTVGLKLHRKGDRSPEPGQT
AHNAK2	AHNAK2	1	2EHR	38.3	QEATEVTLKTEVEAGASGYSVTGGGDQGIFVKQVLKDSSAAKLFNLREGDQLLSTTVFFENIKYEDALKILQYSEPYKVQFKIRRQ
APBA1	MINT1	1	1U37	100	EFKDVFIEKQKGEILGVVIVESGWGSILPTVIIANMMHGGPAEKSGKLNIGDQIMSINGTSLVGLPLSTCQSIIKGLKNQSRVKLNIVR
APBA1	MINT1	2	1U39	100	PPVTTVLIRRPDLRYQLGFSVQNGIICSLMRGGIAERGGVRVGHRIIEINGQSVVATPHEKIVHILSNAVGEIHMKTMPA
APBA2	MINT2	1	1U38	82.83	NCKELQLEKHKGEILGVVVVESGWGSILPTVILANMMNGGPAARSGKLSIGDQIMSINGTSLVGLPLATCQGIIKGLKNQTQVKLNIVS
APBA2	MINT2	2	1Y7N	90.7	PPVTTVLIKRPDLKYQLGFSVQNGIICSLMRGGIAERGGVRVGHRIIEINGQSVVATAHEKIVQALSNSVGEIHMKTMPAA
APBA3	MINT3	1	2YT7	100	DNCREVHLEKRRGEGLGVALVESGWGSLLPTAVIANLLHGGPAERSGALSIGDRLTAINGTSLVGLPLAACQAAVRETKSQTSVTLSIVHS
APBA3	MINT3	2	2YT8	100	PVTTAIIHRPHAREQLGFCVEDGIICSLLRGGIAERGGIRVGHRIIEINGQSVVATPHARIIELLTEAYGEVHIKTMPAATYRLLTGQ
APXL	APXL	1	2EDP	62.2	DGGRLVEVQLSGGAPWGFTLKGGREHGEPLVITKIEEGSKAAAVDKLLAGDEIVGINDIGLSGFRQEAICLVKGSHKTLKLVVKRRSE
ARHGAP21	ARHGAP10	1	2YUY	100	GPKTVTLKRTSQGFGFTLRHFIVYPPESAIQFSYKDEENGNRGGKQRNRLEPMDTIFVKQVKEGGPAFEAGLCTGDRIIKVNGESVIGKTYSQVIALIQNSDTTLELSVMPKD
ARHGAP23	ARHGAP23	1	2YUY	85.07	QGPRTLLLYKSPQDGFGFTLRHFIVYPPESAVHCSLKEEENGGRGGGPSPRYRLEPMDTIFVKNVKEDGPAHRAGLRTGDRLVKVNGESVIGKTYSQVIALIQNSDDTLELSIMPKD
ARHGEF11	PDZ-RHOGEF	1	2DLS	100	TGLVQRCVIIQKDQHGFGFTVSGDRIVLVQSVRPGGAAMKAGVKEGDRIIKVNGTMVTNSSHLEVVKLIKSGAYVALTLLGSS
ARHGEF12	RhoGEF12	1	2OMJ	100	TGLVQRCVIIQKDDNGFGLTVSGDNPVFVQSVKEDGAAMRAGVQTGDRIIKVNGTLVTHSNHLEVVKLIKSGSYVALTVQGRPPGS
Carma1	CARD11	1	1UIT	50	RGPGPSVQHTTLNGDSLTSQLTLLGGNARGSFVHSVKPGSLAEKAGLREGHQLLLLEGCIRGERQSVPLDTCTKEEAHWTIQRCSGPVTLHYKVNHE
Carma2	CARD14	1	1UIT	53.57	RRRPARRILSQVTMLAFQGDALLEQISVIGGNLTGIFIHRVTPGSAADQMALRPGTQIVMVDYEASEPLFKAVLEDTTLEEAVGLLRRVDGFCCLSVKVNTDGYKR
CASK	CASK	1	1KWA	100	RVRLVQFQKNTDEPMGITLKMNELNHCIVARIMHGGMIHRQGTLHVGDEIREINGISVANQTVEQLQKMLREMRGSITFKIVPS
CNKSR1	CNK1	1	1UEW	24	EQKAVLEQVQLDSPLGLEIHTTSNCQHFVSQVDTQVPTDSRLQIQPGDEVVQINEQVVVREERDMVGWPRKNMVRELLREPAGLSLVLKK
CNKSR2	CNKSR2	1	1N7F	31.87	SQSAHLEVIQLANIKPSEGLGMYIKSTYDGLHVITGTTENSPADRCKKIHAGDEVIQVNHQTVVGWQLKNLVNALREDPSGVILTLKKRPQSMLTSAP
CNKSR3	CNKSR3	1	2DLU	28.42	MSQCACLEEVHLPNIKPGEGLGMYIKSTYDGLHVITGTTENSPADRSQKIHAGDEVIQVNQQTVVGWQLKNLVKKLRENPTGVVLLLKKRPTGSFNFTPAP
DEPDC2	PREX2	1	1I92	35.06	HDLKVVENVIAKSLLIKSNEGSYGFGLEDKNKVPIIKLVEKGSNAEMAGMEVGKKIFAINGDLVFMRPFNEVDCFLKSCLNSRKPLRVLVSTKPRE
DEPDC2	PREX2	2	1GQ5	44.83	PLRVLVSTKPRETVKIPDSADGLGFQIRGFGPSVVHAVGRGTVAAAAGLHPGQCIIKVNGINVSKETHASVIAHVTACRKYRRPTKQDSIQWVYNSIESAQEDLQKSHSKP
DEPDC6	DEPDC6	1	1Q3P	34.12	TPGAPYARKTFTIVGDAVGWGFVVRGSKPCHIQAVDPSGPAAAAGMKVCQFVVSVNGLNVLHVDYRTVSNLILTGPRTIVMEVMEELE
DFNB31	WHIRLIN	1	1UEZ	98.85	GEVRLVSLRRAKAHEGLGFSIRGGSEHGVGIYVSLVEPGSLAEKEGLRVGDQILRVNDKSLARVTHAEAVKALKGSKKLVLSVYSAGRIPG
DFNB31	WHIRLIN	2	1UF1	98.92	GDRRSTLHLLQGGDEKKVNLVLGDGRSLGLTIRGGAEYGLGIYITGVDPGSEAEGSGLKVGDQILEVNGRSFLNILHDEAVRLLKSSRHLILTVKDVGRLPRARTTVDETKWIASSR
DFNB31	WHIRLIN	3	1UFX	100	TSTLVRVKKSAATLGIAIEGGANTRQPLPRIVTIQRGGSAHNCGQLKVGHVILEVNGLTLRGKEHREAARIIAEAFKTKDRDYIDFLVTEFN
DLG1	DLG1	1	1ZOK	90	EYEEITLERGNSGLGFSIAGGTDNPHIGDDSSIFITKIITGGAAAQDGRLRVNDCILRVNEVDVRDVTHSKAVEALKEAGSIVRLYVKRRKPVS
DLG1	DLG1	2	2G2L	96.84	SEKIMEIKLIKGPKGLGFSIAGGVGNQHIPGDNSIYVTKIIEGGAAHKDGKLQIGDKLLAVNNVCLEEVTHEEAVTALKNTSDFVYLKVAKPTSMYMND
DLG1	DLG1	3	1TQ3	88.04	DDEITREPRKVVLHRGSTGLGFNIVGGEDGEGIFISFILAGGPADLSGELRKGDRIISVNSVDLRAASHEQAAAALKNAGQAVTIVAQYRPEEYSRFEAKIHDLREQMMNSSISS
DLG2	DLG2	1	1RGR	86	EYEFEEITLERGNSGLGFSIAGGTDNPHIGDDPGIFITKIIPGGAAAEDGRLRVNDCILRVNEVDVSEVSHSKAVEALKEAGSIVRLYVRRRRP
DLG2	DLG2	2	2BYG	100	ETVVEIKLFKGPKGLGFSIAGGVGNQHIPGDNSIYVTKIIDGGAAQKDGRLQVGDRLLMVNNYSLEEVTHEEAVAILKNTSEVVYLKVGKPTTIY
DLG2	DLG2	3	2HE2	100	EPRKVVLHKGSTGLGFNIVGGEDGEGIFVSFILAGGPADLSGELQRGDQILSVNGIDLRGASHEQAAAALKGAGQTVTIIAQYQPEDYARFEAKIH
DLG3	DLG3	1	2I1N	100	DGMFKYEEIVLERGNSGLGFSIAGGIDNPHVPDDPGIFITKIIPGGAAAMDGRLGVNDCVLRVNEVDVSEVVHSRAVEALKEAGPVVRLVVRRRQP
DLG3	DLG3	2	2FE5	100	ETIMEVNLLKGPKGLGFSIAGGIGNQHIPGDNSIYITKIIEGGAAQKDGRLQIGDRLLAVNNTNLQDVRHEEAVASLKNTSDMVYLKVAKPGS
DLG3	DLG3	3	1UM7	100	TREPRKIILHKGSTGLGFNIVGGEDGEGIFVSFILAGGPADLSGELRRGDRILSVNGVNLRNATHEQAAAALKRAGQSVTIVAQYRPEEYSRFESK
DLG4	PSD95	1	1RGR	100	EYEEITLERGNSGLGFSIAGGTDNPHIGDDPSIFITKIIPGGAAAQDGRLRVNDSILFVNEVDVREVTHSAAVEALKEAGSIVRLYVMRRK

