Joint analysis of dynamically correlated networks and coevolved residue clusters : large-scale analysis and methods for predicting the effects of genetic disease associated mutations

Yasaman Karami

UPMC
SORBONNE UNIVERSITÉS

THÈSE DE DOCTORAT DE
l'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité

Informatique

École doctorale Informatique, Télécommunications et Électronique
(Paris)

Présentée par

Yasaman KARAMI

Pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

Joint analysis of dynamically correlated networks and coevolved residue clusters: largescale analysis and methods for predicting the effects of genetic disease associated mutations

soutenue le 18 Novembre 2016

devant le jury composé de :

Mme. Franca Fraternali	Rapporteur de thèse
Mme. Marianne Rooman	Rapporteur de thèse
Mme. Sonia Longhi	Examinateur de thèse
M. Richard Lavery	Examinateur de thèse
M. Jacques Chomilier	Examinateur de thèse
Mme. Alessandra Carbone	Directrice de thèse
Mme. Elodie Laine	Encadrant
M. Serge Amselem	Encadrant

Acknowledgement

I would like to express my endless gratitude to my supervisors, Prof. Alessandra Carbone and Dr. Elodie Laine. I would like to thank you for encouraging my research, continuous support, patience, insightful discussions and more important, all the moral supports you gave me. Your advices have been priceless, I could not have imagined having better supervisors for my PhD study. Besides my supervisors, I would like to show gratitude to my committee: Prof. Franca Fraternali, Dr. Marianne Rooman, Dr. Sonia Longhi, Dr. Richard Lavery, Dr. Jacques Chomilier and Prof. Serge Amselem for giving the pleasure to have them as the jury members of the thesis. I would also like to thank all my friends at the lab, for all the wondeful momnets we had together, all the supports and discussions we had and for being amazing frineds.

A special thanks to my family. Words cannot express how grateful I am to my mother, father and brother for supporting me spiritually and incenting me to strive towards my goal. At the end I would like to express appreciation to my beloved husband, Hamed Khakzad who was always my support in all the moments of difficulty, when there was no one to answer my queries.

Contents

Chapter 1

Résumé en francais

1.1 Introduction

Les protéines régulent les processus biologiques en interagissant avec d'autres molécules (protéines, acides nucléiques, cofacteurs, ...) et en adaptant leur forme et leurs mouvements aux changements environnementaux. Le comportement dynamique des protéines est directement lié à leur fonction (Henzler-Wildman and Kern, 2007; Frauenfelder et al., 1991). Les protéines sont également soumises à des contraintes évolutives et structurelles afin de maintenir et / ou adapter leurs fonctions. Les protéines sont des macromolécules formées par une ou plusieurs chaînes d'acides aminés. Une mutation est un changement dans la séquence d'acides aminés d'une protéine, qui peut affecter sa structure et / ou sa fonction. Les mutations peuvent induire différents phénotypes : elles peuvent être *bénéfiques* (gain de fonction), *neutres* ou *délétères* pour la fonction de la protéine. La plasticité structurale d'une protéine peut être modifiée par des modifications génétiques (par exemple des mutations ponctuelles) qui peuvent induire des effets sur des sites distants de la protéine, provoquant ainsi des maladies. Par conséquent, caractériser les préférences et changements conformationnels des protéines peuvent ouvrir la voie à la conception de médicaments, pour comprendre les mécanismes qui sous-tendent les maladies, et pour les prévenir/traiter.

Dynamique structurale des protéines

La façon dont une protéine se replie en une structure 3D est codée dans sa séquence d'acides aminés. Néanmoins les protéines ne sont pas statiques. Au cours de leur vie, elles subissent des changements conformationnels et s'associent avec des partenaires. L'idée selon laquelle les protéines existent en solution comme un ensemble de conformations est aujourd'hui généralement admise. La dynamique conformationnelle d'une protéine est directement liée à sa fonction. (Henzler-Wildman and Kern, 2007; Frauenfelder et al., 1991). Les simulations moléculaires atomistiques sont une méthode de choix pour explorer l'espace conformationnel d'une protéine. L'accumulation de données de dynamique moléculaire (DM) nécessite le développement de méthodes capables d'extraire des informations biologiques pertinentes et de les visualiser d'une manière globale. Les simulations de DM consistent à simuler (de façon computationnelle) le comportement des protéines en solution. Elles fournissent des informations détaillées sur les fluctua-

tions et les changements conformationnels des protéines et des acides nucléiques, en produisant des *trajectoires* (ensemble des positions de la protéine à des intervalles de temps réguliers).

Dans les simulations en solvant explicite, la protéine est placée dans une boîte d'eau virtuelle et ses mouvements sont simulés typiquement sur 0,01-10 microsecondes et enregistrés. Une taille moyenne pour un système (molécules de protéines et d'eau) à étudier avec des simulations de DM, est de 10^4 à 10^5 atomes et le temps de calcul est de plusieurs jours ou années de CPU. Les simulations DM requièrent l'utilisation des ressources de calcul haute performance.

Les résidus d'une protéine "communiquent" entre eux, ce qui peut résulter en un couplage allostérique (*i.e.* la propagation d'un signal de perturbation entre des sites distincts, pouvant être situés loin dans la séquence ainsi que dans la structure de la protéine) qui module la fonction de la protéine. Des efforts méthodologiques ont été précédemment engagés pour identifier des groupes ou des chaînes de résidus responsables de couplage entre sites protéiques distants (Chiappori et al., 2012; Papaleo et al., 2012; Laine et al., 2012; Raimondi et al., 2013; Pandini et al., 2013; Blacklock and Verkhivker, 2013; McClendon et al., 2014; Invernizzi et al., 2014; Allain et al., 2014). Par exemple, la méthode MONETA (Allain et al., 2014) s'est avérée utile pour identifier les chemins de communications pour les protéines régulées de façon allostérique et pour guider la mutagenèse *in silico* (Laine et al., 2012). MONETA vise à aider l'analyse des données de simulations de DM. Elle permet de mettre l'accent sur des régions ou des résidus spécifiques, à condition que l'utilisateur possède une certaine connaissance du système. Des valeurs fixes sont codées dans l'outil pour la plupart des paramètres, ce qui limite son application et sa flexibilité.

Protéines intrinsèquement désordonnées (PID) Les PIDs sont caractérisées par leur absence de structure tertiaire stable dans des conditions physiologiques *in vitro* (Dunker et al., 2001) ainsi que le fait que leurs séquences partagent certaines propriétés, *i.e.*, elles ont des complexités plus faibles, leur nombre d'acides aminés hydrophobes est plus petit, elles ont plus de résidus polaires ou chargés (Wright and Dyson, 1999; Ishida and Kinoshita, 2007) et leur séquence est faiblement conservée (Brown et al., 2011; Mei et al., 2014). Ces résultats ont abouti à l'élaboration de nombreuses approches basées sur la séquence pour prédire des régions désordonnées dans les protéines (He et al., 2009; Peng and Kurgan, 2012; Deng et al., 2012).

Conservation évolutive and co-evolution

Le degré de variation pour chaque position le long d'un ensemble de séquences homologues peut être très large. On note que plus la variation à une position est faible, plus le degré de conservation est élevé et donc la position biologique est importante. Les régions où les résidus sont identiques ou quasi-identiques parmi toutes les espèces sont appelées *conservées*. Différentes méthodes mesurent le niveau de conservation de chaque position d'un ensemble d'alignement de séquences homologues (alignement de séquences multiples (MSA)). Les analyses basées sur la notion classique de contenu d'information capturent numériquement la variabilité d'un résidu à une position donnée du MSA en fournissant un score numérique global. Ce score représente l'entropie de

l'ensemble des séquences par la combinaison de l'information locale sur les positions d'alignement (Akashi, 1999; Thompson et al., 1999; Duret et al., 2000; Lecompte et al., 2001; Notredame, 2002; Wallace et al., 2005; Watson et al., 2005; Notredame, 2007). Des informations supplémentaires peuvent être considérées, par exemple les propriétés physico-chimiques des résidus et la conservation locale de ces propriétés (voir (Carbone and Dib, 2011) pour une liste de références).

La *co-évolution* dans les protéines correspond à des changements sur des positions différentes qui se produisent en même temps. La co-évolution est le signe d'une dépendance fonctionnelle et/ou structurelle entre les deux positions. Les résidus co-évolués peuvent être impliqués dans des interactions fonctionnelles entre des protéines et des biomolécules (Lichtarge et al., 1996; Engelen et al., 2009; Lichtarge and Wilkins, 2010). Les signaux de co-variation évolutiive ont également été exploités pour prédire avec une grande précision des contacts natifs au sein de structures de protéines (Morcos et al., 2011), d'interactions intermoléculaires (Champeimont et al., 2016) et de communication allostérique intramoleculaire (Sung et al., 2016).

Les approches pour détecter les signaux de coévolution existants, peuvent être divisés en deux groupes principaux: les méthodes statistiques et les méthodes combinatoires. Les analyses de la première catégorie capturent la covariation entre les positions des séquences alignées en mesurant des coefficients de corrélation (Goh et al., 2000; Fares and Travers, 2006), l'information mutuelle (Atchley et al., 2000; Ramani and Marcotte, 2003; Gloor et al., 2005) et la déviation des distributions pour estimer le couplage thermodynamique entre les résidus (Süel et al., 2003b; Dima and Thirumalai, 2006; Weigt et al., 2009; Sadowski et al., 2011). Par exemple, La méthode Statistical Coupling Analysis (SCA) (Lockless and Ranganathan, 1999) calcule la répartition des acides aminés à une position par rapport aux changements à toute autre position et identifie un groupe de résidus coévoluant, appelé secteur. La méthode Direct Coupling Analysis (DCA) (Morcos et al., 2011) est une autre technique statistique pour détecter le signal coévolutif entre les résidus. La force de la méthode réside dans sa capacité à démêler les interactions directes des interactions indirectes. D'un autre côté, des approches combinatoires son proposées pour surmonter les restrictions des approches statistiques (Fryxell, 1996; Pazos and Valencia, 2001). Elles s'appuient sur l'extraction de sous-arbres à partir de l'arbre phylogénétique des séquences associé au MSA. MST et BIS, sont deux approches combinatoires, proposées dans notre laboratoire, pour détecter les résidus coévoluant (Baussand and Carbone, 2009; Dib and Carbone, 2012b). Leur avantage réside dans le fait qu'elles peuvent traiter un petit ensemble de séquences. MST fonctionne mieux sur des ensembles de séquences avec des divergences variables, alors que BIS exige des séquences qui sont hautement conservées.

Sonder le paysage de mutations des protéines La question de savoir comment les variations de séquence d'acides aminés façonnent le paysage conformationnel des protéines et impactent leur fonction est très importante en biologie, et encore loin d'être résolue. Des technologies récemment mises au point, communément nommées "balayage mutationnel profond", permettent d'estimer les conséquences fonctionnelles de chaque changement d'acide aminé possible à chaque position dans une protéine(McLaughlin et al., 2012; Fowler and Fields, 2014; Figliuzzi et al., 2016). Ces développements sont prometteurs et les données produites peuvent être utilisées pour valider **1)** des prédictions *in*

silico et **2)** des méthodes de calcul développées pour prédire les effets des mutations. Plusieurs méthodes ont été mises au point pour prédire l'effet des mutations, par exemple: Polyphen-2 (Adzhubei et al., 2010), SIFT (Ng and Henikoff, 2003), PoPMuSiC (Dehouck et al., 2011), I-Mutant2.0 (Capriotti et al., 2005) and MUpro (Cheng et al., 2006). Un autre exemple est le comptage simple des fréquences d'acides aminés à chaque position pour prédire les résultats phénotypiques des mutations.

Quelques questions biologiques en biologie computationnelle

Les principales questions biologiques que nous abordons dans cette thèse, sont les suivantes: **1)** Quelles sont les positions dans une protéine qui sont très sensibles aux mutations? **2)** Quel est l'effet d'une substitution d'acide aminé particulière? **3)** Quel est le lien entre les contraintes structurales et évolutives? Et beaucoup d'autres questions connexes comme: quelles régions de la protéine sont plus sensibles aux mutations et quelles sont les chaînes de résidus (*chemin*) à travers lesquelles les perturbations pourraient se propager à travers la structure? Nous cherchons à caractériser le paysage mutationnel des protéines à large échelle et de manière systématique à travers l'analyse conjointe des structures et des séquences protéiques. Outre les questions biologiques mentionnées ci-dessus, nous abordons plusieurs défis computationnels : **1)** Conception d'une méthode qui permet l'analyse de la dynamique des protéines à différents niveaux. **2)** Développement d'un outil basé sur la méthode proposée, pour étudier la dynamique conformationnelle des protéines d'une manière systématique à grande échelle. **3)** La capacité de se déplacer entre les différentes représentations de protéines (1D, 2D, 3D et 4D) doit être incorporée dans le procédé.

1.1.1 Approches

Les réseaux de résidus corrélés dynamiquement jouent un rôle crucial dans la propagation de signaux de perturbation comme les mutations. Ces résidus sont réputés pour être très conservés et/ou co-évolués. Cependant, la relation entre évolution des séquences et dynamique structurale a été rarement explorée. Dans ce travail, nous avons exploité les séquences et les structures protéiques pour prédire les effets des mutations et explorer la relation entre les contraintes structurales et évolutives. Nous avons développé des méthodes et des mesures quantitatives pour extraire et décrire des signaux structuraux/dynamiques et évolutifs de façon automatisée et à large échelle. En outre, nous avons développé des mesures pour prédire les effets des mutations basés sur la séquence et la structure/analyse de la dynamique. De plus, nous avons appliqué ces méthodes et exploré la relation séquence-structure-fonction. Nous présentons un résumé des méthodes et leurs applications dans les sections suivantes.

1.2 COMmunication MApping (COMMA)

Nous présentons COMmunication MApping (COMMA), une méthode pour caractériser l'architecture dynamique d'une protéine et cartographier cette informations sur la structure en trois dimensions de la protéine. La méthode proposée est publiée (Karami et al., 2016) et est disponible gratuitement `www.lcqb.upmc.fr/COMMA`. COMMA extrait les informations pertinentes de la dynamique des protéines et les intègre de manière systématique.

La méthode va au-delà des analyses classiques de simulations DM. Elle intègre les propriétés dynamiques et définit des régions protéiques servant de blocs ou d'unités de communication. COMMA fournit également des mesures pour prédire les effets des mutations à grande échelle, identifier les régions/résidus importants dans une protéine, prédire la flexibilité/désordre, et comparer différentes protéines ou différents états d'une même protéine.

1.2.1 Principe de COMMA

Le principe de la méthode COMMA est représenté sur la **Figure 1.1** et son algorithme se déroule comme suit: **1)** extraction de cinq propriétés dynamiques à partir d'un ensemble de conformations: corrélations dynamiques locales, distances minimales, propensions de communication, forces d'interaction non-covalentes et structures secondaires (box 1). **2)** Ces propriétés sont utilisées pour grouper les résidus en (*i*) cliques indépendantes et (*ii*) chemins de communication (boxes 2-3). Les cliques indépendantes sont des clusters de résidus qui présentent des fluctuations atomiques concertées alors que les chemins de communication sont des chaînes non-covalentes de résidus qui se déplacent ensemble. **3)** Les informations obtenues à partir des cliques indépendantes et les chemins de communication sont intégrées dans un graphe, appelé Protein Communication Network (PCN) (box 4). **4)** Les composantes connexes sont extraites de ce graphique pour définir des blocs de communication de la protéine (box 5). **5)** Les chemins de communication qui relient des éléments de structure secondaire différents sont utilisée pour définir des paires de segments communiquants et mesurer la force de l'interaction (box 6). COMMA permet de visualiser les blocs de communication et les paires de segments communiquants en les mappant sur la conformation moyenne des protéines.

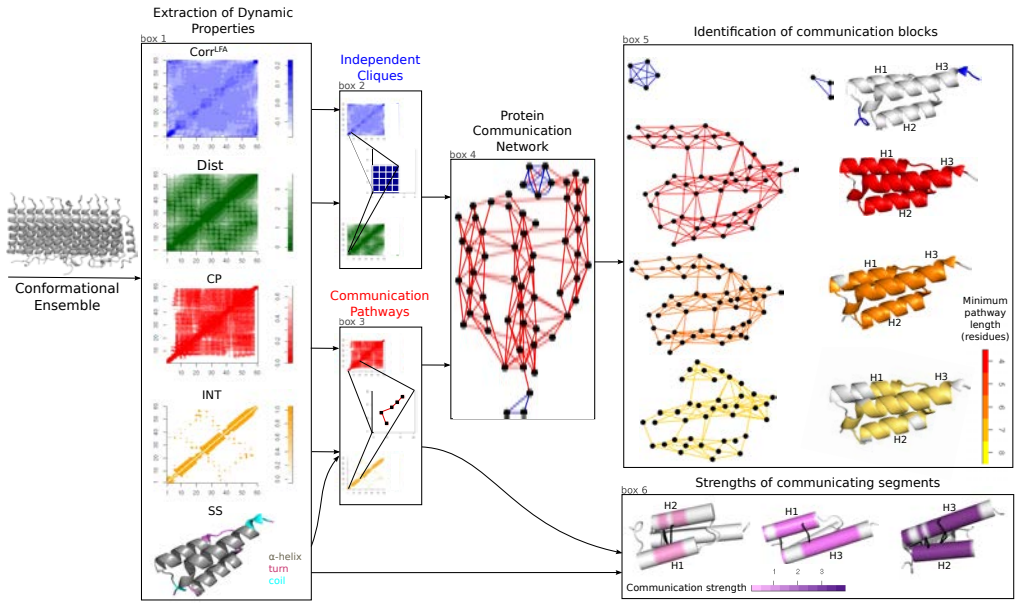

Figure 1.1: **Représentation schématique du flux de travail de COMMA.**

Un certain nombre de méthodes ont été développées précédemment pour caractériser la dynamique conformationnelle des protéines et leur communication inter-résidus. Cependant, ces outils ne considèrent généralement que des corrélations dynamiques ou / et

des interactions non-covalentes, alors que COMMA combine cinq propriétés dynamiques différentes dans un cadre unifié. De plus, COMMA décrit la communication à différents niveaux, à partir de résidus individuels jusqu'à l'ensemble de l'architecture dynamique de la protéine. En particulier, l'identification de la communication des paires d'éléments de structure secondaire est une caractéristique unique de notre méthode. Enfin, COMMA, qui utilise package Python de DMTraj (**?**), ne dépend pas d'un package DM particulier et peut gérer les formats les plus populaires utilisés dans la communauté de la dynamique moléculaire.

1.2.2 Application de COMMA à trois protéines archétypales

Nous avons appliqué COMMA à trois protéines archétypes: (*i*), le domaine B de la protéine staphylococcique A [PDB:1BDD] (résidus 1-60, NMR), une protéine hautement stable, (*ii*) le domaine de liaison à l'ADN de la protéine humain suppresseur de tumeur p53 [PDB:2XWR] (chaîne A, résidus 89-293, 1.68Å résolution), une protéine hautement flexible, (*iii*) la région cytoplasmique du récepteur la tyrosine kinase KIT [PDB:1T45] (résidus 547-935, 1.90Å résolution), une protéine régulée de manière allostérique. Pour chaque protéine, 2 répliques de 50 ns ont été réalisées et COMMA a été appliqué sur les 30 derniers ns (30 000 conformations) de chaque réplique.

KIT Le récepteur tyrosine kinase de type III KIT est impliqué dans des chemins de signalisation cruciales pour la croissance, la différenciation et la survie cellulaires (Lemmon and Schlessinger, 2010; Edling and Hallberg, 2007; Qiu et al., 1988). La mutation de l'aspartate en position 816 en une valine conduit à l'activation constitutive du récepteur et est associé au développement de mastocytoses et de tumeurs stromales gastro-intestinales (Orfao et al., 2007; Miettinen M, 2002). Il a été montré expérimentalement que la mutation induit des effets à longue distance qui produisent un changement dans l'équilibre conformationnel de la kinase (Gajiwala et al., 2009). COMMA a été appliqué à la région cytoplasmique de KIT (331 résidus) sauvage et mutée D816V. Une analyse conformationnelle similaire avait été effectuée précédemment avec MONETA 2.0 (Laine et al., 2011a).

Les blocs de communication (*CBs*) trouvés dans KIT peuvent être classés en fonction des informations structurales et dynamiques utilisées pour les identifier. Dans le type sauvage, COMMA détecte 8 CBs^{clique}, qui représentent des régions de la protéine ayant des fluctuations élevées, concertées à l'intérieur du CB^{clique} mais indépendantes vis-à-vis du reste de la protéine. 3 CBs^{path} sont détectés en considérant des chaînes de résidus (chemins) liés de proche en proche par des interactions non-covalentes et se déplaçant ensemble. Un CB^{path} constitué uniquement par des longs chemins ($\geq$ 6 résidus), et formant ainsi le "noyau" de la communication à longue portée dans KIT, représente plus d'un tiers de la protéine.

Les blocs de communication identifiés par COMMA dans les types sauvage et muté ont été comparés. COMMA a détecté 3 CBs^{path} de longue portée dans le mutant au lieu d'1 dans le type sauvage. La mutation induit une refonte complète des *CBs* de KIT, caractérisé par une réorganisation de la hiérarchie entre les résidus communicant à longue portée et à courte portée et par la fusion de plusieurs CBs^{clique}. Soulignons que la position 816 est située dans un *CB* dont une partie des résidus correspondent à des résidus impliqués dans un réseau allostérique chez la kinase Src (Foda et al., 2015).

Les résultats de COMMA ont été comparés à ceux obtenus avec MONETA 2.0. MONETA identifie des segments dynamiques indépendants et des chemins de communication à partir de simulations DM tout-atome (Allain et al., 2014), qui sont semblables aux cliques indépendantes et aux chemins de communication identifiées par COMMA (**Figure 1.1**, boxes 2 and 3). Cependant, COMMA exploite ces composants pour une analyse plus approfondie (**Figure 1.1**, boxes 4, 5 and 6) d'une manière qui est complètement différent de MONETA (Allain et al., 2014). Les composants de MONETA sont sensiblement différents des blocs de communication identifiés par COMMA et MONETA ne décrit pas les connexions entre eux. De cette comparaison, il est clair que COMMA apporte des informations supplémentaires sur la définition et la disposition des unité dynamiques de la protéine, par rapport à MONETA.

MONETA a précédemment permis de mettre en évidence un chemin de communication crucial dans le type sauvage de KIT qui relie deux régions, la boucle d'activation et le segment juxta-membranaire (JMR), à travers le résidu D792 de la boucle catalytique (Laine et al., 2012). Le chemin est interrompu par la mutation D816V. Dans la représentation par COMMA du type sauvage de KIT, tous les résidus qui participent à ce chemin sont contenus dans le CB^{path} de longue portée. En revanche, dans le mutant, D792 est impliquée dans des chemins de communication plus courts par rapport au type sauvage, et aucun chemin ne va de D792 au JMR. Les résultats de COMMA sont donc en accord avec ceux de MONETA. De plus, en détectant des blocs de communication, COMMA permet d'identifier d'autres chemins longs qui sont interrompus dans le mutant. En particulier, le fait que le long CB^{path} dans le type sauvage est divisé en deux plus petits CBs^{path} dans le mutant correspond à une rupture de communication entre N655 et les résidus I653, H651 et K807. Fait intéressant, ces résidus ont été précédemment identifiés comme formant un réseau d'interactions appelé 'frein moléculaire' crucial pour la stabilité de la conformation inactive des tyrosine kinases (Chen et al., 2007). Par conséquent, l'analyse de COMMA permis de mettre en évidence un effet délétère de la mutation activatrice D816V sur ce 'frein moléculaire' qui n'avait pas été détectée auparavant.

Comparaison des protéines A et p53 Le domaine B de la protéine A (BdpA) et le domaine de liaison à l'ADN (DBD) de p53 représentent deux protéines archétypales en termes de stabilité thermodynamique et cinétique. Alors que le second se déplie juste au-dessus la température physiologique (Bullock et al., 1997), le premier se replie de manière rapide et stable (Lei et al., 2008). En outre, BdpA se compose de trois hélices tandis que DBD contient principalement des feuillets β. Nos analyses des deux protéines montrent des résultats très différents. COMMA détecte deux très petits CBs^{clique} dans BdpA, correspondant aux deux extrémités et représentant 13% des résidus protéiques. En revanche, les CBs^{clique} identifiés dans DBD de la p53 représentent près de 60% de la protéine. Ils englobent tous les résidus impliqués dans l'interaction avec l'ADN, qui adoptent des conformations variables dans les structures expérimentales de la protéine (Lukman et al., 2013). COMMA permet également de caractériser l'évolution de CBs^{path} en fonction de longueur minimale des chemins de communication. Le noyau de communication de BdpA, défini sur la base des chemins très longs ($\geq$ 8 résidus), comprend toute l'hélice H3 et certains résidus de H1 et H2 (**Figure 1.1**, box 5, en jaune). Ceci est cohérent avec les données expérimentales montrant que l'hélice H3 est la plus stable parmi les trois (Bai et al., 1997). p53 DBD présente un comportement dynamique remarquablement différent,

avec un noyau de communication composé de résidus provenant de différents brins β qui constituent le premier feuillet β. Filtrer progressivement les chemins de communication les plus courts exclut d'abord les boucles qui encadrent les brins β, puis les extrémités des brins β du CB. Notez que la longueur des chemins ne dépend pas de la longueur des brins β, *i.e.* de longs brins β ne présentent pas des chemins nécessairement plus longs. Ces observations sur BdpA et p53 DBD montrent que COMMA est utile pour comparer des protéines de natures très différentes d'une manière simple.

1.3 Mutations de l'hormone de croissance associées à des maladies génétiques

L'hormone de croissance (GH) est composée de 4 hélices et régule une grande variété de processus physiologiques, y compris la croissance et la différenciation des muscles, des os, du cartilage et des cellules (Sundstrom et al., 1996). La régulation de la croissance humaine normale est initiée par la liaison de GH à son récepteur (GHR), avec la stoechiométrie 1: 2, GH se lie à 2 sous-unités identiques du récepteur (de Vos et al., 1992). Bien que les 2 molécules de récepteur soient identiques, les sites de liaison de GH sont différents. GH est composé de 4 hélices et trois d'entre elles sont impliquées dans l'interaction avec GHR. Les hélices H1 et H4 forment le premier site de liaison, tandis que H1 et H3 forment le second site de liaison. Le premier a une forte affinité pour GHR, tandis que le second a une faible affinité.

Les deux sites de liaison sont couplés de manière allostérique, et il a été montré précédemment que le site 2 peut être modifié par des mutations au niveau le site 1 (Walsh et al., 2004). Des mutations ponctuelles du complexe GH-GHR peuvent causer des maladies génétiques telles que la petite taille (Petkovic et al., 2010). Nous avons eu accès à un groupe de ces mutants grâce à notre collaborateur, Serge Amselem (Service de Génétique et d'Embryologie Médicales, UMR S933 INSERM/UPMC, Hôpital Armand-Trousseau) et à partir de cette liste, nous avons sélectionné 2 mutations pathologiques, L124R et R183H. Dans ce chapitre, nous démontrons l'impact de ces 2 mutations sur le comportement dynamique du complexe. Ensuite, nous caractérisons le couplage allostérique entre les 2 sites de liaison, et étudions l'impact des mutations sur ce couplage. En outre, nous établissons un lien entre la communication allostérique dans le complexe et les signaux de co-évolution.

1.3.1 Effets des mutations révélés par l'analyse classique de DM

Nous rapportons ici un résumé de l'analyse DM du complexe GH-GHR. Des résultats similaires ont été observés à partir de l'analyse de GH seul, mais ils ne sont pas présentés ici pour éviter la redondance. Pour le complexe GH-GHR nous avons effectué 2 répliques de 100 ns de simulations DM, puis pour chaque réplique, nous avons analysé les 70 dernières ns. Pour le monomère GH nous avons effectué 2 répliques de 50 simulations ns DM et analysé les 30 dernières ns de chacun.

Tout d'abord, nous avons analysé les fluctuations atomiques des types sauvage et mutés (MU^{L124R} et MU^{R183H}). Les résultats démontrent que le complexe de type sauvage est plus rigide que les deux mutants. Pour les deux mutants, une région en boucle (sur les

résidus L128 à R134) qui fait face aux 2 positions de mutation, présente des fluctuations plus élevées par rapport au WT. D'autre part, les effets à longue portée de ces mutations se traduisent aussi par une plus grande flexibilité dans certaines régions de boucle au niveau des récepteurs par rapport au WT.

Les conformations DM moyennes des 2 répliques du complexe WT et des mutants, MU^{L124R} et MU^{R183H} ont été superposées. Pour WT et MU^{L124R}, les répliques moyennes sont bien superposées dans la région de l'hormone, alors que des différences à longue portée sont visibles dans les régions de boucle des deux récepteurs. Pour WT et MU^{R183H}, les structures moyennes sont presque superposées, cependant des régions en boucle de GH et sur les deux récepteurs, adoptent des positions très différentes moyennes chez MU^{R183H}. Dans ces régions, les répliques WT affichent des profils très similaires alors que les répliques mutantes se déplacent dans deux directions différentes. Ce comportement suggère également des effets de longue portée de la mutation dans les récepteurs.

Les liaisons hydrogène entre les positions de mutation et leurs résidus voisins ont été étudiées pour WT, MU^{L124R} et MU^{R183H}. L'analyse a révélé l'affaiblissement du réseau d'interactions autour de L124R en MU^{L124R} et également autour de R183H en MU^{R183H}, alors que la mutation L124R génère de nouvelles interactions entre H1 et H4, reliant les deux sites de liaison.

Nous avons étudié l'ensemble des interactions au niveau des 2 sites de liaison. Il est frappant qu'un plus grand nombre de résidus sont impliqués dans les interactions au site 1 (31 résidus) avec une force d'interaction moyenne de 76%, par rapport au site 2 qui n'implique que 19 résidus de GH avec une force d'interaction moyenne de 71%. Cette observation est en accord avec l'affinité de liaison inférieure du site 2 (Walsh et al., 2004). Une légère diminution globale de la force d'interaction est rapportée au cours des 70 ns de simulation DM pour les 2 répliques de MU^{L124R} par rapport à WT sur les 2 sites de liaison, tandis que la baisse est plus marquée au site 2. D'autre part, pour les 2 répliques de MU^{R183H}, le long de 70 ns de DM simulations la force d'interaction est légèrement réduite sur les 2 sites de liaison, mais la baisse est plus marquée sur le site 1.

1.3.2 Effets des mutations sur la communication du complexe

Nous avons appliqué une analyse COMMA pour étudier GH monomère et le complexe GH-GHR, mais nous rapportons ici un résumé des résultats de l'analyse de GH-GHR WT et MUs (Figure 1.2). Afin de comparer WT avec MUs (MU^{L124R} et MU^{R183H}), les chemins de communication avec au moins 4 résidus dans les sites de liaison ont été extraits (**Figure 1.2**). Les CBs^{path} sont colorés sur la structure et les chemins à des sites de liaison sont présentés dans des lignes noires. Les chemins dans le premier site de liaison, entre la GH et R1 ne sont présents que dans le WT, alors que les chemins dans le second site de liaison, entre GH et R2 ne sont présents que dans les mutants. Dans MU^{L124R} seulement H1 de GH communique avec R2, alors que MU^{R183H} démontre plus la communication dans le site 2, entre H1 et H3 de GH et R2. Une telle augmentation des chemins pourrait être liée à l'augmentation de la force d'interaction au site 2 pour MU^{R183H}.

La réorganisation des blocs de MU^{L124R} par rapport au WT correspond à l'augmentation de la couverture du plus grand bloc (**Figure 1.2a** en rouge). Ce bloc de MU^{L124R} couvre plusieurs blocs du WT (colorés en rose foncé, jaune, marron et rose sur WT). Pour MU^{R183H}, des chemins nouvellement formées dans le mutant sont détectés à l'intérieur de

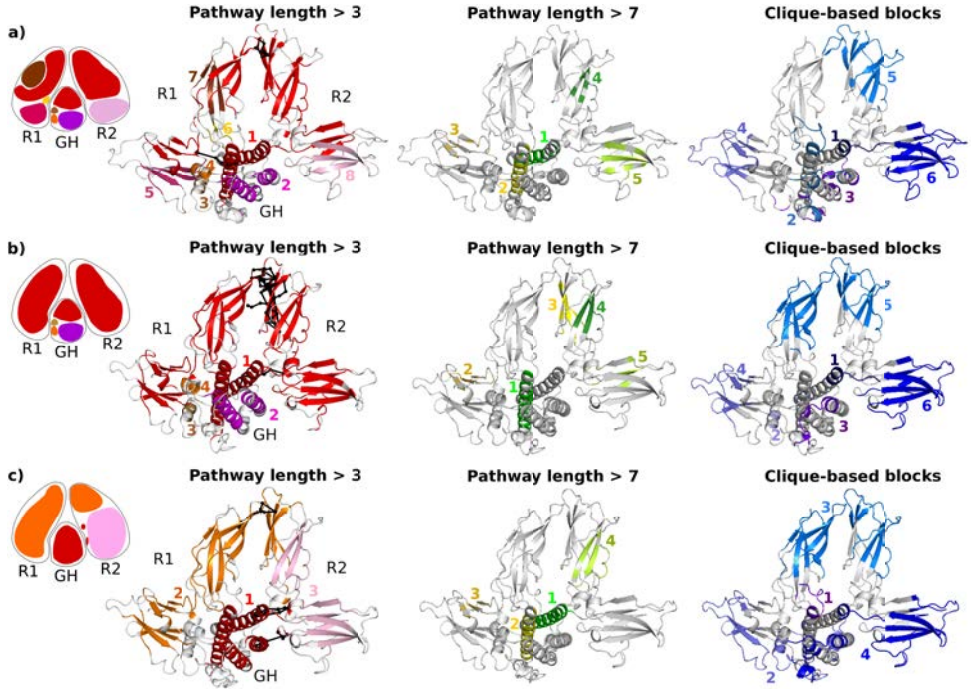

Figure 1.2: **Blocs de communication pour WT, MU^{L124R} and MU^{R183H} GH-GHR complexe.** Blocs de communication des chemins avec au moins 4 résidus pour a) WT, b) MU^{L124R} and c) MU^{R183H} complexes. La représentation schématique des blocs sont représentés sur la gauche. Toutes les chemins au niveau des sites de liaison sont représentés en traits noirs sur la structure du WT, MU^{L124R} et MU^{R183H}.

GH (reliant les blocs rouges et violets à l'intérieur de GH sur WT) et R1 (assemblage de blocs roses jaunes, bruns et rouges sombres à l'intérieur de R1 sur WT).

1.3.3 Analyse de coévolution de GH et de GH-GHR

Nous avons effectué une analyse de coévolution du complexe GH-GHR. Après avoir effectué l'alignement multiple de séquences (MSA) sur l'ensemble des séquences homologues de GH-GHR, les résidus de co-évolution ont été détectés en utilisant BIS et regroupés avec CLAG. Par conséquent, 9 groupes différents ont été détectés. L'observation importante concerne les chemins détectées au niveau du premier site de liaison (entre la GH et GHR1), lorsqu'ils passent à travers les résidus qui appartiennent aux premier et second clusters de résidus coévoluant. Les résidus coévoluant du premier groupe sont détectés sur les hélices de GH, et sur Rec1D1 et Rec2D2 (**figures 1.3** sphères vertes). D'autre part, dans le groupe 2, les 4 résidus coévoluant sont détectés sur l'hormone et les deux récepteurs (**figures 1.3** sphères rouges). Les chemins du site 1 traversent à travers ces résidus coévoluant et relie D171 sur H4 de GH à R203 sur R1 (**figures 1.3**). Cette observation met en évidence l'importance de D171 sur GH et R203 sur R1, car ils sont directement liés par très peu de chemins de liaison GH à GHR, alors qu'ils appartiennent à deux groupes de résidus coévoluant différents.

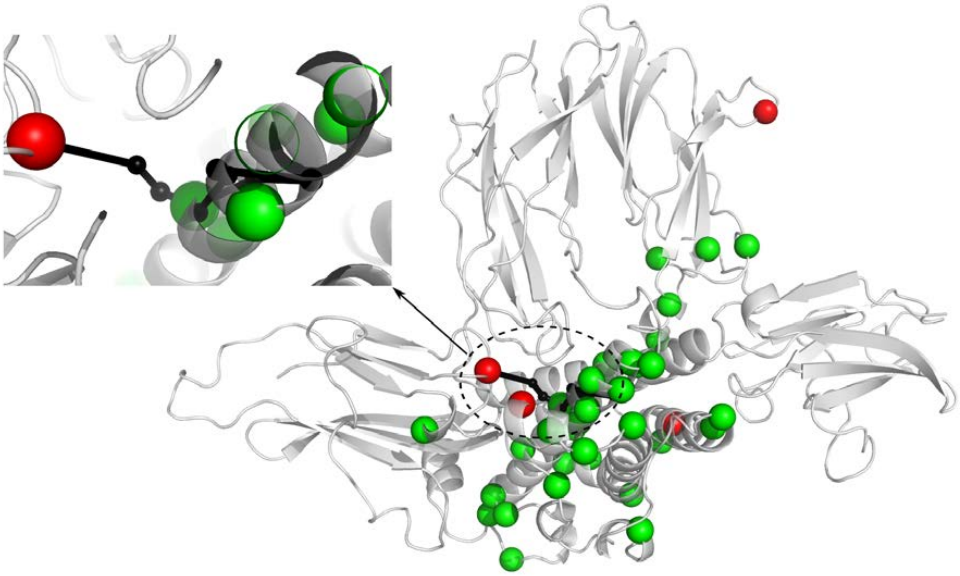

Figure 1.3: **Les chemins reliant les résidus coévoluant.** Les résidus coévoluant qui appartiennent à des groupes 1 et 2 sont présentés dans les sphères et colorés en vert et rouge, respectivement. Les chemins qui communiquent à site1 sont représentés par des lignes noires, ils relient les deux pôles.

1.4 Infostérie des protéines

Nous proposons une méthode pour détecter les points chauds mutationnels délétères et caractériser les positions des résidus qui sont bénéfiques pour la fonction de la protéine. Nous identifions, en étudiant la dynamique conformationnelle, les régions de la protéine qui sont cruciales pour la diffusion de l'information au sein de la structure de la protéine et définissions ces régions formellement. Le nouveau concept de infostérie, de 'info' - information - et 'stérique' - arrangement de résidus dans l'espace - est introduit.

PDZ3, est le troisième domaine PDZ de la protéine synaptique PSD-95. PDZ3 se lie à la protéine de liaison à PDZ (CRIPT) riche en cystéines, ce qui permet à PSD-95 de s'associer au cytosquelette. Le ligand de PDZ3 est le peptide C-terminal dérivé de la CRIPT (TKNYKQTSV). Des technologies récemment mises au point ("deep mutational scanning") permettent d'estimer les conséquences fonctionnelles de chaque changement d'acide aminé unique possible à chaque position dans une protéine (Fowler and Fields, 2014). Ce balayage a été appliqué à un domaine PDZ dans le contexte cellulaire (McLaughlin et al., 2012). Les auteurs ont montré qu'il y avait un bon chevauchement entre l'ensemble des 20 positions affichant la plus haute sensibilité à la mutation et un réseau physiquement contigu de résidus coévoluant détectés à partir d'un alignement multiple de séquences d'homologues de PDZ (sector) (Lockless and Ranganathan, 1999).

1.4.1 Architecture dynamique du complexe PDZ-CRIPT

Nous avons effectué 5 répliques de 20 ns simulations DM pour PDZ-CRIPT et appliqué COMMA pour analyser les 15 dernières ns de chaque réplique (**Figure 1.4**). L'organisation de la façon dont l'information est transmise à travers le complexe PDZ-CRIPT a révélé

que le ligand est presque entièrement intégrée dans la communication de la PDZ3 et peut être divisée en deux parties: les résidus C-terminaux se déplacent avec 2 feuillets β de PDZ3 comme un corps rigide, tandis que les résidus N-terminaux sont plus flexibles et fluctuent de concert avec la boucle L2. Il a également révélé que les résidus PDZ3 encerclant le ligand ne se déplacent pas ou fluctuent tous ensemble, mais se rapportent à 2 régions différentes de la protéine (de couleur rouge/rose sur **Figure 1.4a** et marine/bleu sur **Figure 1.4b**).

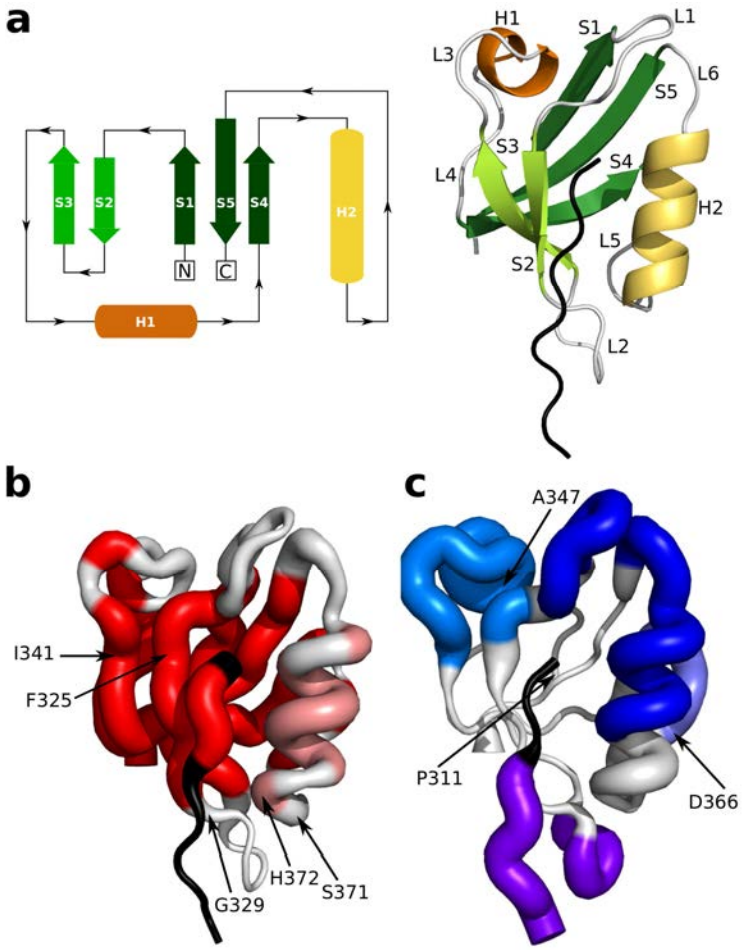

Figure 1.4: **Blocs de communication (CB) identifiés par COMMA dans de type sauvage PDZ3.** La protéine est représenté sous la forme d'un dessin. **(a)** 2 CBs^{path} sont détectées, de couleur rouge et magenta. **(b)** 4 CBs^{clique} sont détectés, dans différents tons bleus. La taille de la saucisse reflète la tendance de chaque résidu à être détecté dans un CB. Les résidus dont les substitutions ont été étudiés sont étiquetés.

Nous avons également étudié l'impact de 8 mutations dont les effets ont été rapportés dans (McLaughlin et al., 2012): P311W (bénéfique), D366A, S371A and F325A (neutre), I341A, H372A, G329A and A347F (délétère). Ils ont été choisis de manière à couvrir différents endroits dans le domaine PDZ et pour représenter les différents phénotypes des mutations. Des simulations DM des complexes mutés (5 répliques de 20 ns chacune) ont été effectuées et les analyses classiques n'ont pas révélé de changements drastiques dans leurs structures ou leurs mouvements (**Figure 1.5a**). En revanche, l'analyse COMMA révèle des différences frappantes dans la communication des mutants par rapport au type

sauvage (**Figure1.5b-c**).

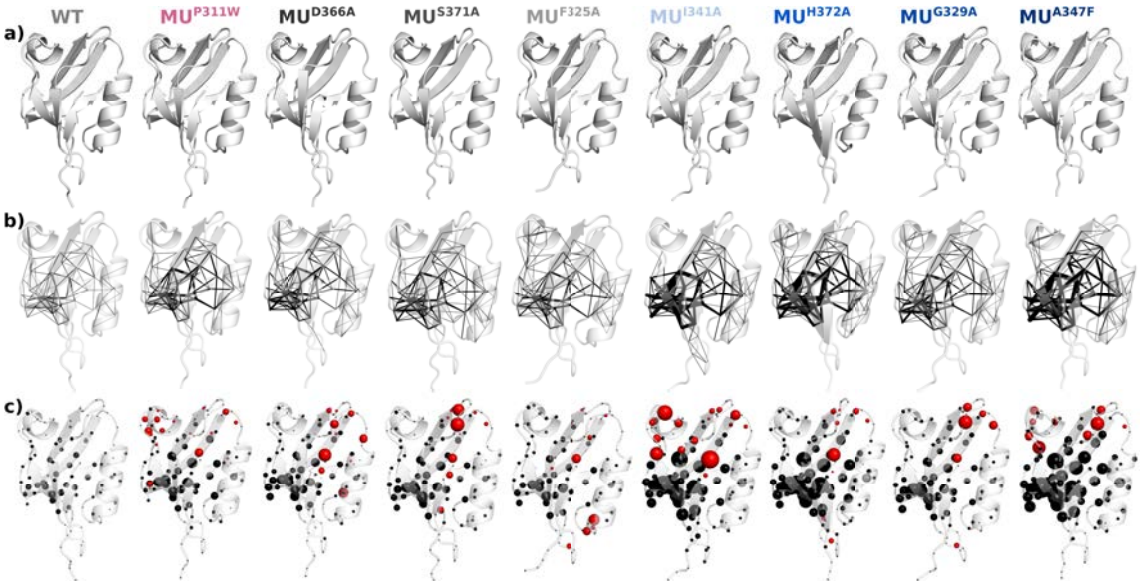

Figure 1.5: **analyse COMMA pour le complexe de PDZ-CRIPT. (a)** Conformations DM moyenne du complexe de type sauvage (WT) et de 8 mutants. **(b)** Les chemins de communication (> 3 résidus) détectés par COMMA sont mappés sur la conformation moyenne et affichés sous forme de lignes noires. L'épaisseur de chaque segment est proportionnelle au nombre de chemins reliant les deux résidus.**(c)** Les résidus traversés par au moins un chemin de communication (> 3 résidus) sont affichés sous forme de sphères noires, centrées sur leurs atomes C-α. La taille de chaque sphère est proportionnelle au nombre de chemins passant par le résidu.

1.4.2 Caractériser l'effet de mutations uniques

Pour quantifier les différences entre les mutants et le type sauvage, d'abord, nous avons étudié les chemins qui traversent chaque résidu du WT et MUs. Cette analyse a montré que, en définissant de nouvelles mesures reflétant la façon dont les résidus communiquent les uns avec les autres, nous avons été en mesure d'évaluer les effets des mutations délétères. COMMA a révélé des différences frappantes entre les mutants et le type sauvage, même si les systèmes ont été simulés sur des échelles de temps relativement courtes. Plus précisément, les mutations délétères induisent une augmentation importante du nombre et de la longueur des chemins de communication, ce qui entraîne des CBs^{path} élargies et un plus grand nombre de résidus fortement connectés. La concentration de chemins accrue pourrait être interprétée comme une rigidité accrue du complexe.

1.4.3 Prédire les points sensibles aux mutations

Notre caractérisation de l'infostérie du type sauvage a révélé que la plupart des positions sensibles aux mutations correspondent à des résidus servant de ponts critiques soit entre la protéine et le peptide, ou un CB^{path} et un CB^{clique}, ou deux éléments de structure secondaire/motifs distincts. Fait intéressant, certains résidus peuvent jouer des rôles multiples. Cette analyse a également démontré que, en exploitant des simulations d'un

seul état conformationnel du complexe PDZ-CRIPT type sauvage DM, sans comprendre les changements conformationnels induits par toutes les mutations, nous pouvons prédire 80% des hotspots délétères avec une précision de 80%.

1.4.4 Prédire les points sensibles aux mutationsen utilisant l'analyse de séquence

Nous avons effectué une analyse de séquence sur PDZ3 pour prédire ses hotspots mutationnels. Nous avons utilisé la matrice de valeurs ΔE mesurées expérimentalement rapportées dans (McLaughlin et al., 2012) comme notre référence. En raison de la disponibilité des données expérimentales qui révèle le paysage mutationnelle de PDZ3, nous avons été en mesure d'examiner l'efficacité et la précision de nos résultats. Notre analyse a conduit à la détection de positions sensibles avec une grande précision. Au-delà de la détection de ces positions, les signaux de conservation sur un site unique, associés à des termes d'interaction par paires entre chaque position et ses voisins structuraux, peuvent être utilisés pour prédire les résultats phénotypiques de l'ensemble des 20×83 substitutions possibles. Une corrélation de 0.51 a été obtenue avec l'ensemble de matrice expérimentale ΔE valeurs. Cette corrélation est très bonne, considérant le bruit contenu dans les données expérimentales. De plus, il est nettement meilleur que ce qui est obtenu à partir des méthodes plus sophistiquées (SIFT, Polyphen-2, PoPMuSiC, I-Mutant2.0 et MUpro).

1.5 Désordre dans les "coiled-coils"

Les "coiled-coils" sont des motifs d'oligomérisation omniprésents chez les protéines, où jusqu'à 7 hélices alpha amphipatiques (les nombres les plus courants d'hélices étant 2 et 3) s'entrelacent ensemble de manière similaire aux fibres d'une corde. Les "coiled-coils" les plus courants ont une orientation "gauche". Ils présentent un motif de séquence spécifique composé de sept résidus *abcdefg* où *a* et *d* sont hydrophobes et les autres résidus sont apolaires. La stabilisation des "coiled-coils" est principalement due à des interactions hydrophobes (Isoleucines, leucines et Valines). Nous concentrons notre étude sur les domaines phosphatases de multimérisation (PDM) de deux virus, à savoir le virus Measles (MeV) et le virus Nipah (NiV) qui adoptent une structure à enroulement en spirale (orientation "gauche") en solution.

Le virus Measles (MeV) est un simple brin, non segmenté du virus négatif qui appartient à la famille des Paramyxoviridae et il est encapsidé par les monomères de la nucléoprotéine (N) (Blocquel et al., 2014). Plusieurs structures cristallines ont été résolues pour le domaine P de multimérisation (PDM) de MeV. Dans MeV PDM, il y a une rupture dans la réplique du motif abcdefg autour de L342 qui conduit à l'apparition d'un coude à cette position. Par conséquent, les positions L339 à L342 forment une hélice 3_10 qui oriente K343 vers l'extérieur. Les structures diffèrent au niveau de la région C-terminale. Dans une structure (code PDB: 3ZDO) (Communie et al., 2013), cette région est ordonnée alors qu'elle est absente dans l'autre structure (Code PDB: 4BHV) (Blocquel et al., 2014).

Le virus Nipah (NiV) est un agent pathogène humain nouvellement émergé dans la famille des Paramyxoviridae (Eaton et al., 2007) et aucun vaccin à usage humain n'a encore été développé (Broder, 2012). Le domaine P de multimérisation (PDM) de NiV couvre les résidus 470-578. La structure cristalline de NiV PDM a été résolue comme un long "coiled-coil" tétramérique parallèle (Bruhn et al., 2014), dans lequel l'extrémité N-terminale de chaque monomère forme un capuchon à 2 hélices. Il y a un coude ("kink") formé à la position Pro 544 au milieu de chaque hélice.

Nous sommes intéressés à étudier 2 questions biologiques principales. Nous souhaitons tout d'abord prédire la région désordonnée de PDM des deux virus mentionnés. Les données cristallographiques sont contradictoires et nos collaborateurs, Sonia Longhi (Université d'Aix-Marseille) et Denis Gerlier (Ecole Normale Supérieure de Lyon), ont réalisés des expériences qui suggèrent que la partie C-terminale de PMD est intrinsèquement désordonnée. D'autre part, les virus de Paramyxoviridae ont été montrées pour former des trimères ou des tétramères. En particulier, les PDM des virus Measle (MeV) et Nipah (NiV) ont été cristallisés comme tétramères. Nos collaborateurs ont démontré une forte preuve expérimentale obtenue à partir d'études SAXS et spectroscopie UV lointain pour l'existence d'une forme trimérique de NiV PDM en solution (Blocquel et al., 2013). Par conséquent, il y a un fort intérêt pour savoir quele est la forme la plus stable pour les NiV PDM et MeV PDM, le trimère ou le tétramère.

1.5.1 Analyse de COMMA

Nous avons produit 2 répliques de 50 ns de simulations DM pour chacun des MeV et NiV PDM. Ensuite COMMA a été appliqué aux 30 dernières ns de chaque réplicat pour extraire les blocs de communication pour chacun des systèmes (**Figure 1.6**). Il convient de souligner que toutes les interactions non-covalentes squelette-squelette sont ignorées pour l'analyse de COMMA. Bien que, dans chaque système, les hélices ont une séquence identique, l'analyse de COMMA révèle un comportement différent pour chacune. Même le long des hélices, des comportements différents sont observés. Par exemple, la seconde moitié de la chaîne A en MeV PDM est détecté comme un CB^{path} séparé, tandis qu'un CB^{clique} est détecté au niveau du coude à cette même région (ils ont un chevauchement de 6 résidus). Aussi, la deuxième moitié de la chaîne B de NiV PDM, est détecté comme un CBs^{path} séparé. En outre, le regroupement des hélices est différent entre la PDM des deux virus.

Nous avons examiné si ces observations, en particulier la détection d'une partie désordonnée C-terminale, pourrait être reproduit pour toute structure "coiled-coil". Nous avons analysé la protéine RhcC de *marinus Staphylothermus* qui a été résolu comme un tétramère "coiled-coil" d'orientation "droite" (**Figure 1.6c,f,i**). Les résultats suggèrent un comportement différent de ce que nous avons observé pour les "coiled-coils" gauches. Aucun des CBs^{clique} identifiés au niveau des extrémités N- et C-terminaux du tétramère n'englobe les quatre hélices, ce qui suggère l'absence d'une région désordonnée sur la structure de RhcC. Ce que nous avons observé pour MeV et NiV PDM, ne se reproduit pas. L'absence d'un noyau de communication unique à travers les hélices est visible ici, où chaque hélice affiche un rôle indépendant. Une des raisons du contraste observé entre les deux types de "coiled-coils", pourrait être l'absence de coude dans RhcC.

Deux répliques de 50 ns de simulations DM ont également été produits pour chacun

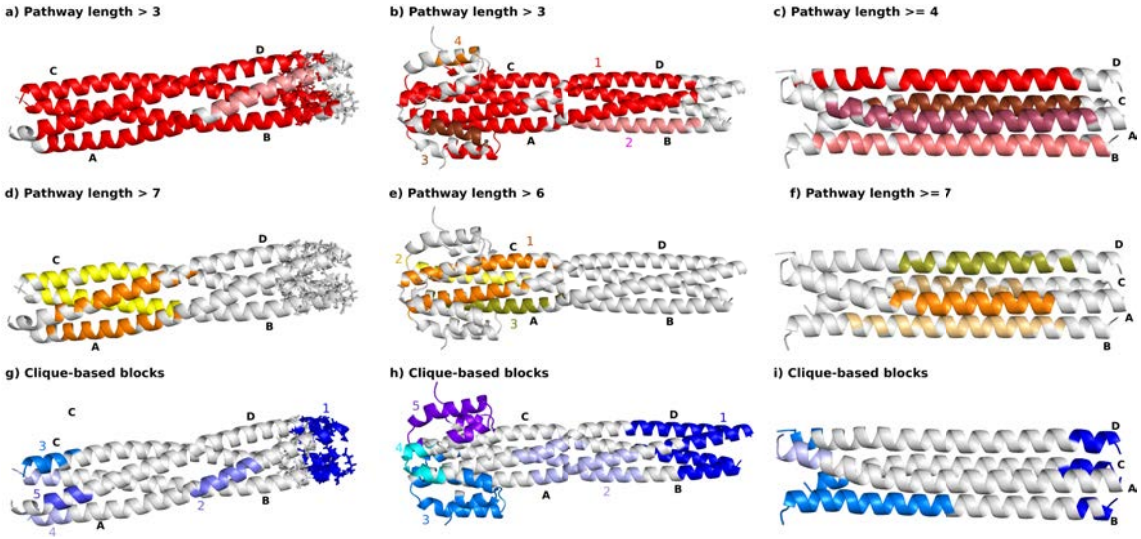

Figure 1.6: **Blocs de communication de MeV PDM, NiV PDM et RhcC identifiés par COMMA.** Les blocs de communication sont mappés sur la conformation DM moyenne. Ces blocs sont obtenus en tenant compte des chemins d'une longueur égale ou supérieure à 4 (a, b et c), 8 (d) et 7 (e et f) les résidus. Les blocs de communication basé sur les cliques sont colorés dans des tons bleus (g, h et i). Les résidus connus impliqués dans la région de désordre de MeV PDM sont représentés par des bâtons.

des monomères MeV PDM, NiV PDM et RhcC. Dans les simulations des deux MeV et NiV PDM, le déroulement de la C-term a été observée. Les hélices simples sont fortement courbées lors de la simulation avec le kink servant de point d'articulation. La structure moyenne révèle l'apparition d'une courbure importante avec un haut degré de flexion pour tous les trois systèmes. Les résultats indiquent que MeV et NiV PDM n'adoptent pas de conformations monomèriques stables en solution et que leurs parties C-terminales sont intrinsèquement désordonnées. Nous pouvons interpréter ces résultats comme une transition de l'état désordonné à l'état "pas-si-ordonné" lors de la formation du tétramère pour les deux PDMs.

Nous avons comparé nos résultats avec trois programmes basés sur la séquence: Coils server, IUPred et ANCHOR, en utilisant leurs paramètres par défaut. Les trois méthodes ne sont pas en mesure de détecter des régions flexibles et instables des "coiled-coils" de manière aussi précise que COMMA. Fait intéressant, COMMA fournit également un moyen de distinguer les deux comportements différents. La présence d'un seul CB^{clique} sur les quatre chaînes représentent la probabilité d'une région désordonnée, alors que l'existence d'un autre type de CBs^{clique} fournir indice sur la région flexible pour la dynamique des bobines enroulées.

1.5.2 Contrôle de la flexibilité/communication par des mutations

Pour sonder la stabilité du tétramère MeV PDM, nous avons muté 4 résidus hydrophobes en position *a* du motif répété en acides aminés chargés négativement: V315D, L322D, V346D et I353D. Les deux premières positions sont situées avant le "kink" (première moitié), tandis que les deux dernières positions sont situées après (seconde moitié). Des simulations DM ont été réalisées pour chaque mutation, puis nous avons appliqué l'analyse

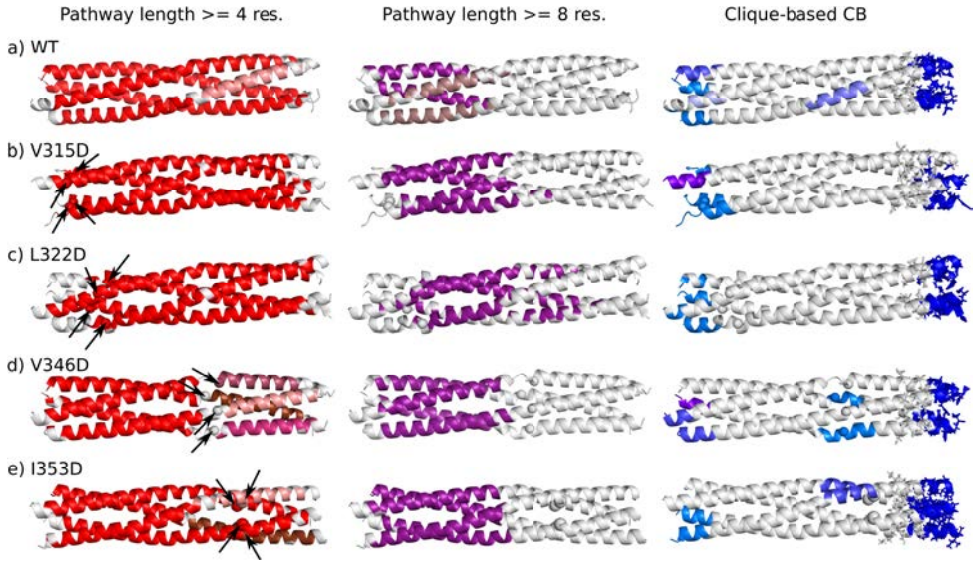

Figure 1.7: **Etude des mutants de communication MeV PDM.** Blocs de communication identifiés par COMMA pour le **a)** de type sauvage et des mutations au niveau de la première moitié des hélices, **b)** V315D et **c)** L322D) et des mutations au niveau de la seconde moitié des hélices, **d)** I353D et **e)** V346D, sont mappés sur la conformation DM moyenne. Les CBs^{path} sont obtenus en tenant compte des interactions impliquant des chaînes latérales et des chemins d'une longueur égale ou supérieure à 4 et 8 résidus. Les résidus connus impliqués dans la région de trouble sont présentés comme des bâtons. Les blocs de communication basée clique-identifiés par COMMA sont colorées dans des tons bleus.

COMMA à l'ensemble des conformations obtenues pour chaque mutation (**Figure 1.7**).

Les mutations de la première moitié conduisent à l'augmentation du nombre de communications directes dans la seconde moitié et à l'établissement d'une communication entre les deux moitiés, qui à son tour a permis aux deux moitiés de communiquer à travers la structure. D'autre part, les mutations de la deuxième moitié conduit à la fragmentation des CBs et diminution de la communication entre les deux moitiés et à une augmentation significative de la communication à l'intérieur de la première moitié. La comparaison des résultats représentent une séparation entre les deux moitiés des hélices avant et après le coude (”kink”). Le N-term est devenu plus stable et rigide, tandis que le C-term est plus flexible. La caractérisation expérimentale des mutants pourrait aider à valider ces hypothèses.

1.6 Conclusion

Dans cette thèse, nous avons présenté COMMA, une méthode pour décrire et comparer les architectures dynamiques de différentes protéines ou différentes variantes de la même protéine. COMMA extrait propriétés dynamiques de ensembles conformationnels pour identifier les *chemins de communication*, des chaînes de résidus liés par des interactions stables qui se déplacent ensemble, et *cliques indépendantes*, des groupes de résidus qui fluctuent de manière concertée. Les chemins et les cliques sont utilisés pour définir des *blocs de communication*. Le terme 'communication' se réfère à la façon dont l'information est transmise à travers la structure de la protéine. L'originalité de la méthode réside dans

le fait qu'elle représente les deux modes de communication différents, grâce à l'utilisation des chemins et des cliques. Elle fournit une description de l'infostérie d'une protéine qui va au-delà des notions de chaîne, domaine et structure/motif secondaire, et au-delà des mesures classiques de la façon dont une protéine se déplace et/ou change de forme.

Nous avons montré l'efficacité de notre approche pour fournir des idées mécanistiques sur les effets des mutations délétères en identifiant les résidus qui jouent un rôle clé dans la propagation de ces effets, à travers différentes études de cas. En outre COMMA révèle un lien entre les clusters de coévoluant résidus et les réseaux de corrélations dynamiques. Il permet de comparer les différents types de communication se produisant entre les résidus et de hiérarchiser les différentes régions d'une protéine en fonction de l'efficacité de leur communication. En outre, nous avons présenté une approche pour exploiter les séquences et les dynamiques structurelles pour prédire un paysage mutationnel. La discussion des exemples, a révélé comment l'étude de la conservation apporte des idées importantes sur le rôle de chaque mutation. Notre méthode proposée pour étudier la dynamique des protéines, peut détecter des régions de protéines qui sont sujettes à des troubles ou des réarrangements conformationnels substantiels. En outre, l'analyse COMMA des "coiled-coils" nous a permis de proposer des mutations qui régulent leur stabilité.

Une analyse plus poussée peut être appliquée pour améliorer les méthodes proposées. **(1)** La mise en place automatique des seuils utilisés dans COMMA a besoin d'être modifié pour l'étude des grands complexes. **(2)** Le lien entre les clusters de coévoluant résidus et les réseaux de positions corrélées dynamiquement doit encore être étudiée plus. **(3)** Nous avons proposé une hypothèse de mutations pour réguler la stabilité des "coiled-coils", l'étude expérimentale de ces mutations peut ajouter plus de preuves à nos résultats.

Chapter 2

Introduction

Contents

In this chapter we explain the basics of biology that are necessary to understand the following chapters and discuss the biological questions that we try to answer.

2.1 Some biological questions in computational biology

Proteins perform many biological functions, among which are chemical reaction catalysis, cell structuring, signal transduction and gene expression. They regulate biological processes by interacting with other molecules (proteins, nucleic acids, cofactors, ...) and by adapting their shape and motions to environmental changes. Protein conformational dynamics are directly linked to protein functions (Henzler-Wildman and Kern, 2007; Frauenfelder et al., 1991). For example calmodulin is a protein that adopts completely different conformations depending on the number of calcium ions bound to it and the proteins (cellular partners or proteins from pathogens) it interacts with (**Figure 2.1**) (Liddington, 2002; Vetter and Leclerc, 2003). Proteins are also subject to evolutionary and structural constraints to maintain and/or adapt their functions.

Proteins are macromolecules formed by one or several chains of amino acids. The structure of proteins is organized according to four different hierarchical levels: **primary structure** represents the sequence of amino acids forming one protein, **secondary structures** are the set of local structures with regular repeats that are stabilized by hydrogen bonds formed between the backbone atoms of the amino acid residues (α-helices,

β-sheets, turns, etc.) in the protein, **tertiary structure** is the arrangement of secondary structures in 3D space that is stabilized by non-bonded interactions, salt bridges, hydrogen bonds, disulphide bonds, etc. and **quaternary structure** is the arrangement of domains/chains within a protein or proteins within a macromolecular structure.

Proteins can be viewed as 1-dimensional (1D) sequences of amino acids or as 2-dimensional (2D) arrangement of secondary structure elements or as highly dynamic 3-dimensional (3D) structures, interacting with each other to form complexes (quaternary structure). The analysis of the sequences of evolutionary related proteins enables to identify conservation and coevolution signals. Amino acid residues highly conserved through evolution are likely to be very important for the function of the protein. Pairs or groups of residues that coevolve, i.e. they change at the same time during evolution, are likely to cooperate or collectively contribute to the function of the protein. The analysis of secondary structures in 2D space, enables to detect locally stable structure element. Biologically pertinent information can also be extracted from the description of the positions of protein atoms in 3D space (protein tertiary structure) and their displacements along time (protein motions). Proteins change their shape (conformation) in response to environmental conditions. Such changes can be studied by simulating/characterizing the dynamical behaviour of proteins in solution. Moreover, the interactions between domains/chains/proteins can be studied from the analysis of complexes. Accounting for the changes of proteins through time and/or the interactions of proteins between each other, the study of complexes adds some levels of complexity to their description and permits to better understand how they perform their functions.

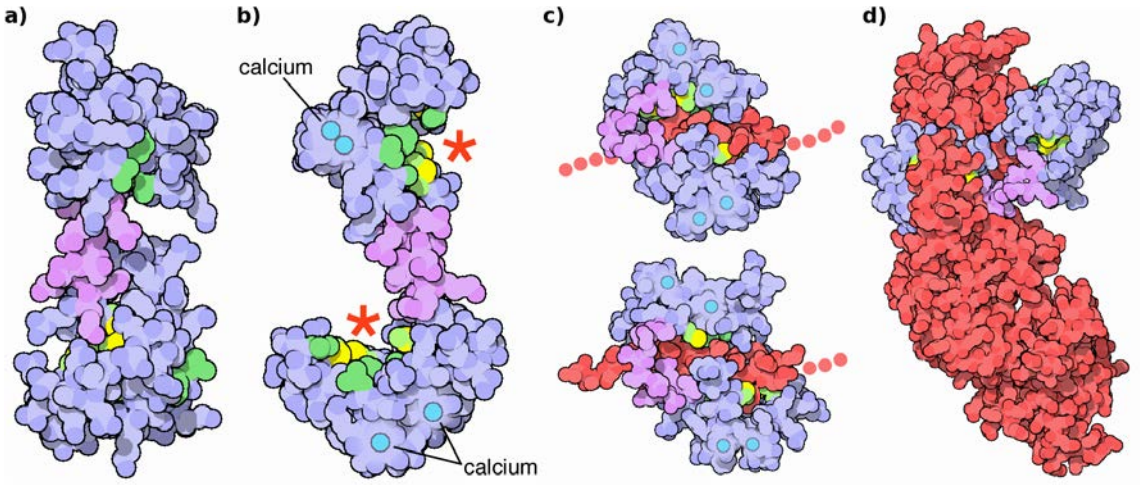

Figure 2.1: **Calmodulin in different environmental conditions.** Calmodulin is a calcium sensor and adopts different conformations depending on the level of calcium ion bound to it or the interacting proteins (figure from (Berman et al., 2000)). The figure represents: **a)** Calmodulin structure without calcium (Hitoshi Kuboniwa et al., 1995), **b)** Calmodulin after binding to calcium (Chattopadhyaya et al., 1992), **c)** Calmodulin bound to two different enzymes shown in red (calmodulin-dependent protein kinase II-alpha at the top (Wall et al., 1997) and myosin light chain kinase on the bottom (Ikura et al., 1992)) and **d)** Calmodulin bound to a toxin (the edema factor toxin from the anthrax bacteria) (Drum et al., 2002).

A **mutation** is a change in the amino acid sequence that may affect the structure and/or function of the proteins. Mutations may have different phenotypic outcomes: they can be

beneficial (gain-of-funtion), **neutral** or **deleterious** for the function of the protein.

A deleterious mutation may alter the structural plasticity of a protein and induces effects at distant protein sites, thereby provoking diseases. For example growth hormone (GH) is a protein responsible for human growth and point mutations of this protein may cause genetic diseases such as short stature (Petkovic et al., 2010). Another example is given by the receptor tyrosine kinase KIT, where a cancer mutation located in the activation loop induces structural changes in a region distant by more than 15 Å (**Figure 2.2**). Therefore, any finding of the protein conformational preferences and changes may lead the way to designing drugs, recognizing the function of inheriting contagious diseases and hopefully preventing them.

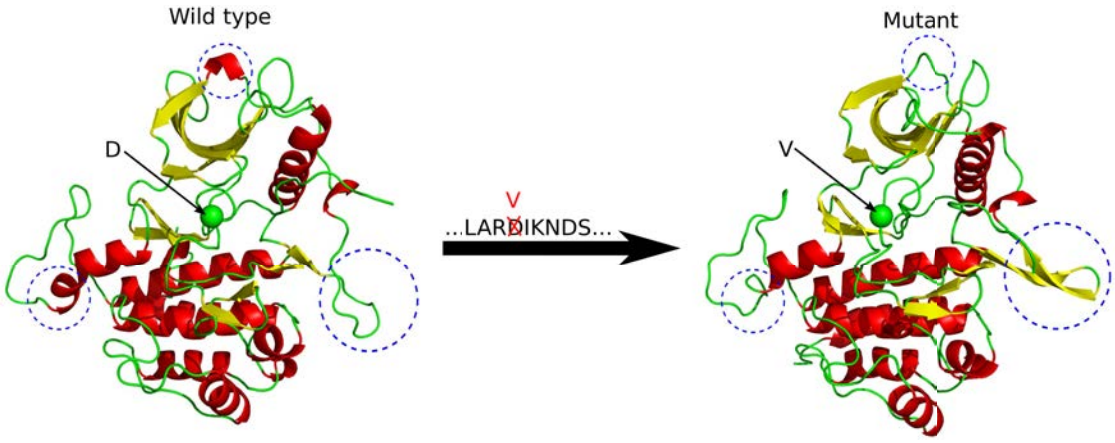

Figure 2.2: **The effect of a mutation.** The change in one amino acid on the sequence of KIT, effects the structure. The mutation position is shown with a green sphere and the affected regions are represented by dotted blue circles on the structure.

Although experimental methods such as X-ray crystallography and Nucleic Magnetic Resonance (NMR) can provide high resolution atomic details of individual protein structures, it is difficult to determine experimentally the atomic structures of proteins at large-scale and systematically assess the effects of mutations on the structure and/or dynamical behavior of a protein (not all the mutations induce effects on the structure). It is true that these techniques have improved dramatically in recent years, but it is equally true that the study of the mutational landscape is drastically difficult. Therefore, the development of computational prediction methods to analyze the effect of mutations in a large scale is extremely important to help understand the molecular mechanisms of biological systems.

The main biological questions that we address in this thesis, are the following: **1)** What are the positions in a protein that are highly sensitive to mutations? **2)** What is the effect of a particular amino acid substitution? **3)** What is the link between the structural and the evolutionary constraints? And many other related questions like: which regions of the protein are more sensitive to mutations and what are the chains of residues (*pathways*) through which the perturbations could propagate across the structure? We wish to characterize the mutational landscape of a protein at large scale or in a systematic way through the joint analysis of protein structure/sequence.

In addition to the biological questions mentioned above, we address the following computational challenges in this thesis: **1)** Design of a computational method that enables the analysis of protein dynamics at different levels. **2)** Development of a computational

tool based on the proposed method, to study the conformational dynamics of proteins in a systematic way at large scale. **3)** The ability to move between different protein representations (1D, 2D, 3D and 4D) has to be embedded in the method. This feature enables to extract relevant information from each representation and take advantage of them in the joint analysis of structure/sequence and also the analysis of protein dynamics.

2.2 Background

2.2.1 Protein structures

The building blocks of proteins are **amino acids**. Amino acids are small molecules, that are present in all living organisms. They consist of carbon, hydrogen, nitrogen and phosphor atoms. There are 20 different amino acids, they share a common scaffold comprised of one acidic part (carboxylic acid, $COOH$) and one basic part (amine, NH_2), along with a side chain that is specific to each amino acid (**Figure 2.3**). They are encoded in two ways, by a three letter code or a single letter code (A or ALA for Alanine, C or CYS for Cysteine, etc.). Within proteins, amino acids are linked together by a peptidic bond and they are called *residues*, as they lose their acid group when binding together (Berk et al., 2000).

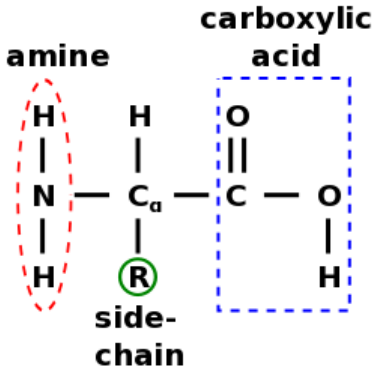

Figure 2.3: **The general structure of amino acids.** Amino acids are made from one amine group, one carboxylic acid and different side-chains.

The way a protein folds into a 3D structure is encoded in its amino acid sequence. Nevertheless proteins are not static. During their lifetime, they undergo conformational changes and associate with partners (small molecules, other proteins, DNAs, RNAs, ...). The view according to which proteins exist in solution as an ensemble of conformations is nowadays generally accepted. Likewise, the necessity to characterize the dynamical behaviour of a protein to improve our comprehension of its functioning in the cell is increasingly acknowledged. The amino acid composition, structural context and possible biochemical modifications (*e.g.* post-translational modification like phosphorylation) will determine to what extent a protein segment or region will fold, be more and less flexible, amenable to conformational changes and/or subject to disorder.

2.2.2 Conformational dynamics

Protein conformational dynamics are directly linked to protein functions (Henzler-Wildman and Kern, 2007; Frauenfelder et al., 1991). They are sensitive to environmental changes, point mutations, ligand binding and post-translational biochemical modifications (Tsai et al., 2008; Kern and Zuiderweg, 2003; Weber, 1972). Atomistic molecular simulation is a method of choice to explore a protein's conformational space. It has become increasingly popular with the recent advances in computational power, force field accuracy and sampling algorithm development (Piana et al., 2014; Dror et al., 2012). The accumulation of molecular dynamics (MD) data calls for the development of methods able to extract pertinent biological information and visualise it in a comprehensive way.

Molecular dynamics MD simulations consist in simulating (computationally) the behaviour of proteins in solution. They provide detailed information on the fluctuations and conformational changes of proteins and nucleic acids, by producing **trajectories** (ensembles of snapshots of the protein taken at regular time intervals). Physical movements of atoms and molecules are approximated by using the Newton's equation of motion (classical mechanics). MD was first introduced by Alder and Wainwright in the late 1950's (Alder and Wainwright, 1957, 1959) to study the interactions of hard spheres. Their studies paved the way for further improvements. The first MD simulation was carried out by Rahman in 1964 using liquid argon (Rahman, 1964), the first MD simulation of a realistic system using liquid water was performed in 1974 (Stillinger and Rahman, 1974) and it was in 1977 that MacCammon et al. performed the first MD simulations of a protein (McCammon et al., 1977).

In explicit solvent simulation, the protein is placed in a virtual box of water and its motions are simulated typically over 0.01-10 microseconds and recorded. An average size for a system (protein and water molecules) to be studied with MD simulations, is 10^4 to 10^5 atoms and the computational time is several CPU-days or CPU-years. For example a protein of medium size (PDZ domain), comprised of 83 residues, contains 1238 atoms and each of this atom has 3 degrees of freedom (cartesian coordinates x, y, z). The motions of the protein depends on the interactions between all atoms of the protein and also between the protein atoms and the solvent (water molecules) surrounding it, consequently in total the system has 42072 degrees of freedom. The computational time needed to apply 20-ns MD simulations over this system with 14024 atoms (with pmemd of Amber package (Case et al., 2012)), is about 440 hours (or more than 18 days) of CPU-time. MD simulations are CPU intensive, consequently the use of high performance computing resources is necessary and the advances in parallel algorithms allow the simulations to be distributed among several CPUs.

Numerous studies have used all-atom MD simulations in explicit solvent to successfully characterize mutation-induced changes on the structure, internal dynamics and thermodynamic stability of proteins, and predict their functional implications (see (Liu and Nussinov, 2008; Dixit and Verkhivker, 2009; Laine et al., 2011b; Calhoun and Daggett, 2011; Couve et al., 2014; Chauvot de Beauchene et al., 2014; Da Silva Figueiredo Celestino Gomes et al., 2014; Kamaraj and Bogaerts, 2015; Saladino and Gervasio, 2016; Lu et al., 2016) for a non-exhaustive list of references). Increasing computational resources now permit to simulate mutated systems on time-scales that are functionally relevant (sev-

eral tens of microseconds). Still, the complete description of a protein's conformational landscape is far beyond reach. In addition, identifying the meaningful protein properties to be recorded in the simulations is not trivial and extracting pertinent biological information often require some expert knowledge on the system studied.

Allosteric coupling The binding of a ligand at one site of a protein can have effect at long distance on another binding site of the protein (Monod and Jacob, 1961; Monod et al., 1965). This phenomenon is referred to as allostery and was first described by Changeux 50 years ago (Changeux, 1961). Such propagation of a perturbation signal between distinct sites, possibly located far away in the sequence and structure of the protein, is modulated by "communication" between residues. An example of allosteric pathways responsible for a transition between tensed and relaxed states in heamoglobin is shown here (Motlagh et al., 2014). Experimental evidence has demonstrated that protein residues "communicate" either through stable **non-covalent interactions** (Monod et al., 1965) or via changes in their local **atomic fluctuations** (Schrank et al., 2009). Previous methodological efforts were engaged toward the identification of clusters or chains of residues mediating long-range communication in proteins (Chiappori et al., 2012; Papaleo et al., 2012; Laine et al., 2012; Raimondi et al., 2013; Pandini et al., 2013; Blacklock and Verkhivker, 2013; McClendon et al., 2014; Invernizzi et al., 2014; Allain et al., 2014) and most of these methods construct a graph to represent the protein.