DLG4	PSD95	2	1QLC	100	AEKVMEIKLIKGPKGLGFSIAGGVGNQHIPGDNSIYVTKIIEGGAAHKDGRLQIGDKILAVNSVGLEDVMHEDAVAALKNTYDVVYLKVAKPSNA
DLG4	PSD95	3	1TQ3	100	DIPREPRRIVIHRGSTGLGFNIVGGEDGEGIFISFILAGGPADLSGELRKGDQILSVNGVDLRNASHEQAAIALKNAGQTVTIIAQYKPEEYSRFEAK
DLG5	DLG5	1	2I1N	38.71	ETEVVEFERETEDIDLKALGFDMAEGVNEPCFPGDCGIFVTKVDKGSIADGRLRVNDWLLRINDVDLINKDKKQAIKALLNGEGAINMVVRRRKSLG
DLG5	DLG5	2	1KWA	39.19	GKVVTPLHINLSGQKDSGISLENGVYAAAVLPGSPAAKEGSLAVGDRIVAINGIALDNKSLNECESLLRSCQDSLTLSLLKVFPQSS
DLG5	DLG5	3	1UIT	100	GERRKDRPYVEEPRHVKVQKGSEPLGISIVSGEKGGIYVSKVTVGSIAHQAGLEYGDQLLEFNGINLRSATEQQARLIIGQQCDTITILAQYNPHVHQLSSHSRS
DLG5	DLG5	4	1UM7	31.46	DANKKTLEPRVVFIKKSQLELGVHLCGGNLHGVFVAEVEDDSPAKGPDGLVPGDLILEYGSLDVRNKTVEEVYVEMLKPRDGVRLKVQYRPEEFTKAKGLPGDS
DVL1	DVL1	1	2KAW	98.95	NIVTVTLNMERHHFLGISIVGQSNDRGDGGIYIGSIMKGGAVAADGRIEPGDMLLQVNDVNFENMSNDDAVRVLREIVSQTGPISLTVAK
DVL2	DVL2	1	2REY	97.85	TMSLNIITVTLNMEKYNFLGISIVGQSNERGDGGIYIGSIMKGGAVAADGRIEPGDMLLQVNDMNFENMSNDDAVRVLRDIVHKPGPIVLTVAKCWDPSP
DVL3	DVL3	1	2REY	95.6	SLNIITVTLNMEKYNFLGISIVGQSNERGDGGIYIGSIMKGGAVAADGRIEPGDMLLQVNEINFENMSNDDAVRVLREIVHKPGPITLTVAKCWDPSPRG
FLJ10324	KIAA1849	1	1UM1	100	YVFTVELERGPSGLGMGLIDGMHTHLGAPGLYIQTLLPGSPAAADGRLSLGDRILEVNGSSLLGLGYLRAVDLIRHGGKKMRFLVAKSDVETAKKIHFRTPPL
FRPD1	FRMPD1	1	2EDV	96.47	PVRHTVKIDKDTLLQDYGFHISESLPLTVVAVTAGGSAHGKLFPGDQILQMNNEPAEDLSWERAVDILREAEDSLSITVVRCTSGVPKSS
FRPD2	FRMPD2	1	2FNE	34.12	GREIVRVTLKRDPHRGFGFVINEGEYSGQADPGIFISSIIPGGPAEKAKTIKPGGQILALNHISLEGFTFNMAVRMIQNSPDNIELIISQSK
FRPD2	FRMPD2	2	1VJ6	61.45	SAGEIYFVELVKEDGTLGFSVTGGINTSVPYGGIYVKSIVPGGPAAKEGQILQGDRLLQVDGVILCGLTHKQAVQCLTGPGQVARLVLERRVPRS
FRPD2	FRMPD2	3	3PDZ	40.7	TDGPKFEVKLKKNANGLGFSFVQMEKESCSHLKSDLVRIKRLFPGQPAEENGAIAAGDIILAVNGRSTEGLIFQEVLHLLRGAPQEVTLLLCRP
FRPD3	FRMPD3	1	2EDV	35	EQLPAEILRQVTVHRDPIYGFGFVAGSERPVVVRSVRPGGPSENKLLAGDQIVAINEEDVSEAPRERLIELIRSAKEFIVLTVLHTHQS
G2SYN	SNTG2	1	1Z86	52.17	NRRTVTLRRQPVGGLGLSIKGGSEHNVPVVISKIFEDQAADQTGMLFVGDAVLQVNGIHVENATHEEVVHLLRNAGDEVTITVEYLREAP
GIPC1	GIPC1	1	3GGE	65.12	KGQRKEVEVFKSEDALGLTITDNGAGYAFIKRIKEGSVIDHIHLISVGDMIEAINGQSLLGCRHYEVARLLKELPRGRTFTLKLTEPRKAFDMISQ
GIPC2	GIPC2	1	3GGE	97.75	KGIEKEVNVYKSEDSLGLTITDNGVGYAFIKRIKDGGVIDSVKTICVGDHIESINGENIVGWRHYDVAKKLKELKKEELFTMKLIEPKKAFE
GIPC3	GIPC3	1	3GGE	61.63	RGETKEVEVTKTEDALGLTITDNGAGYAFIKRIKEGSIINRIEAVCVGDSIEAINDHSIVGCRHYEVAKMLRELPKSQPFTLRLVQPKRAFDMIGQ
GOPC	GOPC	1	2DC2	98.85	KKSQGVGPIRKVLLLKEDHEGLGISITGGKEHGVPILISEIHPGQPADRCGGLHVGDAILAVNGVNLRDTKHKEAVTILSQQRGEIEFEVVYVAPE
GORASP1	GRASP65	1	3RLE	70	MGLGVSAEQPAGGAEGFHLHGVQENSPAQQAGLEPYFDFIITIGHSRLNKENDTLKALLKANVEKPVKLEVFNMKTMRVREVEVVPSNMWGGQGLLGASVRFCSFRRASEQVWHVLDVEPSSPAALAG LRPYTDYVVGSDQILQESEDFFTLIESHEGKPLKLMVYNSKSDSCREVTVTPNAAWGGEGSLGCGIGYGYLHRIPTQPPS
GORASP2	GRASP55	1	3RLE	100	GMGSSQSVEIPGGGTEGYHVLRVQENSPGHRAGLEPFFDFIVSINGSRLNKDNDTLKDLLKANVEKPVKMLIYSSKTLELRETSVTPSNLWGGQGLLGVSIRFCSFDGANENVWHVLEVESNSPAALAGL RPHSDYIIGADTVMNESEDLFSLIETHEAKPLKLYVYNTDTDNCREVIITPNSAWGGEGSLGCGIGYGYLHRIPTRPFE
GRASP	GRASP	1	2PNT	95.7	EQQRKVLTLEKEDNQTFGFEIQTYGLHHREEQRVEMVTFVCRVHESSPAQLAGLTPGDTIASVNGLNVEGIRHREIVDIIKASGNVLRLETLYGTSIR
GRID2IP	GRID2	2	2DLS	33.75	GPGGARRTVRVYKGNKSFGFTLRGHGPVWIESVLPGSPADNAALKSGDRILFLNGLDMRNCSHDKVVSMLQGSGAMPTLVVEEGLVPFASDSDSLDSPN
GRID2IP	GRID2	1	2KV8	44	MATTATPATNQGWPEDFGFRLGGSGPCFVLEVAKGSSAHAGGLRPGDQILEVEGLAVGGLSRERLVRLARRCPRVPPSLGVLPAPDG
GRIP1	GRIP1	1	2DC2	98.95	EFKGSTVVELMKKEGTTLGLTVSGGIDKDGKPRVSNLRQGGIAARSDQLDVGDYIKAVNGINLAKFRHDEIISLLKNVGERVVLEVEYELP
GRIP1	GRIP1	2	2JIL	98.94	SVIFRTVEVTLHKEGNTFGFVIRGGAHDDRNKSRPVVITCVRPGGPADREGTIKPGDRLLSVDGIRLLGTTHAEAMSILKQCGQEAALLIEYDVS
GRIP1	GRIP1	3	1V62	60.22	DSVATASGPLLVEVAKTPGASLGVALTTSMCCNKQVIVIDKIKSASIADRCGALHVGDHILSIDGTSMEYCTLAEATQFLANTTDQVKLEILPHHQTRLALKG
GRIP1	GRIP1	4	1V5Q	99.01	QVVHTETTEVVLTADPVTGFGIQLQGSVFATETLSSPPLISYIEADSPAERCGVLQIGDRVMAINGIPTEDSTFEEASQLLRDSSITSKVTLEIEFDVAESVIP
GRIP1	GRIP1	5	1P1D	97.73	ESVIPSSGTFHVKLPKKHNVELGITISSPSSRKPGDPLVISDIKKGSVAHRTGTLELGDKLLAIDNIRLDNCSMEDAVQILQQCEDLVKLKIRKDED
GRIP1	GRIP1	6	1N7F	98.91	SSGAIIYTVELKRYGGPLGITISGTEEPFDPIIISSLTKGGLAERTGAIHIGDRILAINSSSLKGKPLSEAIHLLQMAGETVTLKIKKQTDAQSASS
GRIP1	GRIP1	7	1M5Z	98.8	SPTPVELHKVTLYKDSDMEDFGFSVADGLLEKGVYVKNIRPAGPGDLGGLKPYDRLLQVNHVRTRDFDCCLVVPLIAESGNKLDLVISRNP
GRIP2	GRIP2	1	2QT5	81.61	EEFRGITVVELIKKEGSTLGLTISGGTDKDGKPRVSNLRPGGLAARSDLLNIGDYIRSVNGIHLTRLRHDEIITLLKNVGERVVLEVEYELPP
GRIP2	GRIP2	2	2JIL	64.89	ENNPRIISKTVDVSLYKEGNSFGFVLRGGAHEDGHKSRPLVLTYVRPGGPADREGSLKVGDRLLSVDGIPLHGASHATALATLRQCSHEALFQVEYDVATPDTVANASG
GRIP2	GRIP2	3	1V62	100	DTVANASGPLMVEIVKTPGSALGISLTTTSLRNKSVITIDRIKPASVVDRSGALHPGDHILSIDGTSMEHCSLLEATKLLASISEKVRLEILPVPQSQRPLR
GRIP2	GRIP2	4	1X5R	94	GGQIVHTETTEVVLCGDPLSGFGLQLQGGIFATETLSSPPLVCFIEPDSPAERCGLLQVGDRVLSINGIATEDGTMEEANQLLRDAALAHKVVLEVEFDVAES
GRIP2	GRIP2	5	1P1D	88.64	DVAESVIPSSGTFHVKLPKKRSVELGITISSASRKRGEPLIISDIKKGSVAHRTGTLEPGDKLLAIDNIRLDNCPMEDAVQILRQCEDLVKLKIRKDED
GRIP2	GRIP2	6	1N7F	89.41	TTGAVSYTVELKRYGGPLGITISGTEEPFDPIVISGLTKRGLAERTGAIHVGDRILAINNVSLKGRPLSEAIHLLQVAGETVTLKIKKQLDR
GRIP2	GRIP2	7	1M5Z	71.08	PTPLEMHKVTLHKDPMRHDFGFSVSDGLLEKGVYVHTVRPDGPAHRGGLQPFDRVLQVNHVRTRDFDCCLAVPLLAEAGDVLELIISRKP

HTRA1	HTRA1	1	2YTW	100	TESHDRQAKGKAITKKKYIGIRMMSLTSSKAKELKDRHRDFPDVISGAYIIEVIPDTPAEAGGLKENDVIISINGQSVVSANDVSDVIKRESTLNMVVRRGNEDIMITVIPEEIDP
HTRA2	HTRA2	1	2PZD	100	SGSQRRYIGVMMLTLSPSILAELQLREPSFPDVQHGVLIHKVILGSPAHRAGLRPGDVILAIGEQMVQNAEDVYEAVRTQSQLAVQIRRGRETLTLYVTPEVTE
HTRA3	HTRA3	1	2P3W	100	KKRFIGIRMRTITPSLVDELKASNPDFPEVSSGIYVQEVAPNSPSQRGGIQDGDIIVKVNGRPLVDSSELQEAVLTESPLLLEVRRGNDDLLFSIAPEVVM
HTRA4	HTRA4	1	2YTW	45.28	HQMKGKAFSNKKYLGLQMLSLTVPLSEELKMHYPDFPDVSSGVYVCKVVEGTAAQSSGLRDHDVIVNINGKPITTTTDVVKALDSDSLSMAVLRGKDNLLLTVIPETIN
IL16	nIL16	1	2FNE	52.63	QASVISNIVLMKGQAKGLGFSIVGGKDSIYGPIGIYVKTIFAGGAAAADGRLQEGDEILELNGESMAGLTHQDALQKFKQAKKGLLTLTVRTRLTAP
IL16	nIL16	2	2DLU	35.71	STAKPNYRIMVEVSLQKEAGVGLGIGLCSVPYFQCISGIFVHTLSPGSVAHLDGRLRCGDEIVEISDSPVHCLTLNEVYTILSHCDPGPVPIIVSRHPDPQVSEQQLKE
IL16	nIL16	3	1X6D	100	KQLDGIHVTILHKEEGAGLGFSLAGGADLENKVITVHRVFPNGLASQEGTIQKGNEVLSINGKSLKGTTHHDALAILRQAREPRQAVIVTRKLTPE
IL16	nIL16	4	1I16	100	MPDLNSSTDSAASASAASDVSVESTAEATVCTVTLEKMSAGLGFSLEGGKGSLHGDKPLTINRIFKGAASEQSETVQPGDEILQLGGTAMQGLTRFEAWNIIKALPDGPVTIVIRRKSLQSKETTAAGDS
InaDl	INADL	1	2DB5	100	KLGNEDFNSVIQQMAQGRQIEYIDIERPSTGGLGFSVVALRSQNLGKVDIFVKDVQPGSVADRDQRLKENDQILAINHTPLDQNISHQQAIALLQQTTGSLRLIVAREPVHTKSSTS
InaDl	INADL	2	2DLU	100	PETVCWGHVEEVELINDGSGLGFGIVGGKTSGVVVRTIVPGGLADRDGRLQTGDHILKIGGTNVQGMTSEQVAQVLRNCGNSVRMLVARDPAGDISVT
InaDl	INADL	3	2DMZ	99	SLFETYNVELVRKDGQSLGIRIVGYVGTSHTGEASGIYVKSIIPGSAAYHNGHIQVNDKIVAVDGVNIQGFANHDVVEVLRNAGQVVHLTLVRRKTSSSTSPLEPPSDRGT
InaDl	INADL	4	2HE2	45.45	DTQIADDAELQKYSKLLPIHTLRLGVEVDSFDGHHYISSIVSGGPVDTLGLLQPEDELLEVNGMQLYGKSRREAVSFLKEVPPPFTLVCCRRLFDDEASVDEPRR
InaDl	INADL	5	2D92	100	DDGELALWSPEVKIVELVKDCKGLGFSILDYQDPLDPTRSVIVIRSLVADGVAERSGGLLPGDRLVSVNEYCLDNTSLAEAVEILKAVPPGLVHLGICKPLVEDNEEE
InaDl	INADL	6	2EHR	100	PNFSHWGPPRIVEIFREPNVSLGISIVGGQTVIKRLKNGEELKGIFIKQVLEDSPAGKTNALKTGDKILEVSGVDLQNASHSEAVEAIKNAGNPVVFIVQSLSSTPRVIP
InaDl	INADL	7	2DAZ	97.92	DAFTDQKIRQRYADLPGELHIIELEKDKNGLGLSLAGNKDRSRMSIFVVGINPEGPAAADGRMRIGDELLEINNQILYGRSHQNASAIIKTAPSKVKLVFIRNEDAVNQMAVTP
InaDl	INADL	8	2DM8	98.94	PATCPIVPGQEMIIEISKGRSGLGLSIVGGKDTPLNAIVIHEVYEEGAAARDGRLWAGDQILEVNGVDLRNSSHEEAITALRQTPQKVRLVVYRDEAHYRDEENLE
InaDl	INADL	9	2QG1	72.84	EIFPVDLQKKAGRGLGLSIVGKRNGSGVFISDIVKGGAADLDGRLIQGDQILSVNGEDMRNASQETVATILKCAQGLVQLEIGRLR
InaDl	INADL	10	2IWP	61.54	EPRTVEINRELSDALGISIAGGRGSPLGDIPVFIAMIQASGVAARTQKLKVGDRIVSINGQPLDGLSHADVVNLLKNAYGRIILQVVADTN
INTU	PDZD6	1	2HE2	33.96	KEQLKLLEVLVGIIHQTKWSWRRTGKQGDGERLVVHGLLPGGSAMKSGQVLIGDVLVAVNDVDVTTENIERVLSCIPGPMQVKLTFENAYDVKRETSHPRQK
LAP2	ERBIN	1	1N7T	100	GSHMGHELAKQEIRVRVEKDPELGFSISGGVGGRGNPFRPDDDGIFVTRVQPEGPASKLLQPGDKIIQANGYSFINIEHGQAVSLLKTFQNTVELIIVREVSS
LDB3	ZASP	1	1WJL	98.81	MSYSVTLTGPGPWGFRLQGGKDFNMPLTISRITPGSKAAQSQLSQGDLVVAIDGVNTDTMTHLEAQNKIKSASYNLSLTLQKSKR
LIMK1	LIMK1	1	2YUB	41.94	PGSHLPHTVTLVSIPASSHGKRGLSVSIDPPHGPPGCGTEHSHTVRVQGVDPGCMSPDVKNSIHVGDRILEINGTPIRNVPLDEIDLLIQETSRLLQLTLEHDPHDTLGHGLGP
LIMK2	LIMK2	1	2YUB	94.79	QEQLPYSVTLISMPATTEGRRGFSVSVESACSNYATTVQVKEVNRMHISPNNRNAIHPGDRILEINGTPVRTLRVEEVEDAISQTSQTLQLLIEHDPVSQRLDQLR
LIN7A	VELI1	1	2DKR	90.36	SEGHSHPRVVELPKTDEGLGFNVMGGKEQNSPIYISRIIPGGVAERHGGLKRGDQLLSVNGVSVEGEHHEKAVELLKAAKDSVKLVVRYTPK
LIN7B	VELI2	1	2DKR	97.59	SEGHAHPRVVELPKTDEGLGFNIMGGKEQNSPIYISRVIPGGVADRHGGLKRGDQLLSVNGVSVEGEQHEKAVELLKAAQGSVKLVVRYTPR
LIN7C	VELI3	1	2DKR	91.57	SEGHSHPRVVELPKTEEGLGFNIMGGKEQNSPIYISRIIPGGIADRHGGLKRGDQLLSVNGVSVEGEHHEKAVELLKAAQGKVKLVVRYTPK
LMO7	LMO7	1	2EAQ	100	QFSDMRISINQTPGKSLDFGFTIKWDIPGIFVASVEAGSPAEFSQLQVDDEIIAINNTKFSYNDSKEWEEAMAKAQETGHLVMDVRRYGK
LNX1	LNX1	1	2DM8	45.78	PRLYHLIPDGEITSIKINRVDPSESLSIRLVGGSETPLVHIIIQHIYRDGVIARDGRLLPGDIILKVNGMDISNVPHNYAVRLLRQPCQVLWLTVMREQKFRSRNNGQAPD
LNX1	LNX1	2	2VWR	62.2	DAYRPRDDSFHVILNKSSPEEQLGIKLVRKVDEPGVFIFNVLDGGVAYRHGQLEENDRVLAINGHDLRYGSPESAAHLIQASERRVHLVVSRQVRQRSPD
LNX1	LNX1	3	3B76	100	TITCHEKVVNIQKDPGESLGMTVAGGASHREWDLPIYVISVEPGGVISRDGRIKTGDILLNVDGVELTEVSRSEAVALLKRTSSSIVLKALEVKEYE
LNX1	LNX1	4	2FNE	41.86	RCLYNCKDIVLRRNTAGSLGFCIVGGYEEYNGNKPFFIKSIVEGTPAYNDGRIRCGDILLAVNGRSTSGMIHACLARLLKELKGRITLTIVSWPGT
LNX2	LNX2	1	2DM8	54.79	PLSLPEGEITTIEIHRSNPYIQLGISIVGGNETPLINIVIQEVYRDGVIARDGRLLAGDQILQVNNYNISNVSHNYARAVLSQPCNTLHLTVLRERRFGNRAHN
LNX2	LNX2	2	2VWR	98.85	REEIFQVALHKRDSGEQLGIKLVRRTDEPGVFILDLLEGGLAAQDGRLSSNDRVLAINGHDLKYGTPELAAQIIQASGERVNLTIARPGKPQP
LNX2	LNX2	3	3B76	55.88	TQCVTCQEKHITVKKEPHESLGMTVAGGRGSKSGELPIFVTSVPPHGCLARDGRIKRGDVLLNINGIDLTNLSHSEAVAMLKASAASPAVALKALEVQIVEE
LNX2	LNX2	4	2FNE	48.61	PSTLHSCHDIVLRRSYLGSWGFSIVGGYEENHTNQPFFIKTIVLGTPAYYDGRLKCGDMIVAVNGLSTVGMSHSALVPMLKEQRNKVTLTVICWPGS
LRRC7	LAP1	1	1MFG	60	EQFCVRIEKNPGLGFSISGGISGQGNPFKPSDKGIFVTRVQPDGPASNLLQPGDKILQANGHSFVHMEHEKAVLLLKSFQNTVDLVIQRELT
MAGI1	MAGI1	1	2EEH	50	MSKVIQKKNHWTSRVHECTVKRGPQGELGVTVLGGAEHGEFPYVGAVAAVEAAGLPGGGEGPRLGEGELLLEVQGVRVSGLPRYDVLGVIDSCKEAVTFKAVRQGG
MAGI1	MAGI1	2	2I04	100	KPFFTRNPSELKGKFIHTKLRKSSRGFGFTVVGGDEPDEFLQIKSLVLDGPAALDGKMETGDVIVSVNDTCVLGHTHAQVVKIFQSIPIGASVDLELCRGYPLPFDPDDPNTSLVTSVAILDKEP
MAGI1	MAGI1	3	1UJV	97.53	QPELITVHIVKGPMGFGFTIADSPGGGGQRVKQIVDSPRCRGLKEGDLIVEVNKKNVQALTHNQVVDMLVECPKGSEVTLLVQRGGLP
MAGI1	MAGI1	4	1UEP	97.65	DYQEQDIFLWRKETGFGFRILGGNEPGEPIYIGHIVPLGAADTDGRLRSGDELICVDGTPVIGKSHQLVVQLMQQAAKQGHVNLTVRRKVVSGP