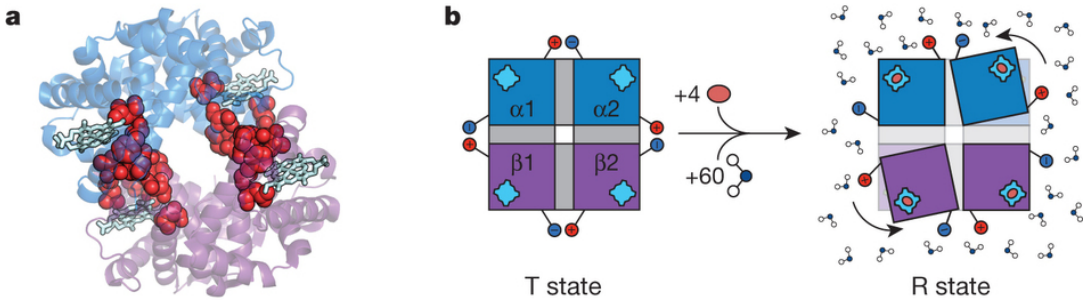

Figure 2.4: **Allostery in heamoglobin. a)** Proposed allosteric pathways that are responsible for the transition from tensed form (T) to relaxed form (R) are represented on the structure of tetrameric haemoglobin with red spheres and light blue sticks show the haem groups (Süel et al., 2003a). **b)** The proposed allosteric transition of tetrameric haemoglobin by Perutz (Perutz and TenEyck, 1972; Perutz et al., 1998). The α-subunits and β-subunits are shown for each state with blue and purple colors, respectively and the haem groups are shown in light blue. The salt bridges (shown in red and blue spheres, representing positive and negative charges, respectively) hold the molecule in T state. The transition from T to R state happens upon binding of oxygen (orange oval), which leads to release of salt bridges and rotation of subunits by 15°. This is along with binding of 60 water molecules to the R state that brings the equilibrium (Colombo et al., 1992). (Figure is taken from (Motlagh et al., 2014).)

Graph representation In order to develop computational tools to analyse conformational dynamics of proteins, the representation of a protein as a graph is useful as it unravels more easily and readily its properties at the atomic or residue level. Typically, each node of the graph represents one residue of the protein and the edges represent non-covalent interactions that stabilise the protein three-dimensional structure (Boede et al.,

2007; Vishveshwara S, 2009). Information about the dynamical behaviour of the protein can also be integrated in several ways. For example, the edges can be constructed and weighted based on the persistence values of the interactions computed over a conformational ensemble instead of their presence/absence in a static structure (Tiberti et al., 2014). Other types of dynamic properties can be taken into consideration, such as dynamical correlations between residues (Seeber et al., 2011; Bhattacharyya et al., 2013; Skjærven et al., 2014). Alternatively, every conformation of a MD trajectory can be represented by a contact graph and the evolution of the graphs can be analysed over time to detect important structure-changing events of the graphs (Wriggers et al., 2009).

MONETA (Laine et al., 2012) A number of previously developed methods are dedicated to the analysis of the dynamical behaviour of proteins and their inter-residue communication (Skjærven et al., 2014; Schrank et al., 2009; Tiberti et al., 2014; Raimondi et al., 2013; McClendon et al., 2014). In particular, the method MONETA proved useful to identify communication routes in allosterically regulated proteins and to guide *in silico* mutagenesis (Allain et al., 2014). MONETA is intended to assist the analysis of MD simulation data in a manually-guided way. It enables to focus on specific protein regions or residues provided that the user has some prior knowledge of the system. Fixed values are encoded in the tool for most of the parameters, which limits its applicability and flexibility.

2.2.3 Intrinsic disorder

Intrinsically disordered proteins (IDP) and **intrinsically disordered regions (IDR)** in proteins are characterized by lack of stable tertiary structure under physiological conditions in vitro (Dunker et al., 2001). Previous studies demonstrated that IDPs and IDRs sequences share some properties, *i.e.*, they have lower complexities, lower number of hydrophobic amino acids and more polar or charged residues (Wright and Dyson, 1999; Ishida and Kinoshita, 2007), and low sequence conservation (Brown et al., 2011; Mei et al., 2014). These findings resulted in the development of the numerous sequence-based computational approaches to predict disordered regions in proteins (He et al., 2009; Peng and Kurgan, 2012; Deng et al., 2012).

The interest in protein intrinsic disorder has grown over the recent years, as its prevalence in the proteome and its contribution to protein function has become more and more evident. However, it remains unclear whether the physico-chemical properties associated to intrinsic disorder are fundamentally different from those associated with flexibility and conformational plasticity (Uversky and Dunker, 2010). Alternatively, one can think of a continuous scale ranging from well-ordered stable states to completely unfolded states and spanning a variety of degrees of flexibility or disorder. The rapidly growing body of structural data available in the Protein Data Bank (Berman et al., 2000) provides a means to infer the position of protein regions on such a scale. Specifically, previous studies have exploited the redundancy of the PDB, *i.e.* the fact that 2 or more models of the same protein are available, which is the case of a majority of proteins in the PDB (Uversky and Dunker, 2010). At one end of the spectrum, residues that adopt very similar conformations in all PDB structures are considered as stable and well-ordered. At the other end, residues that are missing from all PDB structures are considered as intrinsically

disordered (Bloomer et al., 1978; Bode et al., 1978). An interesting concept was introduced to describe the residues lying in between, that of ambiguous or dual-personality fragments/regions which are present in some PDBs and missing in others. There can be several possible reasons explaining such discrepancies between PDB structures: different crystallization conditions (Lian, 1998; Sidote and Hoffman, 2003), different space groups (*e.g.* a solvent-exposed loop may be stabilized by crystal packing), different conformational states (*e.g.* active/inactive in the case of enzymes) (Frimpong et al., 2010), the presence/absence of cofactors, ligands, biochemical modifications, etc (more references in (Uversky and Dunker, 2010)). These differences highlight the fact that X-ray crystal structures – which form the vast majority of the PDB – are static snapshots representing stable states of the protein that were captured by crystallization among others that are populated in solution (Harauz et al., 2009). By contrast, NMR models can give insights on the dynamics of the protein and NMR techniques have been developed in recent years to characterize intrinsically disordered proteins (Mizutani et al., 2008; Kobe et al., 2008; Bahadur and Zacharias, 2008).

Tools to predict disorder In this section we present three web-based tools to predict disordered region from protein sequence.

- Coils server (Lupas et al., 1991). This program measures the probability of a sequence to form coiled-coil conformations. It takes a sequence as input and compares it with a database of known parallel two-stranded coiled-coils and measures a score based on the similarity. Consequently, it compares the score with the distribution of scores for globular and coiled-coils proteins to obtain the probability of forming coiled-coil conformation.

- IUPred (Dosztanyi et al., 2005). The program predicts the set of intrinsically disordered regions from the protein sequence. The idea behind is to estimate the pairwise energy content. The amino acids in globular proteins are able to form great number of favorable interactions, while intrinsically disordered proteins do not have the potential to form enough favorable interactions, due to their lack of stable structure.

- ANCHOR (Dosztanyi et al., 2009). The program is based on the same energy estimation approach as in IUPred and predicts the disordered binding regions. A large set of disordered proteins experience a transition from disorder to order when binding to a structured partner. The idea behind ANCHOR is to look for disordered segments that are not able to fold on their own through forming sufficient favorable interactions and presumably will gain stabilizing energy by interacting with a globular protein partner.

2.2.4 Evolutionary conservation and co-evolution

Conservation Homologous sequences are the sets of sequences that have evolved along billions of years, they are the results of evolution and natural selection and share the same common ancestor. Every set of ancestral sequences are generated through mutation, insertion and deletion. The degree of variation in every position along a set of homologous sequences can be very diverse, the smaller the variation at a position, the higher the degree

of conservation and hence the more biologically important the position. The study of conservation rate for positions within homologous sequences can help to predict regions that are more important and perturbations in those regions are more likely to be deleterious (see (Carbone and Dib, 2011) for a list of references).

Different methods are proposed to measure the conservation level of every position from a set of aligned homologous sequences (multiple sequence alignment (MSA)). Analysis have been based on the classical notion of information content, and captured numerically the residue variability in a single position of the MSA, by providing a global numerical score. This score represents the entropy of the set of sequences through the combination of local information on alignment positions (Akashi, 1999; Thompson et al., 1999; Duret et al., 2000; Lecompte et al., 2001; Notredame, 2002; Wallace et al., 2005; Watson et al., 2005; Notredame, 2007). Additional information have been also considered, for example physico-chemical properties of residues and local preservation of those properties along the MSA (see (Carbone and Dib, 2011) for a list of references).

Several methods are proposed to analyse the phylogenetic tree topology and the evolutionary distances between a family of homologous sequences, in order to extract signals of conservation and identify positions that are conserved at different levels of an evolutionary tree (Faith, 1992; Mihalek et al., 2004; Landau et al., 2005; Sankararaman et al., 2009; Carbone, 2014). A phylogenetic tree can be associated with the evolution of a group of species, where leaves correspond to species in the group and the internal nodes represent their ancestors. In most of the cases, species that are positioned close to each other share the same biological behavior, whereas in the case of long branches, phylogenetically close species may represent different behaviours (Carbone, 2014). For the conserved protein domains, it is shown that phylogenetically close species share conserved patterns in homologous sequences and for proteins that have diverged sequence identity (more than 50%), we can still find homologous sequences that display specific conserved patterns, but this is more likely in distant species than in phylogenetically close ones (Carbone, 2014). Consequently, the topology of the phylogenetic tree plays a very important role to identify conservation along the tree.

Conservation could be used as a method to predict mutational effects (de Juan et al., 2013). Assuming that a protein in all the species performs the same function, however it possesses changes on the sequences at some extent, one can infer the neutral effect of such mutations. On the other hand, if different functions are observed among species, then it suggests the deleterious effect of the mutations, because of the induced changes of protein function.

Coevolution Coevolution in proteins refers to the cases where changes on different positions, along the sequences, happen at the same time. Coevolution is the sign of a functional and/or structural dependency between two positions, for example direct non-covalent interactions, but it can be other things, like propagation of signals. In the case of residue pairs forming non-covalent interactions, a change in one of the residues to another amino acid with different physico-chemical properties, can break the interaction. In order to maintain the interaction, the other residue involved in the pair has to be changed in an acceptable manner. Such simultaneous changes may result in equal or even larger fitness and be accepted by the natural selection. Nevertheless, in a large number of cases, the structural constraints acting on conserved and/or coevolved residues remain to be identi-

fied. A toy example of coevolution and conservation is represented in Figure 2.5.

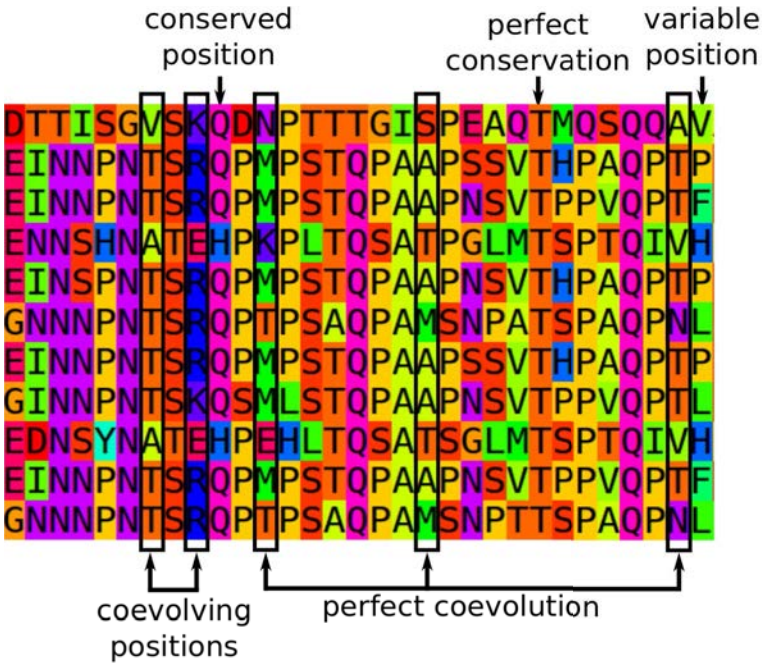

Figure 2.5: **Toy example for conservation and coevolution patterns.**

The first method to investigate correlated changes of amino acids along homologous sequences has been introduced about twenty years ago (Pollock and Taylor, 1997). A large number of methods have been proposed to investigate the evolutionary constraints in proteins through the analysis of sequences. The set of coevolving residues detected by these methods, are usually close in the three-dimensional structure (Lockless and Ranganathan, 1999; Baussand and Carbone, 2009; Jones et al., 2012; Morcos et al., 2013), they are shown to contain approximately a third of protein residues and form connected networks. Moreover, residues conserved through evolution can be involved in functional interactions between proteins and biomolecules (Lichtarge et al., 1996; Pupko et al., 2002; Lichtarge and Sowa, 2002; Glaser et al., 2003; Cheng et al., 2005; Innis, 2007; Engelen et al., 2009; Lichtarge and Wilkins, 2010). In addition, due to the availability of experimental data, the crucial role of such coevolving residues were proved for a few protein complexes in the allosteric mechanism (Lockless and Ranganathan, 1999; Kuriyan, 2004; Baussand and Carbone, 2009), to maintain short paths in network communication and to mediate signalling (del Sol et al., 2006, 2007). Signals of evolutionary covariation have also been exploited to predict with high accuracy native contacts within protein structures (Morcos et al., 2011), inter-molecular interactions (Champeimont et al., 2016) and intramolecular allosteric communication (Sung et al., 2016).

The existing approaches to capture coevolution signals, can be divided into two main groups: statistical methods and combinatorial methods. Analysis of the first category are based on capturing covariation between positions of the aligned sequences by measuring correlation coefficients (Goh et al., 2000; Fares and Travers, 2006), mutual information (Atchley et al., 2000; Ramani and Marcotte, 2003; Gloor et al., 2005) and deviance of distributions to estimate the thermodynamic coupling between residues (Lockless and Ranganathan, 1999; Süel et al., 2003b; Dima and Thirumalai, 2006; Weigt et al., 2009;

Sadowski et al., 2011; Morcos et al., 2011). Some of the statistical approaches use also the phylogenetic tree for the analysis of sequences with similar degree of covariation (Yeang and Haussler, 2007). The second group of sequence-based approaches to detect coevolving residues, are based on sequence counting and the use of phylogenetic trees, in order to overcome the restrictions of statistical approaches (Fryxell, 1996; Pazos and Valencia, 2001; Baussand and Carbone, 2009; Dib and Carbone, 2012b). In those methods the distance tree is extracted from the phylogenetic tree and the analysis are done based on the combinatorics of distance subtrees.

MST (Baussand and Carbone, 2009) and BIS (Dib and Carbone, 2012b), are two combinatorial approaches, proposed in our lab, to detect coevolving residues. In both methods, the degree of conservation and coevolution is measured by constructing the distance tree from the phylogenetic tree of MSA. Their advantage lies in the fact that they can measure coevolution over a small set of sequences. MST performs better over a set of sequences with variable divergence, while BIS requires sequences that are highly conserved.

An example of coevolution is represented in Figure 2.6, where we aligned the homologous sequences of Growth Hormone (GH). A subset of the multiple sequence alignment is shown here, where positions 158 and 165 display simultaneous changes. At position 158 several switches between Y and S are present, whereas at position 165, we have D and H. But the changes are all concurrent, every time there is a Y at position 158, position 165 is D. On the other hand, if position 158 is S, there is a H at position 165, but it never happens that at positions 158 and 165 we observe Y and H at the same time, nor S and D.

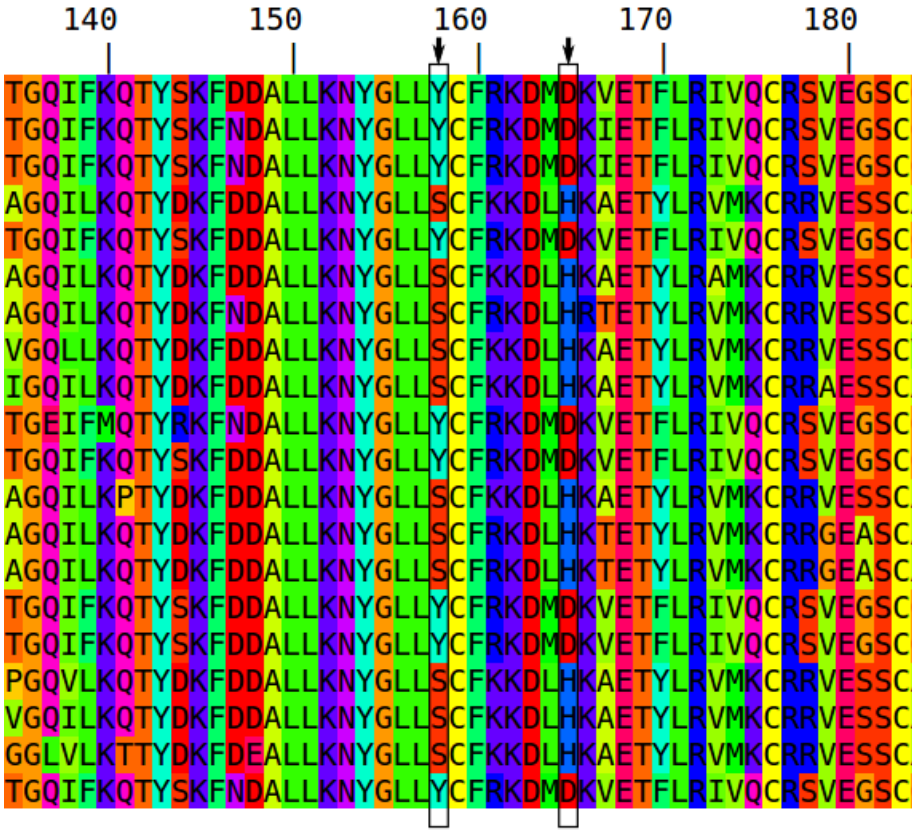

Figure 2.6: **An extract of an alignment of homologous sequences of Growth Hormone.**

We present a short summary of state-of-the-art methods to extract the sequence conservation and coevolution signals, then we will focus on how these signals can help us to detect mutational hotspots in proteins. Several methods are proposed to detect coevolution across sequences, but here we will give a concise summary for some of them that are used in the following chapters.

SCA (Lockless and Ranganathan, 1999) The Statistical Coupling Analysis (SCA) method, is one of the first proposed coevolution detection methods. SCA measures the statistical coupling between residues, in other words it computes the distribution of amino acids at one position with respect to the changes at any other position. From several Multiple Sequence Alignments (MSA), SCA quantitatively measures the statistical energy coupling which is the degree of coevolution among the residues of the studied protein. The given MSA should contain at least a set of 100 divergent sequences for good results, identical residues will not add additional information and will only bias the results. In addition, SCA identifies a group of coevolving residues, called sector. The sector detected for PDZ domian, by using SCA was shown to have significant functional importance for the binding of the protein to its cognate ligand (McLaughlin Jr et al., 2012).

DCA (Morcos et al., 2011) Direct Coupling Analysis (DCA) method, is another statistical technique to detect the coevolutionary signal between residues. The strength of the method is in its ability to disentangle the direct interactions from indirect interactions. For example if residue A interacts with residue B and B interacts with residue C, DCA can enable us to differentiate between the indirect correlation of A and C and the direct interaction of A with B and B with C. But DCA needs a large number of diverged sequences (more than 1000), to detect the coupling between residues, which is a weak point in case of animal sequences. DCA was shown to predict efficiently the mutational landscape of proteins, where authors reported the linear correlation between the experimental data of deep sequencing and predicted values for four different proteins, TEM1, PDZ3, RRM and β-glucosidase (Figliuzzi et al., 2016).

MST (Baussand and Carbone, 2009) Maximal SubTree method (MST) identifies networks of positions that represent coevolution/conservation through the study of phylogenetic tree of homologous sequences of a protein family. The underlying method is to extract distance trees from multiple sequence alignments, analyse the combinatorial of the subtrees and apply clustering with Cluster Aggregation (CLAG) (Dib and Carbone, 2012a) (**Figure 2.7**). MST can be applied to the sequences of protein families with variable divergence. MST is efficient even with few number of sequences, as well as low sequence identity.

BIS (Dib and Carbone, 2012b) Blocks In Sequences (BIS) method considers phylogenetic trees and multiple sequence alignments and extracts conserved and coevolved positions. The specific feature of BIS, is its ability to find blocks of residues that are consecutive along the sequence and that represent coevolution patterns. Nevertheless, it also allows to detect single residues. Then the CLAG clustering (Dib and Carbone, 2012a) is applied to the blocks, in order to extract clusters of coevolving blocks. The mentioned

procedure in shown in figure 2.7. BIS can be applied to protein families that are highly conserved or represented by few sequences. The method is combinatorial in nature, it measures the regularity of a pattern (its "perfection") and the distance from this regularity, with respect to minimal changes, induced by a few mutations. While, statistical methods extract coevolution signals by measuring how distant two amino acids distributions are from noise. BIS have been exploited to extract the coevolving residues for the prediction of protein-protein interactions network of Hepatitis C Virus with high accuracy (Champeimont et al., 2016).

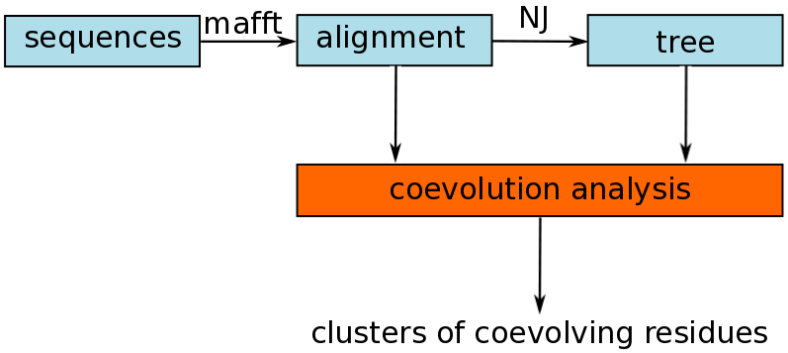

Figure 2.7: **The underlying procedure of MST and BIS.**

2.2.5 Probing proteins mutational landscape

The question of how amino acid sequence variations (re-)shape the conformational landscape of proteins and impact their function is one of outstanding importance in biology, yet far from being resolved. This can be explained by several reasons.

First, systematically assessing the phenotypic outcomes of protein sequence changes is very challenging, both experimentally and computationally. In part, this is due to the combinatorial explosion arising from considering all possible substitutions for single to multiple-point mutant variants. Another difficulty resides in the design of the experiment: what should be measured as phenotypic outcome? For instance, disease-associated mutations can impair protein function in various ways, by destabilizing the structural stability of the protein, by shifting the equilibrium of conformation populations, or by modulating the binding affinity of the protein for its cellular partner(s), to name a few. These effects are difficult to probe directly and unambiguously. Second, the unprecedented breadth of data now accessible though deep sequencing is not always obvious to interpret in terms of protein structure and function.

Recently developed technologies, commonly designated as deep mutational scanning, enable to estimate the functional consequences of every possible single amino acid change at every position in a protein (McLaughlin et al., 2012; Fowler and Fields, 2014; Figliuzzi et al., 2016). These developments are promising and the produced data can be used to validate *in silico* predictions. Several methods have been developed for predicting the effect of mutations, and here we describe a number of them briefly.

Independent position conservation Independent model (IND) is a very straightforward sequence-based approach to predict mutational phenotypic outcomes, based on

counting sequences from a MSA to have a measure of conservation at each position, called Independent model. In this model, the effect of substituting the amino acid a^0, present in the wild-type sequence at the position i, by another amino acid a is estimated as:

$$\Delta E^{pred}_{i(a^0 \to a)} = log\left(\frac{\#(a_i)}{\#(a_i^0)}\right) \tag{2.1}$$

where $\#(a_i)$ (resp. $\#(a_i^0)$) is the number of sequences where a (resp. a^0) occurs at position i. Intuitively, if fewer sequences have a in position i compared to a^0, replacing the latter by the former should be deleterious. This formula, that is called independent model, represents the contribution of every amino acid at each position and reflects the conservation effects. Similar approach was used to predict the phenotypic outcome of mutations for four proteins (Figliuzzi et al., 2016). By measuring the overall Pearson correlation, authors showed about R^2 =39% of the variability in the experimental data of deep sequencing for TEM1 protein, is explained by the independent model.

Polyphen-2 (Adzhubei et al., 2010) Polymorphism Phenotyping v2, is a tool to predict the possible effect of a mutation on the structure and/or function of a human protein. The method is based on using straightforward physical and comparative considerations. It analyses information from conservation of the MSA, solvent accessibility of the residue in the 3D structure (if available), annotations on the sequence in UniProt (binding, active site, lipid and metal) and secondary structures.

SIFT (Ng and Henikoff, 2003) A web-based tool that predicts the impact of an amino acid substitution on protein function based on sequence homology. PSI-BLAST (Altschul et al., 1997) is used to generate the MSA for each protein from closely related sequences, furthermore predictions are based on the conservation degree obtained over the MSA.

PoPMuSiC (Dehouck et al., 2011) Another web-based tool to predict mutational effects. The method is based on evaluating stability change of a protein due to a single amino acid substitution, under the basis of protein's structure. Stability change is measured as a linear combination of 13 statistical potentials to highlight the coupling between protein sequence and structure descriptors of the wild-type and mutant, the variation of amino acid volume due to the mutation (2 terms for wild-type and mutant) and an independent term. The potentials are extracted from a database of known protein structures and represent the correlation between different sequence or structure descriptors (Dehouck et al., 2006). In addition, a neural network is used to adjust the proportionality coefficients according to the the solvent accessibility of the mutated residue.

I-Mutant2.0 (Capriotti et al., 2005) A web-server based on support vector machine and the input vector consists of 42 elements, including the mutated and mutant amino acid, their structural environment, the temperature and pH. It has the ability to consider information from protein structure and/or sequence. After training/testing with a cross validation procedure, I-Mutant2.0 represented an accuracy of 80% for the stability predictions using the structure information and 77% using sequence information.

MUPro (Cheng et al., 2006) A support vector machine web-server to predict stability changes induced by single mutations based on information from both sequence and structure. 84% of accuracy is reported for the method, when considering only the stability changes. On the other hand, the accuracy of predictions that are only based on sequence analysis, are reported to be close to the accuracy obtained using 3D information. Consequently, MUpro overcomes other predictors that are based on structure information.

2.3 Approaches

Networks of dynamically correlated residues play a crucial role in propagating mutations perturbation signals. These residues are expected to display high degrees of conservation and/or coevolution. Understanding the role of disease-related mutations on the link between coevolution and dynamical correlation can help decipher the molecular mechanisms of mutation-induced allosteric deregulation. However the relationship between sequence evolution and structural dynamics has been seldom explored yet. In this work, we have exploited both protein sequences and structures to predict mutational effects and to explore the relationship between structural and evolutionary constraints. We have developed methods and quantitative measures to extract and describe structural/dynamical and evolutionary signals in an automated way and at large scale.

Specifically, we have developed a computational tool, COmmunication MApping to analyze MD simulations and describe the dynamical architecture of proteins. Furthermore, we developed metrics to predict the mutational effects based on sequence and structure/dynamics analysis. Moreover, we applied those methods to different systems and explored the sequence-structure-function relationship. We characterized the effect of genetic disease associated mutations on the structural stability of growth hormone and its interaction with its receptor. We proposed an approach to identify critical residues in proteins and predict the mutational outcome of mutations. We applied it to the third PDZ domain of PSD95 in complex with its cognate ligand. Finally, we characterized the dynamical behaviour of coiled-coils proteins from two viruses to identify regions prone to disorder and predicted the effect of mutations on the stability of the protein oligomer and disorder content.

2.4 Organization of the thesis

COMMA method We present a method, COMmunication MApping (COMMA), to describe the dynamical architecture of a protein starting from a conformational ensemble representing a micro-state and typically generated by MD simulations. COMMA extracts dynamic features, namely dynamical correlations, distances, secondary structures and non-covalent interactions. Then, it integrates them in a graph theoretic framework, where it identifies *communication blocks*, which do not necessarily correspond to domains or groups of secondary structure elements. The term 'communication' refers to the way information is transmitted throughout the protein structure. COMMA is a fully automated tool with broad applicability. Then, we show the utility and capabilities of COMMA by applying it to three archetypal proteins, namely protein A, the tyrosine kinase KIT and the tumour suppressor p53. Our method permits to compare in a direct way the dynam-

ical behaviour either of proteins with different characteristics or of the same protein in different conditions. It is useful to identify residues playing a key role in protein allosteric regulation and to explain the effects of deleterious mutations in a mechanistic way.

Growth hormone in complex with its receptor We present a consensus analysis of dynamically correlated and coevolved residue networks of growth hormone (GH) in complex which its receptors (GHR). The mutants of GH-GHR are involved in human genetic diseases. We had access to a group of these mutants through the collaborations that we have with Serge Amselem (Service de Génétique et d'Embryologie Médicales, UMR S933 INSERM / UPMC, Hôpital Armand-Trousseau). We have examined the impact of two disease-related mutations on the allosteric communication of growth hormone. COMMA provided hints on how the mutation affects the dissociation of the hormone and its receptors and enables us to detect the key pathways on the structure of the wild-type and mutants. In addition the comparison of wild-type and mutant showed a rewiring of Communication Pathways linking coevolved residues. Characterizing the dynamical behavior of proteins provides a means for physical understanding of coevolution signals. Understanding the role of disease-related mutations on the link between coevolution and dynamical correlation can help decipher the molecular mechanisms of mutation-induced allosteric deregulation.

Mutational hotspot of PDZ domain We introduce that pertinent information can be extracted by COMMA from the structural dynamics of the wild-type PDZ3-CRIPT peptide complex to identify the highly deleterious positions with very high accuracy and provide a physical interpretation of their sensitivity to mutations. Moreover, we propose a protocol to predict the effects of specific amino acid substitutions and show that it enables to distinguish deleterious mutations from neutral and beneficial ones. We will show that, although the global shape from the molecular dynamics simulations of the wild-type complex and of the mutants are indistinguishable, COMMA is able to reveal significant differences in the communication between the protein residues. Consequently, we show that even in the absence of mutation-induced conformational changes, meaningful information is contained in and can be retrieved from the arrangement of residues in space and their atomic fluctuations. We may refer to this property of proteins dynamical behavior as "infostery", from 'info' - information - and 'steric' - arrangement of residues in space -. In addition, we propose an original approach to predict mutational effects based on protein sequence and define a score derived from sequence analysis and structural information to predict the phenotypic outcomes of the mutations. The predictive power of the score is equivalent to or higher than more sophisticated state-of-the-art methods for predicting mutational outcome.

Disordered proteins In the final chapter, we show that COMMA can detect protein regions that are prone to disorder or substantial conformational rearrangements, without requiring the input MD trajectory to actually sample the unfolded states of these regions. Our collaborators, Sonia Longhi (Université d'Aix-Marseille) and Denis Gerlier (Ecole Normale Supérieure de Lyon) performed significant project concerning disordered coiled-coils and we had access to their experimental data. Consequently, we were able to compare COMMA results with the experimental data. On the other hand, we discuss a

hypothesis to control the stability of coiled-coils and propose mutations that modulate the stability.

Chapter 3

Method

Contents

In this chapter, we present the methods that we proposed to solve the general questions, discussed in previous section, that we are trying to answer. We will introduce our proposed method to dissect the dynamical architecture of proteins. We will show how the study of both sequences and conformational changes can help us to explain the effects of deleterious mutations and to identify residues playing a key role in protein allosteric regulation in a mechanistic way. The proposed method is published (Karami et al., 2016) and is freely available to the community at `www.lcqb.upmc.fr/COMMA`.

3.1 COMmunication MApping (COMMA)

Characterizing a protein's motions and conformational changes can help predict the functional outcomes of mutations and the molecular mechanisms underlying diseases. Molecular dynamics (MD) simulations provide a way to probe the dynamical behavior of proteins in solution. However, it is sometimes difficult to determine what is important in the simulation and what property or measure should be recorded. The accumulation of MD data calls for the development of methods able to extract pertinent biological information and visualise it in a comprehensive way. Consequently, In recent years, methodological efforts have been made toward the definition of new measures to describe the dynamical architecture of proteins.

We are interested to develop a method that aims toward the extraction of pertinent information from the dynamics of proteins and integrate them in a systematic way. Therefore, the starting data are conformational ensembles, typically generated by MD conformations, but it could also be a set of X-ray or NMR structures. We propose a method that goes beyond the classical MD analysis, integrates the dynamical properties and defines the dynamical architecture of a protein by introducing communication modules. Our proposed method provides a measure to predict the effects of mutations at large scale, to identify the regions/residues important in a protein, to predict flexibility/disorder, to compare different proteins or state of a protein in a straightforward way and many other applications.

In this section, we discuss the proposed method to dissect the dynamical architecture of proteins. The present work builds up on previous efforts to propose a systematic dissection of protein architectures from a dynamical perspective. We provide Communication Mapping (COMMA), a method for analysing molecular dynamics-based communication in proteins and for mapping this information onto protein three-dimensional structures.

3.1.1 COMMA workflow

The workflow of the COMMA method is depicted on **Figure 3.1**. COMMA requires as input a conformational ensemble representing the protein of interest. Typically, the method is intended to analyse all-atom MD trajectories, but it is not restricted to this type of data. The analysis can also be performed on conformations obtained from another sampling method or on experimentally determined structures. The order of the input conformations does not influence the results. The ensemble can be divided into several sets, for example corresponding to several replicates of an MD simulation. COMMA algorithm proceeds as follows:

1. It analyses the conformational ensemble and extracts five residue-based dynamic properties: local dynamical correlations, minimum distances, communication propensities, non-covalent interaction strengths and secondary structures (box 1).

2. These properties are used to group residues into (*i*) independent cliques and (*ii*) communication pathways (boxes 2-3). Independent cliques are clusters of residues that display concerted atomic fluctuations while communication pathways are non-covalent chains of residues that move together (see below).

3. The information obtained from the independent cliques and the communication pathways is integrated in a graph, called Protein Communication Network (PCN) (box 4).

4. Connected components are extracted from this graph to define protein communication blocks (box 5).

5. The communication pathways that link different secondary structure elements are used to define communicating segment pairs and measure the strength of the interaction (box 6).

COMMA allows to visualise communication blocks and communicating segment pairs by mapping them onto the protein average conformation.

3.1.2 Extraction of dynamic properties

COMMA defines several measures that reflect the dynamic properties of the query protein. These measures are computed from each input set of conformations. Four measures are defined for pairs of residues and provide 4 distinct matrices. A fifth measure, which is new compared to MONETA, evaluates the likeliness of a residue to belong to a secondary structure.

Local dynamical correlations. Principal Component Analysis (PCA) is used to describe the atomic fluctuations of a protein through eigenvectors or modes. These modes are linear combinations of degrees of freedom. Starting from n PCA modes, describing the protein's essential dynamics (*i.e.* explaining 80% of the total atomic fluctuations), we apply a statistical technique called Local Feature Analysis (LFA) (Zhang and Wriggers, 2006). LFA computes residual correlations $Corr^{LFA}(i, j)$ between residues i an j as:

$$Corr^{LFA}(i, j) = \sum_{d=1}^{3} \sum_{r=1}^{n} \Psi_r(i_d)\Psi_r(j_d) \tag{3.1}$$

where d is the (x, y, z)-coordinate index of each Cα atom in a residue and Ψ_r is the PCA r^{th} eigenvector. The $Corr^{LFA}$ matrix is characterised by sparse correlation patterns (see on **Figure 3.1**). The LFA formalism identifies a set of n *seed* residues that are highly fluctuating and representative of these correlation patterns.

Minimum distances. The minimum distance d_{ij}^{min} between two residues i and j is defined as the smallest distance between any pair of atoms (a_i, a_j) belonging to residues i and j respectively, averaged over the set of conformations.

Communication propensities. We evaluate the communication propensity $CP(i, j)$ of residues i and j as the variance of the inter-residue distance (Chennubhotla and Bahar, 2007):

$$CP(i, j) =< (d_{ij} - \bar{d}_{ij})^2 > \tag{3.2}$$

where d_{ij} is the distance between the Cα atoms of residues i and j and $\bar{d}_{ij}$ is the mean value computed over the set of conformations. Intuitively, the smaller the variance, the

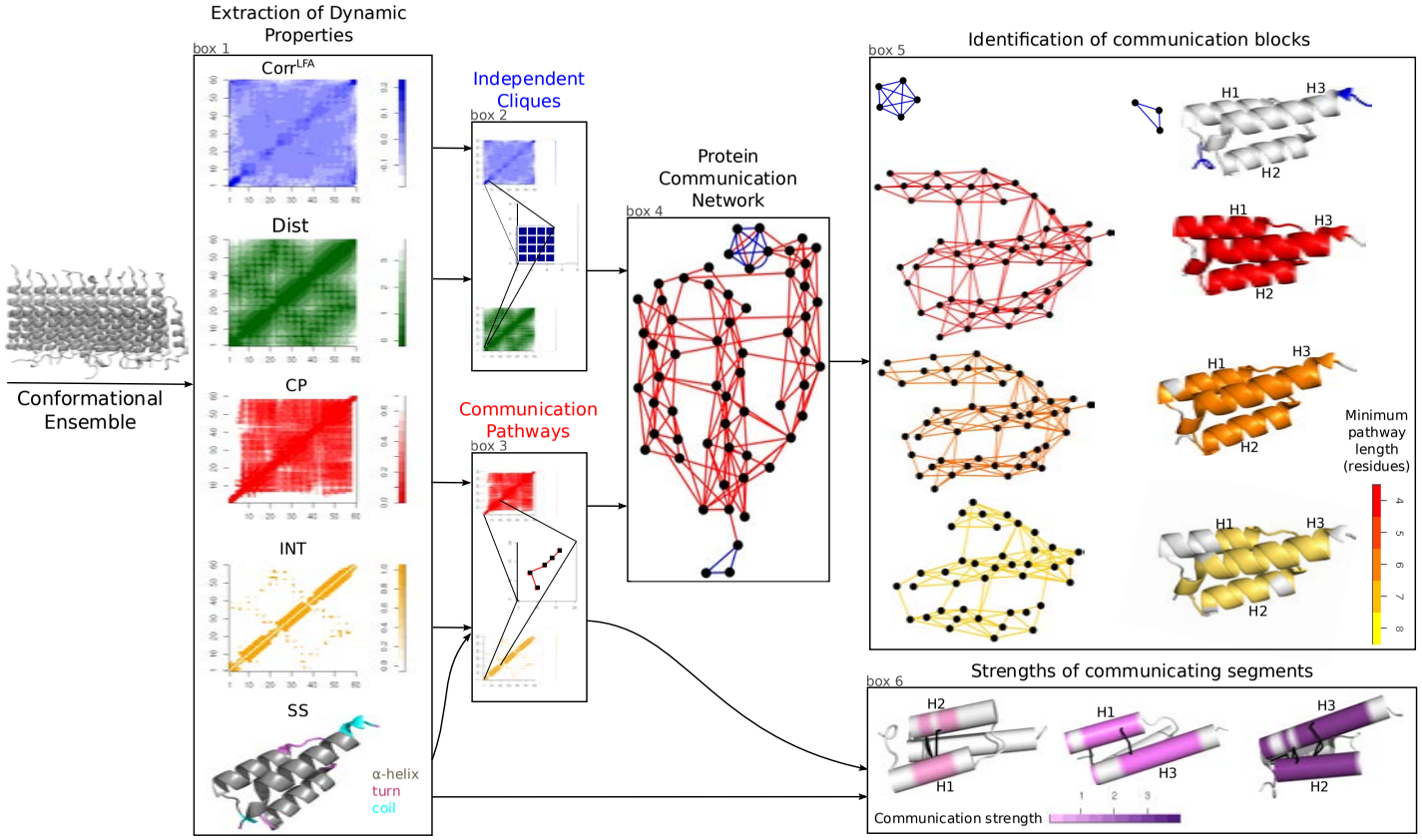

Figure 3.1: **Schematic representation of COMMA workflow.** Starting from one or several MD trajectories, COMMA computes matrices of residue-based dynamic properties: local dynamical correlations (Corr^{LFA}), minimum distances (Dist), communication propensities (CP), non-covalent interaction strengths (INT) and secondary structures (SS). Local dynamical correlations and minimum distances are used to identify independent cliques while communication propensities, non-covalent interaction strengths and secondary structures are used for communication pathway detection. A coloured graph, called Protein Communication Network (PCN), is constructed from independent cliques (blue edges) and communication pathways (red edges). The graph is analysed and two groups of communication blocks are extracted. The first group is made of CBs^{clique} (blue cliques in PCN), and the second group is made of CBs^{path} (subgraphs of the red PCN) where pathways have bounded length. In the schema, three different CBs^{path} are displayed, corresponding to a minimal path length of 4 (red), 8 (orange) and 9 (yellow) respectively. Communication pathways are also used to detect pairs of communicating segments, which are portions of secondary structure elements. Residues belonging to pathways that cross two secondary structures are coloured. For each pair of segments, the communication strength of the interaction is evaluated, on a scale of strengths going from low (pink) to strong (violet) strength. The segments and their interaction strength between H1-H2, H1-H3 and H2-H3 helices are shown.

more efficient the communication. Consequently, small values of $CP(i, j)$ are indicative of efficient signal transmission between residues i and j.