MAGI1	MAGI1	5	1UEW	70.1	GSGVVSTVVQPYDVEIRRGENEGFGFVIVSSVSRPEAGTTFGNACVAMPHKIGRIIEGSPADRCGKLKVGDRILAVNGCSITNKSHSDIVNLIKEAGNTVTLRIIPGDESSN
MAGI1	MAGI1	6	1WFV	96.59	QEQDFYTVELERGAKGFGFSLRGGREYNMDLYVLRLAEDGPAERCGKMRIGDEILEINGETTKNMKHSRAIELIKNGGRRVRLFLKRGDGSVP
MAGI2	MAGI2	1	2EEH	34.21	MSKSLKKKSHWTSKVHESVIGRNPEGQLGFELKGGAENGQFPYLGEVKPGKVAYESGSKLVSEELLLEVNETPVAGLTIRDVLAVIKHCKDPLRLKCVKQGG
MAGI2	MAGI2	2	1UEQ	100	KPLFTRDASQLKGTFLSTTLKKSNMGFGFTIIGGDEPDEFLQVKSVIPDGPAAQDGKMETGDVIVYINEVCVLGHTHADVVKLFQSVPIGQSVNLVLCRGYPLPFDPEDPANS
MAGI2	MAGI2	3	1UJV	100	QAELMTLTIVKGAQGFGFTIADSPTGQRVKQILDIQGCPGLCEGDLIVEINQQNVQNLSHTEVVDILKDCPIGSETSLIIHRG
MAGI2	MAGI2	4	1UEP	98.89	PDYKELDVHLRRMESGFGFRILGGDEPGQPILIGAVIAMGSADRDGRLHPGDELVYVDGIPVAGKTHRYVIDLMHHAARNGQVNLTVRRKVLCG
MAGI2	MAGI2	5	1UEW	96.08	SLQTSDVVIHRKENEGFGFVIISSLNRPESGSTITVPHKIGRIIDGSPADRCAKLKVGDRILAVNGQSIINMPHADIVKLIKDAGLSVTLRIIPQEELNSPTS
MAGI2	MAGI2	6	1WFV	100	QDFDYFTVDMEKGAKGFGFSIRGGREYKMDLYVLRLAEDGPAIRNGRMRVGDQIIEINGESTRDMTHARAIELIKSGGRRVRLLLKRGT
MAGI3	MAGI3	1	1X6D	43.33	MSKTLKKKKHWLSKVQECAVSWAGPPGDFGAEIRGGAERGEFPYLGRLREEPGGGTCCVVSGKAPSPGDVLLEVNGTPVSGLTNRDTLAVIRHFREPIRLKTVKPGKVINKDLR
MAGI3	MAGI3	2	1UEQ	73.96	TRDPSQLKGVLVRASLKKSTMGFGFTIIGGDRPDEFLQVKNVLKDGPAAQDGKIAPGDVIVDINGNCVLGHTHADVVQMFQLVPVNQYVNLTLCRGYPLPDDSEDP
MAGI3	MAGI3	3	1UJV	58.33	SQPELVTIPLIKGPKGFGFAIADSPTGQKVKMILDSQWCQGLQKGDIIKEIYHQNVQNLTHLQVVEVLKQFPVGADVPLLILRG
MAGI3	MAGI3	4	1UEP	63.33	EDKPPNTKDLDVFLRKQESGFGFRVLGGDGPDQSIYIGAIIPLGAAEKDGRLRAADELMCIDGIPVKGKSHKQVLDLMTTAARNGHVLLTVRRKIFYGEK
MAGI3	MAGI3	5	1UEW	63.73	QEPYDVVLQRKENEGFGFVILTSKNKPPPGVIPHKIGRVIEGSPADRCGKLKVGDHISAVNGQSIVELSHDNIVQLIKDAGVTVTLTVIAEEEHH
MAGI3	MAGI3	6	1WFV	62.64	NQNLGCYPVELERGPRGFGFSLRGGKEYNMGLFILRLAEDGPAIKDGRIHVGDQIVEINGEPTQGITHTRAIELIQAGGNKVLLLLRPGT
MAGIX	JM10	1	2DJT	98.85	SQASGHFSVELVRGYAGFGLTLGGGRDVAGDTPLAVRGLLKDGPAQRCGRLEVGDVVLHINGESTQGLTHAQAVERIRAGGPQLHLVIRRPLET
MAST1	MAST1	1	2W7R	73.68	RSPITIQRSGKKYGFTLRAIRVYMGDTDVYSVHHIVWHVEEGGPAQEAGLCAGDLITHVNGEPVHGMVHPEVVELILKSGNKVAVTTTPFEN
MAST2	MAST2	1	2W7R	74.74	RPPIIIHRAGKKYGFTLRAIRVYMGDSDVYTVHHMVWHVEDGGPASEAGLRQGDLITHVNGEPVHGLVHTEVVELILKSGNKVAISTTPLEN
MAST3	MAST3	1	2W7R	74.74	RPPIVIHSSGKKYGFSLRAIRVYMGDSDVYTVHHVVWSVEDGSPAQEAGLRAGDLITHINGESVLGLVHMDVVELLLKSGNKISLRTTALEN
MAST4	MAST4	1	2W7R	100	QPIVIHSSGKNYGFTIRAIRVYVGDSDIYTVHHIVWNVEEGSPACQAGLKAGDLITHINGEPVHGLVHTEVIELLLKSGNKVSITTTPFENTS
MPDZ	MUPP1	1	2O2T	100	DEFDQLIKNMAQGRHVEVFELLKPPSGGLGFSVVGLRSENRGELGIFVQEIQEGSVAHRDGRLKETDQILAINGQALDQTITHQQAISILQKAKDTVQLVIARGSLPQLVS
MPDZ	MUPP1	2	2DLU	69.23	HSNPVHWQHMETIELVNDGSGLGFGIIGGKATGVIVKTILPGGVADQHGRLCSGDHILKIGDTDLAGMSSEQVAQVLRQCGNRVKLMIARGAIEERT
MPDZ	MUPP1	3	2IWN	100	EESETFDVELTKNVQGLGITIAGYIGDKKLEPSGIFVKSITKSSAVEHDGRIQIGDQIIAVDGTNLQGFTNQQAVEVLRHTGQTVLLTLMRRGMKQE
MPDZ	MUPP1	4	2OPG	34.62	NYEIVVAHVSKFSENSGLGISLEATVGHHFIRSVLPEGPVGHSGKLFSGDELLEVNGITLLGENHQDVVNILKELPIEVTMVCCRRT
MPDZ	MUPP1	5	2D92	60.47	QAPLAMWEAGIQHIELEKGSKGLGFSILDYQDPIDPASTVIIIRSLVPGGIAEKDGRLLPGDRLMFVNDVNLENSSLEEAVEALKGAPSGTVRIGVAKPLPLSPEE
MPDZ	MUPP1	6	2K20	39.71	QNVSKESFERTINIAKGNSSLGMTVSANKDGLGMIVRSIIHGGAISRDGRIAIGDCILSINEESTISVTNAQARAMLRRHSLIGPDIKITYVPAEHLEEFKISLGQQS
MPDZ	MUPP1	7	2EHR	98.95	TAYSNWNQPRRVELWREPSKSLGISIVGGRGMGSRLSNGEVMRGIFIKHVLEDSPAGKNGTLKPGDRIVEVDGMDLRDASHEQAVEAIRKAGNPVVFMVQSIINRP
MPDZ	MUPP1	8	2DAZ	77.08	DKEDEFGYSWKNIRERYGTLTGELHMIELEKGHSGLGLSLAGNKDRSRMSVFIVGIDPNGAAGKDGRLQIADELLEINGQILYGRSHQNASSIIKCAPSKVKIIFIRNKDAVNQMAV
MPDZ	MUPP1	9	2DAZ	31.17	PTVTTSDAAVDLSSFKNVQHLELPKDQGGLGIAISEEDTLSGVIIKSLTEHGVAATDGRLKVGDQILAVDDEIVVGYPIEKFISLLKTAKMTVKLTIHAENPDSQAVPS
MPDZ	MUPP1	10	2OPG	100	PGCETTIEISKGRTGLGLSIVGGSDTLLGAIIIHEVYEEGAACKDGRLWAGDQILEVNGIDLRKATHDEAINVLRQTPQRVRLTLYRDEAPYK
MPDZ	MUPP1	11	2QG1	98.8	DTLTIELQKKPGKGLGLSIVGKRNDTGVFVSDIVKGGIADADGRLMQGDQILMVNGEDVRNATQEAVAALLKCSLGTVTLEVGRIKAGP
MPDZ	MUPP1	12	2IWP	100	QGLRTVEMKKGPTDSLGISIAGGVGSPLGDVPIFIAMMHPTGVAAQTQKLRVGDRIVTICGTSTEGMTHTQAVNLLKNASGSIEMQVVAGGD
MPDZ	MUPP1	13	2FNE	100	PPQCKSITLERGPDGLGFSIVGGYGSPHGDLPIYVKTVFAKGAASEDGRLKRGDQIIAVNGQSLEGVTHEEAVAILKRTKGTVTLMVLS
MPP1	MPP1	1	2EV8	100	VRLIQFEKVTEEPMGITLKLNEKQSCTVARILHGGMIHRQGSLHVGDEILEINGTNVTNHSVDQLQKAMKETKGMISLKVIPNQQ
MPP2	MPP2	1	2E7K	100	DAVRMVGIRKTAGEHLGVTFRVEGGELVIARILHGGMVAQQGLLHVGDIIKEVNGQPVGSDPRALQELLRNASGSVILKILPSYQE
MPP3	MPP3	1	1VA8	43.02	DNIDEDFDEESVKIVRLVKNKEPLGATIRRDEHSGAVVVARIMRGGAADRSGLVHVGDELREVNGIAVLHKRPDEISQILAQSQGSITLKIIPATQEED
MPP4	MPP4	1	1VA8	44	PDNIPESEEAMRIVCLVKNQQPLGATIKRHEMTGDILVARIIHGGLAERSGLLYAGDKLVEVNGVSVEGLDPEQVIHILAMSRGTIMFKVVPVSDPPVNS
MPP5	MPP5	1	1VA8	100	TDERVYESIGQYGGETVKIVRIEKARDIPLGATVRNEMDSVIISRIVKGGAAEKSGLLHEGDEVLEINGIEIRGKDVNEVFDLLSDMHGTLTFVLIPSQ
MPP6	MPP6	1	2E7K	74.36	DAIRILGIHKRAGEPLGVTFRVENNDLVIARILHGGMIDRQGLLHVGDIIKEVNGHEVGNNPKELQELLKNISGSVTLKILPS
MPP7	MPP7	1	1VA8	45.35	DPVLPPMPEDIDDEEDSVKIIRLVKNREPLGATIKKDEQTGAIIVARIMRGGAADRSGLIHVGDELREVNGIPVEDKRPEEIIQILAQSQGAITFKIIPGSKEET
MYO18A	MYO18A	1	1GQ4	40	TLRELELQRRPTGDFGFSLRRTTMLDRGPEGQACRRVVHFAEPGAGTKDLALGLVPGDRLVEINGHNVESKSRDEIVEMIRQSGDSVRLKVQP