Non-covalent interaction strengths. We consider as non-covalent interactions hydrogen(H)-bonds and hydrophobic contacts, detected using the HBPLUS algorithm (McDonald and Thornton, 1994). H-bonds are detected between donor (D) and acceptor (A) atoms that satisfy the following geometric criteria: (*i*) maximum distances of 3.9Å for D-A and 2.5Å for H-A, (*ii*) minimum value of 90° for D-H-A, H-A-AA and D-A-AA angles, where AA is the acceptor antecedent. Hydrophobic contacts are identified with an inter-atomic distance lower than 3.9Å. The detected non-covalent interactions are then classified as backbone-backbone, backbone-side chain and side chain-side chain. For a given interaction type, an interaction strength matrix INT is computed, where each entry (i, j) describes the percentage of conformations in which at least one non-covalent interaction is formed between some pair of atoms (a_i, a_j) in residues i and j.

Secondary structures. Secondary structures are defined from the backbone torsion angles of the protein by using the DSSP algorithm (Kabsch and Sander, 1983). Three persistence values p_α, p_β and p_{turn} are computed for each residue. They reflect the percentage of conformations in which the residue is in a α-helix, a β-sheet or a turn, respectively. The secondary structure type that has the highest persistence value is assigned to the residue.

3.1.3 Identification of independent cliques and communication pathways

By combining the measures described above, COMMA identifies groups of residues that mediate communication across the protein structure, namely independent cliques and communications pathways. The computation is performed on each input set of conformations. These components are similar to the independent dynamic segments and communication pathways identified by MONETA. What is new in COMMA is the automated set up of pertinent values for the parameters depending on the system studied (see *Parameters*).

Independent cliques

It can happen that two seeds detected by LFA are very close in the sequence (distant by less than 6 residues). In that case, only the seed with the highest fluctuations is retained. The $Corr^{LFA}$ matrix is characterised by dense correlation patterns around every seed identified by LFA analysis. COMMA defines independent cliques as protein regions that correspond to these patterns. Each seed is extended into an independent clique S of residues by means of an extension algorithm that progressively adds residues in such a way that: (i) have a minimum distance smaller than 3.7Å and (ii) display concerted atomic fluctuations, indicated by high local dynamical correlations, that is the mean correlation value computed over S must be higher than a threshold (Laine et al., 2012):

$$\frac{1}{|S|}\sum_{i,j\epsilon S} Corr^{LFA}(i, j) \geq Corr^{LFA}_{cut} \tag{3.3}$$

The set up of $Corr^{LFA}_{cut}$ is explained below (see *Parameters*). The extension algorithm terminates when no more residue can be added. At the beginning of the iteration, S is made by the starting seed. We obtain $k \leq n$ independent cliques, where n is the initial number of seeds. Notice that the algorithm identifying the independent cliques uses information coming from the local dynamical correlation and the minimum distance matrices.

Communication pathways

Any two residues i and j are considered to communicate efficiently if their communication propensity is below a threshold, $CP(i, j) \leq CP_{cut}$. They form stable non-covalent interaction(s) if their interaction strength is higher than a threshold, $INT(i, j) \geq INT_{cut}$. The set up of the parameters CP_{cut} and INT_{cut} is explained below (see *Parameters*). Starting from a given residue, the algorithm implemented in COMMA generates a tree of paths that satisfies the following conditions (Laine et al., 2012): two consecutive residues in a path (*i*) are not adjacent in the sequence, (*ii*) form stable non-covalent interaction(s) and (*iii*) communicate efficiently. We ask that all residues in a path communicate efficiently with each other by transitivity. Notice that the algorithm identifying the pathway-based edges uses the communication propensity and the interaction strength matrices, and also the secondary structure information, that plays a role for the set up of CP_{cut} (see *Parameters*).

3.1.4 Construction of a protein communication network

Independent cliques and communication pathways are used to construct a Protein Communication Network (PCN) that reflects the way information is transmitted across the protein 3D structure. A PCN(N,E) is a coloured graph defined by nodes N that correspond to the residues of the protein and edges E that connect dynamically correlated residues. Two types of edges are constructed:

1. **Clique-based edges:** two vertices representing residues i and j are connected by a clique-based edge if they belong to the same independent clique and if $Corr^{LFA}(i, j) \geq Corr^{LFA}_{cut}$.

2. **Pathway-based edges:** two vertices representing residues i and j are connected by a pathway-based edge if they are consecutive in some communication pathway.

The PCN is constructed by considering the union of all independent cliques and all communication pathways detected from every input set of conformations. Let us stress that MONETA 2.0 (Allain et al., 2014) also provides a graph representing the protein, but it uses communication pathways and covalent bonds to construct it and the criteria employed are markedly different from those employed by COMMA to construct the PCN.

3.1.5 Extraction of communication blocks and communicating segment pairs

COMMA final outputs consist in dynamics-based decompositions of the query protein 3D structure. Two types of decompositions are produced. The protein is divided into: (*i*) communication blocks defined from the PCN, (*ii*) communicating segment pairs defined

from secondary structure elements and communication pathways. These two notions are completely new compared to MONETA.

Communication blocks

Connected components in an undirected graph are isolated subgraphs. COMMA extracts connected components from the constructed PCN by using depth-first search (DFS) and defines protein communication blocks. Different types of communication blocks are defined, namely CBs^{clique} and CBs^{path}. CBs^{clique} are directly extracted by considering all clique-based edges. Different kinds of CBs^{path} are defined, either by considering all but very short (≤ 3 residues) pathways, or by considering pathways longer than a fixed number of residues. An interesting threshold is given by MPL_{cut} as defined below (see *Parameters*).

Communicating segment pairs

COMMA detects pairs of protein segments that are part of secondary structure elements (SSEs) and that are linked by communication pathways. A SSE is constituted by residues (at least three) that adopt the same secondary structure type. First the algorithm identifies all SSEs contained in the protein structure. Then, it computes, for each pair (A,B) of SSEs: (i) the proportion PR_{AB} (resp. PR_{BA}) of residues from A (resp. B) that are linked by at least a communication path to some residue from B (resp. A), (ii) the number of pairs of residues (i^A, j^B) of A and B that are consecutive in a communication path, $Cont_{AB}$. The residues of A and B that are linked by at least a communication path constitute a communicating segment pair. The communication strength between the two segments defined from A and B is calculated as:

$$S_{AB} = PR_{AB} * PR_{BA} * Cont_{AB} \tag{3.4}$$

3.1.6 Visualisation

COMMA is interfaced with PyMoL (DeLano, 2002) to permit the visualisation of the communication blocks and the communicating segment pairs by mapping them on the protein average conformation. COMMA produces PyMoL files (.pml extension) that enable the following representations:

- **Communication blocks:** the residues involved in communication blocks are coloured accordingly. Residues that are not detected in a communication block are coloured in white. Non-covalent interactions between blocks are shown as thick black lines.
- **Communicating segment pairs:** given a pair of SSEs, the residues involved in the communicating segments in these SSEs are highlighted in colours. Pathways-based edges linking residues in the two segments are shown as thick black lines.

3.1.7 Parameters

COMMA uses several parameters and allows the user to tune them depending on the question asked and on the system studied. However, to allow for a large-scale application

of the method, we have implemented automated procedures to set up default values for all parameters.

$Corr^{LFA}_{cut}$. We define the LFA correlation threshold $Corr^{LFA}_{cut}$ to delimit protein regions of concerted atomic fluctuations. $Corr^{LFA}_{cut}$ is chosen such that 5% of the values in the $Corr^{LFA}$ matrix are higher than $Corr^{LFA}_{cut}$ (**Figure 3.2A**).

CP_{cut}. We define a cutoff CP_{cut} to determine whether the communication between two residues is efficient. The strategy employed to set the value of CP_{cut} is inspired from (Dixit and Verkhivker, 2011). Intuitively, neighbouring residues in the sequence forming well-defined secondary structures are expected to communicate efficiently with each other. First, we evaluate the proportion p_{ss} of residues that are in an α-helix, a β-sheet or a turn in more than half of the conformations. Then for every residue i, we compute a modified communication propensity $MCP(i)$ as:

$$MCP(i) = \frac{1}{8} \sum_{\substack{j=i-4 \\ j\neq i; 1\leq j\leq N}}^{i+4} CP(i,j) \tag{3.5}$$

where N is the total number of residues. CP_{cut} is chosen such that the proportion p_{ss} of MCP values are lower than CP_{cut} (**Figure 3.2B**). Any two residues i and j for which $CP(i,j) < CP_{cut}$ are considered to communicate efficiently.

INT_{cut}. We define a threshold value INT_{cut} to filter out non-covalent interactions that are not relevant. For this, an adjacency graph is constructed from the INT matrix by considering different cutoff values, ranging from 0.25 to 1, by increments of 0.05, and the size of the largest connected component is computed (**Figure 3.2A**). INT_{cut} is the largest interaction strength for which the size of the largest component is maximal (Brinda and Vishveshwara, 2005) (**Figure 3.2C**).

MPL_{cut}. We define a threshold MPL_{cut} to discriminate between short and long paths. For this, connected components are extracted from subgraphs of the PCN. The subgraphs are defined by considering pathway-based edges that are derived from communication pathways comprising at least n residues, n ranging from 4 to 8. MPL_{cut} is chosen as the minimum path length for which we observe the largest reduction of the size of the largest connected component (**Figure 3.2D**).

3.1.8 Related tools

As noted in the introduction (*Chapter 1, Introduction*), a number of previously developed methods are proposed to analyse the dynamical conformation of proteins and their inter-residue communication. These tools however typically consider only dynamical correlations or/and non-covalent interactions, whereas COMMA combines four different dynamical properties in a unified framework (**Table 3.1**). Moreover COMMA describes communication at different levels, from individual residues to the whole dynamical architecture of the protein. In particular, the identification of communicating pairs of secondary

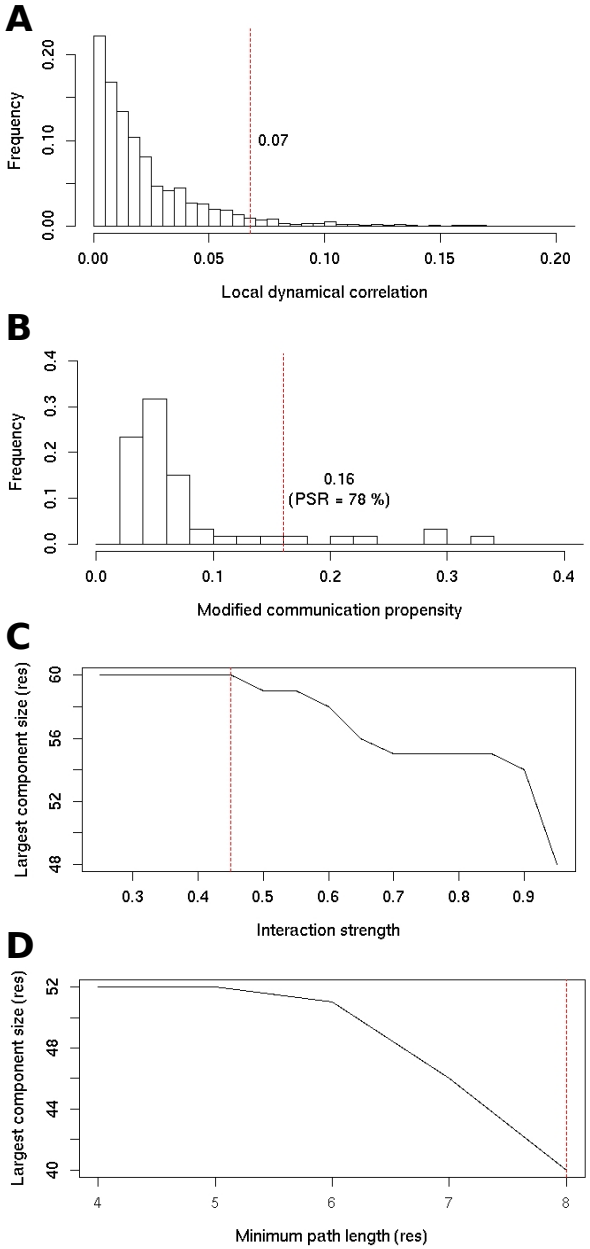

Figure 3.2: **Parameters for Protein A. (A)** Distribution of the local dynamical correlation ($Corr_{LFA}$) values. **(B)** Distribution of the communication propensity (CP) values. **(C)** Size of the largest connected component (in residues) extracted from the adjacency graph constructed based on non-covalent interaction strengths. **(D)** Size of the largest connected component (in residues) extracted from the PCN by considering communication pathways with different minimum lengths.

structure elements is a unique feature of our method (**Table 3.1**). Finally, COMMA, which uses MDTraj Python package (McGibbon et al., 2014), does not depend on a particular MD package and can handle most popular formats used in the protein structural dynamics community.

	COMMA	Bio3D (Skjærven et al., 2014)	GSATools (Schrank et al., 2009)	PyInteraph (Tiberti et al., 2014)	PSN-ENM (Raimondi et al., 2013)	Taylor *et al.* (McClendon et al., 2014)
Software availability	✓	✓	✓	✓	-	-
Open source	✓	✓	✓	✓	-	-
Dependencies	MDtraj, Eigen and Numpy python packages	R, Muscle	GNU, Scientific Library, GROMACS	Python, Pymol	-	-
Programming language	C++, Python	R	C	Python	-	-
Input trajectory formats	AMBER, GROMACS, NAMD, CHARMM...	GROMACS (.dcd)	GROMACS	AMBER, GROMACS, NAMD, CHARMM...	-	-
Dynamical properties:						
non-covalent interactions	✓	-	-	✓	✓	✓
inter-residue distances	✓	-	-	-	-	✓
secondary structures	✓	-	-	-	-	-
dynamical correlations	✓(PCA, LFA, CP)	✓(ENM-NMA, PCA)	✓(between frames)	-	✓(ENM-NMA)	✓(MI)
Description levels:						
residue	✓	-	✓	✓	✓	✓
secondary structure	✓	-	-	-	-	-
region/domain	✓	✓	✓	-	-	✓
protein	✓	✓	✓	✓	✓	✓
Outputs:						
protein network	✓	-	✓	✓	✓	✓
communicating regions	✓(pathway- and CBs^{clique})	✓(dynamic domain, correlation network)	✓(functional fragments)	-	-	✓(communities)
communicating segment pairs	✓	-	-	-	-	-
functional domains	-	-	✓	-	-	✓
pathways	✓	-	✓	✓	✓	✓

Table 3.1: **Comparison between different methods to analyse the dynamical behaviour of proteins and their inter-residue communication.** The technical characteristics and functionalities of COMMA and of five state-of-the-art methods are reported. The PSN-ENM method (Raimondi et al., 2013) and the method proposed by Taylor *et al.* (McClendon et al., 2014) are not implemented as software.

3.2 Application of COMMA on three archetypal proteins

Here, we have applied COMMA on three following case studies to illustrate its capabilities.

3.2.1 Molecular dynamics simulations

We applied the COMMA method to three archetypal proteins: (*i*) the B domain of staphylococcal protein A [PDB:1BDD] (residues 1-60, NMR), a highly stable protein, (*ii*) the DNA-binding domain of the human tumour suppressor protein p53 [PDB:2XWR] (chain A, residues 89-293, 1.68Å resolution), a highly flexible protein, (*iii*) the cytoplasmic region of the receptor tyrosine kinase KIT [PDB:1T45] (residues 547-935, 1.90Å resolution), an allosterically regulated protein. The following molecular dynamics protocol was applied to all studied systems. More details on the MD trajectories of the wild-type KIT and its oncogenic mutant D816V can be found in (Laine et al., 2011a).

Set up of the systems The 3D coordinates for the studied proteins were retrieved from the Protein Data Bank (PDB) (Berman et al., 2000). All crystallographic water molecules and other non-protein molecules were removed. The structure of the DNA-binding domain of P53 contains a bound zinc ion. At physiological temperature, Zn^{2+} rapidly dissociates from the protein and the resulting Zn^{2+}-free P53 is folded and stable (Butler and Loh, 2007, 2003). Consequently, we removed the zinc ion from the initial PDB structure and simulated P53 in the apo form. The mutated form of KIT was generated by *in silico* substitution of the aspartate (D) in position 816 into a valine (V) using MODELLER 9v7 (Marti-Renom et al., 2000). All models were prepared using the LEAP module of AMBER 12 (Case et al., 2012), with the ff12SB forcefield parameter set: (*i*) hydrogen atoms were added, (*ii*) Na^+ or Cl^- counter-ions were added to neutralise the systems charge, (*iii*) the solute was hydrated with a cuboid box of explicit TIP3P water molecules with a buffering distance up to 10Å. The environment of the histidines was manually checked and they were consequently protonated with a hydrogen at the ϵ nitrogen. The details of structure preparation and solvent models are given in **Table 3.2**.

Minimisation, heating and equilibration The systems were minimised, thermalised and equilibrated using the SANDER module of AMBER 12. The following minimisation procedure was applied: (*i*) 10,000 steps of minimisation of the water molecules keeping protein atoms fixed, (*ii*) 10,000 steps of minimisation keeping only protein backbone fixed to allow protein side chains to relax, (*iii*) 10,000 steps of minimisation without any constraint on the system. Heating of the system to the target temperature of 310 K was performed at constant volume using the Berendsen thermostat (Berendsen et al., 1984) and while restraining the solute C_α atoms with a force constant of 10 kcal/mol/Å^2. Thereafter, the system was equilibrated for 100 ps at constant volume (NVT) and for further 100 ps using a Langevin piston (NPT) (Loncharich et al., 1992) to maintain the pressure. Finally the restraints were removed and the system was equilibrated for a final 100-ps run. Backbone deviations obtained after equilibration are smaller than 1.3 Å (**Table 3.2**).

	Protein A	P53	KIT WT	KIT MU
Total charge of counter-ions	+2	-3	+1	0
Water box dimensions (Å^3)	75x49x49	76x73x64	77x73x81	77x73x81
Number of water molecules	4 174	8 215	13 195	13 197
Total number of atoms	13 477	27 755	44 870	44 879
Deviation after equilibration (Å)	0.83	0.65	1.29	1.13

Table 3.2: **MD preparation and equilibration details.** The counter-ions employed to neutralize the systems are Na^+ and Cl^-. Root mean square deviations were computed on the backbone atoms of the equilibrated conformations versus the initial template.

Production of the trajectories For every protein, 2 replicates of 50 ns, with different initial velocities, were performed in the NPT ensemble using the PMEMD module of AMBER 12. The temperature was kept at 310 K and pressure at 1 bar using the Langevin piston coupling algorithm. The SHAKE algorithm was used to freeze bonds involving hydrogen atoms, allowing for an integration time step of 2.0 fs. The Particle Mesh Ewald method (PME) (Darden et al., 1993) was employed to treat long-range electrostatics. The coordinates of the system were written every ps. Standard analyses of the MD trajectories were performed with the *ptraj* module of AMBER 12.

Stability of the trajectories The simulations of wild-type and mutated KIT were previously shown to have good stability (Laine et al., 2011a). To assess the stability of the B domain of protein A and of the DNA-binding domain of p53, the Cα atoms root mean square deviation (RMSD) from the equilibrated structure, the stability of secondary structures and the radius of gyration were recorded along each 50-ns MD simulation replicate (**Figure 3.3 and Figure 3.4**). The B domain of protein A deviates by no more than 2.2Å (**Figure 3.3A**) from the equilibrated structure and has an average radius of gyration of 10.5±0.1Å (**Figure 3.3D**). p53 DNA-binding domain displays RMSD values in the range 1.5-3.0Å (**Figure 3.4A**) and its radius of gyration values 16.6±0.1Å (**Figure 3.4D**). Secondary structure profiles are highly stable for both replicates of both proteins (**Figure 3.3B-C and Figure 3.4B-C**). Overall, the evolution of RMSD, secondary structure and radius of gyration shows that protein A and p53 are stable over the 50-ns runs. The systems are fully relaxed after 20 ns (**Figure 3.3A and Figure 3.4A**). Consequently, COMMA was applied on the last 30 ns of every replicate. COMMA input sets for the three study cases are made of 30,000 conformations.

Convergence of the trajectories To evaluate the convergence of the dynamic properties extracted by COMMA, a convergence analysis (Lyman and Zuckerman, 2006) was applied to the MD trajectories of the studied systems. The analysis comprises two steps: (*i*) a set of reference conformations are identified, (*ii*) all MD conformations from the trajectory are clustered into corresponding reference groups. Each reference conformation is first picked up randomly and the conformations distant by less than an arbitrary cutoff *r* are binned with it. Then the trajectory is split in two halves and conformations from each half are grouped based on their RMSD from each reference conformation. If the

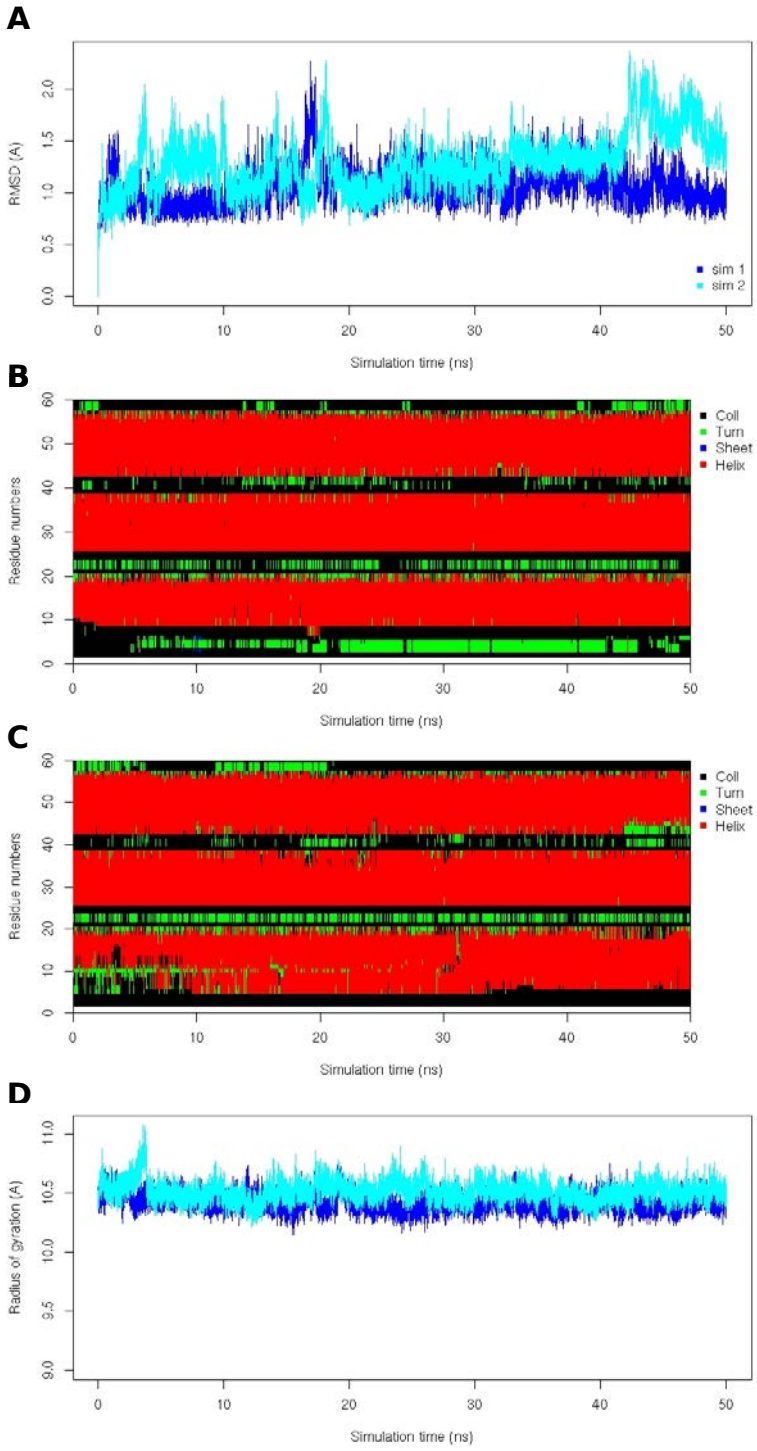

Figure 3.3: Stability analysis of the B domain of Protein A over 50 ns of MD simulations. A) RMS deviations from the equilibrated structure, computed on the Cα atoms. The first and second MD replicates are in blue and cyan. B,C) Secondary structures recorded over simulation time for the first (B) and second (C) MD replicates. D) Radius of gyration.

simulation has converged, then each reference cluster should be populated equally from both halves of the trajectory.

The RMSD was computed on the Cα atoms and the cutoff r was empirically chosen

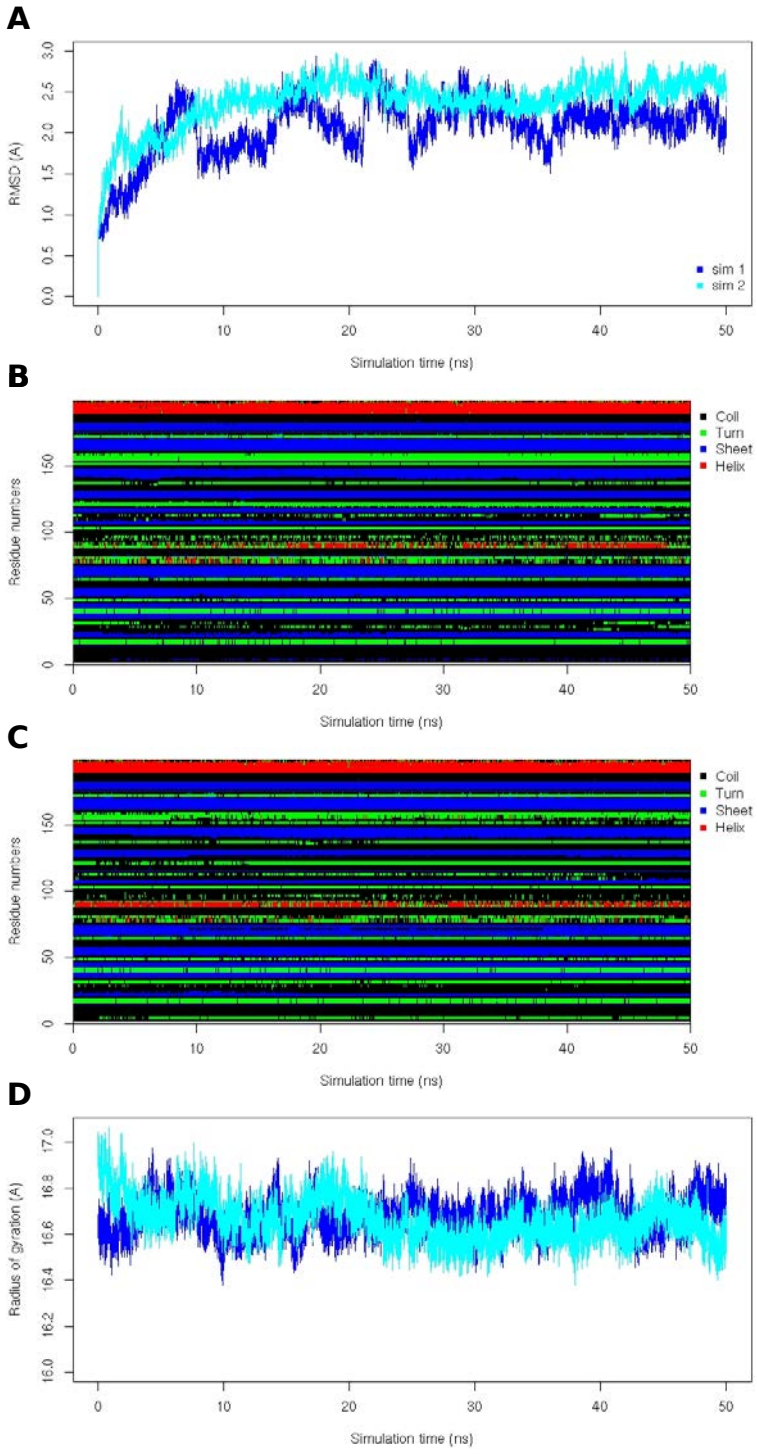

Figure 3.4: Stability analysis of P53 DNA-binding domain over 50 ns of MD simulations. A) RMS deviations from the equilibrated structure, computed on the Cα atoms. The first and second MD replicates are in blue and cyan. B,C) Secondary structures recorded over simulation time for the first (B) and second (C) MD replicates. D) Radius of gyration.

so as to get a reasonable number of representative MD conformations, typically between 2 and 7. To reduce the bias resulting from the random choices of the references, the process was repeated 5 times for each analyzed trajectory. The convergence quality of

each simulation was measured using a convergence criterion c defined as (Chauvot de Beauchene et al., 2014):

$$c = 1 - \left(\frac{1}{5}\sum_{k=1}^{5}\frac{\#(\text{lone reference conformations})}{\#(\text{reference conformations})}\right) \tag{3.6}$$

A lone reference conformation is a reference conformation that is not visited in one half of the trajectory (less that 1% of the frames in the corresponding reference group). The convergence criterion c is comprised between 0 and 1; a value of 1 corresponds to an optimal convergence. All trajectories show good to very good convergence, with values of c ranging between 0.6 and 0.9 (**Table 3.3**). This indicates that the conformational sampling furnished by the last 30 ns of each productive MD run is sufficient to apply COMMA.

	Protein A		P53		KIT WT		KIT MU	
Replicate	1	2	1	2	1	2	1	2
Cutoff (Å)	1.5	2.2	2.3	2.3	2.5	2.5	2.5	2.5
# reference conformations	2-3	2-6	2-4	2-5	4-7	5-7	2-4	5-6
Convergence criterion c	0.9	0.8	0.9	0.9	0.6	0.4	0.7	0.6

Table 3.3: **Convergence analysis of the two MD replicates of each studied system.** The analysis was applied 5 times on the last 30 ns of every productive run. The convergence criterion c was calculated as described in Methods.

3.2.2 Communication blocks in KIT protein and its oncogenic mutant

Presentation of KIT

KIT is a receptor tyrosine kinase of type III implicated in signalling pathways crucial for cell growth, differentiation and survival (Lemmon and Schlessinger, 2010; Edling and Hallberg, 2007; Qiu et al., 1988). The mutation of the aspartate located in position 816 to a valine leads to the constitutive activation of the receptor and is associated to mastocytoses and gastrointestinal stromal tumours (Orfao et al., 2007; Miettinen M, 2002). It was shown experimentally that the mutation induces long-range effects that lead to a shift in the conformational equilibrium of the kinase away from the auto-inhibited state, resulting in a 536-fold increased activation rate (Gajiwala et al., 2009). COMMA was applied to the cytoplasmic region of KIT (331 residues), starting from 2 replicates of 50-ns MD simulations of the wild-type and D816V-mutated proteins (Laine et al., 2011a) (see *Methods*). The method identified 11 (resp. 9) communication blocks in the wild type (resp. mutant) (Table 3.4). These blocks reflect the way information is transmitted across the protein structure (see *Methods*). They were mapped onto the average MD conformations of the wild-type and mutated proteins for visualisation (**Figure 3.5A**). They were also used to derive schematic representations of the two proteins (**Figure 3.5B**).

Wild type													
name	A	B	C	D	E	F	G	H	I	J	K	-	-
size (res.)	22	11	32	16	14	13	11	127	160	9	4	-	-
Mutant													
name	-	B'	C'	D'	-	-	G'	H'	I'	J'	-	L'	M'
size (res.)	-	12	20	18	-	-	10	86	186	8	-	35	66
Overlap (%)	-	96	65	76	-	-	95	80	87	71	-	-	-

Table 3.4: **Mapping of communication blocks between wild-type KIT and the D816V mutant.** The overlap o_{ij} between two blocks B_i and B_j, identified in the wild type and in the mutant, is evaluated as: $o_{ij} = 2 * \#(B_i \cap B_j)/(\#(B_i) + \#(B_j))$. Two blocks are defined as counterparts, namely X and X' if: (*i*) X' (resp. X) yields the maximum overlap with X (rest. X') over all blocks in the mutant (resp. wild-type) protein; (*ii*) the overlap is greater than 60%.

Decomposition of KIT dynamical architecture

KIT communication blocks can be classified according to the structural and dynamical information used to identify them. In the wild type (**Figure 3.5A-B**, on top), blocks A to G (in blue tones) were obtained from independent cliques (see *Methods*). These blocks represent protein regions whose internal dynamics are independent from each other and from the rest of the protein. Blocks H (in red), I (in green), J (in lime green) and K (in dark green) were obtained from communication pathways, *i.e.* chains of dynamically correlated residues stabilised by non-covalent interactions (see *Methods*). Blocks I, J and K were identified by considering all but very short paths while block H comprises only long paths ($\geq$ 6 residues).

Different types of connections are established between blocks (**Figure 3.5A-B**), namely, from the strongest to the weakest: (*a*) inclusion, *e.g.* block H is included in block I, (*b*) overlap, *e.g.* blocks D and I share some residues in common, (*c*) contact, *e.g.* some residues from blocks B and I are adjacent in the sequence, (*d*) interaction, *e.g.* some residues in blocks A and C form a stable H-bond or hydrophobic contact. We observed that two blocks that share residues or contact each other (types *a, b, c*) are also connected by non-covalent interactions (type *d*).

The architecture of KIT is composed of a core of long-range communicating residues forming block H, that represents more than one third of the protein (**Table 3.4**). This core spans the two lobes of the protein and covers most of the enzymatic site (**Figure 3.5A-B**, on top). It is extended by a layer of short-range communicating residues contained in block K and is connected to several much smaller blocks. These small blocks establish few connections between them. However an interconnected set of small blocks (A, C, and J) can be detected, that is constituted by residues from the N-terminal lobe and represents about 20% of the protein.

Comparison of wild-type and mutated KIT

The communication blocks identified by COMMA in wild-type and mutated KIT were compared. The pairs of blocks from the two proteins that are constituted in large part

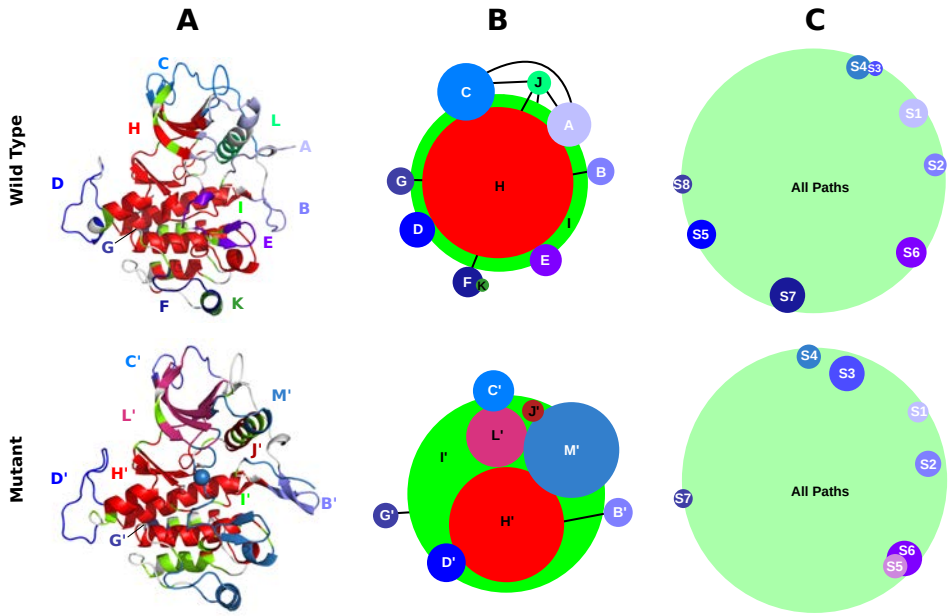

Figure 3.5: **Dynamical architecture of wild-type KIT and the D816V mutant. On top.** Wild-type protein. **At the bottom.** Mutant protein. **On the left.** The communication blocks identified by COMMA are mapped onto the average conformation represented as a cartoon. The mutation site is represented by a sphere (at the bottom). The protein residues are coloured according to the block they belong to and the different blocks are labelled. See Table 3.4 for details on the mapping between the two proteins. **In the middle.** Schematic representations of the proteins depicting the communication blocks identified by COMMA and the connections between them. Each block is represented by a round and is labelled. The larger the number of residues in the block, the larger the size of the round. Overlapping blocks share some residues in common. Contacting blocks are connected by covalent bonds. The black links indicate the presence of stable non-covalent interactions between blocks. Notice that non-covalent interactions are formed between blocks that share some residues in common or contact each other, but they are not displayed for a sake of clarity. **On the right.** Schematic representations of the proteins depicting the results obtained from MONETA. The large round in green include all residues involved in some communication pathway. The smaller blocks in blue tones represent independent dynamic segments. The size of the round depends on the number of residues involved (same scaling as for COMMA results).

by the same residues were identified (**Table 3.4**). Overall, the composition of the blocks and their connections can vary substantially upon mutation (**Figure 3.5B**). Specifically, block M' (in sky blue) of the mutant comprises most of the residues constituting blocks A, E and F in the wild type. Let us stress that the mutational position 816 is located in block E of the wild type protein and in block M' of the mutant (indicated as a sphere on **Figure 3.5A**, at the bottom). Interestingly, the protein regions comprised in block M' were recently highlighted as forming an allosteric network in Src kinase (Foda et al., 2015). In addition to these changes, COMMA detected three long-range communication blocks in the mutant (in red tones) instead of one in the wild type. Block H' (in red) is 1.5 times smaller than block H. Some residues from the N-lobe that were included in block H now form the disjoint block L' (in raspberry). The residues forming block J' (in firebrick) communicate at longer range than the residues forming block J in the wild type. These three blocks H', J' and L' are included in block I', which is slightly bigger than

I. Consequently, the mutation induces a complete reshaping of communication blocks in KIT, characterised by a reorganisation of the hierarchy between long-range and short-range communicating residues and the merge of three CBs^{clique}.