NOS1	NOS	1	1B8Q	98.81	QQIQPNVISVRLFKRKVGGLGFLVKERVSKPPVIISDLIRGGAAEQSGLIQAGDIILAVNGRPLVDLSYDSALEVLRGIASETHVVLILRGPE
P-Rex1	PREX1	1	2YT8	30	NKQLRNDFKLVENILAKRLLILPQEEDYGFDIEEKNKAVVVKSVQRGSLAEVAGLQVGRKIYSINEDLVFLRPFSEVESILNQSFCSRRPLRLLVATKAKEIIKIPDQPDT
P-Rex1	PREX1	2	1UF1	39.53	ATKAKEIIKIPDQPDTLCFQIRGAAPPYVYAVGRGSEAMAAGLCAGQCILKVNGSNVMNDGAPEVLEHFQAFRSRREEALGLYQ
PAR6A	PARD6A	1	1RY4	85.11	PETHRRVRLHKHGSDRPLGFYIRDGMSVRVAPQGLERVPGIFISRLVRGGLAESTGLLAVSDEILEVNGIEVAGKTLDQVTDMMVANSHNLIVTVKPANQR
PAR6B	PARD6B	1	1RY4	98.97	DFRPVSSIIDVDILPETHRRVRLYKYGTEKPLGFYIRDGSSVRVTPHGLEKVPGIFISRLVPGGLAQSTGLLAVNDEVLEVNGIEVSGKSLDQVTDMMIANSRNLIITVRPANQR
PAR6G	PARD6G	1	1NF3	87.23	RPVSSIIDVDLVPETHRRVRLHRHGCEKPLGFYIRDGASVRVTPHGLEKVPGIFISRMVPGGLAESTGLLAVNDEVLEVNGIEVAGKTLDQVTDMMIANSHNLIVTVKPANQRN
PARD3	PAR3	1	2DLU	31.08	PNFSLDDMVKLVEVPNDGGPLGIHVVPFSARGGRTLGLLVKRLEKGGKAEHENLFRENDCIVRINDGDLRNRRFEQAQHMFRQAMRTPIIWFHVVPAANKEQYEQLSQS
PARD3	PAR3	2	2OGP	98.86	KKIGKRLNIQLKKGTEGLGFSITSRDVTIGGSAPIYVKNILPRGAAIQDGRLKAGDRLIEVNGVDLVGKSQEEVVSLLRSTKMEGTVSLLVFRQEDA
PARD3	PAR3	3	2K20	97.94	GTREFLTFEVPLNDSGSAGLGVSVKGNRSKENHADLGIFVKSIINGGAASKDGRLRVNDQLIAVNGESLLGKTNQDAMETLRRSMSTEGNKRGMIQLIVARRIS
PARD3B	ALS2CR19	1	2K1Z	27.78	QTELLTSPRTKDTLSDMTRTVEISGEGGPLGIHVVPFFSSLSGRILGLFIRGIEDNSRSKREGLFHENECIVKINNVDLVDKTFAQAQDVFRQAMKSPSVLLHVLPPQNREQYEKS
PARD3B	ALS2CR19	2	2OGP	60	NKNAKKIKIDLKKGPEGLGFTVVTRDSSIHGPGPIFVKNILPKGAAIKDGRLQSGDRILEVNGRDVTGRTQEELVAMLRSTKQGETASLVIARQEG
PARD3B	ALS2CR19	3	2K1Z	97.89	ETSEQLTFEIPLNDSGSAGLGVSLKGNKSRETGTDLGIFIKSIIHGGAAFKDGRLRMNDQLIAVNGESLLGKSNHEAMETLRRSMSMEGNIRGMIQLVILRRPER
PCLO	PICCOLO	1	1UJD	100	NGKTMHYIFPHARIKITRDSKDHTVSGNGLGIRIVGGKEIPGHSGEIGAYIAKILPGGSAEQTGKLMEGMQVLEWNGIPLTSKTYEEVQSIISQQSGEAEICVRLDLNMLSDSEN
PDLIM1	ELFIN	1	2PKT	98.84	MTTQQIDLQGPGPWGFRLVGGKDFEQPLAISRVTPGSKAALANLCIGDVITAIDGENTSNMTHLEAQNRIKGCTDNLTLTVARSEHK
PDLIM2	MYSTIQUE	1	2PA1	100	MALTVDVAGPAPWGFRITGGRDFHTPIMVTKVAERGKAKDADLRPGDIIVAINGESAEGMLHAEAQSKIRQSPSPLRLQLDRS
PDLIM3	ALP	1	1V5L	94.38	MPQTVILPGPAPWGFRLSGGIDFNQPLVITRITPGSKAAAANLCPGDVILAIDGFGTESMTHADAQDRIKAAAHQLCLKIDRGETHLWSPQVSE
PDLIM4	RIL	1	2EEG	98.84	MPHSVTLRGPSPWGFRLVGGRDFSAPLTISRVHAGSKAALAALCPGDLIQAINGESTELMTHLEAQNRIKGCHDHLTLSVSRP
PDLIM5	ENH	1	1RGW	98.81	MSNYSVSLVGPAPWGFRLQGGKDFNMPLTISSLKDGGKAAQANVRIGDVVLSIDGINAQGMTHLEAQNKIKGCTGSLNMTLQRASAAPKPEP
PDLIM7	ENIGMA	1	2Q3G	100	MDSFKVVLEGPAPWGFRLQGGKDFNVPLSISRLTPGGKAAQAGVAVGDWVLSIDGENAGSLTHIEAQNKIRACGERLSLGLSRAQ
PDZD2	AIPC	1	2R4H	36.67	PEMEICTVYLTKELGDTETVGLSFGNIPVFGDYGEKRRGGKKRKTHQGPVLDVGCIWVTELRKNSPAGKSGKVRLRDEILSLNGQLMVGVDVSGASYLAEQCWNGGFIYLIMLRRFKH
PDZD2	AIPC	2	2BYG	43.68	REEVGRIWKMELLKESDGLGIQVSGGRGSKRSPHAIVVTQVKEGGAAHRDGRLSLGDELLVINGHLLVGLSHEEAVAILRSATGMVQLVVASKENSAED
PDZD2	AIPC	3	1X6D	42.5	PWRLIRPSVISIIGLYKEKGKGLGFSIAGGRDCIRGQMGIFVKTIFPNGSAAEDGRLKEGDEILDVNGIPIKGLTFQEAIHTFKQIRSGLFVLTVRTKLVSPSLTPCSTP
PDZD2	AIPC	4	2QG1	34.88	KDRIVMEVTLNKEPRVGLGIGACCLALENSPPGIYIHSLAPGSVAKMESNLSRGDQILEVNSVNVRHAALSKVHAILSKCPPGPVRLVIGRHPN
PDZD2	AIPC	5	1X6D	50	KAQSENEEDVCFIVLNRKEGSGLGFSVAGGTDVEPKSITVHRVFSQGAASQEGTMNRGDFLLSVNGASLAGLAHGNVLKVLHQAQLHKDALVVIKKGMDQPRPSARQE
PDZD2	AIPC	6	1I16	50	RSVAVHDALCVEVLKTSAGLGLSLDGGKSSVTGDGPLVIKRVYKGGAAEQAGIIEAGDEILAINGKPLVGLMHFDAWNIMKSVPEGPVQLLIRKHRNSS
PDZD7	FLJ00011	3	2VRF	47	GELKTVTLSKMKQSLGISISGGIESKVQPMVKIEKIFPGGAAFLSGALQAGFELVAVDGENLEQVTHQRAVDTIRRAYRNKAREPMELVVRVPGPS
PDZD9	c16orf65	1	1WIF	85	HNLSKTQQTKLTVGSLGLGLIIIQHGPYLQITHLIRKGAAANDGKLQPGDVLISVGHANVLGYTLREFLQLLQHITIGTVLQIKVYRDFINIPEEWQE
PDZK1	PDZK1	1	2EDZ	89.25	TSTFNPRECKLSKQEGQNYGFFLRIEKDTEGHLVRVVEKCSPAEKAGLQDGDRVLRINGVFVDKEEHMQVVDLVRKSGNSVTLLVLDGDSYEKAVKTRVDLKELGQ
PDZK1	PDZK1	2	2EEI	100	QPRLCYLVKEGGSYGFSLKTVQGKKGVYMTDITPQGVAMRAGVLADDHLIEVNGENVEDASHEEVVEKVKKSGSRVMFLLVDKETDKRHVEQK
PDZK1	PDZK1	3	2D90	80	PHQPRIVEMKKGSNGYGFYLRAGSEQKGQIIKDIDSGSPAEEAGLKNNDLVVAVNGESVETLDHDSVVEMIRKGGDQTSLLVVDKETDNMYR
PDZK1	PDZK1	4	2EEJ	98.84	KPKLCRLAKGENGYGFHLNAIRGLPGSFIKEVQKGGPADLAGLEDEDVIIEVNGVNVLDEPYEKVVDRIQSSGKNVTLLVCGKKA
PDZK10	FRMPD4	1	2EDV	36.9	ESCQIIPPAPRKVEMRRDPVLGFGFVAGSEKPVVVRSVTPGGPSEGKLIPGDQIVMINDEPVSAAPRERVIDLVRSCKESILLTVIQPYPSPKS
PDZK11	PDZD11	1	1WI2	97.56	NNELTQFLPRTITLKKPPGAQLGFNIRGGKASQLGIFISKVIPDSDAHRAGLQEGDQVLAVNDVDFQDIEHSKAVEILKTAREISMRVRFFP
PDZK2	PDZD3	1	2EDZ	38.64	DPYDPWSLERPRFCLLSKEEGKSFGFHLQQELGRAGHVVCRVDPGTSAQRQGLQEGDRILAVNNDVVEHEDYAVVVRRIRASSPRVLLTVLARHAHDVARAQLGED
PDZK2	PDZD3	2	2EEI	41.38	RPRLCHIVKDEGGFGFSVTHGNQGPFWLVLSTGGAAERAGVPPGARLLEVNGVSVEKFTHNQLTRKLWQSGQQVTLLVAGPEVEEQCR
PDZK2	PDZD3	3	2V90	100	TKPRCLHLEKGPQGFGFLLREEKGLDGRPGQFLWEVDPGLPAKKAGMQAGDRLVAVAGESVEGLGHEETVSRIQGQGSCVSLTVVDPE
PDZK2	PDZD3	4	2HE4	38.82	GSRQCFLYPGPGGSYGFRLSCVASGPRLFISQVTPGGSAARAGLQVGDVILEVNGYPVGGQNDLERLQQLPEAEPPLCLKLAARSLR
PDZK4	PDZD4	1	1WH1	72.34	PQEADRLDELEYEEVELYKSSHRDKLGLMVCYRTDDEEDLGIYVGEVNPNSIAAKDGRIREGDRIIQINGVDVQNREEAVAILSQEENTNISLLVARPESQLAKRWKDSD
PDZK7	PDZD7	1	2EEH	97.67	DIIHSVRVEKSPAGRLGFSVRGGSEHGLGIFVSKVEEGSSAERAGLCVGDKITEVNGLSLESTTMGSAVKVLTSSSRLHMMVRRMGRVP
PDZK7	PDZD7	2	1UF1	50	SDTSSEDGVRRIVHLYTTSDDFCLGFNIRGGKEFGLGIYVSKVDHGGLAEENGIKVGDQVLAANGVRFDDISHSQAVEVLKGQTHIMLTIKETGRYPAYKEMVSEYCWLDRLSNG

PDZK8	PDZD8	1	2EDZ	40.85	KQRSSIKTVELIKGNLQSVGLTLRLVQSTDGYAGHVIIETVAPNSPAAIADLQRGDRLIAIGGVKITSTLQVLKLIKQAGDRVLVYYERPVGQSNQGAVLQDN
PDZRN3	LNX3	1	1UHP	100	TKSLTLVLHRDSGSLGFNIIGGRPSVDNHDGSSSEGIFVSKIVDSGPAAKEGGLQIHDRIIEVNGRDLSRATHDQAVEAFKTAKEPIVVQVLRRT
PDZRN3	LNX3	2	1WH1	100	DIHQEMDREELELEEVDLYRMNSQDKLGLTVCYRTDDEDDIGIYISEIDPNSIAAKDGRIREGDRIIQINGIEVQNREEAVALLTSEENKNFSLLIARPELQLDEGWMDDD
PDZRN4	LNX4	1	1UHP	67.03	HRRDGEHKPFTIVLERENDTLGFNIIGGRPNQNNQEGTSTEGIYVSKILENGPADRADGLEIHDKIMEVNGKDLSKATHEEAVEAFRNAKEPIVVQVLRRTPLSR
PDZRN4	LNX4	2	1WH1	75.79	PADADRTEDFEYEEVELCRVSSQEKLGLTVCYRTDDEEDTGIYVSEVDPNSIAAKDGRIREGDRILQINGEDVQNREEAVALLSNDECKRIVLLVARPEIQLDEGWLEDE
PICK1	PICK1	1	2GZV	100	TVPGKVTLQKDAQNLIGISIGGGAQYCPCLYIVQVFDNTPAALDGTVAAGDEITGVNGRSIKGKTKVEVAKMIQEVKGEVTIHYNKLQADPKQ
PPP1R9A	NRB1	1	1WF8	100	KLELFPVELEKDEDGLGISIIGMGVGADAGLEKLGIFVKTVTEGGAAQRDGRIQVNDQIVEVDGISLVGVTQNFAATVLRNTKGNVRFVIGREKP
PPP1R9B	NRB2	1	1WF8	100	RLELFPVELEKDSEGLGISIIGMGAGADMGLEKLGIFVKTVTEGGAAHRDGRIQVNDLLVEVDGTSLVGVTQSFAASVLRNTKGRVRFMIGRER
PRX	PRX	1	2EHR	33.33	MEARSRSAEELRRAELVEIIVETEAQTGVSGINVAGGGKEGIFVRELREDSPAARSLSLQEGDQLLSARVFFENFKYEDALRLLQCAEPYKVSFCLKRTVPTGDLALRPGTVSG
PSCDBP	PSCDBP	1	2Z17	100	RSSSLSDFSWSQRKLVTVEKQDNETFGFEIQSYRPQNQNACSSEMFTLICKIQEDSPAHCAGLQAGDVLANINGVSTEGFTYKQVVDLIRSSGNLLTIETLNGT
PSMD9	PSMD9	1	3GDV	30.77	KEKQARDMAEAHKEAMSRKLGQSESQGPPRAFAKVNSISPGSPASIAGLQVDDEIVEFGSVNTQNFQSLHNIGSVVQHSEGKPLNVTVIRRGEKHQLRLVPTRWAGKGLLGCNIIPLQR
PTPN13	PTPL1	1	1UM7	47.56	SSPEREITLVNLKKDAKYGLGFQIIGGEKMGRLDLGIFISSVAPGGPADLDGCLKPGDRLISVNSVSLEGVSHHAAIEILQNAPEDVTLVISQPKEKISKVPST
PTPN13	PTPL1	2	1OZI	100	KPGDIFEVELAKNDNSLGISVTVLFDKGGVNTSVRHGGIYVKAVIPQGAAESDGRIHKGDRVLAVNGVSLEGATHKQAVETLRNTGQVVHLLLEKGQSP
PTPN13	PTPL1	3	2OGP	37.78	TEENTFEVKLFKNSSGLGFSFSREDNLIPEQINASIVRVKKLFPGQPAAESGKIDVGDVILKVNGASLKGLSQQEVISALRGTAPEVFLLLCRPPPG
PTPN13	PTPL1	4	2BYG	40.7	EDFELEVELLITLIKSEKGSLGFTVTKGNQRIGCYVHDVIQDPAKSDGRLKPGDRLIKVNDTDVTNMTHTDAVNLLRAASKTVRLVIGRVLELPR
PTPN13	PTPL1	5	1X6D	31.76	PMLPHLLPDITLTCNKEELGFSLCGGHDSLYQVVYISDINPRSVAAIEGNLQLLDVIHYVNGVSTQGMTLEEVNRALDMSLPSLVLKATRNDLPVVPSS
PTPN3	PTPH1	1	2VPH	70.97	DSYLVLIRITPDEDGKFGFNLKGGVDQKMPLVVSRINPESPADTCIPKLNEGDQIVLINGRDISEHTHDQVVMFIKASRESHSRELALVIRRR
PTPN4	PTPN4	1	2VPH	100	DNLVLIRMKPDENGRFGFNVKGGYDQKMPVIVSRVAPGTPADLCVPRLNEGDQVVLINGRDIAEHTHDQVVLFIKASCERHSGELMLLVRP
RAPGEF2	PDZGEF1	1	1WI2	50	NIACAAKAKRRLMTLTKPSREAPLPFILLGGSEKGFGIFVDSVDSGSKATEAGLKRGDQILEVNGQNFENIQLSKAMEILRNNTHLSITVKTNLFVFKE
RAPGEF6	PDZGEF2	1	2EEH	54.41	KAKWRQVVLQKASRESPLQFSLNGGSEKGFGIFVEGVEPGSKAADSGLKRGDQIMEVNGQNFENITFMKAVEILRNNTHLALTVKTN
RGS12	RGS12	1	1TQ3	35.48	RAGEASKRPLPGPSPPRVRSVEVARGRAGYGFTLSGQAPCVLSCVMRGSPADFVGLRAGDQILAVNEINVKKASHEDVVKLIGKCSGVLHMVIAEGVGRFESCSSDEEGGLYEGKGWLK
RGS3	RGS3	1	1WHD	100	SERRYRQITIPRGKDGFGFTICCDSPVRVQAVDSGGPAERAGLQQLDTVLQLNERPVEHWKCVELAHEIRSCPSEIILLVWRMVPQ
RHPN1	RHOPHILIN1	1	1VAE	46.97	AKNRWRLVGPVHLTRGEGGFGLTLRGDSPVLIAAVIPGSQAAAAGLKEGDYIVSVNGQPCRWWRHAEVVTELKAAGEAGASLQVVSLLPSSRLPSLGDR
RHPN2	RHOPHILIN2	1	1VAE	96.47	SANKRWTPPRSIRFTAEEGDLGFTLRGNAPVQVHFLDPYCSASVAGAREGDYIVSIQLVDCKWLTLSEVMKLLKSFGEDEIEMKVVSLLDSTSSMHNKSAT
RIMS1	RIM1	1	2CSS	98.9	PVTWQPSKEGDRLIGRVILNKRTTMPKDSGALLGLKVVGGKMTDLGRLGAFITKVKKGSLADVVGHLRAGDEVLEWNGKPLPGATNEEVYNIILESKSEPQVEIIVSRP
RIMS2	RIM2	1	1WFG	100	HSHSDKHPVTWQPSKDGDRLIGRILLNKRLKDGSVPRDSGAMLGLKVVGGKMTESGRLCAFITKVKKGSLADTVGHLRPGDEVLEWNGRLLQGATFEEVYNIILESKPEPQVELVVSRPIGD
SCRIB	SCRIB	1	1X5Q	100	EPARIEEEELTLTILRQTGGLGISIAGGKGSTPYKGDDEGIFISRVSEEGPAARAGVRVGDKLLEVNGVALQGAEHHEAVEALRGAGTAVQMRVWRE
SCRIB	SCRIB	2	1WHA	100	RQRHVACLARSERGLGFSIAGGKGSTPYRAGDAGIFVSRIAEGGAAHRAGTLQVGDRVLSINGVDVTEARHDHAVSLLTAASPTIALLLERE
SCRIB	SCRIB	3	1IU2	45.26	PYPVEEIRLPRAGGPLGLSIVGGSDHSSHPFGVQEPGVFISKVLPRGLAARSGLRVGDRILAVNGQDVRDATHQEAVSALLRPCLELSLLVRRDP
SCRIB	SCRIB	4	1UJU	100	RRDPAPPGLRELCIQKAPGERLGISIRGGARGHAGNPRDPTDEGIFISKVSPTGAAGRDGRLRVGLRLLEVNQQSLLGLTHGEAVQLLRSVGDTLTVLVCDGFEASTD
SDCBP	SYNTENIN1	1	1YBO	100	GIREVILCKDQDGKIGLRLKSIDNGIFVQLVQANSPASLVGLRFGDQVLQINGENCAGWSSDKAHKVLKQAFGEKITMTIRDRPFERT
SDCBP	SYNTENIN1	2	1YBO	100	ERTITMHKDSTGHVGFIFKNGKITSIVKDSSAARNGLLTEHNICEINGQNVIGLKDSQIADILSTSGTVVTITIMPA
SDCBP2	SYNTENIN2	1	1YBO	70.65	KPGVREIHLCKDERGKTGLRLRKVDQGLFVQLVQANTPASLVGLRFGDQLLQIDGRDCAGWSSHKAHQVVKKASGDKIVVVVRDRPFQRT
SDCBP2	SYNTENIN2	2	1YBO	69.74	RPFQRTVTMHKDSMGHVGFVIKKGKIVSLVKGSSAARNGLLTNHYVCEVDGQNVIGLKDKKIMEILATAGNVVTLTIIPSVIYEH
SHAN1	SHANK1	1	1Q3P	100	SDYIIKEKTVLLQKKDSEGFGFVLRGAKAQTPIEEFTPTPAFPALQYLESVDEGGVAWRAGLRMGDFLIEVNGQNVVKVGHRQVVNMIRQGGNTLMVKVVMVTRHPD
SHANK2	SHANK2	1	1Q3P	87	SDCIIEEKTVVLQKKDNEGFGFVLRGAKADTPIEEFTPTPAFPALQYLESVDEGGVAWQAGLRTGDFLIEVNNENVVKVGHRQVVNMIRQGGNHLVLKVVTVTRNLDP
SHANK3	SHANK3	1	1Q3O	81	HSDYVIDDKVAVLQKRDHEGFGFVLRGAKAETPIEEFTPTPAFPALQYLESVDVEGVAWRAGLRTGDFLIEVNGVNVVKVGHKQVVALIRQGGNRLVMKVVSVTRKPEE
SHROOM3	SHROOM	1	2EDP	60.98	TKGRYIYLEAFLEGGAPWGFTLKGGLEHGEPLIISKVEEGGKADTLSSKLQAGDEVVHINEVTLSSSRKEAVSLVKGSYKTLRLVVRRDVCTD
SHROOM4	SHROOM4	1	2EDP	98.82	SFQYVPVQLQGGAPWGFTLKGGLEHCEPLTVSKIEDGGKAALSQKMRTGDELVNINGTPLYGSRQEALILIKGSFRILKLIVRRRNAPV
SIPA1	SIPA1	1	2EEH	34.78	CETRELALPRDGQGRLGFEVDAEGFVTHVERFTFAETAGLRPGARLLRVCGQTLPSLRPEAAAQLLRSAPKVCVTVLPPDES

SIPA1L1	SIPA1L1	1	1Q3P	36.84	SKGCESVEMTLRRNGLGQLGFHVNYEGIVADVEPYGYAWQAGLRQGSRLVEICKVAVATLSHEQMIDLLRTSVTVKVVIIPPHDD
SIPA1L2	SIPA1L2	1	1Q3P	36.84	TRGCETVEMTLRRNGLGQLGFHVNFEGIVADVEPFGFAWKAGLRQGSRLVEICKVAVATLTHEQMIDLLRTSVTVKVVIIQPHDD
SIPA1L3	SIPA1L3	1	1Q3P	36.84	TSGWETVDMTLRRNGLGQLGFHVKYDGTVAEVEDYGFAWQAGLRQGSRLVEICKVAVVTLTHDQMIDLLRTSVTVKVVIIPPFED
SLC9A3R1	NHERF	1	1G9O	100	PLPRLCCLEKGPNGYGFHLHGEKGKLGQYIRLVEPGSPAEKAGLLAGDRLVEVNGENVEKETHQQVVSRIRAALNAVRLLVVDPETDEQ
SLC9A3R1	NHERF	2	2JXO	100	EQRELRPRLCTMKKGPSGYGFNLHSDKSKPGQFIRSVDPDSPAEASGLRAQDRIVEVNGVCMEGKQHGDVVSAIRAGGDETKLLVVDRETDE
SLC9A3R2	NHERF2	1	2OCS	100	PRLCRLVRGEQGYGFHLHGEKGRRGQFIRRVEPGSPAEAAALRAGDRLVEVNGVNVEGETHHQVVQRIKAVEGQTRLLVVDQE
SLC9A3R2	NHERF2	2	2JXO	97.62	GPLRELRPRLCHLRKGPQGYGFNLHSDKSRPGQYIRSVDPGSPAARSGLRAQDRLIEVNGQNVEGLRHAEVVASIKAREDEARLLVVDPETDEHFKR
SNTA1	SNT1	1	2PDZ	98.81	QRRRVTVRKADAGGLGISIKGGRENKMPILISKIFKGLAADQTEALFVGDAILSVNGEDLSSATHDEAVQVLKKTGKEVVLEVKYMKD
SNTB1	SNT2B1	1	2VRF	79.45	SNQKRGVKVLKQELGGLGISIKGGKENKMPILISKIFKGLAADQTQALYVGDAILSVNGADLRDATHDEAVQALKRAGKEVLLEVKYMRE
SNTB2	SNTB2	1	1Z86	95.6	PVRRVRVVKQEAGGLGISIKGGRENRMPILISKIFPGLAADQSRALRLGDAILSVNGTDLRQATHDQAVQALKRAGKEVLLEVKFIRE
SNTG1	SNTG1	1	1Z86	43.48	GERTVTIRRQTVGGFGLSIKGGAEHNIPVVVSKISKEQRAELSGLLFIGDAILQINGINVRKCRHEEVVQVLRNAGEEVTLTVSFLKRAP
SNX27	SNX27	1	1Q3P	41.94	GPRVVRIVKSESGYGFNVRGQVSEGGQLRSINGELYAPLQHVSAVLPGGAADRAGVRKGDRILEVNHVNVEGATHKQVVDLIRAGEKELILTVLSVPPHEAD
STXB4	STXBP4	1	1WI4	87.36	EKDPAFQMITIAKETGLGLKVLGGINRNEGPLVYIQEIIPGGDCYKDGRLKPGDQLVSVNKESMIGVSFEEAKSIITGAKLRLESAWEIAFIRQKSDN
SYNJ2BP	SYNJ2BP	1	2JIK	100	DYLVTEEEINLTRGPSGLGFNIVGGTDQQYVSNDSGIYVSRIKENGAAALDGRLQEGDKILSVNGQDLKNLLHQDAVDLFRNAGYAVSLRVQHRLQVQ
SYNP2	SYNPO2	1	2PKT	37.18	GTGDFICISMTGGAPWGFRLQGGKEQKQPLQVAKIRNQSKASGSGLCEGDEVVSINGNPCADLTYPEVIKLMESITDSLQMLIKRPSSG
SYNPO2L	SYNPO2L	1	1RGW	39.74	MGAEEEVLVTLSGGAPWGFRLHGGAEQRKPLQVSKIRRRSQAGRAGLRERDQLLAINGVSCTNLSHASAMSLIDASGNQLVLTVQRLAD
TIAM1	TIAM1	1	2D8I	100	EIEICPKVTQSIHIEKSDTAADTYGFSLSSVEEDGIRRLYVNSVKETGLASKKGLKAGDEILEINNRAADALNSSMLKDFLSQPSLGLLVRTYPELEEGVE
TIAM2	TIAM2	1	2D8I	27.06	YDEIEVFPLNVYDVQLTKTGSVCDFGFAVTAQVDERQHLSRIFISDVLPDGLAYGEGLRKGNEIMTLNGEAVSDLDLKQMEALFSEKSVGLTLIARPPDTKAT
TJP1	ZO1	1	2H2C	98.86	EETAIWEQHTVTLHRAPGFGFGIAISGGRDNPHFQSGETSIVISDVLKGGPAEGQLQENDRVAMVNGVSMDNVEHAFAVQQLRKSGKNAKITIRRKKKVQ
TJP1	ZO1	2	2RCZ	98.86	PTKVTLVKSRKNEEYGLRLASHIFVKEISQDSLAARDGNIQEGDVVLKINGTVTENMSLTDAKTLIERSKGKLKMVVQRDE
TJP1	ZO1	3	3TSV	100	DGILRPSMKLVKFRKGDSVGLRLAGGNDVGIFVAGVLEDSPAAKEGLEEGDQILRVNNVDFTNIIREEAVLFLLDLPKGEEVTILAQKKKDVYRRIVESDVG
TJP2	ZO2	1	2CSJ	95.92	MEELIWEQYTVTLQKDSKRGFGIAVSGGRDNPHFENGETSIVISDVLPGGPADGLLQENDRVVMVNGTPMEDVLHSFAVQQLRKSGKVAAIVVKRPRKVQVAALQ
TJP2	ZO2	2	3E17	100	RGRPGPIGVLLMKSRANEEYGLRLGSQIFVKEMTRTGLATKDGNLHEGDIILKINGTVTENMSLTDARKLIEKSRGKLQLVVLRDSQQT
TJP2	ZO2	3	1UF1	41.67	EDEAIYGPNTKMVRFKKGDSVGLRLAGGNDVGIFVAGIQEGTSAEQEGLQEGDQILKVNTQDFRGLVREDAVLYLLEIPKGEMVTILAQSRADVYRD
TJP3	ZO3	1	2CSJ	53.76	MEELTIWEQHTATLSKDPRRGFGIAISGGRDRPGGSMVVSDVVPGGPAEGRLQTGDHIVMVNGVSMENATSAFAIQILKTCTKMANITVKRPRRIHLPATKASPSSPGR
TJP3	ZO3	2	3E17	57.89	QMKPVKSVLVKRRDSEEFGVKLGSQIFIKHITDSGLAARHRGLQEGDLILQINGVSSQNLSLNDTRRLIEKSEGKLSLLVLRDRGQ
TJP3	ZO3	3	1UF1	45.76	EDRGYSPDTRVVRFLKGKSIGLRLAGGNDVGIFVSGVQAGSPADGQGIQEGDQILQVNDVPFQNLTREEAVQFLLGLPPGEEMELVTQRKQDIFWKMVQSRVGDSFYIRT
TX1B3	TIP1	1	3DJ1	100	QPVTAVVQRVEIHKLRQGENLILGFSIGGGIDQDPSQNPFSEDKTDKGIYVTRVSEGGPAEIAGLQIGDKIMQVNGWDMTMVTHDQARKRLTKRSEEVVRLLVTRQSLQKAVQQSMLS
USH1C	HARMONIN	1	1UF1	35.37	DQLTPRRSRKLKEVRLDRLHPEGLGLSVRGGLEFGCGLFISHLIKGGQADSVGLQVGDEIVRINGYSISSCTHEEVINLIRTKKTVSIKVRHIGLIPVKSSPDE
USH1C	HARMONIN	2	2KBS	100	KEKKVFISLVGSRGLGCSISSGPIQKPGIFISHVKPGSLSAEVGLEIGDQIVEVNGVDFSNLDHKEAVNVLKSSRSLTISIVAAAGRELFMT
USH1C	HARMONIN	3	1V6B	96	SMFTPEQIMGKDVRLLRIKKEGSLDLALEGGVDSPIGKVVVSAVYERGAAERHGGIVKGDEIMAINGKIVTDYTLAEAEAALQKAWNQGGDWIDLVVAVCPPKEYDDE