Comparison with other classifications

The definition of KIT communication blocks provided by COMMA can be compared with the definition of KIT regulatory regions reported in the literature (Jr., 2005; Griffith et al., 2004; Nolen et al., 2004; Huse and Kuriyan, 2002). Blocks B, C, D, E, F and L partially match the JM-Switch (JMS), the JM-Zipper (JMZ), the kinase insert domain (KID), the A(ctivation)-loop, the substrate-binding platform (helix G) and the C-helix respectively (**Figure 3.6A**). Block A contains the JM-Proximal (JMP) and the glycine-rich loop (P-loop). The blocks can also be evaluated based on the flexibility profile of the residues they contain. CBs^{path} tend to contain rather rigid residues while CBs^{clique} are highly flexible (**Figure 3.6B**). From a secondary structure perspective, residues in CBs^{path} tend to form stable secondary structures whereas residues in CBs^{clique} are in solvent-exposed loops (**Figure 3.6C**). We observed that these trends are general among the proteins we studied. These observations show that the identification of communication blocks by COMMA correlates positively with protein residue classifications based on the literature, on rigidity/flexibility or on secondary structures. Furthermore, COMMA enables to go beyond such classifications by providing a more precise dissection of the protein's dynamical architecture.

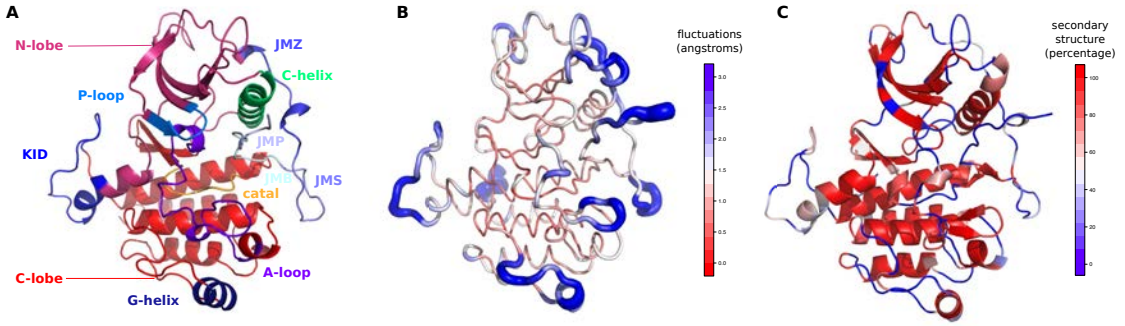

Figure 3.6: Classifications of KIT residues based on the literature, the rigidity/flexibility and the secondary structures. The crystallographic structure 1T45 (**A**) or the MD average conformation (**B-C**) of KIT cytoplasmic region is represented as a cartoon. (**A**) KIT regulatory regions defined in the literature are highlighted by colours: the JM-Proximal (JMP, residues 547-552) in light blue, the JM-Binder (JMB, residues 553-559) in light cyan, the JM-Switch (JMS, residues 560-570) in slate, the JM-Zipper (JMZ, residues 571-581) in sky blue, the glycine-rich loop (P-loop, residues 596-601) in marine, the C-helix (residues 631-647) in lime green, the catalytic loop (catal, residues 790-797) in orange, the activation loop (A-loop, residues 810-835) in purple and the helix G (residues 877-885) in dark blue. The N-terminal lobe is coloured in raspberry, the kinase insert domain (KID) in blue and the C-terminal lobe in red. (**B**) The per-residue atomic fluctuations (in Å) computed over the MD conformational ensemble are indicated by colours ranging from red (low fluctuations) through white to blue (high fluctuations) and by the size of the tube (the larger the tube the higher the fluctuations). (**C**) The percentages of conformations in which every residue is in a secondary structure element (either α-helix, β-sheet or turn) are indicated by colours ranging from blue (low percentage) through white to blue (high percentage).

Comparison with MONETA

COMMA results were compared to those obtained with MONETA 2.0 (**Figure 3.5C**). MONETA identifies independent dynamic segments and communication pathways from all-atom MD simulations (Allain et al., 2014), which are similar to the independent cliques and communication pathways identified by COMMA (**Figure 3.1**, boxes 2 and 3). However, COMMA exploits these components for further analysis (**Figure 3.1**, boxes 4, 5 and 6) in a way that is completely different from MONETA (Allain et al., 2014). **Figure 3.5C** depicts schematic representations of the dynamic segments and communication pathways detected by MONETA in KIT. The green round corresponds to the ensemble of residues involved in some path (representing 90% of the protein). The rounds in blue tones represent dynamic segments. These components are substantially different from the communication blocks identified by COMMA (**Figure 3.5B**) and MONETA does not characterise the connections between them. From this comparison, it is clear that COMMA brings additional information on the definition and arrangement of the protein's dynamical architecture building blocks, compared to MONETA.

MONETA previously permitted to put in evidence a crucial communication pathway in wild-type KIT that links the A-loop and the JMS through residue D792 from the catalytic loop (Laine et al., 2012). The path was disrupted upon D816V mutation. In COMMA representation of wild-type KIT (**Figure 3.5**, on top), all residues participating in this path are contained in the long-pathway based block H (in red), from D792 in the catalytic loop to V559 in the JMS. By contrast, in the mutant (**Figure 3.5**, at the bottom), D792 is contained in the CB^{path} I' (in green) but not in block H' (in red), indicating that this residue is involved in shorter communication pathways compared to the wild type, and that no pathway goes from D792 to the JMS. COMMA results are thus in agreement with those obtained by using MONETA. Moreover, by identifying communication blocks, COMMA enables to pinpoint other long pathways that are interrupted in the mutant. Specifically, the fact that the long-CB^{path} H in the wild type is divided in H' and L' in the mutant is associated to a disruption of the communication between residue N655 and residues I653, H651 and K807. Interestingly, these residues were shown to form a network of interactions (called 'molecular brake') crucial for the stability of the inactive conformation of tyrosine kinases (Chen et al., 2007). Consequently, COMMA analysis permits to put in evidence a deleterious effect of the activating D816V mutation on this 'molecular brake' which was not previously detected.

3.2.3 Communicating segment pairs in Protein A

Protein A

The B domain of protein A (BdpA) from *Staphylococcus aureus* is a small α-helical protein. It comprises 60 residues arranged in three helices, namely H1 (residues 10-19), H2 (residues 25-37) and H3 (residues 42-56), linked by two turns, namely T1 (residues 20-24) and T2 (residues 38-41). The fast-folding kinetics of protein A have been extensively characterised through experiments and computer simulations (Lei et al., 2008; Sato et al., 2006, 2004; Vu et al., 2004; Bai et al., 1997), enabling to establish the following statements: (*i*) the isolated H3 has a higher stability and helical content compared to the two other helices, (*ii*) H2 and H3 form a stable or marginally stable intermediate, (*iii*) H1 is

docked in the rate limiting step.

Dynamical architecture of Protein A

COMMA was used to identify communicating segment pairs in BdpA (60 residues). For this, we performed 2 replicates of 50-ns MD simulations, starting from an average nuclear magnetic resonance (NMR) structure (see *Methods*). By analysing the MD trajectories, COMMA detected five stable secondary structure elements (SSEs) in the protein: three α-helices formed by residues 5-18, 25-37 and 39-55 and two turns formed by residues 2-4 and 56-59. We focus here on the three α-helices, which match well the experimentally-defined helices H1, H2 and H3. Three pairs of communicating segments were identified between H1/H2, H1/H3 and H2/H3 (**Figure 3.1**, box 6). The communication strengths (computed as the product of the proportions of residues involved in communication pathways linking the two segments multiplied by the number of pairs of residues directly linked by a pathway, see *Methods*) for these pairs are 0.5, 1.1 and 4.1 respectively. The significantly higher strength of the segment pair corresponding to H2/H3 is the result of a larger number of residues involved in the communication and a larger number of direct links (5 versus 2 and 3, shown as black lines on **Figure 3.1**, box 6). Let us remind that a direct link is a pair of residues from the two communicating segments that are consecutive in a communication path (see *Methods*). Moreover, one can observe that the communicating segments of H1 cover a significantly smaller portion of the helix compared to the segments of H2 and H3. The communication blocks identified in protein A also show that the residues of H1 are involved in shorter paths compared to H2 and H3 (**Figure 3.1**, box 5). These observations are in agreement with the experimental evidence that H1 docks to a stable assembly of H2 and H3 during the folding process. Let us stress that this result could not be obtained by simply analysing non-covalent interactions along the MD trajectories: there are 8, 4 and 8 interactions for the H1/H2, H1/H3 and H2/H3 pairs. This emphasises the importance of the notions of communication propensity and communication pathways in our analysis.

3.2.4 The role of pathway length and interaction type in p53 communication

Presentation of P53

The tumour suppressor p53 is a transcription factor regulating a wide range of genes involved in DNA repair, apoptosis, senescence and metabolism (Li et al., 2012; Vousden and Prives, 2009; Vogelstein et al., 2000). The p53 protein plays a crucial role in conserving the stability of the genome and preventing genomic mutation (Strachan T, 1999). The loss of p53 tumour suppressor function is associated with cancer (Lu et al., 2009). The sequence of p53 can be divided into an N-terminal transactivation domain, a DNA-binding core domain (DBD), a tetramerisation domain and a C-terminal regulatory domain (Okorokov and Orlova, 2009). The DBD is intrinsically unstable and thus highly susceptible to oncogenic mutations (Canadillas et al., 2006). The three-dimensional structure of the DBD comprises two antiparallel β-sheets, characteristic of the immunoglobulin-like β-sandwich fold (**Figure 3.7A**, topology diagram on the left). In total, it contains 11 β-strands and 2 α-helices linked by flexible loops (**Figure 3.7A**, see labels on the right). The

dynamical architecture of p53 DBD (199 residues) was characterised by COMMA, starting from 2 replicates of 50-ns MD simulations (see *Methods*). We investigated the evolution of the pathway-based communication blocks identified by COMMA when varying the minimum length of the pathways considered and the type of non-covalent interactions used to construct them (**Figure 3.7**).

Hierarchical description of p53 communication

The ensemble of all but very short ($\leq$ 3 residues) communication pathways identified in p53 yielded one communication block (**Figure 3.7B**, in red), representing about 50% of the protein residues. This block comprises the 11 β-strands of the protein, some residues from the loops that frame them and a portion of the helix H2. The edges of the corresponding subgraph show that communication pathways go along individual β-strands (the nodes coloured in the same grey tone belong to the same β-strand) and also cross them. The edges linking different β-strands reflect well the interactions that stabilise the two β-sheets of the protein. Filtering out pathways smaller than 6 residues yields a communication block twice as small (**Figure 3.7C**, in orange). The β-strands S1, S3 and S8 that form the first β-sheet (**Figure 3.7A**, in pink) are completely absent from the block, as well as helix H2. The block is further reduced by two times when keeping only very long ($\geq$ 8 residues) pathways (**Figure 3.7D**, in lime green). Only a portion of the second β-sheet, composed of S4, S7, S9 and S10 (**Figure 3.7A**, in red), remain in the block. This region can be viewed as the communication core of the protein.

Influence of non-covalent interaction type

Secondary structure units (*e.g.* β-sheets) are stabilised by H-bonds formed between backbone atoms (*e.g.* from parallel or anti-parallel β-strands). We analysed the impact of disregarding information from these interactions on p53 DBD communication. Only interactions involving side chain atoms were retained to construct communication pathways and the corresponding communication blocks were extracted (**Figure 3.7E-G**). The obtained subgraphs show a significantly reduced number of edges linking different β-strands. This result is expected owing to the nature of β-sheets. More surprisingly, however, the smaller number of edges minimally impacts the communication within each β-sheet. This indicates that numerous interactions are established within the β-sheets, other than backbone-backbone H-bonds. By contrast, the loss of these interactions is determinant for the communication between the two β-sheets and results in each of them being detected as an isolated communication block (**Figure 3.7E**, in red and pink). Two communication blocks are also detected when pathways smaller than 6 residues are filtered out (**Figure 3.7F**, in orange and yellow-orange), instead of one with all interactions (**Figure 3.7C**). This is due to backbone-backbone interactions being lost within S10 and between S10 and S9. The communication core of the protein, obtained from very long pathways (**Figure 3.7G**), is slightly smaller than when considering all interactions (**Figure 3.7D**), due to missing interactions involving S7.

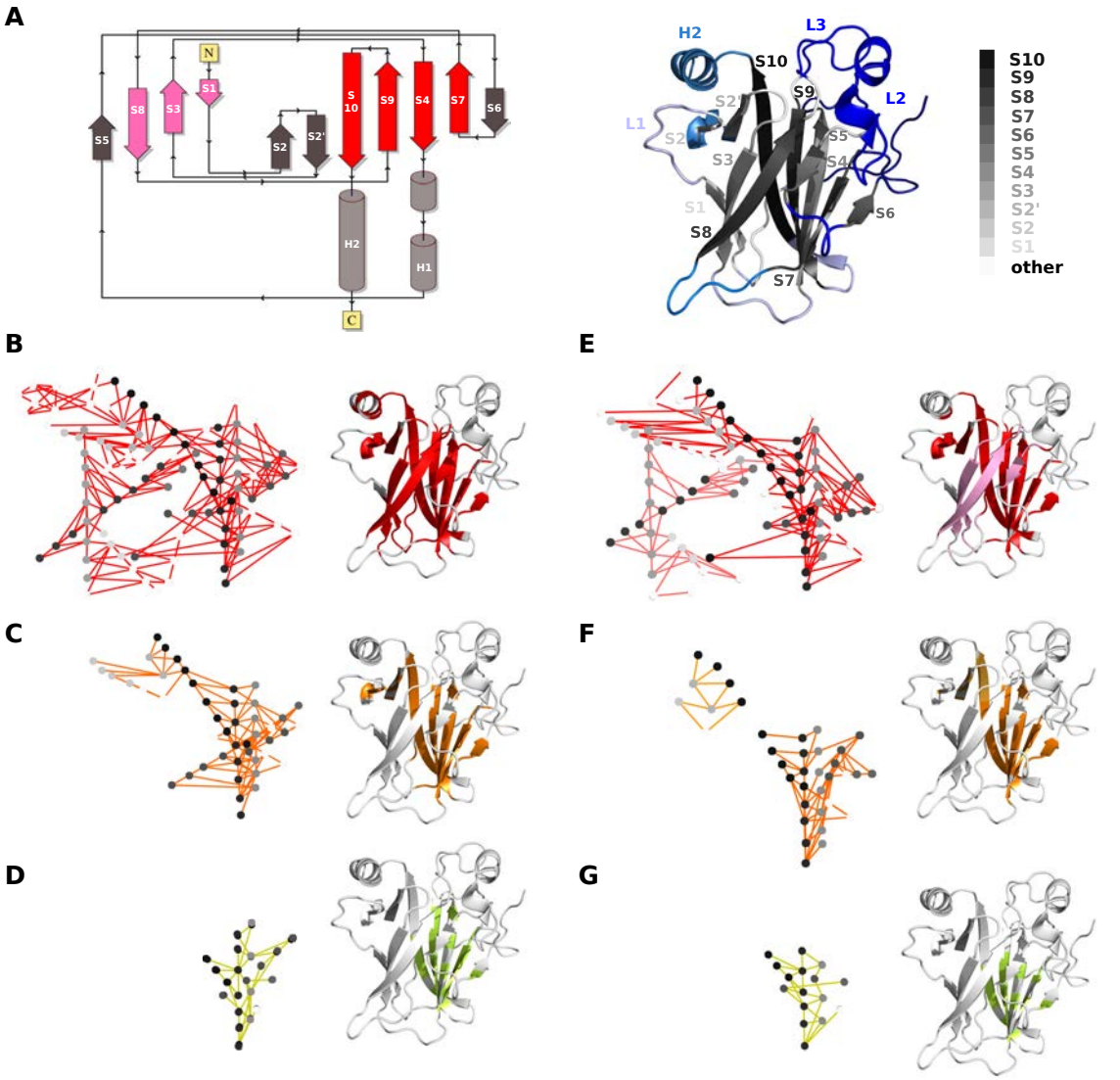

Figure 3.7: **Influence of pathway length and interaction type on P53 DBD communication.** **(A)** 2D topology diagram (on the left) and 3D structure (on the right) of p53 DBD. The diagram was taken from PDBsum (de Beer et al., 2014) and the colours were modified to put in evidence the S-type immunoglobulin-like fold of p53 DBD: the first β-sheet is in pink, the second β-sheet is in red. The 3D structure (average MD conformation) is represented as a cartoon, where the 11 β-strands of the protein are coloured in grey tones and labelled. The clique-based communication blocks identified by COMMA are colored in blue tones. **(B-G)** Pathway-based communication blocks identified by COMMA by using information from all non-covalent interactions (**B-D**) or only interactions involving side chains (**E-G**), and by considering only pathways longer than 3 (**B,E**), 5 (**C,F**) or 7 (**D,G**) residues. The communication blocks are represented as subgraphs in the PCN (on the left) and are mapped on the average MD conformation (on the right). The edges on the subgraphs and the residues on the 3D structure are coloured according to the communication blocks they belong to. The nodes in the subgraphs are coloured in grey tones, indicating the β-strand they belong to.

3.2.5 Comparison of protein A and p53

The B domain of protein A and p53 DBD represent two archetypal proteins in terms of thermodynamic and kinetic stability. While the latter unfolds at just above physiological temperature (Bullock et al., 1997), the former presents fast and stable folding (Lei et al., 2008). Moreover, BdpA is composed of three helices while p53 DBD mainly contains β-sheets. Consistently, our analyses of the two proteins show very different results. COMMA identified 2 very small clique-based communication blocks in BdpA, corresponding to the two extremities and representing 13% of the protein residues. By contrast, the clique-based communication blocks identified in p53 DBD represent almost 60% of the protein (**Figure 3.7A**, on the right and in blue tones). They encompass all residues involved in the interaction with DNA, namely the loops L1, L2 and L3 and the helix H2, which adopt variable conformations in the available experimental structures of p53 DBD (Lukman et al., 2013). COMMA also enabled to characterise the evolution of pathway-based communication blocks when varying the minimum communication pathway length. The communication core of BdpA, defined based on very long ($\geq$ 8 residues) pathways, comprises full-length helix H3 and some residues from H1 and H2 (**Figure 3.1**, box 5, in yellow). This is consistent with experimental evidence showing that H3 is the most stable helix among the three (Bai et al., 1997). p53 DBD presents a strikingly different dynamical behaviour, with a communication core composed of residues from different β-strands that form the first β-sheet (**Figure 3.7D**). Progressively filtering out communication pathways with increasing length results in residues, first from the loops that frame the β-strands, then from the extremities of the β-strands, to be excluded from the communication block (**Figure 3.7B-D**). Notice that the length of the pathways does not depend on the length of the β-strands, *i.e.* longer β-strands do not exhibit longer paths. These observations on BdpA and p53 DBD support the utility of COMMA to compare proteins of very different natures in a straightforward way.

3.2.6 The importance of the conformational sampling

The results obtained from COMMA directly depend on the extent and quality of sampling in the input conformational ensemble. In the case of MD trajectories, the user must carefully check that they have converged before proceeding through COMMA analysis. In the present work, we have performed COMMA analysis on the conformational ensemble generated during the last 30 ns of two 50-ns MD replicates for each studied system. We have assessed the stability of the studied systems in the chosen force field description (**Figure 3.3A** and **Figure 3.4A**) and the convergence of the MD trajectories (**Table 3.3**). We have also applied COMMA to the single trajectories and have obtained similar results (**Table 3.5** and **Table 3.6**). This indicates that our results are reproducible and robust to limited variations of the conformational ensemble. Another important aspect is the number of input conformations. In order to get statistically significant results, in particular for the principal component analysis, the number of conformations shall in principle be larger that the number of degrees of freedom of the system studied. In the examples of application reported here, we have characterised the internal dynamics of three proteins on relatively short simulation times (replicates of 50 ns). Consequently, we have illustrated how COMMA can reveal the dynamical dimension of a 3D structure representing a particular macrostate of the protein. Nevertheless, the utility of COMMA is not limited to

such type of analysis and the tool can be applied to atomistic simulations sampling large conformational changes.

	CB^{path}					CB^{clique}
	$l \geq 4$	$l \geq 5$	$l \geq 6$	$l \geq 7$	$l \geq 8$	All
P53						
all	1	1	1	1	1	5
sim1	3 (55)	3 (61)	1 (84)	1 (75)	1 (88)	8 (78)
sim2	1 (95)	1 (82)	1 (87)	1 (80)	1 (59)	8 (91)
Protein A						
all	1	1	1	1	1	2
sim1	1 (100)	1 (100)	1 (98)	1 (100)	1 (98)	2 (100)
sim2	1 (88)	1 (82)	1 (82)	1 (80)	1 (78)	2 (88)
KIT WT						
all	3	1	1	1	1	7
sim1	3 (94)	1 (93)	1 (88)	1 (96)	1 (96)	6 (76)
sim2	2 (92)	1 (95)	2 (90)	1 (87)	1 (80)	6 (58)
KIT MU						
all	1	2	3	2	2	5
sim1	2 (76)	2 (70)	2 (31)	1 (52)	1 (51)	5 (90)
sim2	1 (96)	2 (94)	3 (90)	2 (98)	2 (100)	6 (68)

Table 3.5: **Reproducibility of communication blocks over the MD replicates of each studied system.** The numbers of pathway-based and clique-based communication blocks (CBs) identified by COMMA when applied to the whole conformational ensemble and to the individual MD trajectories are indicated. For CBs^{path}, different minimum pathway lengths l (in residues) are considered. The overlap (in percentages of residues) between the CBs identified from the first (resp. second) MD replicate and those identified from the whole conformational ensemble are indicated in parentheses.

3.3 Conclusion

We provide to the community a fully automated tool for analysing conformational ensembles of proteins. The power of the COMMA method resides in the fact that it computes a number of dynamic properties of a protein at the residue level and integrates them in a unified framework to dissect the protein dynamical architecture by identifying its building blocks and the connections between them. COMMA permits to enrich the knowledge of a protein structure by bringing precise, complete and synthetic information on/from its internal dynamics. Moreover, the automatic set up of the parameters implemented in COMMA allows for an adapted modelling of the system under study and to contrast the roles of the different protein regions. COMMA can advantageously complement classical

	all	sim1	sim2
H1-H2	0.5	0.1	0.3
H1-H3	1.1	1	0.3
H2-H3	4.1	4.1	2.6

Table 3.6: **Reproducibility of communication strengths between secondary structure elements (SSEs) in Protein A.** The communication strengths (computed as the product of the proportions of residues involved in communication pathways linking the two segments multiplied by the number of pairs of residues directly linked by a pathway, see *Methods*) between pairs of helices (H1, H2, H3) are reported for the whole conformational ensemble and for each individual replicate. The order of communication strengths is the same in the three analyses.

analyses of protein structures and simulations and help look at proteins as dynamical biological objects with a new eye. On the other hand, COMMA introduces new measures and new algorithms, with respect to MONETA, to dissect a protein's architecture building blocks. It integrates different types of structural and dynamical information in a unified graph representing the protein. It detects communication blocks and communicating segments pairs from this graph, which are new concepts representing groups of residues or protein regions that mediate short- and long-range communication.

This analysis illustrates how COMMA can help dissect a protein 3D structure from a dynamical perspective and characterise the effect of a deleterious mutation on the structural dynamics of a protein. The information provided by COMMA was found in agreement with the previous findings on KIT allosteric communication. It further allows a more systematic assessment of the differences between two proteins or two states of the same protein and permits to pinpoint with high precision regions or residues instrumental in the establishment or alteration of the protein communication.

This analysis unveiled the hierarchical roles played by the different structural units (*i.e.* β-sheets) of the p53 DBD in the protein's dynamical architecture. Specifically, the residues constituting the first β-sheet communicate at shorter range than those constituting the second β-sheet. Furthermore, it showed the preponderant role of backbone-backbone interactions in establishing communication between the two β-sheets. These results illustrate how COMMA can be employed to contrast different protein regions from a dynamical point of view and to investigate the molecular determinants of protein communication at a precise level.

Chapter 4

Genetic disease-associated mutations in growth hormone

Contents

4.1 Biological context

Growth Hormone (GH) is a four helix bundle that regulates a wide variety of physiological processes, including growth and differentiation of muscle, bone, and cartilage cells (Sundstrom et al., 1996). The regulation of normal human growth is initiated by the binding of GH to its receptor (GHR), with stoichiometry 1:2, GH binds to two identical subunits of the receptor (de Vos et al., 1992). Although the two receptor molecules are identical, the GH binding sites are different **Figure 4.1**).

GH is composed of four helices and three of them are involved in the interaction with GHR. Helices H1 and H4 form the first binding site, while H1 and H3 form the second binding site. The former has high affinity for GHR, while the latter has low affinity (**Figure 4.1**).

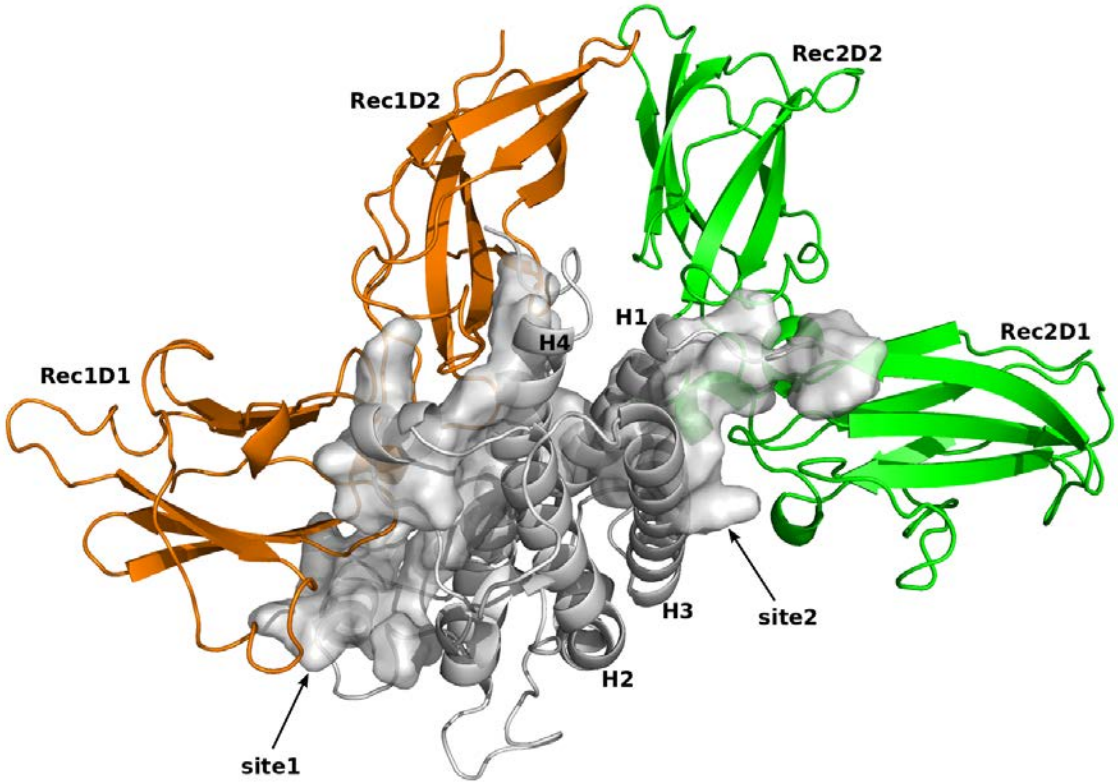

Figure 4.1: **Crystallographic structure of the GH-GHR complex**
Domains of GH-GHR complex are shown here with different colors. GH has one domain shown in gray, comprised of four helices (H1, H2, H3 and H4). First and second receptors are colored in orange and green, respectively and their domain are labelled on the structure. Rec1D1 and Rec1D2 indicate two domains of the first receptor molecule (residues E32-P131 and D132-P234). Rec2D1 and Rec2D2 indicate two domains of the second receptor molecule (residues E32-P131 and D132-P234). The surface of site1 and site2 GH are displayed.

The two binding sites are allosterically coupled and it was shown previously that site2 can be altered by mutations at site1 (Walsh et al., 2004). Point mutations of the GH-GHR complex may cause genetic diseases such as short stature (Petkovic et al., 2010). We had access to a group of these mutants from our collaborator, Serge Amselem (Service de Génétique et d'Embryologie Médicales, UMR S933 INSERM / UPMC, Hôpital Armand-Trousseau) and from that list we selected two disease causing mutations, L124R and R183H. In this chapter we demonstrate the impact of these two disease-related mutations on the dynamical behaviour of the complex. Then we characterize the allosteric coupling between the two binding sites, define residues that are instrumental in this coupling and represent the impact of the mutations on this coupling. Furthermore, we establish a link between the allosteric communication in the complex and coevolution signals.

4.2 Methods

4.2.1 Coevolving clusters

The set of sequences for GH-GHR were extracted following this protocol: a PSI-BLAST (Altschul et al., 1997) search (3 iterations, e-value $< 10^{-5}$) was performed with the sequence of GH (PDB code: 1HWG, chains A, B and C) as the query. A set of 19 homologous sequences were retrieved, with the average sequence identity of 68%. The sequences were aligned with ClustalW (Larkin et al., 2007). It has to be mentioned that for the complex, only one sequence per species was selected.

We extracted the network of coevolving residues from the set of homologous sequences of GH-GHR complex. After performing Multiple Sequence Alignment (MSA) on the given dataset, co-evolving residues were detected using Blocks In Sequences (BIS) (Dib and Carbone, 2012b) and clustered with Cluster Aggregation (CLAG) (Dib and Carbone, 2012a). BIS is the most suitable method for the coevolution analysis of growth hormone, because it can be applied to protein families that are highly conserved or represented by few sequences.

4.2.2 Molecular dynamics simulations

The following molecular dynamics protocol was applied to all studied systems.

Set up of the systems The 3D coordinates for the studied proteins were retrieved from the Protein Data Bank (PDB) (Berman et al., 2000). The PDB id of 1HWG was used for the simulations of monomer (chain A) and complex (chains A, B and C). Number of residues for the Hormone and its receptors are 191 and 237. Also there are some missing residues, residues T148-D153, F191 of GH and residues V54-G62 of GHR. All crystallographic water molecules and other non-protein molecules were removed and the missing residues were modelled with MODELLER 9v7 (Fiser et al., 2000). Terminal residues of the receptors were not modelled (residues F1-K31 for both GHR1 and GHR2 and residues Q235-S237 for the GHR1). Although receptors are identical, they don't have equal lengths after the modelling. There are 8 disulphide bonds for the GH-GHR complex, two inside the hormone (C53,C165) and (C182,C189) and three within each receptor (C38,C48), (C83,C94) and (C108,C122). Those bridges were kept for the molecular dynamics simulation. The two mutated forms of GH and GHGHR, MU^{L124R} and MU^{R183H}, were generated by *in silico* substitution of the leucine (L) in position 124 into a arginine (R) and *in silico* substitution of the arginine (R) in position 183 into a histidine (H), respectively using MODELLER 9v7 (Marti-Renom et al., 2000).

All models were prepared using the LEAP module of AMBER 12 (Case et al., 2012), with the ff12SB forcefield parameter set: (*i*) hydrogen atoms were added, (*ii*) Na^+ or Cl^- counter-ions were added to neutralise the systems charge, (*iii*) the solute was hydrated with a cuboid box of explicit TIP3P water molecules with a buffering distance up to 10Å. The environment of the histidines was manually checked and they were consequently protonated with a hydrogen at the ϵ nitrogen.

Minimisation, heating and equilibration The systems were minimised, thermalised and equilibrated using the SANDER module of AMBER 12. The following minimisation procedure was applied: (*i*) 10,000 steps of minimisation of the water molecules keeping protein atoms fixed, (*ii*) 10,000 steps of minimisation keeping only protein backbone fixed to allow protein side chains to relax, (*iii*) 10,000 steps of minimisation without any constraint on the system. Heating of the system to the target temperature of 310 K was performed at constant volume using the Berendsen thermostat (Berendsen et al., 1984) and while restraining the solute C_α atoms with a force constant of 10 kcal/mol/Å^2. Thereafter, the system was equilibrated for 100 ps at constant volume (NVT) and for further 100 ps using a Langevin piston (NPT) (Loncharich et al., 1992) to maintain the pressure. Finally the restraints were removed and the system was equilibrated for a final 100-ps run.

Production of the trajectories Two replicates of 50ns MD simulations were generated for the monomer (GH) and two replicates of 100 ns for the complex (GH-GHR), with different initial velocities, were performed in the NPT ensemble using the PMEMD module of AMBER 12. The temperature was kept at 310 K and pressure at 1 bar using the Langevin piston coupling algorithm. The SHAKE algorithm was used to freeze bonds involving hydrogen atoms, allowing for an integration time step of 2.0 fs. The Particle Mesh Ewald method (PME) (Darden et al., 1993) was employed to treat long-range electrostatics. The coordinates of the system were written every ps. Standard analyses of the MD trajectories were performed with the *ptraj* module of AMBER 12. We applied the same protocol of MD simulations to the mutated forms of the monomer and complex (GH and GHGHR).

Stability of the trajectories To assess the stability of the complex, the all-atom root mean square deviation (RMSD) from the equilibrated structure were recorded along each 100-ns MD simulation replicate (**Figure 4.2**). The mean RMSD value for the fist replicate of the WT is 2.97+/-0.31 Å, for the second replicate of the WT mean value is 3.14+/-0.26 Å, respectively. The mean RMSD value for the fist replicate of the MU^{L124R} is 4.08+/-0.5 Å and 3.14+/-0.3 Å for the second replicate. Mean value for the first replicate of MU^{R183H} is 3.4+/-0.43 Å and for the second replicate is 3.81+/-0.49 Å. According to those values, the WT has globally slightly smaller conformational drift compared to the mutants.

The stability of the hormone alone, was also analysed (RMSD and fluctuations). The details are not reported here to prevent the redundancy of the plots. Based on the RMSD profiles, we retained the last 70 ns of the GH-GHR complex and the last 30 ns of GH monomer simulations for further analysis.

Detection of H-bond network around mutation position The H-bond network in the neighborhood of the mutations position to measure the local effects of every mutation. All the residues within 10Å of the mutation point were selected and all the H-bonds within these residues were detected, using the HBPLUS algorithm (McDonald and Thornton, 1994). As described in the previous chapter *Chapter 2, Methods*, in HBPLUS, H-bonds are detected between donor (D) and acceptor (A) atoms that satisfy the following geometric criteria: (*i*) maximum distances of 3.9Å for D-A and 2.5Å for H-A, (*ii*) minimum value of 90° for D-H-A, H-A-AA and D-A-AA angles, where AA is the acceptor antecedent.

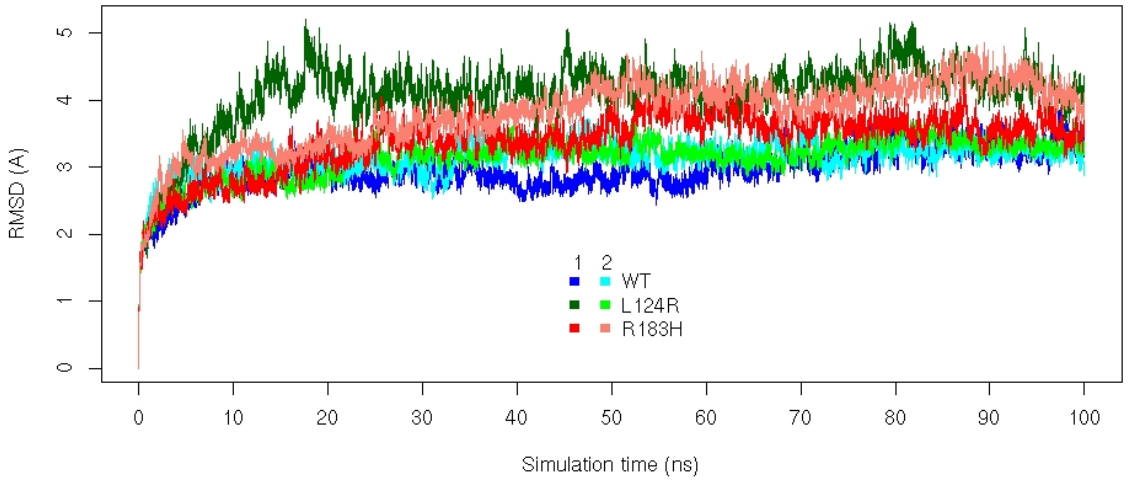

Figure 4.2: **RMSD for the Growth Hormone Complex**
For each replicate of the WT, MU^{L124R} and MU^{R183H} complexes, RMSD over all atoms (computed from the initial structure) are shown for 100 ns MD simulations. Colors correspond to the two replicates (1 and 2) of the WT and mutants.

Then we measured the interaction strength of every pair i and j, that represents the percentage of conformations in which at least one H-bond is formed between some pair of atoms (a_i, a_j) in residues i and j.

4.2.3 COMMA analysis

The details of COMMA were explained (*Chapter 2, Methods*), but as brief summary, COMMA extracts the dynamical properties at the residue level from the conformational ensembles and employs them to identify communication routes (*pathways*) between residues. It defines *communication blocks*, that are groups of residues with high communication propensity and strong non-covalent interactions and maps this information on the structure of the protein. COMMA is useful to identify residues playing a key role in protein allosteric regulation and to explain the effects of deleterious mutations in a mechanistic way. We applied COMMA to extract communication blocks of wild-type, MU^{L124R} and MU^{R183H} of the GH-GHR complex and GH monomer.

4.3 Effects of the mutations revealed by classical MD analysis

Here we report the MD analysis of the GH-GHR complex. Similar results were observed from the analysis of GH alone, but they are not reported here to avoid redundancy. For every GH-GHR complex we analysed 2 replicates of 70 ns and for each GH monomer we analysed 2 replicates of 30 ns MD simulations.

4.3.1 Atomic fluctuations

We analysed the atomic fluctuations of the wild-type and mutants (MU^{L124R} and MU^{R183H}) of GH-GHR complex. The difference between atomic fluctuations of the mutant from wild type (MU - WT), computed over backbone atoms and averaged by residues, are mapped onto the complex structure (**Figure 4.3**). Mutation positions (124 and 183) are pointed out on the structures by gray spheres indicating Ca atoms. Color ranges are from blue to red, indicating changes from rigidity to flexibility. Results demonstrate that wild-type complex is more rigid than the two mutants (**Figure 4.3**).

MU^{L124R} displays higher fluctuations compared to the WT, around residues L128-R134, forming GH_Loop1 that faces mutation position (**Figure 4.3 a**). In addition, residues Q29-E39 forming the C-terminal of Helix1 and GH_Mini_Helix1 display higher fluctuations in MU^{L124R} complex compared to the WT. In the region of Rec1Loop1, residues T51-T58 and L61-Q65 and in the region of Rec2Loop2, residues H55-G62, R70-E75 and I103-Y107 represent higher flexibility in the mutant.

Considering MU^{R183H} (**Figure 4.3 b**), residues L128-R134, forming GH_Loop1 that faces mutation position, display higher fluctuations in the mutant compared to the WT. Three other regions display higher fluctuations in MU^{R183H} complex compared to the WT: **1)** Residues S150-D154 forming GH_Loop2, **2)** residues P37-T50 forming GH_Mini_Helix1 and **3)** residues E32-V54 and G62-T69 in the region of Rec1Loop1. On the other hand, although residues E32 to P133, in the region of Rec2Loop1, globally have higher fluctuations in mutant, WT has locally higher flexibility and fluctuation around the residues V54-G62, E75-W80 and P84-W104, in Rec2Loop1.

Average MD conformation for the two replicates of the WT and mutant complexes, MU^{L124R} and MU^{R183H} are superimposed (**Figures 4.4** and **4.5**, respectively). Considering the WT and MU^{L124R}, the average replicates are well superimposed in the hormone region, whereas long-range differences were reported in the loop regions of the two receptors, Rec1Loop1 and Rec2Loop1 (**Figure 4.4b** and **c**). The difference is more significant in the region of Rec1Loop1.

For the MU^{R183H}, the average structures are almost superimposed, however region GH_Loop1 (figure 4.5b), GH_Loop2 (figure 4.5c), the loops on the two receptors Rec1Loop1 (starting from residue D52 to I64, figure 4.5d) and Rec2Loop1 (residues from E34 to P133, figure 4.5e), adopt very different average positions in MU^{R183H} complex. In those regions, WT replicates display very similar profiles while mutant replicates are moving in two different directions. This behavior also suggests the long range effects of mutation in propagating signals through the receptors. Rec1Loop1 is also more rigid in the mutant (**Figure 4.3b**), this suggests that Rec1Loop1 adopts one or the other position and remains in it while the WT oscillate between the two.