Bibliography

[1] Sudol M (1998) From Src Homology domains to other signaling modules: proposal of the 'protein recognition code'. Oncogene 17: 1469–1474.

[2] Bork P, Downing AK, Kieffer B, Campbell ID (1996) Structure and distribution of modules in extracellular proteins. Q Rev Biophys 29: 119–167.

[3] Diella F, Haslam N, Chica C, Budd A, Michael S, et al. (2008) Understanding eukaryotic linear motifs and their role in cell signaling and regulation. Front Biosci 13: 6580–6603.

[4] Letunic I, Doerks T, Bork P (2012) SMART 7: recent updates to the protein domain annotation resource. Nucleic Acids Res 40: D302–D305.

[5] Gould CM, Diella F, Via A, Puntervoll P, Gemünd C, et al. (2010) ELM: the status of the 2010 eukaryotic linear motif resource. Nucleic Acids Res 38: D167–D180.

[6] Harris BZ, Lim WA (2001) Mechanism and role of PDZ domains in signaling complex assembly. J Cell Sci 114: 3219–3231.

[7] Chen Q, Niu X, Xu Y, Wu J, Shi Y (2007) Solution structure and backbone dynamics of the AF-6 PDZ domain/Bcr peptide complex. Protein Sci 16: 1053–1062.

[8] Doyle DA, Lee A, Lewis J, Kim E, Sheng M, et al. (1996) Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ. Cell 85: 1067–1076.

[9] Tonikian R, Zhang Y, Sazinsky SL, Currell B, Yeh JH, et al. (2008) A specificity map for the PDZ domain family. PLoS Biol 6: e239.

[10] Stiffler MA, Chen JR, Grantcharova VP, Lei Y, Fuchs D, et al. (2007) PDZ domain binding selectivity is optimized across the mouse proteome. Science 317: 364–369.

[11] Chen JR, Chang BH, Allen JE, Stiffler MA, MacBeath G (2008) Predicting PDZ domain-peptide interactions from primary sequences. Nat Biotechnol 26: 1041–1045.

[12] Feng W, Wu H, Chan LN, Zhang M (2008) Par-3-mediated junctional localization of the lipid phosphatase PTEN is required for cell polarity establishment. J Biol Chem 283: 23440–23449.

[13] Petit CM, Zhang J, Sapienza PJ, Fuentes EJ, Lee AL (2009) Hidden dynamic allostery in a PDZ domain. Proc Natl Acad Sci USA 106: 18249–18254.

[14] Mostarda S, Gfeller D, Rao F (2012) Beyond the Binding Site: The Role of the beta2 - beta3 Loop and Extra-Domain Structures in PDZ Domains. PLoS Comput Biol 8: e1002429.

[15] Nomme J, Fanning AS, Caffrey M, Lye MF, Anderson JM, et al. (2011) The Src Homology 3 domain is required for junctional adhesion molecule binding to the third PDZ domain of the scaffolding protein ZO-1. J Biol Chem 286: 43352–43360.

[16] Luck K, Travé G (2011) Phage display can select over-hydrophobic sequences that may impair prediction of natural domain-peptide interactions. Bioinformatics 27: 899–902.

[17] Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009) Jalview version 2–a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189–1191.

[18] Charbonnier S, Nominé Y, Ramírez J, Luck K, Chapelle A, et al. (2011) The structural and dynamic response of MAGI-1 PDZ1 with noncanonical domain boundaries to the binding of human papillomavirus E6. J Mol Biol 406: 745–763.

[19] Luck K, Fournane S, Kieffer B, Masson M, Nominé Y, et al. (2011) Putting into practice domain-linear motif interaction predictions for exploration of protein networks. PLoS One 6: e25376.

[20] Alberts B, Johnson A, Lewis J, Raff M, Roberts K, et al. (2007) Molecular Biology of the Cell. Garland Science, 5th edition.

[21] Lodish H, Berk A, Kaiser CA, Krieger M, Scott MP, et al. (2007) Molecular Cell Biology. W. H. Freeman, 6th edition.

[22] Tan CSH, Pasculescu A, Lim WA, Pawson T, Bader GD, et al. (2009) Positive selection of tyrosine loss in metazoan evolution. Science 325: 1686–1688.

[23] Janin J, Wodak SJ (2002) Protein modules and protein-protein interaction. Introduction. Adv Protein Chem 61: 1–8.

[24] Hegyi H, Bork P (1997) On the classification and evolution of protein modules. J Protein Chem 16: 545–551.

[25] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921.

[26] Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, et al. (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Res 40: D261–D270.

[27] Baker M (2012) Proteomics: The interaction map. Nature 484: 271–275.

[28] Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, et al. (2000) A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. Nature 403: 623–627.

[29] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, et al. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci USA 98: 4569–4574.

[30] Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature 440: 631–636.

[31] Bandyopadhyay S, Chiang CY, Srivastava J, Gersten M, White S, et al. (2010) A human MAP kinase interactome. Nat Methods 7: 801–805.

[32] Hu H, Columbus J, Zhang Y, Wu D, Lian L, et al. (2004) A map of WW domain family interactions. Proteomics 4: 643–655.

[33] Sanger F, Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 94: 441–448.

[34] Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74: 5463–5467.

[35] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.

[36] Copley RR, Ponting CP, Schultz J, Bork P (2002) Sequence analysis of multidomain proteins: past perspectives and future directions. Adv Protein Chem 61: 75–98.

[37] Campbell ID, Baron M (1991) The structure and function of protein modules. Philos Trans R Soc Lond B Biol Sci 332: 165–170.

[38] Sadowski I, Stone JC, Pawson T (1986) A noncatalytic domain conserved among cytoplasmic protein-tyrosine kinases modifies the kinase function and transforming activity of Fujinami sarcoma virus P130gag-fps. Mol Cell Biol 6: 4396–4408.

[39] Mayer BJ, Hamaguchi M, Hanafusa H (1988) A novel viral oncogene with structural similarity to phospholipase c. Nature 332: 272–275.

[40] Mayer BJ, Jackson PK, Baltimore D (1991) The noncatalytic src homology region 2 segment of abl tyrosine kinase binds to tyrosine-phosphorylated cellular proteins with high affinity. Proc Natl Acad Sci U S A 88: 627–631.

[41] Cicchetti P, Mayer BJ, Thiel G, Baltimore D (1992) Identification of a protein that binds to the sh3 region of abl and is similar to bcr and gap-rho. Science 257: 803–806.

[42] Ren R, Mayer BJ, Cicchetti P, Baltimore D (1993) Identification of a ten-amino acid proline-rich sh3 binding site. Science 259: 1157–1161.

[43] Gibson TJ (2009) Cell regulation: determined to signal discrete cooperation. Trends Biochem Sci 34: 471–482.

[44] Davey NE, Roey KV, Weatheritt RJ, Toedt G, Uyar B, et al. (2012) Attributes of short linear motifs. Mol Biosyst 8: 268–281.

[45] Fuxreiter M, Tompa P, Simon I (2007) Local structural disorder imparts plasticity on linear motifs. Bioinformatics 23: 950–956.

[46] Via A, Gould CM, Gemnd C, Gibson TJ, Helmer-Citterich M (2009) A structure filter for the eukaryotic linear motif resource. BMC Bioinformatics 10: 351.

[47] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The protein data bank. Nucleic Acids Res 28: 235–242.

[48] Neduva V, Linding R, Su-Angrand I, Stark A, de Masi F, et al. (2005) Systematic discovery of new recognition peptides mediating protein interaction networks. PLoS Biol 3: e405.

[49] Neduva V, Russell RB (2006) Peptides mediating interaction networks: new leads at last. Curr Opin Biotechnol 17: 465–471.

[50] Cho KO, Hunt CA, Kennedy MB (1992) The rat brain postsynaptic density fraction contains a homolog of the Drosophila discs-large tumor suppressor protein. Neuron 9: 929–942.

[51] Bryant PJ, Watson KL, Justice RW, Woods DF (1993) Tumor suppressor genes encoding proteins required for cell interactions and signal transduction in Drosophila. Dev Suppl : 239–249.

[52] Ponting CP, Phillips C (1995) DHR domains in syntrophins, neuronal NO synthases and other intracellular proteins. Trends Biochem Sci 20: 102–103.

[53] Kennedy MB (1995) Origin of PDZ (DHR, GLGF) domains. Trends Biochem Sci 20: 350.

[54] Ponting CP (1997) Evidence for PDZ domains in bacteria, yeast, and plants. Protein Sci 6: 464–468.

[55] Wang CK, Pan L, Chen J, Zhang M (2010) Extensions of PDZ domains as important structural and functional elements. Protein Cell 1: 737–751.

[56] Velthuis AJWT, Sakalis PA, Fowler DA, Bagowski CP (2011) Genome-wide analysis of PDZ domain binding reveals inherent functional overlap within the PDZ interaction network. PLoS One 6: e16047.

[57] Cabral JHM, Petosa C, Sutcliffe MJ, Raza S, Byron O, et al. (1996) Crystal structure of a PDZ domain. Nature 382: 649–652.

[58] Thompson JD, Gibson TJ, Higgins DG (2002) Multiple sequence alignment using ClustalW and ClustalX. Curr Protoc Bioinformatics Chapter 2: Unit 2.3.

[59] Bliven S, Prlić A (2012) Circular permutation in proteins. PLoS Comput Biol 8: e1002445.

[60] Zhang Y, Appleton BA, Wu P, Wiesmann C, Sidhu SS (2007) Structural and functional analysis of the ligand specificity of the HtrA2/Omi PDZ domain. Protein Sci 16: 1738–1750.

[61] Truschel ST, Sengupta D, Foote A, Heroux A, Macbeth MR, et al. (2011) Structure of the membrane-tethering GRASP domain reveals a unique PDZ ligand interaction that mediates Golgi biogenesis. J Biol Chem 286: 20125–20129.

[62] Korotkov KV, Krumm B, Bagdasarian M, Hol WGJ (2006) Structural and functional studies of EpsC, a crucial component of the type 2 secretion system from Vibrio cholerae. J Mol Biol 363: 311–321.

[63] Runyon ST, Zhang Y, Appleton BA, Sazinsky SL, Wu P, et al. (2007) Structural and functional analysis of the PDZ domains of human HtrA1 and HtrA3. Protein Sci 16: 2454–2471.

[64] Liao DI, Qian J, Chisholm DA, Jordan DB, Diner BA (2000) Crystal structures of the photosystem II D1 C-terminal processing protease. Nat Struct Biol 7: 749–753.

[65] Mühlhahn P, Zweckstetter M, Georgescu J, Ciosto C, Renner C, et al. (1998) Structure of interleukin 16 resembles a PDZ domain with an occluded peptide binding site. Nat Struct Biol 5: 682–686.

[66] Kornau HC, Schenker LT, Kennedy MB, Seeburg PH (1995) Domain interaction between NMDA receptor subunits and the postsynaptic density protein PSD-95. Science 269: 1737–1740.

[67] Kim E, Niethammer M, Rothschild A, Jan YN, Sheng M (1995) Clustering of Shaker-type K+ channels by interaction with a family of membrane-associated guanylate kinases. Nature 378: 85–88.

[68] Gerek ZN, Keskin O, Ozkan SB (2009) Identification of specificity and promiscuity of PDZ domain interactions through their dynamic behavior. Proteins 77: 796–811.

[69] Harris BZ, Lau FW, Fujii N, Guy RK, Lim WA (2003) Role of electrostatic interactions in PDZ domain ligand recognition. Biochemistry 42: 2797–2805.

[70] Songyang Z, Fanning AS, Fu C, Xu J, Marfatia SM, et al. (1997) Recognition of unique carboxyl-terminal motifs by distinct PDZ domains. Science 275: 73–77.