4.3.2 Local H-bond network around the mutation within GH

Hydrogen bond interactions between mutation position (183) and its neighbouring residues on the structure are shown for the WT and MU^{R183H} (**Figure 4.6**). Mutation position (183) interacts through hydrogen bonds with residues Q68, K70 and N72, in addition residues N72, E129, D130, S132 and R134 form another H-bond network in both replicates of WT complex. In contrast, Hbonds are not present in both replicas of MU^{R183H}, except for interacting pair of (N72, D130) in the first replica of MU^{R183H} (4.6 part C) and (E129,

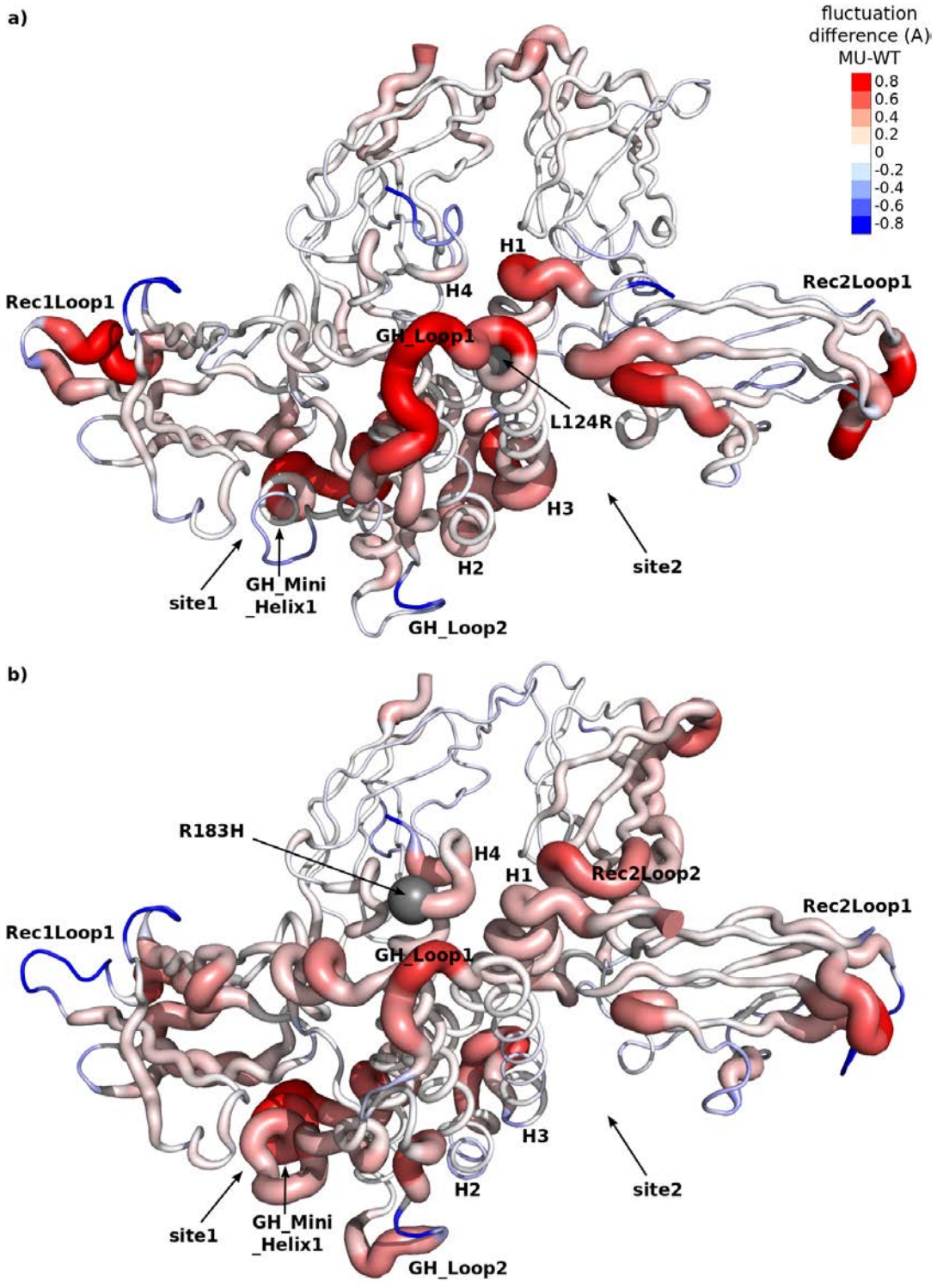

Figure 4.3: **Atomic fluctuations mapped on the complex structure**
Atomic fluctuations over backbone atoms that are averaged by residues are mapped on the average MD conformation of WT and MU complexes. Differences in values are shown through size and color of the cartoons for a) MU^{L124R} and b) MU^{R183H}. Mutant has slightly higher fluctuations and is more flexible.

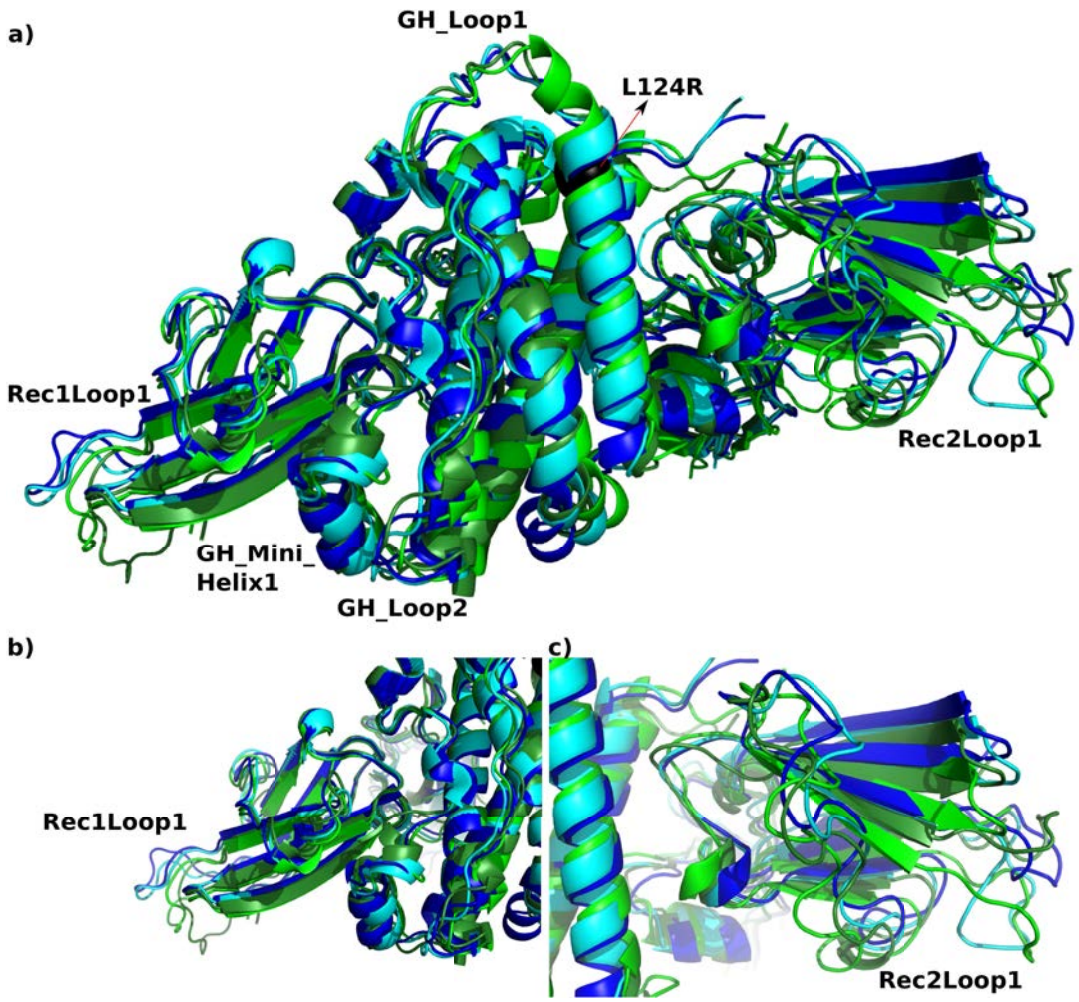

Figure 4.4: **Average structures of the WT and MU^{L124R}**
Average structures for every replicate of the WT and MU^{L124R} are superimposed. Blue and cyan cartoons correspond to the WT while light and dark green to the MU^{L124R}. The two loop regions, Rec1Loop1 and Rec2Loop1, of the mutant receptors adopt very different average positions compared to the WT (**b** and **c**).

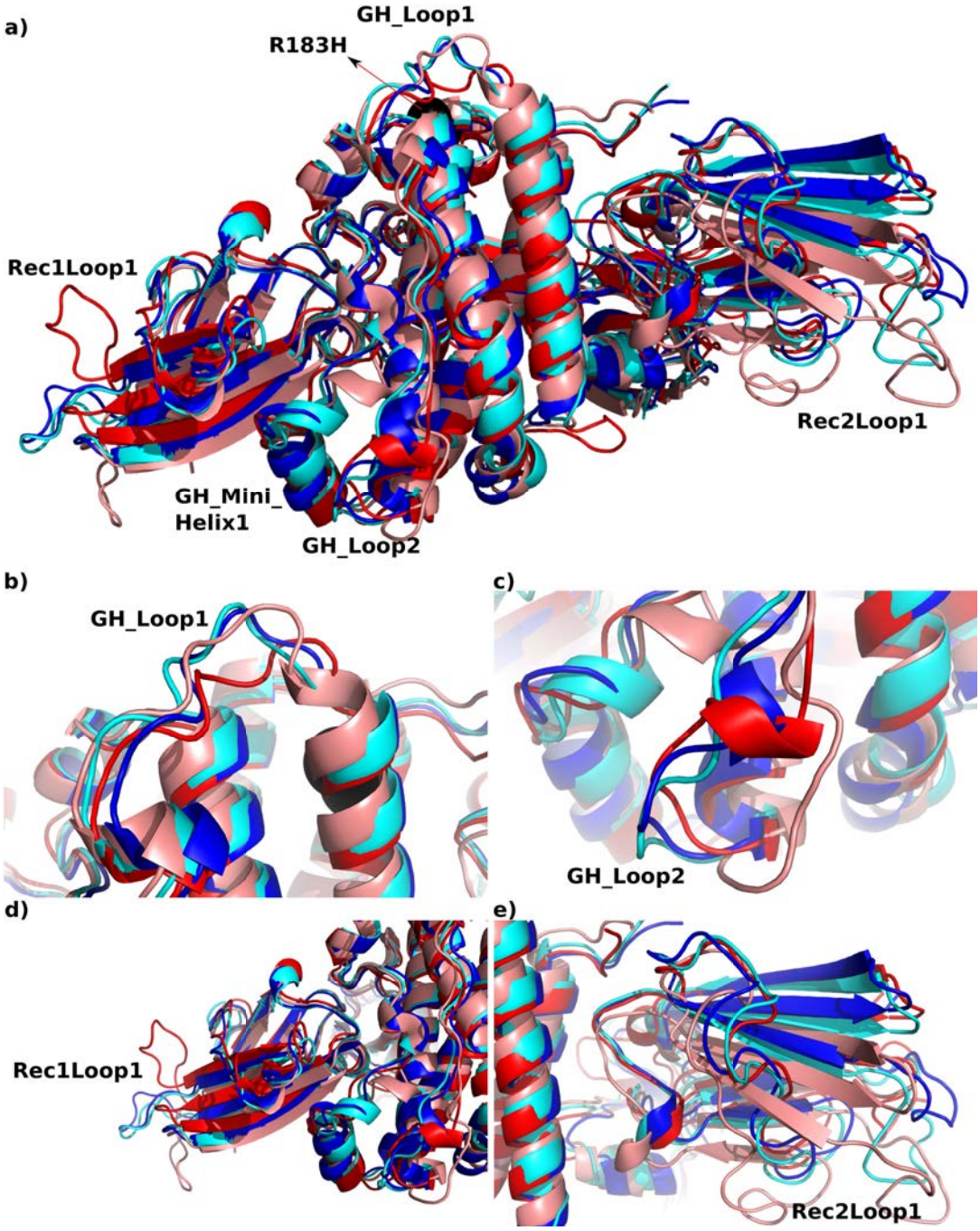

Figure 4.5: **Average structures of the WT and** MU^{R183H}
Average structures for every replicate of the WT and mutant are superimposed. Blue and cyan cartoons correspond to the WT while red and salmon to the mutant. Two loop regions of the hormone are shown where mutant has higher fluctuations (**b** and **c**), in addition to the Rec1Loop1 of first receptor (**d**) and Rec2Loop1 of the second receptor (**e**).

R134) in the second replica (4.6 part D). The interaction network within GH around the mutation is altered/weakened by the mutation.

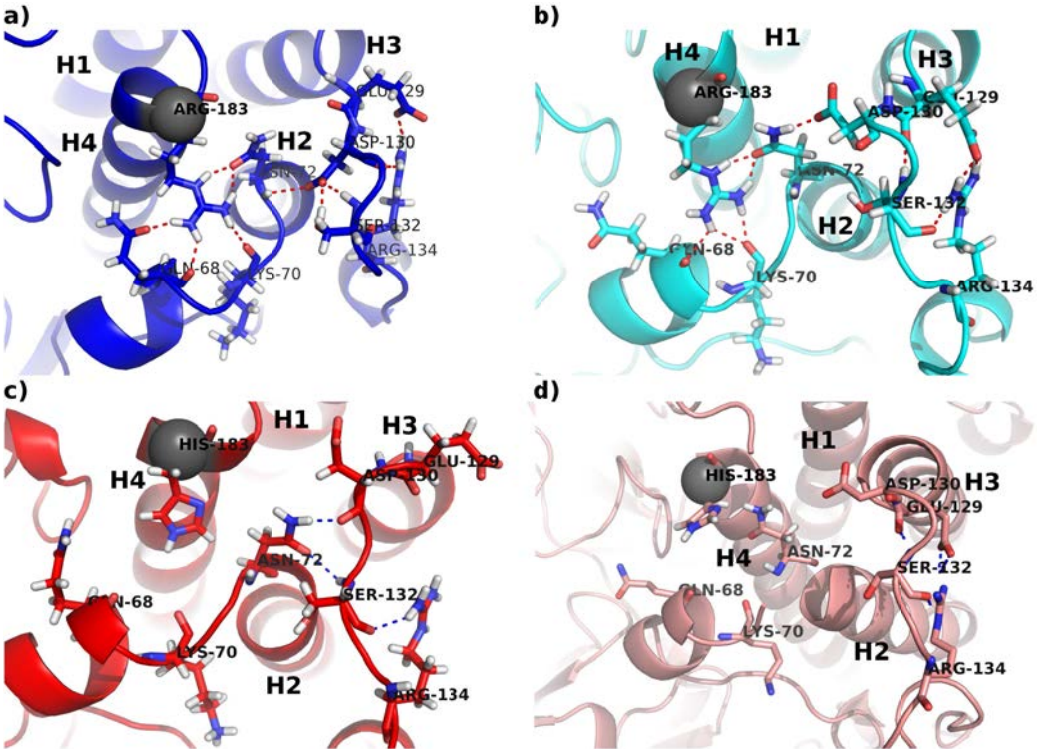

Figure 4.6: **H-bond between residues around the mutation position (183)**
The Hbond network in the neighbourhood of mutation position (183) is shown for the two replicates of wild-type (**a** and **b**) and MU^{R183H} (**c** and **d**) on the GH-GHR complex structures. These structures are the MD conformations that are closest to the average coordinates. Mutation position is represented with a gray sphere on each structure and HBonds are shown with dashed lines. Interactions between R183, K70 and N72 is detected on the WT but not on the MU^{R183H} as the short range effect of the mutation.

Table 4.1 contains quantitative values for supporting the mentioned observations (**Figure 4.6**). R183 has non-bonded interactions with residues Q68, K70 and N72 in the WT replicates, but those interactions are not present or weakened in the replicates of MU^{R183H}. Around the mutation position, other residues are affected by the mutation. Non-bonded interactions between residue pairs of: (N72, D130), (E129, R134), (D130, S132) and (S132, R134) are weakened or removed in the case of MU^{R183H}. As mentioned in figures 4.3 and 4.5 part B, residues L128-R134 forming GH_Loop1 which is facing mutation position, have higher fluctuations in the MU^{R183H} compare to the WT. The changes in interactions around mutation position can cause MU^{R183H} to fluctuate more, because interactions in this region are weakened.

Furthermore, Hydrogen bond interactions in the neighbourhood of position 124 are shown for the WT and MU^{L124R} replicates (**Figure 4.7**). Mutation position (124) interacts through H-Bonds with residues G120, R127 and L128. In addition two residue pairs of I121-M125 and T123-R127 interact through H-bonds for both replicates of WT complex. These H-bonds are not present or weakened in replicates of MU^{L124R}, except for interacting pair of I121-M125. By contrast, residues M125 and L128, interact through H-Bonds

Interaction	Strength			
	WT_1	WT_2	MU_1^{R183H}	MU_2^{R183H}
183-68	67	87		
183-70	91	93		
183-72	100	100		
72-130	66	89	74	
72-132			94	
129-132	39	95		59
129-134	52	99		84
130-132	74			
134-132	52	6	17	26

Table 4.1: **Strength of HBonds detected from MD simulations.**
Hbonds around mutation position (183) are listed here. Values correspond to the strength of the non-bonded interactions (% of MD conformations) over the last 70,000 conformations (we ignored the values below 30%).

only in the replicates of MU^{L124R}. Also due to the mutation, two pairs of H-Bond interactions are recorded between S7-Q181 in first replicate and between F10-Q181 in second replicate of MU^{L124R}. These interactions link H1 to H4. The significance of the observation is that S7 and F10 are in the close neighbourhood of site2, while Q181 is placed very close to site1. Therefore such interactions can help in the allosteric communication between the binding sites. Moreover, interactions between R124 (on H3, very close to the binding site2) and two other residues Q181 and S184 (on H4, very close to site1), are present only in the replicates of the mutant and not in wild type. These two H-bonds, could also support communication between the two binding sites. Other residues involved in the H-Bond network close to the mutation position (**Figure 4.7**), are roughly the same between the replicates of wild-type and MU^{L124R} GH-GHR complex.

Table 4.2 contains quantitative values for supporting the mentioned observations (**Figure 4.7**). Other pairs of interacting H-Bonds are reported in the table, on H1 and H4, while the strength of those interactions is roughly the same between wild type and MU^{L124R}. Consequently, the mutation (L124R), led to the weakening of interaction network around mutation position, whereas it caused new interactions between H1 and H4, connecting the two binding sites.

4.3.3 Interactions between GH and GHR

We studied the set of all interactions at the two binding sites. The average strength of the interactions (over two replicates of MD simulations) are reported for the WT, MU^{L124R} and MU^{R183H} complexes (**Tables 4.3** and **4.4**). Type of the interactions (hbond or hydrophobic) between all atoms of each pair of residues are mentioned, in addition to their strength (% of MD conformations). Strikingly larger number of residues are involved in the interactions at site1 (31 residues) with the average interaction strength of 76%,

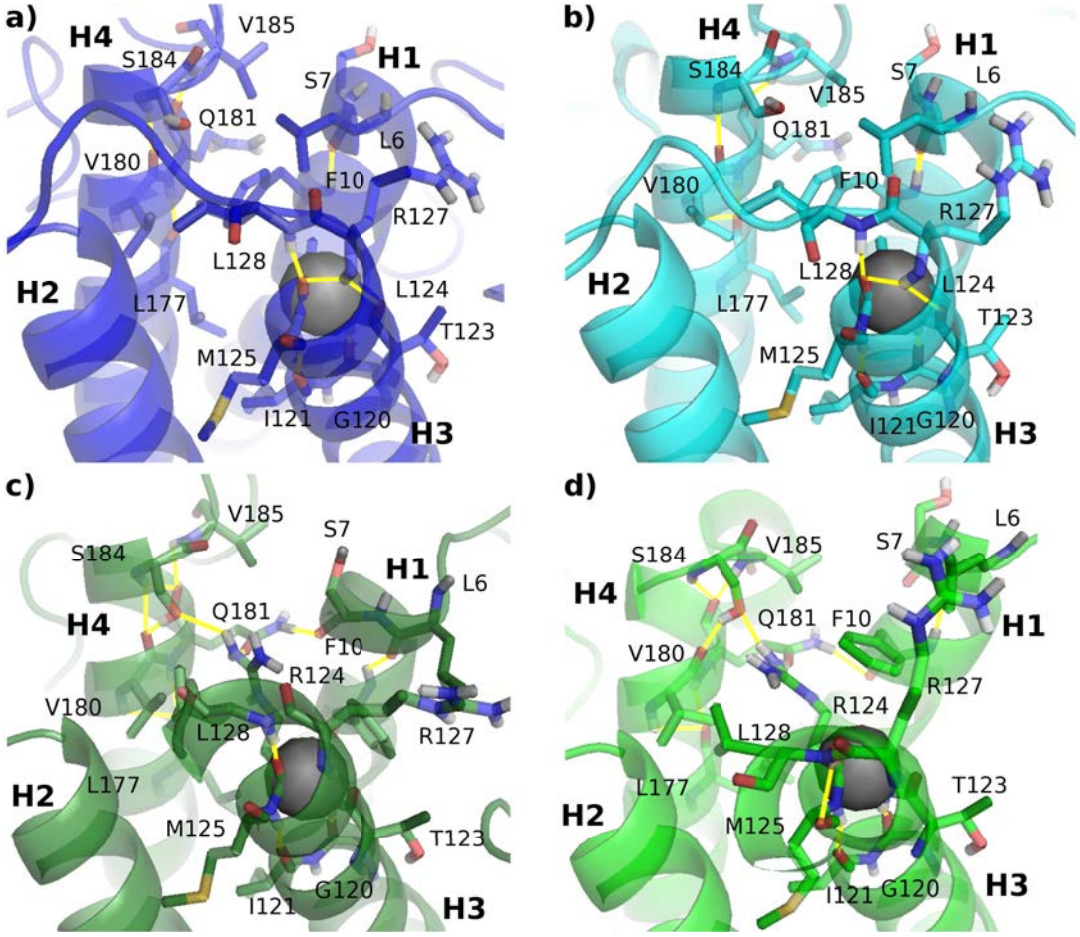

Figure 4.7: **H-bond between residues around the mutation position**
The Hbond network in the neighbourhood of mutation position (124) is shown for the two replicates of wild-type (**a** and **b**) and MU^{L124R} (**c** and **d**) GH-GHR complex. These structures are the MD conformations that are closest to the average coordinates. Mutation position is represented with a gray sphere on each structure and HBonds are shown with yellow lines. R124 interacts with Q181 and S184 in the two replicates of MU^{L124R}, but not in the wild-type complex.

Interaction	Strength			
	WT_1	WT_2	MU_1^{L124R}	MU_1^{L124R}
124-120	98	98	76	86
124-127	59	75	55	
124-128	91	90	43	
124-181			59	49
124-184			59	52
121-125	99	100	98	89
123-127	50	31		
125-128			34	60
6-10	98	99	71	99
7-181			81	
10-181				75
177-181	98	99	96	98
180-184	95	77	100	88
181-185	84	96	78	98
181-184	49	48	67	79

Table 4.2: **Strength of H-Bonds detected from MD simulations.**
H-Bonds around mutation position (124) are reported with the corresponding interaction strength (% of MD conformations) over the last 70,000 conformations.

compared to site2 which involves only 19 hormone residues with an average interaction strength of 71%. This observation is in agreement with the lower binding affinity of site2 (Walsh et al., 2004).

Considering differences in the average interaction strength between WT and MU^{L124R} of more than 10%, at site1, 8 pairs of the interacting residues represent reduction of strength, while 4 other pairs of residues have an increase of strength **Tables 4.3**. On the other hand, at site 2, 9 pairs of interacting residues lose their strength, but 5 other pairs gain the strength. Also, the mutation induces the formation of interaction between D112 and GHR2 (site2), whereas the average interaction strength of residues, T3 and A13 at site2 reduces bellow 30% in the mutant **Table 4.4**. Consequently, an overall slight decrease of the interaction strength is reported over the 70ns of MD simulation for the two replicates of the mutant, at both binding sites. But the decrease is more present at site2.

Considering differences in the average interaction strength between WT and MU^{R183H} of more than 10%, at site1, 11 pairs of residues lose the strength, while 3 pairs of residues have higher strength upon mutation. Due to the mutation of R183H, E56 interacts at site1 with high strength (79%), while the strength of G190 at site1 decreases bellow 30% **Tables 4.3**. At site2, an average decrease (of 10% or more) is reported for 6 pairs of residues, whereas 4 other pairs of interacting residues represent an increase of interaction strength. On the other hand, the mutation lead to the interaction of D112 with GHR2 at site2 (similar to the observation for MU^{L124R}), whereas T123 almost loses the interaction in the mutant **4.4**. Consequently, for the two replicates of mutant (MU^{R183H}), along 70ns of MD simulations the interaction strength is slightly reduced at both binding sites, but the decrease is more present at site1.

4.4 Effects of the mutation on the communication of the complex

GH monomer When considering all pathways (> 3 res), a single CB^{path} is detected by COMMA analysis for WT that contains 127 residues (**Figure 4.8 in red**). For the MU^{L124R} and MU^{R183H} mutants, in addition to a large CB^{path} (in red) that spans 129 and 130 residues of the GH mutants, respectively, a small CB^{path} (in pink) is detected that contains 10 residues in MU^{L124R} and 6 residues in MU^{R183H} (**Figure 4.8**). In WT, when considering long-range pathways of at least 8 residues, three CB^{path} are detected, helices 2 and 3 are communicating in one block (26 res, in green), whereas 22 residues on helix 1 and 25 residues of helix 4 form two separated blocks (in brown and yellow, respectively). On the other hand, in the two mutants, long pathways span across the whole structure and form a unique CB^{path}, consequently strong coupling between helices is observed compared to the WT. The long-range CB^{path} in MU^{L124R} contains 104 residues and in MU^{R183H} contains 96 residues. CBs^{clique} are similar between the three systems of WT and mutants.

GH-GHR complex Then we applied COMMA analysis to study the complex form of GH-GHR. Figure 4.9 represents CBs^{path} and CBs^{clique} for the WT (**a**), MU^{L124R} (**b**) and MU^{R183H} (**c**) of the complex GH-GHR. We report the details of the COMMA blocks.

GH:site1	WT			MU^{L124R}			MU^{R183H}		
	Strength (%)	Hbond	Hydrophobic	Strength (%)	Hbond	Hydrophobic	Strength (%)	Hbond	Hydrophobic
H18	94		*	89	*	*	88	*	*
H21	100	*	*	**80**↓	*	*	100	*	*
Q22	94	*	*	**43**↓		*	90	*	*
F25	77		*	**43**↓		*	75		*
Y28	88	*	*	**54**↓		*	**72**↓	*	*
K41	83	*	*	**50**↓	*	*	**34**↓	*	*
Y42	92	*	*	99		*	**50**↓		*
L45	91		*	84		*	**74**↓		*
Q46	67		*	69		*	**53**↓		*
P48	63		*	**47**↓		*	**31**↓		*
S51	58	*	*	**74**↑	*	*	**30**↓		*
L52	70		*	65		*	**40**↓		*
E56	13		*	15		*	**79**↑	*	*
S62	100	*	*	100	*	*	100	*	*
N63	68		*	75		*	**79**↑	*	*
R64	99		*	100	*	*	99	*	*
T67	92		*	95		*	93		*
Q68	39		*	**82**↑	*	*	44		*
Y164	80		*	87	*	*	75		*
R167	100	*	*	92	*	*	93	*	*
K168	88		*	**100**↑	*	*	**98**↑	*	*
D171	100	*	*	100	*	*	100	*	*
K172	98		*	99		*	98		*
E174	45		*	50	*	*	50	*	*
T175	99	*	*	98	*	*	97	*	*
R178	94		*	88		*	89		*
I179	82		*	90		*	74		*
C182	96		*	96		*	97		*
C189	75		*	**51**↓		*	**36**↓		*
G190	48		*	**25** ↓		*	**23**↓		*
F191	41		*	**74**↑	*	*	**31**↓		*

Table 4.3: **Residues involved in first binding site.**
Residues that are involved in the non-bonded interactions at first binding site are listed. The type of the all-atom interactions (H-Bonds or hydrophobic) and their strength (percentage of MD simulation time) averaged over the two replicates, are reported for wild type and mutants. The changes between WT and mutants that are equal or greater than 10% are represented by upward and downward arrow that correspond to increase and decrease, respectively.

GH:site2	WT			MU^{L124R}			MU^{R183H}		
	Strength (%)	H-Bonds	Hydrophobic	Strength (%)	H-Bonds	Hydrophobic	Strength (%)	H-Bonds	Hydrophobic
F1	70		*	**51**↓	*	*	**52**↓		*
P2	42		*	**58**↑		*	37		*
T3	53		*	**26**↓		*	46		*
I4	65		*	**50**↓		*	**54**↓	*	*
R8	93	*	*	**62**↓	*	*	**59**↓	*	*
L9	86		*	**53**↓		*	83		*
N12	100	*	*	99	*	*	93		*
A13	94		*	**13**↓		*	91		*
L15	87		*	88		*	92		*
R16	99	*	*	100	*	*	100	*	*
H18	64		*	**35**↓		*	**88**↑		*
R19	52	*	*	**92**↑		*	**94**↑	*	*
Q22	48		*	46		*	**84**↑		*
Y103	73		*	80		*	**48**↓		*
D112	0			**65**↑		*	**50**↑	*	*
D116	100	*	*	**40**↓		*	100	*	*
E119	77		*	**93**↑	*	*	**62**↓	*	*
G120	100		*	**48**↓		*	99		*
T123	52		*	**94**↑		*	**14**↓		*

Table 4.4: **Residues involved in second binding site.**
Residues that are involved in the non-bonded interactions at second binding site are listed. The type of the all-atom interactions (H-Bonds or hydrophobic) and their strength (percentage of MD simulation time) averaged over the two replicates, are reported for wild type and mutants. The changes between WT and mutants that are equal or greater than 10% are represented by upward and downward arrow that correspond to increase and decrease, respectively.

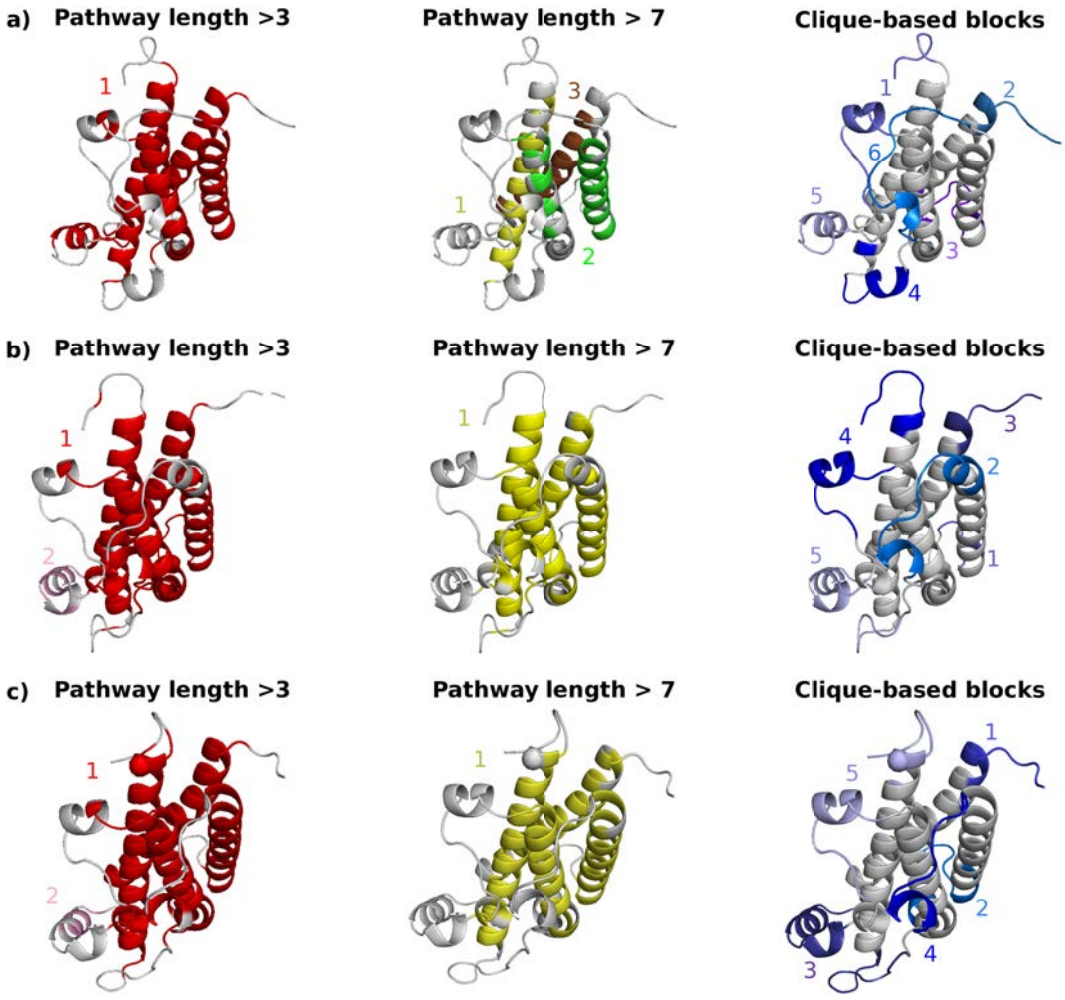

Figure 4.8: **Communication blocks for a) WT, b)** MU^{L124R} **and c)** MU^{R183H} **GH monomer defined based on pathways of length** $.3, > 7$ **residues and independent cliques.**

Considering CBs^{path} with at least 4 residues, one large communication block (207 residues) is detected in WT and 7 smaller blocks ($\leq$ 36 residues). The complex is fragmented inside within the receptor molecules. On the other hand, GH is divided in two and the division does not correspond to site1 and site2, bur rather to *with GHR* or *without GHR*. The largest communication block in WT (in red) encompasses 31% of GH residues, 30% of R1 and 42% of R2. It must be emphasized that there is no direct pathway connecting GH to R2 and the residues inside the red block are communicating through a chain of overlapping pathways that connect: **(1)** residues inside GH, **(2)** GH to R1, **(3)** residues inside R1, **(4)** R1 to R2 and **(5)** residues inside R2 (pathways at the binding sites are shown in black).

Considering CBs^{path} in MU^{L124R}, 4 blocks are defined while the largest block covers most of the residues in R1, R2 and helices H1 and H4 in GH (**Figure 4.9 b**). The organization of blocks is roughly the same as in WT, except that the red block is more extended and covers almost all the receptors. Also direct communication between GH and receptors, happens only at site2.

In MU^{R183H}, communication is reshaped into three blocks: **(1)** 52 % of GH residues (and 3 residues of R2), **(2)** 63% of R1 and 18% of residues in R2 at the binding site between the two receptors and **(3)** 41% of the residues inside R2 (**Figure 4.9 c**). The organisation of the blocks is completely different for MU^{R183H}, GH is completely contained in one block, instead of being splitted. The block also contains some residues from the receptor 2 (site2) and communication is maintained between the 2 receptors (orange block). In addition, GH communicates only through site2. Although the communication is stronger but does not propagate through the receptor molecules. When considering long-range CBs^{path} ($> 7res$), helix1 is not covered by blocks in MU^{L124R} while residues in Rec2D1 are not covered in MU^{R183H}.

6 CBs^{clique} are detected in WT, three blocks on Rec1D1, Rec2D1 and Rec2D2 (3, 1 and 2, respectively), two blocks on GH (4 and 6) and block 5 that covers some residues on H2, H3, H4 and residues I165 to G168 of R1. Clique 2 contains few residues of the R1 in the binding site between the two receptors. In L1234R the three blocks in GH and the other two in Rec1D1 and Rec2D1 are slightly different compared to WT, wheread block 2 is extended and contains residues in Rec1D2 as well as Rec2D2. CBs^{clique} of MU^{R183H} are significantly different from WT and MU^{L124R}. 4 blocks are detected: 1) covers part of residues in Rec2D1 and H1, H2 and H3 in GH, 2) covers part of residues in Rec1D1 and H1, H2 and H4 in GH, 3) and 4) this clique is positioned on mutating residues and its neighbours (C182-F191) on H4 and W169 on R1.

A general observation is the detection of large cliques in the receptors for all the complexes of WT and MUs. Receptors have many loops, compared to GH and those loops represent higher fluctuations compared to other receptor residues, when analysing the fluctuations along MD simulations. Such behavior may explain the fact that cliques are extended in receptors.

In order to compare the WT with the two mutants (MU^{L124R} and MU^{R183H}), communication pathways with at least 4 residues at the binding sites are extracted (**Figure 4.9**). The details on the number of pathways and residues involved in the three systems and binding sites are recorded in table 4.5. CBs^{path} are colored on the structure and pathways are shown in black lines. The pathways in the first binding site, between GH and R1 are only present in WT (**Figure 4.9b**), whereas pathways in the second binding site, between

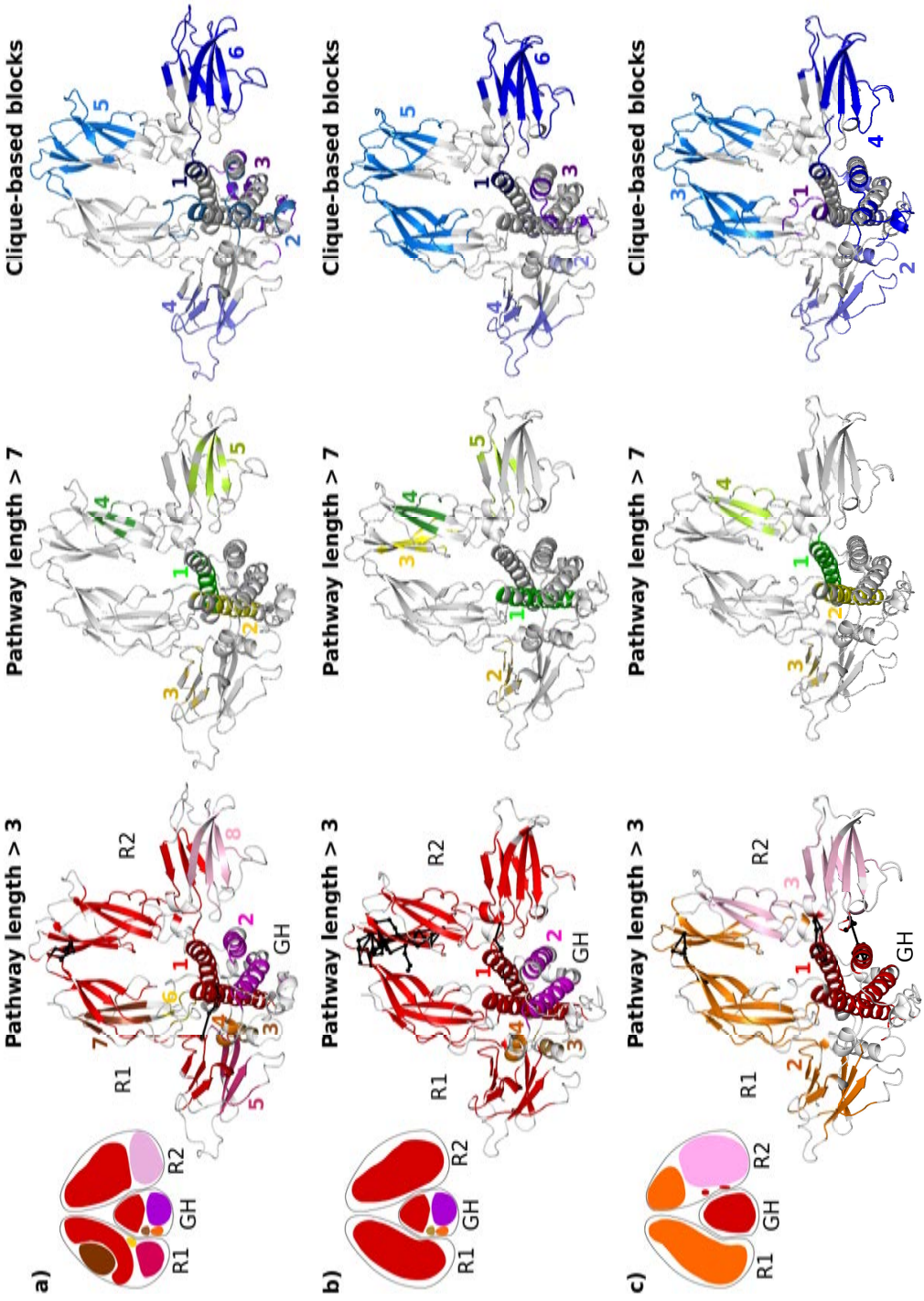

Figure 4.9: **Communication blocks for WT, MU^{L124R} and MU^{R183H} GH-GHR complex.** Communication blocks of pathways with at least 4 residues for a) WT, b) MU^{L124R} and c) MU^{R183H} complexes. The schematic representation of the blocks are depicted on the left. All pathways at the binding sites are shown in black lines on the structure of the WT, MU^{L124R} and MU^{R183H}.

	GH-R1	GH-R2	R1-R2
Wild type			
# Paths	4	0	3
# Residues	5	0	4
MU^{L124R}			
# Paths	0	3	49
# Residues	0	5	21
MU^{R183H}			
# Paths	0	15	3
# Residues	0	13	5

Table 4.5: **Number of paths and residues communicating at binding sites.**
Details of the pathways and residues communicating between hormone and receptors (GH-R1 and GH-R2) and between the two receptors (R1-R2) in WT, MU^{L124R} and MU^{R183H} are listed.

GH and R2 are only present in the mutants. In MU^{L124R} only H1 in GH communicates with R2, whereas MU^{R183H} demonstrates more communications in site2, between H1 and H3 in GH and R2. Such increase of pathways could be related to the increase of the interaction strength at site2 of the MU^{R183H}.

The fact that communication through site1 is present only in WT may be significant for the stability of the interaction of GH within its complex. It may indicate the deleterious effects of the mutations. The analysis of the interaction networks at binding sites (**Tables 4.3** and **4.4**) represented slight reduction of interaction strength at both binding sites for the two mutants. We observe the same behaviour when analysing the communication at binding site1 (**Table 4.5**). On the other hand, increase of communication is reported for the two mutants at site2 (**Table 4.5**). However, a drastic increase is reported for MU^{R183H}. Even though the interaction networks at binding sites (**Tables 4.3** and **4.4**) are not drastically different between WT and mutants, we can clearly see some changes in the communication. Moreover, the changes in the re-shaping of the blocks and the differences between communications at binding sites, reveal the ability of COMMA to capture the differences between the two mutants.

4.4.1 Pathways that correspond to block reshaping

COMMA enables us to detect the set of pathways that correspond to the reshaping of the CBs^{path}. For example there are 2 CBs detected on the structure of R2 in WT, whereas there is only one CB on the R2 of MU^{L124R}. Here we report the subset of pathways that correspond to the reshaping of path CBs, between WT and each mutant:

WT and MU^{L124R} The significant difference between blocks of WT and MU^{L124R} corresponds to the increase in the coverage of largest block (**Figure 4.10** colored in red). This block in MU^{L124R} covers more residues, including portion of residues in Rec1D1, Rec1D2 and Rec2D1 (blocks colored in dark pink, yellow, brown and pink on WT). Set

of new pathways in mutant that correspond to the grouping of all those blocks of WT in mutant are shown with black lines.

WT and MU^{R183H} A further analysis was performed to investigate the set of pathways that result in the reshaping of communication blocks between WT and MU^{R183H} (**Figure 4.11**, green lines). Pathways in WT are detected in R2 and at the first binding site between GH and R1. Newly formed pathways in the mutant are detected inside GH, R1, R2 and at the second binding site between GH and R2.

4.5 Coevolution analysis of GH and GH-GHR complex

4.5.1 Coevolution of the monomer (GH)

The set of homologous sequences were extracted for GH. After performing the sequence alignment, we applied BIS to extract clusters of coevolving residues. For every cluster, the set of hit residues and extensions are reported (**Table 4.6**), while hit residues of the six clusters are colored on the structure of GH (**Figure 4.12**). We analysed each cluster, with respect to the COMMA analysis of the GH-GHR complex. Here we report the results obtained from the analysis of two clusters that are more interesting compared to the rest.

cluster	hit resideus	extensions
1	10,87	86
2	15,116	16,17,117
3	9,99,156	
4	18,72,171,176	16,17,174,175
5	25,184	182,183
6	33,41	40

Table 4.6: **Clusters of coevolving residues for GH.**
Coevolving residues detected by BIS are reported here. The set of hit positions and extensions are grouped separately.

In cluster 1 (F/L10 & L/F87), F10 from H1 belongs to $CB^{path}1$ and its buried side chain is inserted between H1 (Nterm), H3 (C-term) and H4 (C-term). On the other hand, L87 and its neighbouring residue, W86, from H2 and are precisely located at the kink of the helix, L87 is not detected in CB^{path} nor in CB^{clique}, but W86 belongs to $CB^{path}2$. The coevolutionary change of F→L at position 10, may lead to the decrease of the space filled by the residue, while coevolution at position 87 (L→F), would increase the excluded volume and preserve the kink in the context of L10 (H2 could be instrumental in the transmission of signal between the 2 coevolved residues). F10 interacts (91% simulation time) with L124 (a mutational hotspot) from H3, and numerous paths go along H2 or H3 from L124 to W86, which is 100% conserved and covalently linked to L87 (**Figure 4.13a**). It has to be mentioned that in the variant reported by (Walsh et al., 2004) F10

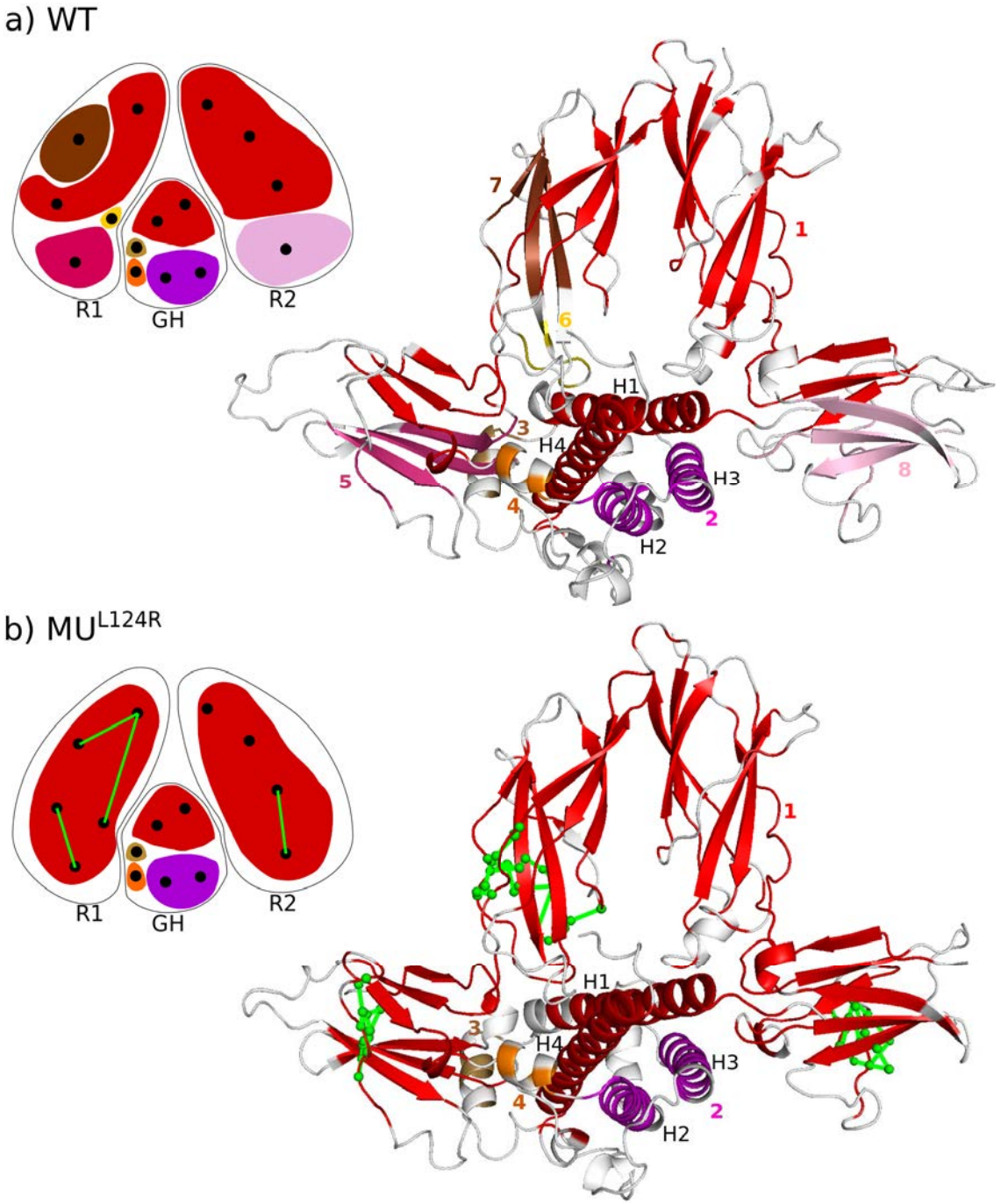

Figure 4.10: **Study of pathways that correspond to the reshaping of the blocks in WT and MU^{L124R} of GH-GHR.**
The set of 8 communication blocks in **a)** WT (pink, red, brown, yellow, dark pink, orange, sand and magenta) and 4 in **b)** MU^{L124R} (pink, orange, magenta and red) are shown on the cartoon representation of the structure. Pathways of at least 4 residues that correspond to the differences between components are shown on the structure, their existence allows the components to collapse. The schematic representations of the blocks in WT and MU^{L124R} are depicted on the left. The green lines represent the connection of separate secondary structures due to the mutation and the black circles highlight secondary structures within CBs.

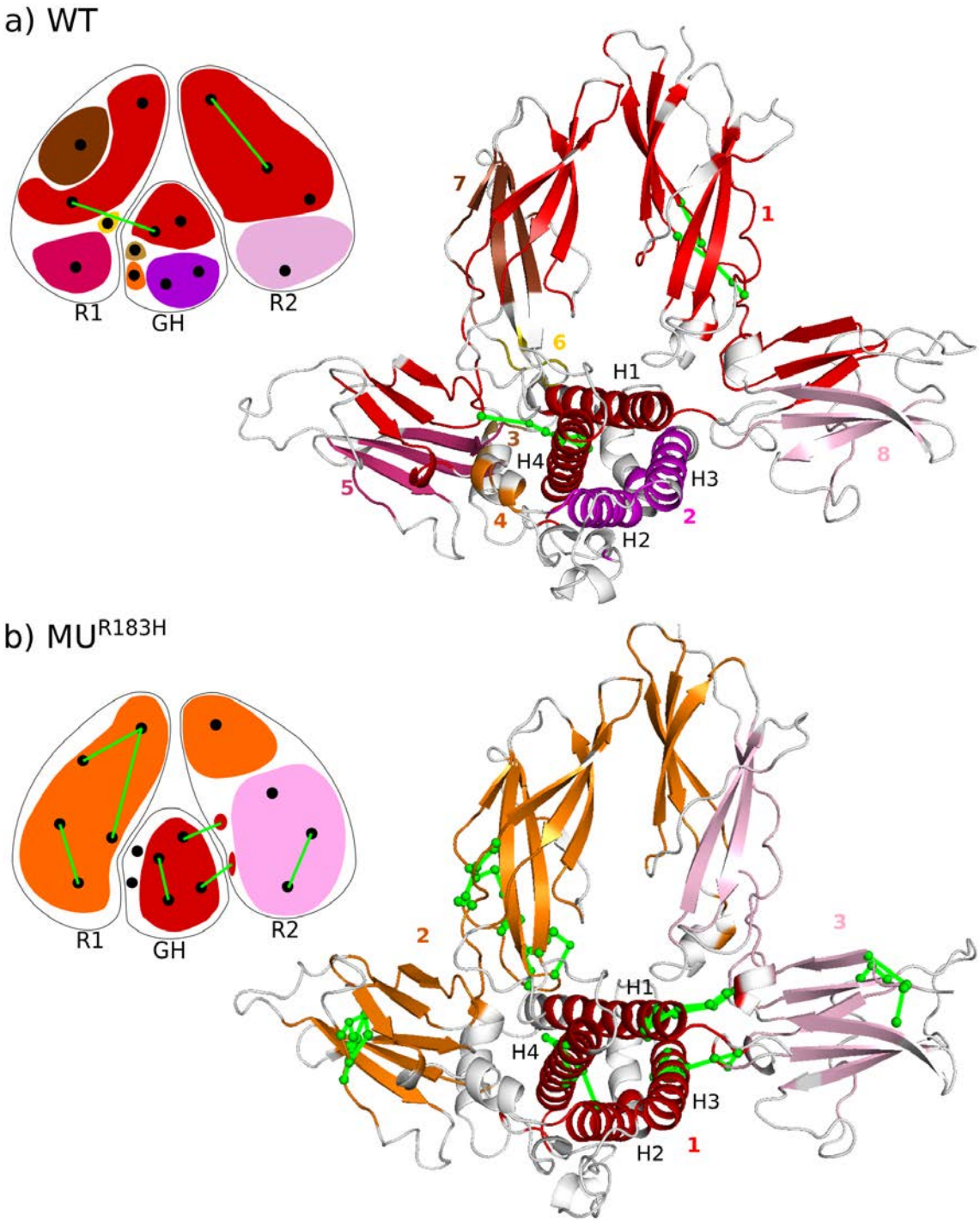

Figure 4.11: **Study of pathways that correspond to the reshaping of the blocks in WT and MU^{R183H} of GH-GHR.**
The set of 8 communication blocks in **a)** WT (pink, red, brown, yellow, dark pink, orange, sand and magenta) and 3 in **b)** MU^{R183H} (pink, orange and red) are shown on the cartoon representation of the structure. Pathways that correspond to the differences between blocks are colored in green, their existence allows the components to collapse. The schematic representations of the blocks in WT and MU^{R183H} are depicted on the left. The green lines represent the connection of separate secondary structures due to the mutation and the black circles highlight secondary structures within CBs.

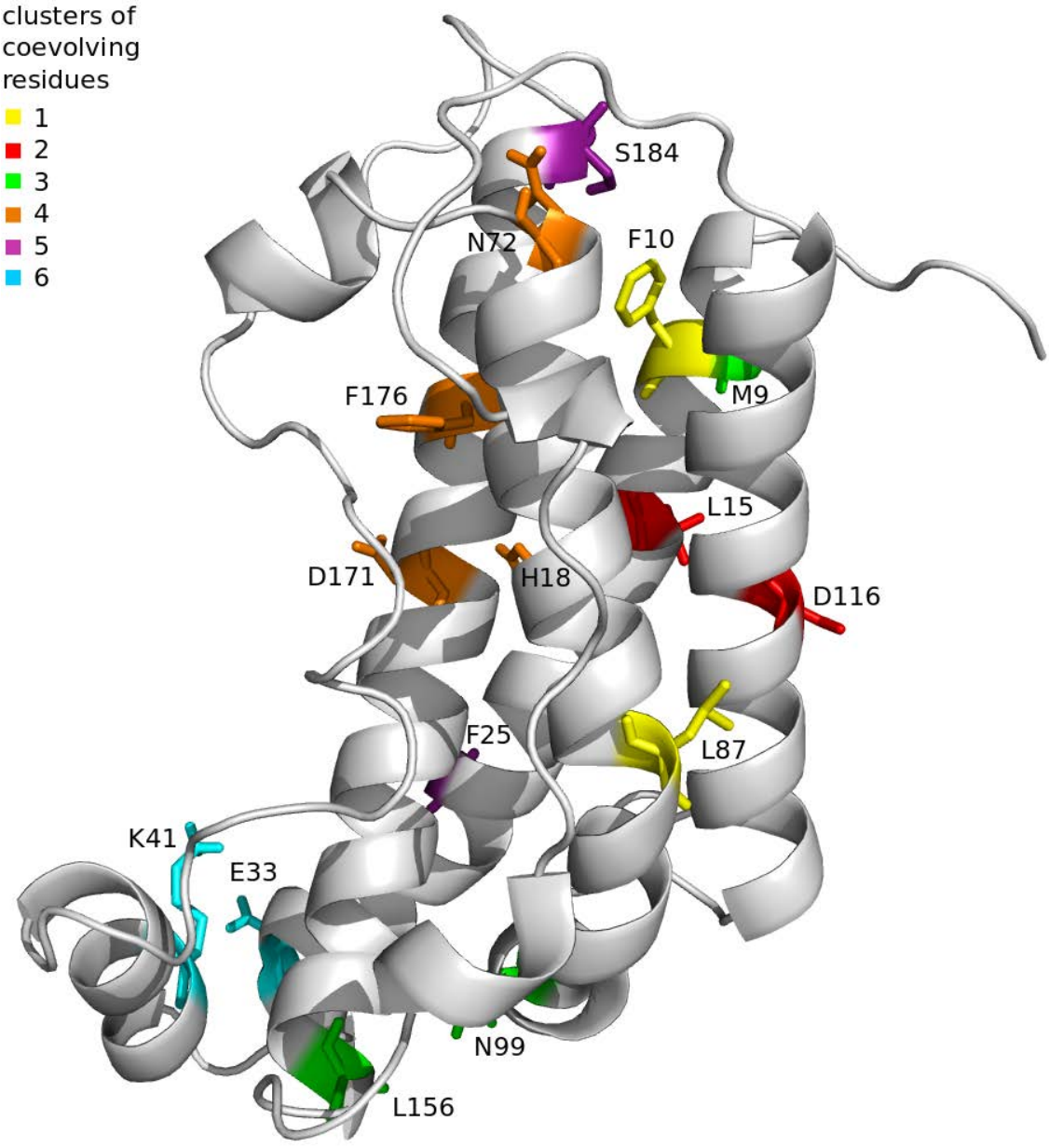

Figure 4.12: **Clusters of coevolved GH residues.** Residues that belong to different coevlution clusters are shown in different colors and with sticks on the structure of GH.

is mutated to A and indeed the region after the kink and the subsequent loop (residues 94-110) adopt a different conformation.

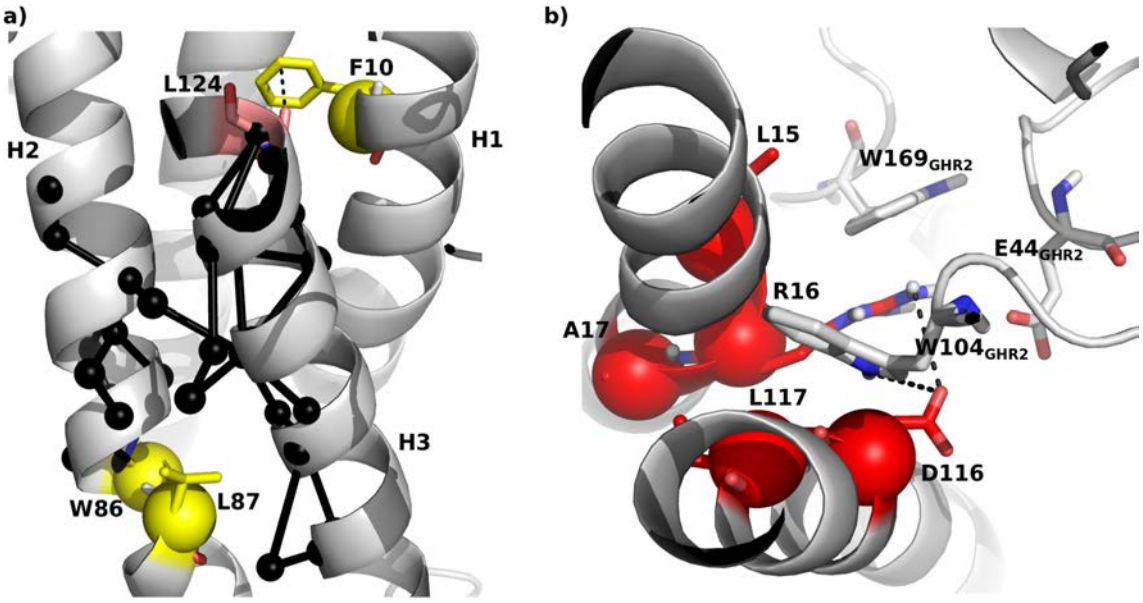

Figure 4.13: **Coevolved residues in clusters 1 and 2.** Coevolved residues in clusters 1 **(a)** and 2 **(b)** are colored in yellow and red, respectively. The pathways detected by COMMA are shown with black lines and the interaction between residues are represented by dashed lines.

In cluster 2 (L/S15 & D/Y116) residues 15-17 from H1 are detected in $CB^{path}1$ and residues 116-117 from H3 are in $CB^{path}8$. D116 interacts with $W104_{GHR2}$ (striking mutational effects were reported for D116A in variant introduced by (Walsh et al., 2004)); R16 is fully conserved, involved in a salt bridge with $E44_{GHR2}$, stacked with $W169_{GHR2}$ and interacts with D116 for roughly 100% of simulation time (**Figure 4.13b**). The possible compensatory effect of D→Y at 116 would be to increase the excluded volume and remove potential H-bonds, whereas L→S would decrease the excluded volume and add potential H-bonds.

4.5.2 Coevolving residues in GH-GHR complex

We performed coevolution analysis of the GH-GHR complex. After performing Multiple Sequence Alignment(MSA) on the set of homologous sequences of GH-GHR, coevolving residues are detected using BIS and clustered with CLAG (**Figure 4.14**). 9 different clusters were detected using BIS, 3 of them include only hormone residues and the rest represent coevolution pattern between hormone and receptor residues. In order to find the link between coevolution and COMMA, we extracted the subset of pathways (with at least 4 residues) that communicate at binding sites. As explained in previous section, there are two pathways at the first binding site and two others are detected between the two receptors, from the COMMA analysis of wild-type GHGHR. The significant observation was the pathways detected at the first binding site (between GH and GHR1), were passing through residues that belong to the first and second coevolving clusters. Coevolving residues of the first cluster are detected on the helices of GH, first Rec1D1 and Rec2D2 (**Figures 4.15 green spheres**). On the other hand, in cluster 2, the 4 coevolving

residues are detected on the hormone, Rec1D1 and Rec2D2 (**Figures 4.15 red spheres**). Pathways at site1 cross through these coevolving residues and connects D171 on H4 of GH to R203 on Rec1D1(**Figures 4.15**). This observation, highlights the importance of D171 on GH and R203 on R1, as they are directly linked by the very few paths connecting GH to GHR, while they belong to two different coevolving clusters.

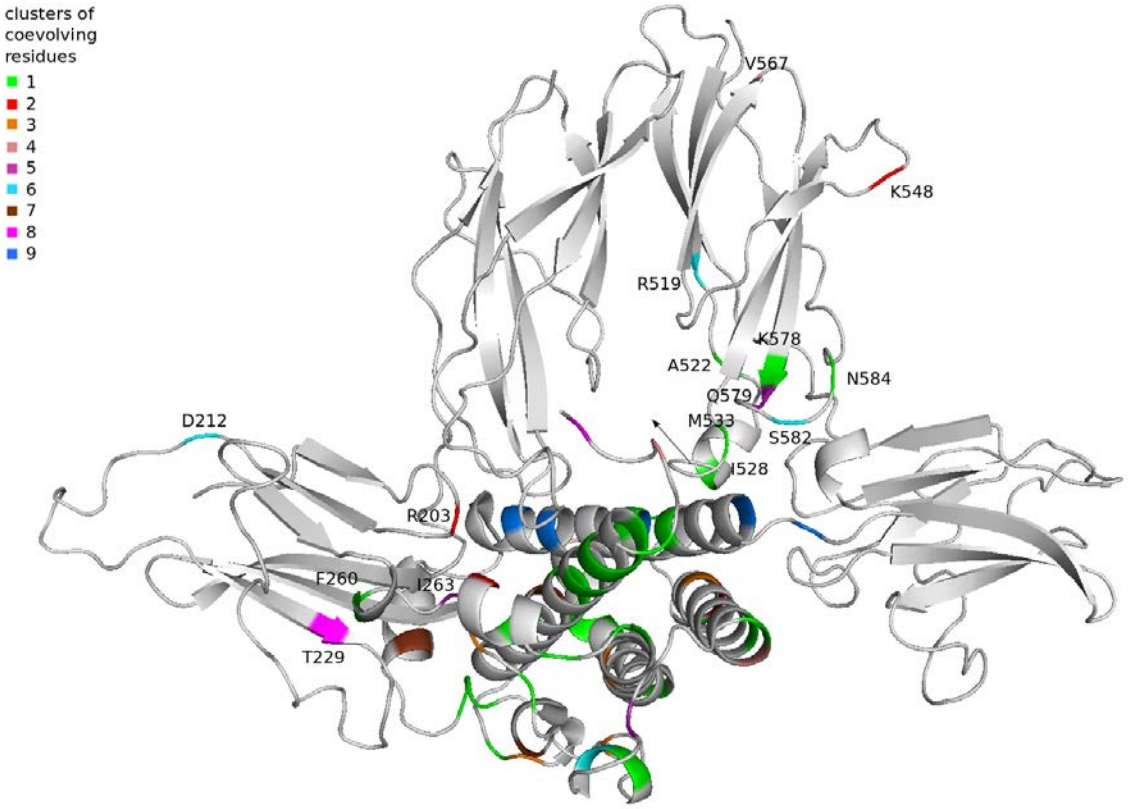

Figure 4.14: **Clusters of coevolved residues for GH-GHR.** Coevolving residues of the GHGHR are colored to highlight different clusters and all residues that belong to the receptors are labelled.

4.6 Conclusions

The results indicate a dynamics-based rewiring of communication network in GH-GHR induced by deleterious mutations. One significant effect of these mutations is the disconnection between GH and R1, new communication routes are formed at second binding site that are locally communicating between GH and few residues of R2. In addition, the study of direct interaction between GH and its receptors are not necessarily the sign of communication. Moreover, in MU^{R183H} all four helices of GH communicate together, while in WT and MU^{L124R} they are splitted in two blocks. COMMA provides hints on how the mutation affects the dissociation of the GH-GHR and enables us to detect the key pathways on the structure of the WT and MUs.

COMMA detected large regions that form cliques in receptor, which overlap with pathway-based blocks to some extent. This observation may indicate that the automatic

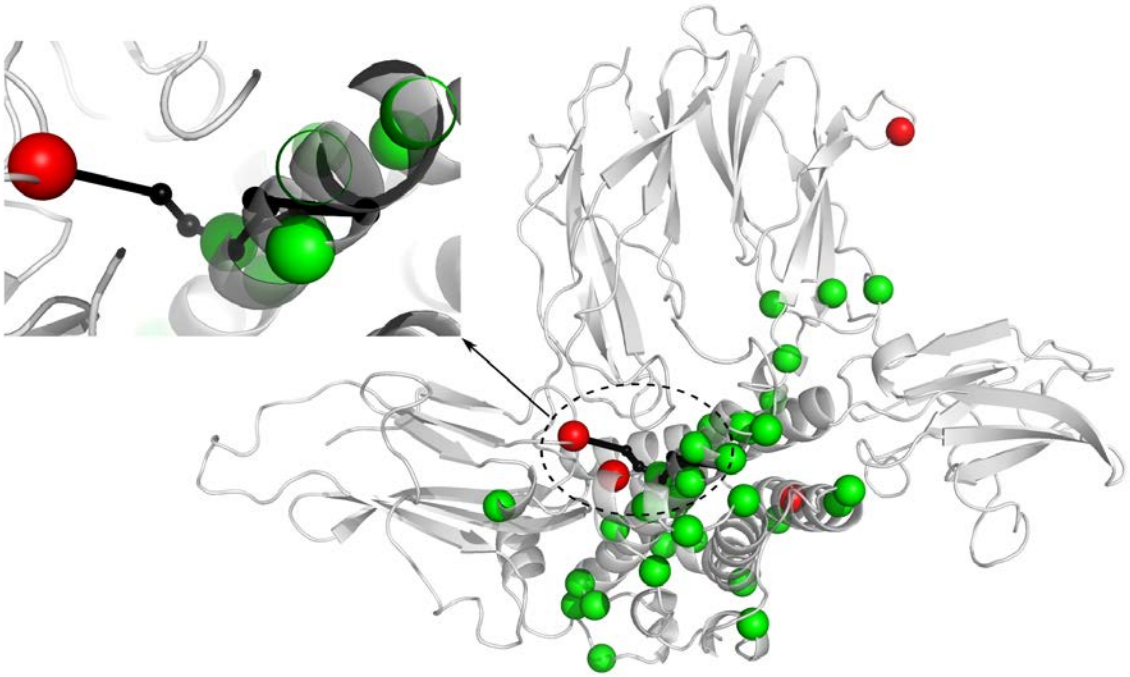

Figure 4.15: **Pathways connecting coevolving residues.** Coevolving residues that belong to clusters 1 and 2 are shown in spheres and colored in green and red, respectively. Pathways that communicate at site1 are shown by black lines, they connect the two clusters.

set up of the thresholds is not adapted for this system and that the definition of the different types of blocks is somewhat unclear. Possibly COMMA has to be be adapted to the study of complexes. One solution might be to analyze the complex piece by piece and not the entire complex.

Coevolving residues of GH and GH-GHR complex were detected using BIS and the link between those residues and pathways detected by COMMA were studied. The analysis of the complex revealed that, the only two detected pathways at the binding site 1, connecting GH to R1, also connect coevolving residues. This observations highlights the importance of joint analysis of coevolution and dynamics, in order to find key residues at the binding sites of GH-GHR complex.

Chapter 5

Protein infostery

Contents

In this chapter we propose a method to detect deleterious mutational hotspots and to characterize residue positions that are beneficial for the function of protein. We identify, by studying conformational dynamics, regions in the protein that are crucial for spreading information within the protein structure and define these regions formally. The new concept of infostery, from 'info' - information - and 'steric' - arrangement of residues in space - is introduced.

5.1 Background

The question of how amino acid sequence variations (re-)shape the conformational landscape of proteins and impact their function is one of outstanding importance in biology. Yet, it is far from being resolved. On the one hand, systematically assessing the phenotypic outcomes of protein sequence changes is very challenging both experimentally and computationally. This is due in part to the combinatorial explosion arising from considering all possible substitutions for single to multiple-point mutant variants. Another difficulty resides in the design of the experiment: what should be measured as phenotypic outcome? Disease-associated mutations can impair protein function in various ways, either by destabilizing the structural stability of the protein, or by shifting the equilibrium of conformation populations, or by modulating the binding affinity of the protein for its cellular partner(s), to name a few. These effects are difficult to probe directly and unambiguously. *In vitro* measurements might not be pertinent in the cellular context while *in vivo* measurements may hide multiple, possibly compensatory, effects. Computational techniques such as molecular dynamics (MD) simulations provide mechanistic details and can lead to very accurate free energy estimations. However, they are very limited in terms of conformational sampling, they do not take into account the cellular environment, and they are generally not applicable to study non-equilibrium processes.

On the other hand, the unprecedented breadth of data now accessible through deep sequencing is not always obvious to interpret in terms of protein structure and function. Conserved residues are generally important for the function of of protein and known to be involved in the interactions between proteins and biomolecules (Lichtarge and Wilkins, 2010; Engelen et al., 2009; Lichtarge et al., 1996). Exploiting the signals of evolutionary covariation has important applications, among which are predicting native contacts within protein structures (Morcos et al., 2011), inter-molecule interactions (Champeimont et al., 2016) and intramolecular allosteric communication (Sung et al., 2016).

Dynamical changes at one site induce local perturbations, along with long-range conformational alteration that is known as the allosteric coupling in protein. In addition, allostery can impact distant sites upon changes in complexes or binding to a ligand. For example, the previous analysis of sequence coevolution demonstrated the presence of evolutionary networks, induced by statistically coupled residues in PDZ domain, that may be important for the allostery (Lockless and Ranganathan, 1999). Those positions are located at binding site and other places on the structure, forming a long-range interaction network.

The term PDZ domain, short for (PSD-95, Discs-large, ZO-I) was first introduced by (Kennedy, 1995) and it represents the name of three first proteins that were shown to share PDZ domain. These proteins are: 1) post-synaptic density protein 95 (PSD-95), which

is a synaptic protein found only in the brain, 2) Drosophila disc large tumor suppressor (Dlg1) and 2) zona occludens 1 (ZO-1) that both play an important role at junctions and in cell signalling complexes.

The most important function of PDZ domains, is described as binding to the C-terminal ends of interacting partners, while participating in signal transduction mechanisms or acting as scaffolding element are mentioned in the literature. Although the sequence is variable among different PDZ domains, the fold is conserved. As an illustration, in PDZ2 that is one family member of PDZ domain, peptide binding results in dynamical and structural regulations of loop regions that are far from the binding site. Whereas, peptide binding in PDZ3 (another family member of PDZ domain), does not induce strong structural and dynamical changes at regions either close of far form the binding site (Papaleo et al., 2012).

PDZ3, is the third PDZ domain of the very well documented brain synaptic protein PDS-95. PDZ3 links to cysteine-rich PDZ-binding protein (CRIPT), which allows PSD-95 to associate with the cytoskeleton. The cognate ligand of PDZ3, is the C-terminal peptide derived from CRIPT (TKNYKQTSV).

Recently developed technologies, commonly designated as deep mutational scanning, enable to estimate the functional consequences of every possible single amino acid change at every position in a protein (Fowler and Fields, 2014). Such scanning was applied to a PDZ domain in cellular context (McLaughlin et al., 2012). In (McLaughlin et al., 2012), the third PDZ domain of the very well documented brain synaptic protein PDS-95 (PDZ3) was used as a model system. The experiment consisted in systematically measuring the effect of single-point mutations on the association of PDZ3 to its cognate ligand, the C-terminal peptide derived from CRIPT (TKNYKQTSV). Based on these measurements, McLaughlin and co-authors (McLaughlin et al., 2012) showed that there was a good overlap between the set of 20 positions displaying the highest sensitivity to mutation (highest impairment of ligand binding, averaged over all possible substitutions) and a physically contiguous network of coevolving residues detected from a multiple sequence alignment of PDZ homologs.

In the present study, we exploit these experimental data to explore the sequence-structure-dynamics-function relationship. First, we show that most of the highly deleterious positions can be detected based on conservation only. We also propose a score derived from sequence analysis and structural information to predict the phenotypic outcomes of the mutations. Second, we demonstrate that pertinent information can be extracted from the structural dynamics of the wild-type PDZ3-CRIPT peptide complex to identify the highly deleterious positions with very high accuracy and provide a physical interpretation of their sensitivity to mutations. Moreover, we propose a protocol to predict the effects of specific amino acid substitutions and show that it enables to distinguish neutral and gain-of-function mutations from deleterious ones. Our approach is based on COMMA (*Chapter 2, Methods*), that is a method to describe and compare the dynamical architectures of different proteins or different variants of the same protein. COMMA goes beyond classical analyses of the behavior of proteins in solution and beyond classical descriptions of proteins based on domains and secondary structures. Specifically, it extracts dynamical properties from conformational ensembles to identify *communication pathways*, *i.e.* chains of residues linked by stable interactions that move together, and *communication cliques*, *i.e.* clusters of residues that fluctuate in a concerted way. Pathways and cliques

PDB code	residue coverage	resolution (Å)
1BE9	302-430	1.82
1BEE	302-402	2.3
1TP3	302-402	1.99
1TP5	302-402	1.54
1TQ3	302-402	1.89

Table 5.1: Crystallized structures of PDZ3.

are used to define communication blocks, which do not necessarily correspond to domains or groups of secondary structure elements. The power of the method is illustrated on **Figure 5.1** for the PDZ3-CRIPT peptide complex. While the average MD conformations of the wild-type form and of two deleterious mutants are indistinguishable (**Figure 5.1a**), COMMA revealed that the communication within the mutants is characterized by more numerous and longer pathways (**Figure 5.1b**) and more highly connected residues (**Figure 5.1c**).

5.1.1 Previous MD simulations of PDZ domain

Crystallized structures for PDZ3 domain are compared in table 5.1 and 1BE9 is shown to have both the highest coverage of the residues and low resolution.

Significant number of the publications on PDZ3, report MD simulations on the structure of 1BE9 (Tiwari and Mohanty, 2013; Kalescky et al., 2014; Murciano-Calles et al., 2014). In a recent study (Kalescky et al., 2014), authors performed MD simulations using CHARMM on 1BE9 with and without the CRIPT ligand (bound and unbound simulations). They have also performed rigid-body MD simulations and proposed this method as a systematic approach to study the effect of every single residue on the dynamics of the whole system and identify key allosteric residues.

In (Murciano-Calles et al., 2014) authors performed MD simulations on 1BE9 using CHARMM22 forcefield and NAMD to study the interaction between the PDZ3 and the consensus hexapeptide KKETAV, the highest affinity binding partner of PSD-95. Authors exploited post-translational modifications methods and claimed that interplay of salt bridges between H3 and L23 has high effect on the binding affinity of PDZ3. Their finding highlights the role of H3 that is only present in PDZ3 among all PDZ domains. Additionally the analysis of the binding to the two different ligands, revealed a similar behaviour in terms of the chemical shift dispersion between the protein and the ligand.

In addition, the dynamical changes of PDZ domains were studied upon binding to the ligand. Two different members of PDZ domains are tyrosine phosphatase PDZ2 and PSD-95 PDZ3. Previous studies revealed that upon binding to the ligand, PDZ2 undergoes strong dynamical changes (Gianni et al., 2011), whereas PDZ3 display no significant structural changes at both ligand binding and distal sites (Chi et al., 2008). The structure of PDZ3 was experimentally determined in two different forms, truncated (*delta*) form and full length. In the short form, the N-terminal first 10 residues and the C-terminal helix, H3 and β sheet are lacking.