[71] Zhang Y, Dasgupta J, Ma RZ, Banks L, Thomas M, et al. (2007) Structures of a human papillomavirus (HPV) E6 polypeptide bound to MAGUK proteins: mechanisms of targeting tumor suppressors by a high-risk HPV oncoprotein. J Virol 81: 3618–3626.

[72] Fournane S, Charbonnier S, Chapelle A, Kieffer B, Orfanoudakis G, et al. (2011) Surface plasmon resonance analysis of the binding of high-risk mucosal HPV E6 oncoproteins to the PDZ1 domain of the tight junction protein MAGI-1. J Mol Recognit 24: 511–523.

[73] Stricker NL, Christopherson KS, Yi BA, Schatz PJ, Raab RW, et al. (1997) PDZ domain of neuronal nitric oxide synthase recognizes novel C-terminal peptide sequences. Nat Biotechnol 15: 336–342.

[74] Tochio H, Zhang Q, Mandal P, Li M, Zhang M (1999) Solution structure of the extended neuronal nitric oxide synthase PDZ domain complexed with an associated peptide. Nat Struct Biol 6: 417–421.

[75] Tyler RC, Peterson FC, Volkman BF (2010) Distal interactions within the par3-VE-cadherin complex. Biochemistry 49: 951–957.

[76] Madsen KL, Beuming T, Niv MY, Chang CW, Dev KK, et al. (2005) Molecular determinants for the complex binding specificity of the PDZ domain in PICK1. J Biol Chem 280: 20539–20548.

[77] Liu Y, Henry GD, Hegde RS, Baleja JD (2007) Solution structure of the hDlg/SAP97 PDZ2 domain and its mechanism of interaction with HPV-18 papillomavirus E6 protein. Biochemistry 46: 10864–10874.

[78] Kimple ME, Siderovski DP, Sondek J (2001) Functional relevance of the disulfide-linked complex of the N-terminal PDZ domain of InaD with NorpA. EMBO J 20: 4414–4422.

[79] Lu J, Li H, Wang Y, Südhof TC, Rizo J (2005) Solution structure of the RIM1alpha PDZ domain in complex with an ELKS1b C-terminal peptide. J Mol Biol 352: 455–466.

[80] Nourry C, Grant SGN, Borg JP (2003) PDZ domain proteins: plug and play! Sci STKE 2003: RE7.

[81] Wiedemann U, Boisguerin P, Leben R, Leitner D, Krause G, et al. (2004) Quantification of PDZ domain specificity, prediction of ligand affinity and rational design of super-binding peptides. J Mol Biol 343: 703–718.

[82] Zhang Y, Yeh S, Appleton BA, Held HA, Kausalya PJ, et al. (2006) Convergent and divergent ligand specificity among PDZ domains of the LAP and zonula occludens (ZO) families. J Biol Chem 281: 22299–22311.

[83] Hillier BJ, Christopherson KS, Prehoda KE, Bredt DS, Lim WA (1999) Unexpected modes of PDZ domain scaffolding revealed by structure of nNOS-syntrophin complex. Science 284: 812–815.

[84] Gee SH, Sekely SA, Lombardo C, Kurakin A, Froehner SC, et al. (1998) Cyclic peptides as non-carboxyl-terminal ligands of syntrophin PDZ domains. J Biol Chem 273: 21980–21987.

[85] Penkert RR, DiVittorio HM, Prehoda KE (2004) Internal recognition through PDZ domain plasticity in the Par-6-Pals1 complex. Nat Struct Mol Biol 11: 1122–1127.

[86] Zhang Y, Appleton BA, Wiesmann C, Lau T, Costa M, et al. (2009) Inhibition of Wnt signaling by Dishevelled PDZ peptides. Nat Chem Biol 5: 217-219.

[87] Lemaire JF, McPherson PS (2006) Binding of Vac14 to neuronal nitric oxide synthase: Characterisation of a new internal PDZ-recognition motif. FEBS Lett 580: 6948–6954.

[88] Wong HC, Bourdelas A, Krauss A, Lee HJ, Shao Y, et al. (2003) Direct binding of the PDZ domain of Dishevelled to a conserved internal sequence in the C-terminal region of Frizzled. Mol Cell 12: 1251–1260.

[89] Lenfant N, Polanowska J, Bamps S, Omi S, Borg JP, et al. (2010) A genome-wide study of PDZ-domain interactions in C. elegans reveals a high frequency of non-canonical binding. BMC Genomics 11: 671.

[90] Werme K, Wigerius M, Johansson M (2008) Tick-borne encephalitis virus NS5 associates with membrane protein scribble and impairs interferon-stimulated JAK-STAT signalling. Cell Microbiol 10: 696–712.

[91] Ellencrona K, Syed A, Johansson M (2009) Flavivirus NS5 associates with host-cell proteins zonula occludens-1 (ZO-1) and regulating synaptic membrane exocytosis-2 (RIMS2) via an internal PDZ binding mechanism. Biol Chem 390: 319–323.

[92] Gfeller D, Butty F, Wierzbicka M, Verschueren E, Vanhee P, et al. (2011) The multiple-specificity landscape of modular peptide recognition domains. Mol Syst Biol 7: 484.

[93] Gallardo R, Ivarsson Y, Schymkowitz J, Rousseau F, Zimmermann P (2010) Structural diversity of PDZ-lipid interactions. Chembiochem 11: 456–467.

[94] Chen Y, Sheng R, Källberg M, Silkov A, Tun MP, et al. (2012) Genome-wide functional annotation of dual-specificity protein- and lipid-binding modules that regulate protein interactions. Mol Cell 46: 226-237.

[95] Brenman JE, Chao DS, Gee SH, McGee AW, Craven SE, et al. (1996) Interaction of nitric oxide synthase with the postsynaptic density protein PSD-95 and alpha1-syntrophin mediated by PDZ domains. Cell 84: 757–767.

[96] Fouassier L, Yun CC, Fitz JG, Doctor RB (2000) Evidence for ezrin-radixin-moesin-binding phosphoprotein 50 (EBP50) self-association through PDZ-PDZ interactions. J Biol Chem 275: 25039–25045.

[97] Im YJ, Park SH, Rho SH, Lee JH, Kang GB, et al. (2003) Crystal structure of GRIP1 PDZ6-peptide complex reveals the structural basis for class II PDZ target recognition and PDZ domain-mediated multimerization. J Biol Chem 278: 8501–8507.

[98] Im YJ, Lee JH, Park SH, Park SJ, Rho SH, et al. (2003) Crystal structure of the Shank PDZ-ligand complex reveals a class I PDZ interaction and a novel PDZ-PDZ dimerization. J Biol Chem 278: 48099–48104.

[99] Iskenderian-Epps WS, Imperiali B (2010) Modulation of Shank3 PDZ domain ligand-binding affinity by dimerization. Chembiochem 11: 1979–1984.

[100] Utepbergenov DI, Fanning AS, Anderson JM (2006) Dimerization of the scaffolding protein ZO-1 through the second PDZ domain. J Biol Chem 281: 24671–24677.

[101] Chen J, Pan L, Wei Z, Zhao Y, Zhang M (2008) Domain-swapped dimerization of ZO-1 PDZ2 generates specific and regulatory connexin43-binding sites. EMBO J 27: 2113–2123.

[102] Wu J, Yang Y, Zhang J, Ji P, Du W, et al. (2007) Domain-swapped dimerization of the second PDZ domain of ZO2 may provide a structural basis for the polymerization of claudins. J Biol Chem 282: 35988–35999.

[103] Chang BH, Gujral TS, Karp ES, BuKhalid R, Grantcharova VP, et al. (2011) A systematic family-wide investigation reveals that 30% of mammalian PDZ domains engage in PDZ-PDZ interactions. Chem Biol 18: 1143–1152.

[104] Richier L, Williton K, Clattenburg L, Colwill K, O'Brien M, et al. (2010) NOS1AP associates with scribble and regulates dendritic spine development. J Neurosci 30: 4796–4805.

[105] Nie J, Li SSC, McGlade CJ (2004) A novel PTB-PDZ domain interaction mediates isoform-specific ubiquitylation of mammalian Numb. J Biol Chem 279: 20807–20815.

[106] Audebert S, Navarro C, Nourry C, Chasserot-Golaz S, Lécine P, et al. (2004) Mammalian Scribble forms a tight complex with the betaPIX exchange factor. Curr Biol 14: 987–995.

[107] Anderson JM (1996) Cell signalling: MAGUK magic. Curr Biol 6: 382–384.

[108] Dobrosotskaya I, Guy RK, James GL (1997) MAGI-1, a membrane-associated guanylate kinase with a unique arrangement of protein-protein interaction domains. J Biol Chem 272: 31589–31597.

[109] Iden S, Collard JG (2008) Crosstalk between small GTPases and polarity proteins in cell polarization. Nat Rev Mol Cell Biol 9: 846–859.

[110] Johnston DS, Ahringer J (2010) Cell polarity in eggs and epithelia: Parallels and diversity. Cell 141: 757–774.

[111] Roh MH, Margolis B (2003) Composition and function of PDZ protein complexes during cell polarization. Am J Physiol Renal Physiol 285: F377–F387.

[112] Funke L, Dakoji S, Bredt DS (2005) Membrane-associated guanylate kinases regulate adhesion and plasticity at cell junctions. Annu Rev Biochem 74: 219–245.

[113] González-Mariscal L, Betanzos A, Avila-Flores A (2000) MAGUK proteins: structure and role in the tight junction. Semin Cell Dev Biol 11: 315–324.

[114] Etienne-Manneville S (2008) Polarity proteins in migration and invasion. Oncogene 27: 6970–6980.

[115] Humbert PO, Grzeschik NA, Brumby AM, Galea R, Elsum I, et al. (2008) Control of tumourigenesis by the Scribble/Dlg/Lgl polarity module. Oncogene 27: 6888–6907.

[116] Etienne-Manneville S (2009) Scribble at the crossroads. J Biol 8: 104.

[117] Liu W, Wen W, Wei Z, Yu J, Ye F, et al. (2011) The INAD scaffold is a dynamic, redox-regulated modulator of signaling in the Drosophila eye. Cell 145: 1088–1101.

[118] Reiners J, Nagel-Wolfrum K, Jürgens K, Märker T, Wolfrum U (2006) Molecular basis of human Usher syndrome: deciphering the meshes of the Usher protein network provides insights into the pathomechanisms of the Usher disease. Exp Eye Res 83: 97–119.

[119] Davey NE, Travé G, Gibson TJ (2011) How viruses hijack cell regulation. Trends Biochem Sci 36: 159–169.

[120] Javier RT, Rice AP (2011) Emerging theme: cellular PDZ proteins as common targets of pathogenic viruses. J Virol 85: 11544–11556.

[121] Weiss RS, Javier RT (1997) A carboxy-terminal region required by the adenovirus type 9 E4 ORF1 oncoprotein for transformation mediates direct binding to cellular polypeptides. J Virol 71: 7873–7880.

[122] Watson RA, Thomas M, Banks L, Roberts S (2003) Activity of the human papillomavirus E6 PDZ-binding motif correlates with an enhanced morphological transformation of immortalized human keratinocytes. J Cell Sci 116: 4925–4934.

[123] Hirata A, Higuchi M, Niinuma A, Ohashi M, Fukushi M, et al. (2004) PDZ domain-binding motif of human T-cell leukemia virus type 1 Tax oncoprotein augments the transforming activity in a rat fibroblast cell line. Virology 318: 327–336.

[124] Thomas M, Narayan N, Pim D, Tomaić V, Massimi P, et al. (2008) Human papillomaviruses, cervical cancer and cell polarity. Oncogene 27: 7018–7030.

[125] Nakagawa S, Huibregtse JM (2000) Human scribble (Vartul) is targeted for ubiquitin-mediated degradation by the high-risk papillomavirus E6 proteins and the E6AP ubiquitin-protein ligase. Mol Cell Biol 20: 8244–8253.

[126] Handa K, Yugawa T, Narisawa-Saito M, Ohno SI, Fujita M, et al. (2007) E6AP-dependent degradation of DLG4/PSD95 by high-risk human papillomavirus type 18 E6 protein. J Virol 81: 1379–1389.

[127] Préhaud C, Wolff N, Terrien E, Lafage M, Mégret F, et al. (2010) Attenuation of rabies virulence: takeover by the cytoplasmic domain of its envelope protein. Sci Signal 3: ra5.

[128] Szwajkajzer D, Carey J (1997) Molecular and biological constraints on ligand-binding affinity and specificity. Biopolymers 44: 181–198.

[129] Fu H, editor (2004) Protein-Protein Interactions. Methods and Applications, volume 261. HUMANA PRESS.

[130] Frank F, Sonenberg N, Nagar B (2010) Structural basis for 5'-nucleotide base-specific recognition of guide rna by human ago2. Nature 465: 818–822.

[131] Quinternet M, Starck JP, Delsuc MA, Kieffer B (2012) Unraveling complex small-molecule binding mechanisms by using simple NMR spectroscopy. Chemistry 18: 3969–3974.

[132] Myszka DG (1999) Improving biosensor analysis. J Mol Recognit 12: 279–284.

[133] Greenspan NS (2010) Cohen's Conjecture, Howard's Hypothesis, and Ptashne's Ptruth: an exploration of the relationship between affinity and specificity. Trends Immunol 31: 138–143.

[134] Eaton BE, Gold L, Zichi DA (1995) Let's get specific: the relationship between specificity and affinity. Chem Biol 2: 633–638.

[135] Carothers JM, Oestreich SC, Szostak JW (2006) Aptamers selected for higher-affinity binding are not more specific for the target ligand. J Am Chem Soc 128: 7929–7937.

[136] Schreiber G, Keating AE (2011) Protein binding specificity versus promiscuity. Curr Opin Struct Biol 21: 50–61.

[137] Kupiec JJ (2010) On the lack of specificity of proteins and its consequences for a theory of biological organization. Prog Biophys Mol Biol 102: 45–52.

[138] Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, et al. (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res 39: D561–D568.

[139] Consortium U (2010) The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res 38: D142–D148.

[140] Dosztányi Z, Csizmok V, Tompa P, Simon I (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21: 3433–3434.

[141] Roey KV, Gibson TJ, Davey NE (2012) Motif switches: decision-making in cell regulation. Curr Opin Struct Biol (in press).

[142] Castagnoli L, Costantini A, Dall'Armi C, Gonfloni S, Montecchi-Palazzi L, et al. (2004) Selectivity and promiscuity in the interaction network mediated by protein recognition modules. FEBS Lett 567: 74–79.

[143] Zarrinpar A, Park SH, Lim WA (2003) Optimization of specificity in a cellular protein interaction network by negative selection. Nature 426: 676–680.

[144] Endy D, Yaffe MB (2003) Signal transduction: molecular monogamy. Nature 426: 614–615.

[145] Ladbury JE, Arold S (2000) Searching for specificity in sh domains. Chem Biol 7: R3–R8.

[146] Bezprozvanny I, Maximov A (2001) Classification of PDZ domains. FEBS Lett 509: 457–462.

[147] Vaccaro P, Brannetti B, Montecchi-Palazzi L, Philipp S, Citterich MH, et al. (2001) Distinct binding specificity of the multiple PDZ domains of INADL, a human protein with homology to INAD from Drosophila melanogaster. J Biol Chem 276: 42122–42130.

[148] Vaccaro P, Dente L (2002) PDZ domains: troubles in classification. FEBS Lett 512: 345–349.

[149] Fuh G, Pisabarro MT, Li Y, Quan C, Lasky LA, et al. (2000) Analysis of PDZ domain-ligand interactions using carboxyl-terminal phage display. J Biol Chem 275: 21486–21491.

[150] Laura RP, Witt AS, Held HA, Gerstner R, Deshayes K, et al. (2002) The Erbin PDZ domain binds with high affinity and specificity to the carboxyl termini of delta-catenin and ARVCF. J Biol Chem 277: 12906–12914.

[151] Skelton NJ, Koehler MFT, Zobel K, Wong WL, Yeh S, et al. (2003) Origins of PDZ domain ligand specificity. Structure determination and mutagenesis of the Erbin PDZ domain. J Biol Chem 278: 7645–7654.

[152] Ernst A, Sazinsky SL, Hui S, Currell B, Dharsee M, et al. (2009) Rapid evolution of functional complexity in a domain family. Sci Signal 2: ra50.

[153] Ernst A, Gfeller D, Kan Z, Seshagiri S, Kim PM, et al. (2010) Coevolution of PDZ domain-ligand interactions analyzed by high-throughput phage display and deep sequencing. Mol Biosyst 6: 1782–1790.

[154] Kurakin A, Swistowski A, Wu SC, Bredesen DE (2007) The PDZ domain as a complex adaptive system. PLoS ONE 2: e953.

[155] Basdevant N, Weinstein H, Ceruso M (2006) Thermodynamic basis for promiscuity and selectivity in protein-protein interactions: PDZ domains, a case study. J Am Chem Soc 128: 12766–12777.

[156] Beuming T, Farid R, Sherman W (2009) High-energy water sites determine peptide binding affinity and specificity of PDZ domains. Protein Sci 18: 1609–1619.

[157] Pan L, Chen J, Yu J, Yu H, Zhang M (2011) The structure of the PDZ3-SH3-GuK tandem of ZO-1 protein suggests a supramodular organization of the membrane-associated guanylate kinase (MAGUK) family scaffold protein core. J Biol Chem 286: 40069–40074.

[158] Feng W, Zhang M (2009) Organization and dynamics of PDZ-domain-related supramodules in the postsynaptic density. Nat Rev Neurosci 10: 87–99.

[159] Snel B, Lehmann G, Bork P, Huynen MA (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res 28: 3442–3444.

[160] Obenauer JC, Yaffe MB (2004) Computational prediction of protein-protein interactions. Methods Mol Biol 261: 445–468.

[161] Deng X, Eickholt J, Cheng J (2012) A comprehensive overview of computational protein disorder prediction methods. Mol Biosyst 8: 114–121.

[162] Sonnhammer EL, Eddy SR, Durbin R (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 28: 405–420.

[163] Schultz J, Milpetz F, Bork P, Ponting CP (1998) SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci U S A 95: 5857–5864.

[164] Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, et al. (2012) The Pfam protein families database. Nucleic Acids Res 40: D290–D301.

[165] Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, et al. (2012) InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res 40: D306–D312.

[166] Sigrist CJA, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, et al. (2010) PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res 38: D161–D166.

[167] Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, et al. (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res 39: D225–D229.

[168] Mi T, Merlin JC, Deverasetty S, Gryk MR, Bill TJ, et al. (2012) Minimotif Miner 3.0: database expansion and significantly improved reduction of false-positive predictions from consensus sequences. Nucleic Acids Res 40: D252–D260.

[169] Obenauer JC, Cantley LC, Yaffe MB (2003) Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res 31: 3635–3641.

[170] Neduva V, Russell RB (2005) Linear motifs: evolutionary interaction switches. FEBS Lett 579: 3342–3345.

[171] Sánchez IE, Beltrao P, Stricher F, Schymkowitz J, Ferkinghoff-Borg J, et al. (2008) Genome-wide prediction of SH2 domain targets using structural information and the FoldX algorithm. PLoS Comput Biol 4: e1000052.

[172] Tonikian R, Xin X, Toret CP, Gfeller D, Landgraf C, et al. (2009) Bayesian modeling of the yeast SH3 domain interactome predicts spatiotemporal dynamics of endocytosis proteins. PLoS Biol 7: e1000218.

[173] Linding R, Jensen LJ, Ostheimer GJ, van Vugt MATM, Jrgensen C, et al. (2007) Systematic discovery of in vivo phosphorylation networks. Cell 129: 1415–1426.

[174] Panni S, Montecchi-Palazzi L, Kiemer L, Cabibbo A, Paoluzi S, et al. (2011) Combining peptide recognition specificity and context information for the prediction of the 14-3-3-mediated interactome in S. cerevisiae and H. sapiens. Proteomics 11: 128–143.

[175] Schultz J, Hoffmüller U, Krause G, Ashurst J, Macias MJ, et al. (1998) Specific interactions between the syntrophin PDZ domain and voltage-gated sodium channels. Nat Struct Biol 5: 19–24.

[176] Schillinger C, Boisguerin P, Krause G (2009) Domain Interaction Footprint: a multi-classification approach to predict domain-peptide interactions. Bioinformatics 25: 1632–1639.

[177] Kalyoncu S, Keskin O, Gursoy A (2010) Interaction prediction and classification of PDZ domains. BMC Bioinformatics 11: 357.

[178] Shao X, Tan CSH, Voss C, Li SSC, Deng N, et al. (2011) A regression framework incorporating quantitative and negative interaction data improves quantitative prediction of PDZ domain-peptide interaction from primary sequence. Bioinformatics 27: 383–390.

[179] Yip KY, Utz L, Sitwell S, Hu X, Sidhu SS, et al. (2011) Identification of specificity determining residues in peptide recognition domains using an information theoretic approach applied to large-scale binding maps. BMC Biol 9: 53.

[180] Knisley D, Knisley J (2011) Predicting protein-protein interactions using graph invariants and a neural network. Comput Biol Chem 35: 108–113.

[181] Hui S, Bader GD (2010) Proteome scanning to predict PDZ domain interactions using support vector machines. BMC Bioinformatics 11: 507.

[182] Zaslavsky E, Bradley P, Yanover C (2010) Inferring PDZ domain multi-mutant binding preferences from single-mutant data. PLoS One 5: e12787.

[183] Niv MY, Weinstein H (2005) A flexible docking procedure for the exploration of peptide binding selectivity to known structures and homology models of PDZ domains. J Am Chem Soc 127: 14072–14079.

[184] Staneva I, Wallin S (2009) All-atom Monte Carlo approach to protein-peptide binding. J Mol Biol 393: 1118–1128.

[185] Encinar JA, Fernandez-Ballester G, Sánchez IE, Hurtado-Gomez E, Stricher F, et al. (2009) ADAN: a database for prediction of protein-protein interaction of modular domains mediated by linear motifs. Bioinformatics 25: 2418–2424.

[186] Gerek ZN, Ozkan SB (2010) A flexible docking scheme to explore the binding selectivity of PDZ domains. Protein Sci 19: 914–928.

[187] Kaufmann K, Shen N, Mizoue L, Meiler J (2010) A physical model for PDZ-domain/peptide interactions. J Mol Model 17: 315-324.

[188] Raveh B, London N, Schueler-Furman O (2010) Sub-angstrom modeling of complexes between flexible peptides and globular proteins. Proteins 78: 2029–2040.

[189] Smith CA, Kortemme T (2010) Structure-based prediction of the peptide sequence space recognized by natural and synthetic PDZ domains. J Mol Biol 402: 460–474.

[190] Boulesteix AL (2010) Over-optimism in bioinformatics research. Bioinformatics 26: 437–439.

[191] Jelizarow M, Guillemot V, Tenenhaus A, Strimmer K, Boulesteix AL (2010) Over-optimism in bioinformatics: an illustration. Bioinformatics 26: 1990–1998.

[192] Ben-Hur A, Noble WS (2006) Choosing negative examples for the prediction of protein-protein interactions. BMC Bioinformatics 7 Suppl 1: S2.

[193] Smialowski P, Pagel P, Wong P, Brauner B, Dunger I, et al. (2010) The Negatome database: a reference set of non-interacting protein pairs. Nucleic Acids Res 38: D540–D544.

[194] Singh MK, Dominy BN (2010) Thermodynamic resolution: How do errors in modeled protein structures affect binding affinity predictions? Proteins 78: 1613-1617.

[195] Li L, Zhao B, Du J, Zhang K, Ling CX, et al. (2011) DomPep–a general method for predicting modular domain-mediated protein-protein interactions. PLoS One 6: e25528.

[196] Giallourakis C, Cao Z, Green T, Wachtel H, Xie X, et al. (2006) A molecular-properties-based approach to understanding PDZ domain proteins and PDZ ligands. Genome Res 16: 1056–1072.

[197] Gerek ZN, Ozkan SB (2011) Change in allosteric network affects binding affinities of PDZ domains: analysis through perturbation response scanning. PLoS Comput Biol 7: e1002154.

[198] Beuming T, Skrabanek L, Niv MY, Mukherjee P, Weinstein H (2005) PDZBase: a protein-protein interaction database for PDZ-domains. Bioinformatics 21: 827–828.

[199] Charbonnier S, Stier G, Orfanoudakis G, Kieffer B, Atkinson RA, et al. (2008) Defining the minimal interacting regions of the tight junction protein MAGI-1 and HPV16 E6 oncoprotein for solution structure studies. Protein Expr Purif 60: 64–73.

[200] Kotelevets L, van Hengel J, Bruyneel E, Mareel M, van Roy F, et al. (2005) Implication of the MAGI-1b/PTEN signalosome in stabilization of adherens junctions and suppression of invasiveness. FASEB J 19: 115–117.

[201] Zhan L, Rosenberg A, Bergami KC, Yu M, Xuan Z, et al. (2008) Deregulation of Scribble promotes mammary tumorigenesis and reveals a role for cell polarity in carcinoma. Cell 135: 865–878.

[202] Simonson SJS, Difilippantonio MJ, Lambert PF (2005) Two distinct activities contribute to human papillomavirus 16 E6's oncogenic potential. Cancer Res 65: 8266–8273.

[203] Hubbard TJP, Aken BL, Ayling S, Ballester B, Beal K, et al. (2009) Ensembl 2009. Nucleic Acids Res 37: D690–D697.

[204] Mount DW (2004) Bioinformatics: Sequence and Genome Analysis. Cold Spring Harbor Laboratory Press. URL `http://www.worldcat.org/isbn/0879697121`.

[205] Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, et al. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 4: P3.

[206] Lupas A, Dyke MV, Stock J (1991) Predicting coiled coils from protein sequences. Science 252: 1162–1164.

[207] Charbonnier S, Zanier K, Masson M, Trav G (2006) Capturing protein-protein complexes at equilibrium: the holdup comparative chromatographic retention assay. Protein Expr Purif 50: 89–101.

[208] Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27: 431–432.

Katja LUCK

Vers une meilleure connaissance de la spécificité des interactions protéiques dans la signalisation cellulaire – les domaines PDZ au centre des approches informatiques et expérimentales

Résumé:

Les domaines PDZ reconnaissent des motifs C-terminaux (PBMs), à l'origine de nombreuses interactions qui sont souvent impliquées dans la régulation de la polarité cellulaire. Dans cette thèse, nous avons étudié divers aspects de la spécificité des interactions PDZ-PBM. Nous avons mis en évidence les faibles performances de deux prédicteurs d'interaction entre PDZs et PBMs, considérés sous leurs formes les plus courtes. Ensuite, nous avons développé des protocoles basés sur les méthodes BIAcore et HoldUp pour valider expérimentalement et à grande échelle des prédicteurs d'interaction PDZ-PBM et pour étudier l'influence du contexte de séquence (comme les séquences flanquantes ou les domaines voisins) des PDZs et des PBMs sur laffinité et la spécificité de leurs interactions. Nous avons identifié des interactions potentielles impliquant les protéines humaines à PDZ MAGI1 et SCRIB soulignant leur implication dans les réseaux de signalisation des protéines G. Une revue de la littérature, combinée avec nos propres résultats, a révélé des mécanismes par lesquels le contexte de séquence influence les affinités et spécificités des interactions impliquant les PDZs. Nous avons discuté ces mécanismes dans une revue publiée. Les connaissances obtenues à partir de cette thèse pourront influencer positivement de futures études sur les interactions PDZ-PBM, en particulier, et sur les interactions domaine-motif linéaire en général.

Mots-clés: PDZ, interaction protéique, spécificité, contexte de séquence, prédiction.

Résumé en anglais:

PDZ domains recognise C-terminal PDZ-binding motifs (PBMs) thereby mediating protein interactions that are often involved in cell polarity regulation. In this thesis, we studied under various aspects the specificity of PDZ–PBM interactions. We identified weak performances of two published predictors for interactions between core PDZ domains and short PBMs. Next, we developed protocols based on BIAcore and HoldUp to experimentally validate on a large scale predicted PDZ–PBM interactions and to study the influence of sequence context (e.g. flanking regions or neighbouring domains) of PDZs and PBMs on their interaction affinity and specificity. We identified new potential interactions involving the human PDZ proteins MAGI1 and SCRIB underpinning their implication in G protein signalling pathways. A literature survey combined with our own findings reveal structural mechanisms, by which sequence context influences PDZ interaction affinities and specificities. We have discussed those in a published review. Insights gained from this thesis may positively impact future studies on PDZ–PBM interactions in particular and on domain–linear motif interactions in general.

Keywords: PDZ, protein interaction, specificity, sequence context, prediction.