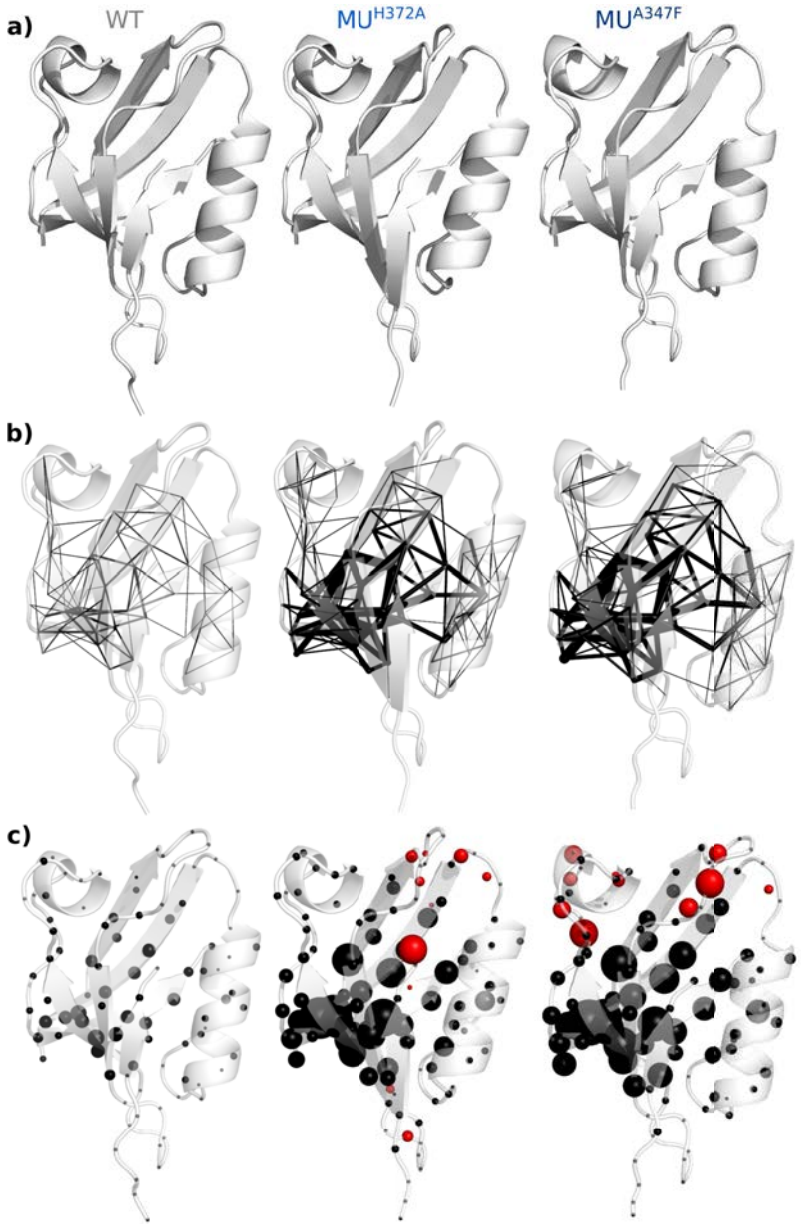

Figure 5.1: **COMMA analysis for PDZ3-CRIPT peptide complex. (a)** Conformations averaged over 5 replicates of 15-ns MD simulations of the wild-type complex (WT) and two deleterious mutants (MU^{H372A} and MU^{A347F}). **(b)** Communication pathways (longer than 3 residues) detected by COMMA are mapped onto the averaged conformation and displayed as black lines. The thickness of each segment is proportional to the number of pathways linking the two residues. **(c)** The residues crossed by at least one communication pathway (longer than 3 residues) are displayed as black spheres, centered on their C-α atoms. The size of each sphere is proportional to the number of pathways crossing the residue.

The allosteric behaviour of these two PDZ domains were studied in (Morra et al., 2014). Authors performed MD simulations of 400ns for PDZ2 (PDB: 3LNY) in complex with RAGEF2 C-terminal peptide, PDZ3 bound state (PDB:1BE9) in complex with CRIPT (sequence: KQTSV), PDZ3 unbound state (PDB: 1BFE) and truncated PDZ3 in both bound and unbound states. The study of atomic fluctuations revealed the overall

higher flexibility in the unbound state of PDZ2, whereas significantly few changes of fluctuations were reported for PDZ3 upon binding. Authors reported stronger interactions of H372 with the ligand in PDZ3 compared to PDZ2. Authors exploited an energy decomposition method to extract the essential interactions responsible for the stability of the protein. The comparison of the energy eigenvector profile obtained by energy decomposition method, revealed the higher impact of ligand binding in PDZ2 compared to PDZ3. One hypothesis for the observed difference, points to their structural differences (additional C-terminal helix H3 and β sheet compared to PDZ2). Their analysis of fluctuations, rearrangement of the binding site and energy profiles led to the conclusion that the truncated PDZ3 has an intermediate dynamic behavior between PDZ2 and PDZ3. Allosteric regions were detected as residues with the energy anti-correlated fluctuations. Furthermore the higher stability of PDZ3 is reported due to the existence of H3 with the most stabilizing energy.

5.1.2 Experimental data

Previously in (Lockless and Ranganathan, 1999), authors defined sector as a group of coevolving amino acids that are crucial for the structure and function of the protein. The sector in PDZ3, connects the ligand-binding pocket with an allosteric site on the opposite surface. On the other hand, the same authors performed high-throughput experiments to study the complete mutational landscape of PDZ3 (McLaughlin et al., 2012). In this work, every residue of PDZ3 was mutated to every other possible amino acids to perform the deep sequencing experiment, where they quantitatively measured the ability of the PDZ3 to bind its cognate ligand. Mutation-induced changes in binding affinity were indirectly estimated by measuring the frequencies of mutated alleles in a bacterial population where cells were classified based on their content of PDZ3-CRIPT peptide complex (assessed by eGFP levels). The authors showed that there was a good overlap between the set of positions displaying the highest sensitivity to mutation (highest impairment of ligand binding, averaged over all possible substitutions) and a physically contiguous network of coevolving residues (residues detected previously as sector). From their analysis, 20 residues out of 83 bring loss of function and their functional effect are greater than 2σ. On the other hand, 15 amino acids out of those 20 residues are sector positions (sensitivity of 75%).

Among those hotspot positions, a sub set of positions (L323, F325, I327 and L379) tolerate substitutions only to the most chemically conservative amino acids and a set of buried residues positioned outside the direct spatial environment of the ligand (G329, G330, I336, A347, L353, V362 and A375), show significant sensitivity to mutation. The largest average mutational effect comes from position G329 and H372, which tolerates essentially no other substitution. Authors mentioned that Proline is the most unfavorable substitution from their analysis, followed by Asp, Glu, Lys, Arg and by tryptophan that hast he largest side chain in terms of volume. The least average perturbations are induced from the substitutions to Ala and Cys.

5.1.3 Choice of mutation

PDZ3 in complex with the CRIPT ligand (PDB code: 1BE9) was considered as the studied structure. In addition to the wild-type PDZ3, three subset of its mutants with different mutational effects (loss-of-function, neutral and gain-of-function mutants) were selected based on the experiment of (McLaughlin Jr et al., 2012). The functional cost of the selected positions is reported (Table 5.2).

position	average ΔE	mutation	ΔE
329	-1.102	G329A	-1.4
347	-0.21	A347F	-1.4
372	-1.08	H372A	-1.34
341	-0.03	I341A	-0.6
325	-0.55	F325A	-0.03
371	0.02	S371A	0
366	0.05	D366A	0.07
311	0.22	P311W	0.3

Table 5.2: **Studied mutations.** The average ΔE and ΔE per mutation is reported for the studied mutations.

Loss-of-function mutants

The following four mutants with deleterious mutational effects were chosen: G329A, A347F, H372A and I341A. Residues at positions 329, 372 and I341 were shown to interact with the ligand, whereas 347 is far from the neighbouring proximity of the ligand. 347 is detected as switch residue, while 372 detected as wire residue (Kalescky et al., 2014). Position 372 determines the ligand specificity and 347 is shown to be coupled to key residue H372 (Lockless and Ranganathan, 1999).

Neutral mutants

A set of three mutations with neutral functional effects were selected: F325A, S371A and D366A. Both positions of 325 and 371 are within the close proximity of the ligand, whereas D366A is positioned far from the ligand, on the other side of the structure.

Gain-of-function mutants

P311W has positive mutational effects (gain of function) and is placed far from the neighbouring proximity of the ligand.

5.2 Molecular dynamics simulations

The following molecular dynamics protocol was applied to all studied systems.

Set up of the systems. The 3D coordinates of PDZ3 in complex with its cognate ligand, a C-terminal peptide derived from CRIPT, were retrieved from the Protein Data Bank (Berman et al., 2000) (PDB code: 1BE9, residues 302 to 430, 1.82 Å resolution (Doyle et al., 1996)). All crystallographic water molecules and other non-protein molecules were removed. The CRIPT peptide (sequence: TKNYKQTSV) is truncated in the PDB structure (sequence: KQTSV). The missing residues and side chains were modeled using MODELLER 9v7 (Marti-Renom et al., 2000). The mutated forms of PDZ were generated by *in silico* substitutions using Rosetta Backrub (Smith and Kortemme, 2008). All systems were prepared with the LEAP module of AMBER 12 (Case et al., 2012), using the ff12SB forcefield parameter set: (*i*) hydrogen atoms were added, (*ii*) the solute was hydrated with a cuboid box of explicit TIP3P water molecules with a buffering distance up to 10Å, (*iii*) Na^+ and Cl^- counter-ions were added to reproduce physiological salt concentration (150 nM solution of potassium chloride). PDZ domain contains 2 histidines, whose protonation states were determined so as to locally optimize the hydrogen-bond network: (*i*) a hydrogen was assigned to the ϵ-nitrogen of H317 and (*ii*) hydrogen was assigned to the δ-nitrogen of H372.

In total, 15 models were built and simulated (**Table 5.3**): (**1-9**) wild-type and all mutated forms of full-length PDZ3 (residues 302 to 430) in complex with the ligand, (**10-11**) wild-type and H372A-mutated full-length PDZ3 (residues 302 to 430) in free form (ligand removed), (**12-13**) wild-type and H372A-mutated truncated PDZ3 (residues 311 to 393) in complex with the ligand, (**14-15**) wild-type and H372A-mutated truncated PDZ3 (residues 311 to 393) in free form (ligand removed).

Minimization, heating and equilibration. The systems were minimized, thermalized and equilibrated using the SANDER module of AMBER 12. The following minimization procedure was applied: (*i*) 10,000 steps of minimization of the water molecules keeping protein atoms fixed, (*ii*) 10,000 steps of minimization keeping only protein backbone fixed to allow protein side chains to relax, (*iii*) 10,000 steps of minimization without any constraint on the system. Heating of the system to the target temperature of 310 K was performed at constant volume using the Berendsen thermostat (Berendsen et al., 1984) and while restraining the solute C_α atoms with a force constant of 10 $kcal/mol/Å^2$. Thereafter, the system was equilibrated for 100 ps at constant volume (NVT) and for further 100 ps using a Langevin piston (NPT) (Loncharich et al., 1992) to maintain the pressure. Finally the restraints were removed and the system was equilibrated for a final 100 ps run.

Production of the trajectories. 5 (models 1-9) or 2 (models 10-15) replicates of 20 ns (**Table 5.3**), with different initial velocities, were performed for each system. The simulations were realized in the NPT ensemble using the PMEMD module of AMBER 12. The time step was set to 2.0 fs. The temperature was kept at 310 K and pressure at 1 bar using the Langevin piston coupling algorithm. The SHAKE algorithm was used to freeze bonds involving hydrogen atoms, allowing for an integration time step of 2.0 fs. The Particle Mesh Ewald (PME) method (Darden et al., 1993) was employed to treat long-range electrostatics. The coordinates of the system were written every ps. Standard analyses of the MD trajectories were performed with the *ptraj* module of AMBER 12.

Model number		Number of replicates	Waterbox dimensions (in Å^3)
	Complex, full length (124 res.)		
1	WT	5	$72 \times 68 \times 58$
2	P311W	5	$72 \times 68 \times 58$
3	D366A	5	$72 \times 68 \times 58$
4	S371A	5	$71 \times 68 \times 58$
5	F325A	5	$72 \times 68 \times 58$
6	I341A	5	$72 \times 68 \times 58$
7	H372A	5	$72 \times 68 \times 58$
8	G329A	5	$72 \times 68 \times 58$
9	A347F	5	$72 \times 68 \times 58$
	Free PDZ, full length (115 res.)		
10	WT	2	$72 \times 68 \times 58$
11	H372A	2	$72 \times 68 \times 58$
	Complex, truncated form (92 res.)		
12	WT	2	$56 \times 67 \times 55$
13	H372A	2	$55 \times 67 \times 55$
	Free PDZ, truncated form (83 res.)		
14	WT	2	$56 \times 67 \times 55$
15	H372A	2	$55 \times 59 \times 55$

Table 5.3: **Computational details of performed MD simulations.** The duration of each replicate was 20 ns.

Stability of the trajectories. To assess the stability of the PDZ domain and its mutants, the all-atom root mean square deviation (RMSD) was recorded along each 20-ns MD simulation replicate (**Figure 5.2** and **5.3**). The simulations of full-length PDZ3 in complex with its ligand (models 1-9) are very stable (**Figure 5.2**). The PDZ domain (residues 311-393) deviates by 1.58-2.59 Å from the equilibrated structure, on average. The CRIPT peptide remains in interaction with the protein all along the simulations, at a minimum distance lower than 2 Å. The conformational drift of the ligand (RMSD values in the range 1.99-6.78 Å) is mainly due to the N-terminal residues (numbered -8 to -5). No significant difference can be observed between the wild-type and mutated forms of the full-length complex (**Figure 5.2**). The PDZ domain also shows high stability in the simulations of the models 10-15 (**Figure 5.3**). The average deviations of wild-type and H372A-mutated PDZ domain are in the range of 1.65-2.67 Å for full-length free form (models 10-11), 1.85-2.41 Å for truncated complexed form (models 12-13) and 1.88-2.62 Å for truncated free form (models 14-15). All systems are fully relaxed after 5 ns. Consequently, the last 15 ns of each replicate were retained for subsequent analyses.

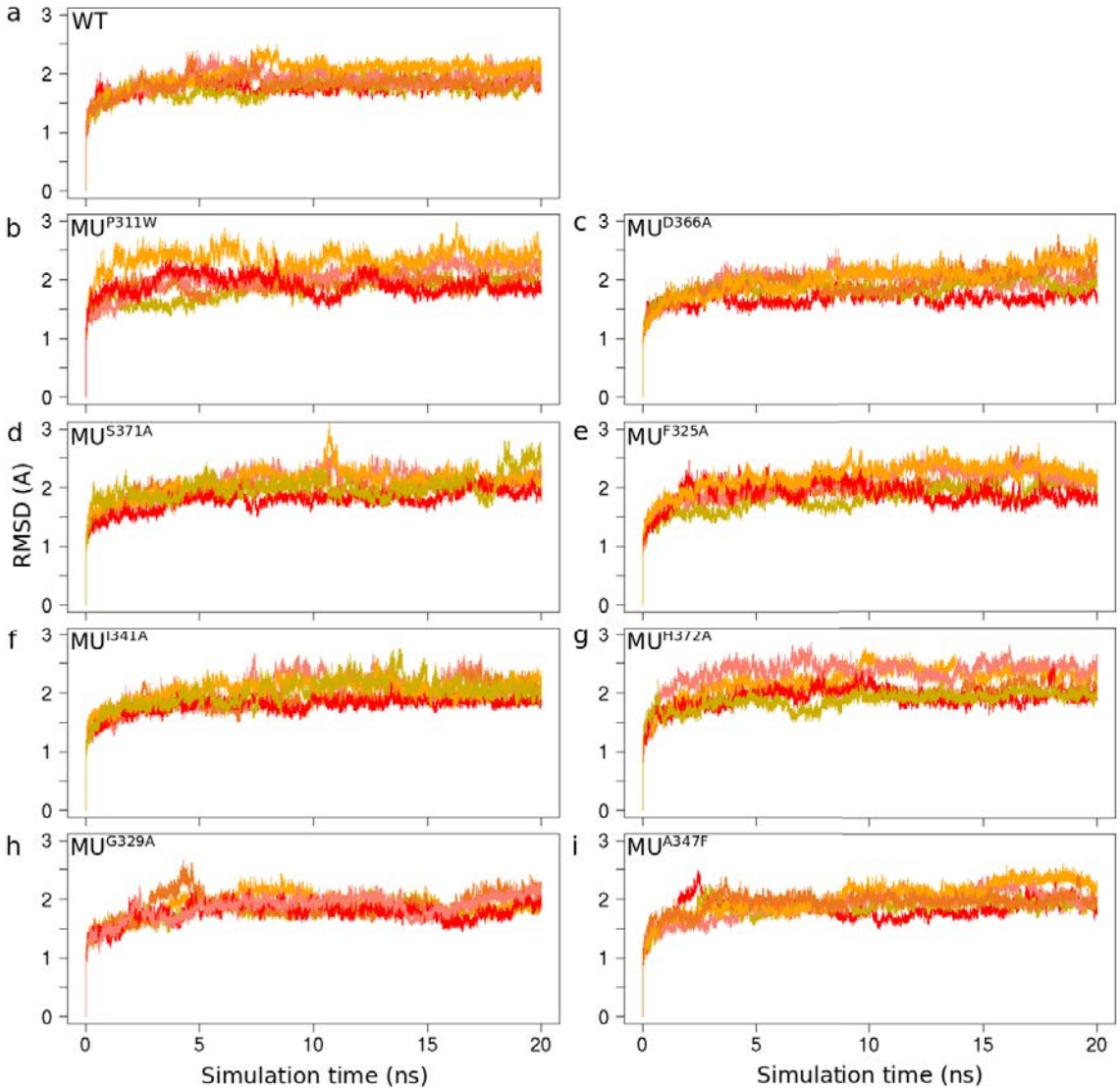

Figure 5.2: **Root mean square deviation for PDZ domain and its mutants.** All-atom RMSD recorded along the simulations of the full-length PDZ3 domain (residues 302-430) in complex with the CRIPT peptide (models 1-9). The initial frame is taken as the reference, for each replicate. The RMSD is computed on residues 311 to 393. Each curve is colored differently and corresponds to one replicate: **(a)** WT, **(b)** MU^{P311W} , **(c)** MU^{D366A} , **(d)** MU^{S371A}, **(e)** MU^{F325A}, **(f)** MU^{I341A}, **(g)** MU^{H372A}, **(h)** MU^{G329A} and **(i)** MU^{A347F}.

Fluctuations and secondary structures. The by-residue RMS fluctuations of the PDZ domain (residues 311 to 393) were computed for all MD simulations. The fluctuations were measured with respect to the average structure in every replicate. The simulations of full-length PDZ3 in complex with its ligand (models 1-9) displayed fluctuations within the range of 0.5Å-4.54Å. The atomic fluctuations profiles are very similar for models 10-15 (between 0.5Å and 3.95Å). Secondary structures were assigned with DSSP (Kabsch and Sander, 1983). They are very stable along all MD simulations.

Solvent accessibility. The solvent accessible relative surface areas (rsa) of the residues in the wild-type PDZ3 and the mutants were recorded using NACCESS (Hubbard and

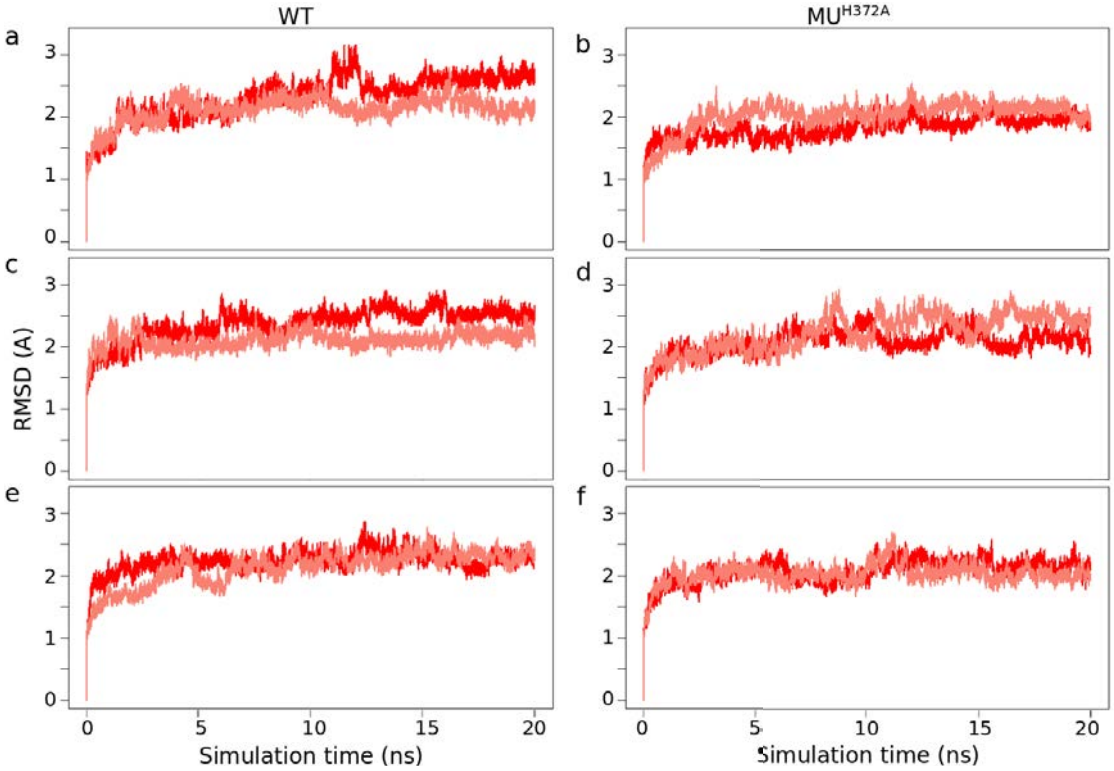

Figure 5.3: **Root mean square deviation for wild-type and H372A-mutated PDZ domains.** All atom RMSD recorded along the simulations of: **(a)** full-length WT in free form (model 10), **(b)** full-length MU^{H372A} in free form (model 11), **(c)** truncated WT in free form (model 14), **(d)** truncated MU^{H372A} in free form (model 15), **(e)** truncated WT in complex with the ligand (model 12), **(f)** truncated MU^{H372A} in complex with the ligand (model 13), The RMSD is computed on residues 311 to 393. Each curve is colored differently and corresponds to one replicate.

Thornton, 1992-6) along the MD simulations. For each residue, the *rsa* was averaged over the replicates of 15-ns MD simulation. The set of residues with $0\% \leq RSA \leq 25\%$ are considered as buried residues. Moreover the RSA was also measured over the X-ray structure as a reference.

5.3 COMMA analysis

COMMA was applied to extract communication blocks of PDZ3 in complex with its cognate ligand. As described above, classical analyses (RMSD, RMSF, secondary structure...) of the MD simulations showed that the truncated and free forms of PDZ3 (models 10-15) behave very similarly to the full-length PDZ3 in complex with its cognate ligand. Consequently, we focused on the latter to study and predict the impact of single-point mutations, and we applied COMMA on the last 15 ns of every replicate (75,000 conformations) of models 1-9.

5.3.1 The confidence for COMMA blocks

COMMA uses a number of thresholds that are automatically set depending on the system studied (*Chapter 2, Methods*). In order to assess the confidence of communication blocks (both pathway-based and clique-based), sensitivity of the residues were analysed when considering different sets of thresholds. The communication propensity threshold, CP_{cut}, used to define pathways was varied from 60% to 80% quantile by 5%. Consequently COMMA was applied to defined CBs^{path} with different CP_{cut}. Residue sensitivity to be included in those blocks was measured. The pathway-based confidence value represents the number of times a given residue was included in CB^{path}, divided by the total number of CP_{cut} considered (5 values).

The same procedure was applied to assess the confidence of CBs^{clique} by changing the correlation threshold, $Corr^{LFA}_{cut}$, used to delimit protein regions of concerted atomic fluctuations. COMMA was applied with different $Corr^{LFA}_{cut}$, changing from 85% to 95% quantiles by 1% to measure the residue sensitivity to be included in CBs^{clique}. The clique-based confidence was assigned to residues based on the number of times they are detected in CB^{clique} divided by the total number of $Corr^{LFA}_{cut}$ (11 values).

5.3.2 Algorithm for picking up isolated *direct contacts* with varying thresholds

All pairs of communicating residues are shown on a dot plot matrix. Different types of communication are highlighted by colors on this matrix, where black dots correspond to direct communication between every pair, gray represents the set of communicating residues that their distance along the sequence is within the range of 4 residues and other colors (such as red and magenta) are used to highlight indirect communications between pairs of residues and the selected color corresponds to different communication blocks that they belong to.

Isolated direct contacts are one or group of neighbouring residue pairs that communicate directly. Residues in every pair of isolated direct contacts, are far along the sequence and belong to different secondary structures. In few cases groups of isolated dots are surrounded by at most 4 other indirect communicating pairs, those are few exceptions considered in the analysis. Isolated dots represent the set of important residues on the structure that link different secondary structures. There are only few short-length pathways crossing through them. Hence they are not surrounded by indirect communications on the matrix, as well as on the structure. Therefore the whole functionality/stability of the structure depends on the pathways crossing them. Their absence can cause negative effects on the flow of signals along the structure.

The communication propensity threshold, CP_{th}, used to define pathways was varied from 60% to 80% quantile by 5%. COMMA was applied by using varied CP_{th} to detect the set of isolated direct contacts. A network was constructed from the set of all detected isolated direct contacts, where nodes correspond to residues involved in such contacts and edges connect the residues in every pair. A weight was assigned to each edge that account for the number times a given contact was observed in WT or MUs. Furthermore a filter was applied in order to remove all the exposed residues. At the end, single nodes that are not linked to any other nodes, were removed from the network. Connected components

of the network were defined and represented on the structure by different colors.

5.4 Dynamical architecture of PDZ-CRIPT peptide complex

In this section, we explored the dynamical behavior of the PDZ3 bound to its cognate ligand for the WT and all 8 studied MUs.

5.4.1 Wild-type complex

The complex between PDZ3 (residues 301 to 415) and the C-terminal CRIPT peptide (TKNYKQTSV) was simulated in explicit solvent (five replicates of 20 ns) and the trajectories were analyzed using COMMA. In the complex, the ligand forms an additional β-strand in the groove between the β-strand S2 (residues 325 to 330) and the helix H2 (residues 372 to 380) of PDZ3 (**Figure 5.4a**). COMMA identified 2 pathway-based and 4 clique-based communication blocks (CBs). CBs partition the protein structure according to the way information is transmitted across it. As one can observe (**Figure 5.4b-c**), they do not correspond to domains but go across domains. They also go beyond chains: pathway-based and CBs^{clique} contain residues of the PDZ domain and of the ligand.

On the one hand, residues in a CB^{path} (CB^{path}) are linked by stable non-covalent interactions and move together. The biggest CB^{path} (**Figure 5.4b**, in red) encompasses most of the 2 β-sheets of the PDZ domain (42 residues) and the C-terminal residues -2 and -1 of the ligand. The second CB^{path} (8 residues, **Figure 5.4b**, in pink) corresponds to the helix H2. On the other hand, residues in a CB^{clique} (CB^{clique}) display high concerted atomic fluctuations (*Chapter 2, Methods*). The major part of the ligand binding pocket is involved in a CB^{clique} (**Figure 5.4c**, in blue). Another CB^{clique} (in marine) comprises helix H1 (residues 341-350) and one residue, G324, from S2, which is in interaction with the ligand. The N-terminal residues (-8 to -3) of the ligand form a CB^{clique} (in purple) with the loop L2 from PDZ. Finally, the smallest CB^{clique} (in slate) is comprised by residues 363-366 from loop L5.

To assess the robustness of the results, we considered a consensus over different values of the thresholds used to detect CBs (see *COMMA analysis*). The propensity (between 0 and 1) of each residue of the complex to be detected in a CB was computed (**Figure 5.4**, size of the sausage). 85% of residues detected in CBs^{path} and 98% of those detected in CBs^{clique} have a propensity value of 1. This indicates that the confidence in the detection is high. We also simulated the free PDZ3 domain, without the ligand. The propensities of the protein residues in the free form to be in a CB^{path} or CB^{clique} are very similar to those computed for the complex.

The organisation of the way information is transmitted across the PDZ3-CRIPT peptide complex revealed that the ligand is almost fully integrated in the communication of the PDZ domain and can be divided in two parts: the C-terminal residues move with the 2 β-sheets of PDZ as a rigid body while the N-terminal residues are more flexible and fluctuate in concert with the loop L2. It also revealed that the PDZ residues encircling the ligand do not move or fluctuate all together, but relate to 2 different protein regions (colored in red/pink on **Figure 5.4a** and marine/blue on **Figure 5.4b**) .

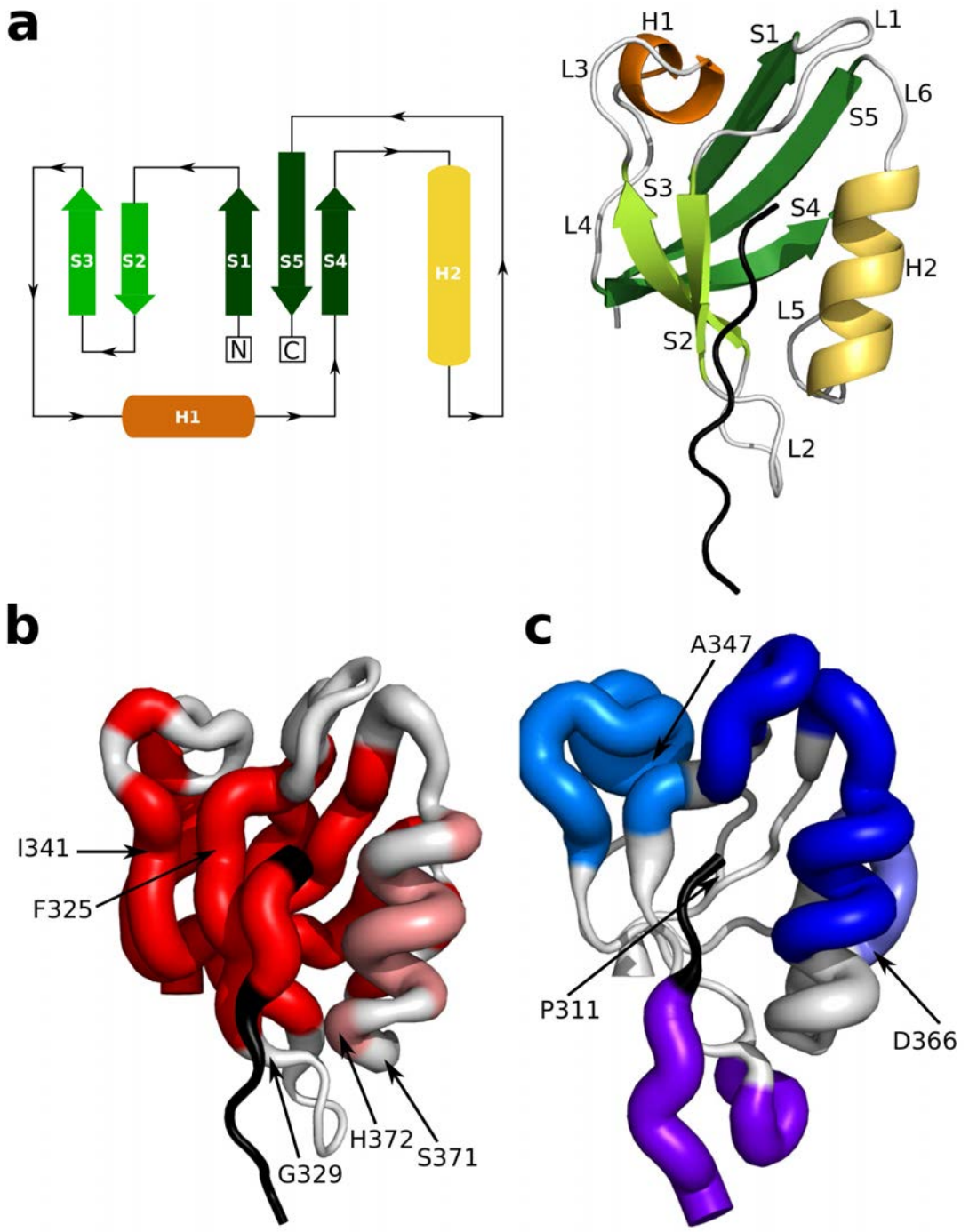

Figure 5.4: **Communication blocks (CBs) identified by COMMA in wild-type PDZ.** The protein is represented as a cartoon. **(a)** 2 CBs^{path} are detected, colored in red and magenta. **(b)** 4 CBs^{clique} are detected, colored in different blue tones. The size of the sausage reflects the propensity of each residue to be detected in a CB. The residues whose substitutions were studied are labelled.

5.4.2 Mutant complexes

We studied the impact of 8 mutations whose effects were reported in (McLaughlin et al., 2012): P311W (beneficial), D366A, S371A and F325A (neutral), I341A, H372A, G329A

and A347F (deleterious). They were chosen so as to span different locations in the PDZ domain (**Figure 5.4** and **Figure 5.5**) and to represent different mutational outcomes (**Table 5.4**). MD simulations of the mutated complexes (5 replicates of 20 ns each) were performed and classical analyses did not reveal any drastic changes in their structures or movements (**Figure 5.1a** and **5.6a**). They displayed RMS deviation profiles similar to that of the wild type (**Figure 5.2**) and their secondary structures remained stable. By contrast, COMMA analysis revealed striking differences in the communication of the mutants compared to the wild type (**Figure 5.1b-c**, **5.6b-c** and **Figure 5.7**).

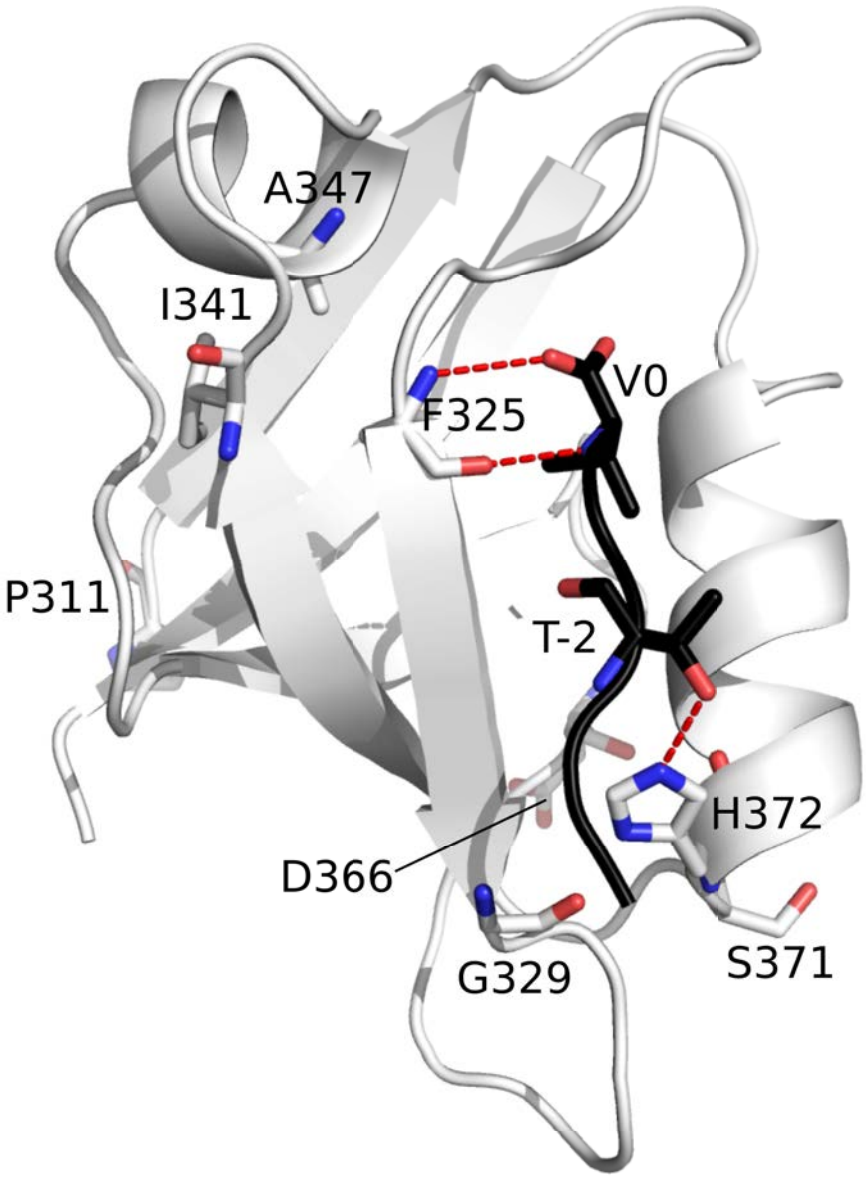

Figure 5.5: **Localization of the studied mutations in the PDZ3-CRIPT peptide complex.** The protein PDZ3 and the ligand (PDB code: 1BE9) are displayed as cartoons colored in white and black respectively. The residues whose mutations were studied and residues from the ligand with which they interact are shown as sticks. Hydrogen-bonds are indicated as dashed red lines.

Mutation	**experimental ΔE (in kcal/mol)** (McLaughlin et al., 2012)	Gain of pathways (> 3 residues)	Number of residues losing communications
P311W	0.31	753	21
D366A	0.07	780	11
S371A	0.02	766	10
F325A	-0.03	450	15
I341A	-0.64	1800	15
H372A	-1.34	2126	8
G329A	-1.36	1000	6
A347F	-1.42	2312	10

Table 5.4: **Studied mutations.** The experimental measurements are given, along with the number of communication pathways (> 3 residues) gained compared to the wild-type complex and the number of residues losing communications.

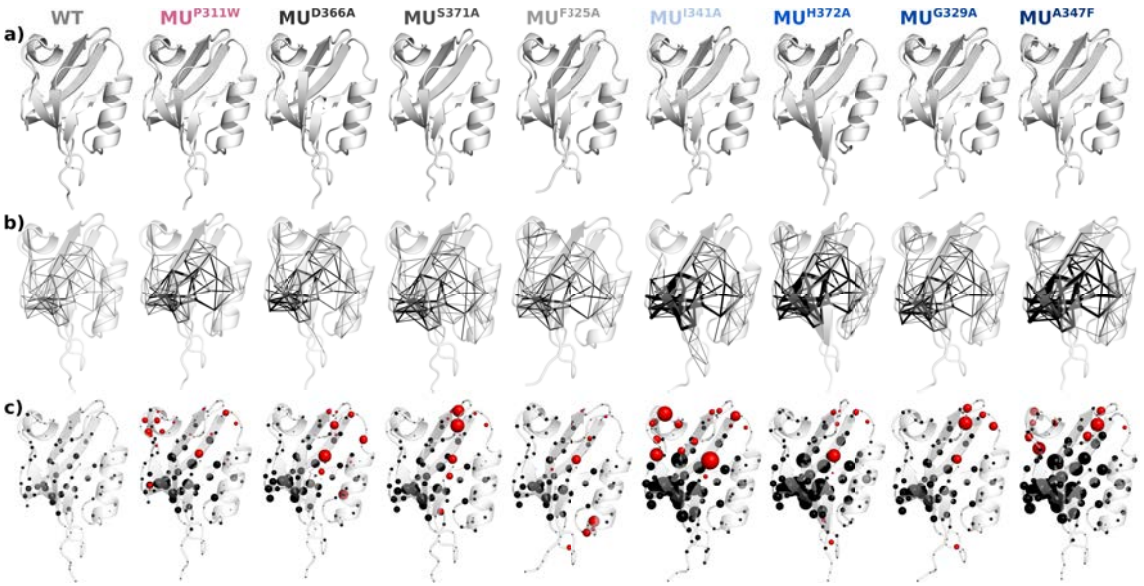

Figure 5.6: **COMMA analysis for PDZ3-CRIPT peptide complex. (a)** Conformations averaged over 5 replicates of 15-ns MD simulations of the wild-type complex (WT) and eight mutants. **(b)** Communication pathways (> 3 residues) detected by COMMA are mapped onto the averaged conformation and displayed as black lines. The thickness of each segment is proportional to the number of pathways linking the two residues. **(c)** The residues crossed by at least one communication pathway (> 3 residues) are displayed as black spheres, centered on their C-α atoms. The size of each sphere is proportional to the number of pathways crossing the residue.

5.4.3 Matrix of properties from COMMA analysis

The effect of mutations were measured based on ΔE values proposed for every substitution (McLaughlin et al., 2012). Here we measured different variations of ΔE to highlight those that are deleterious or beneficial for the ligand binding of PDZ3. First the following groups of physico-chemical classes were considered: 1) hydrophobic: $\{V, I, L, M, F, W, A, P, G\}$, 2) negatively charged: $\{D, E\}$, 3) positively charged: $\{K, R\}$, 4) polar: $\{C, Y, H, N, S, T, Q\}$, to introduce $mean^{*}\Delta E$. For every position that belong to one class, the average is mea-

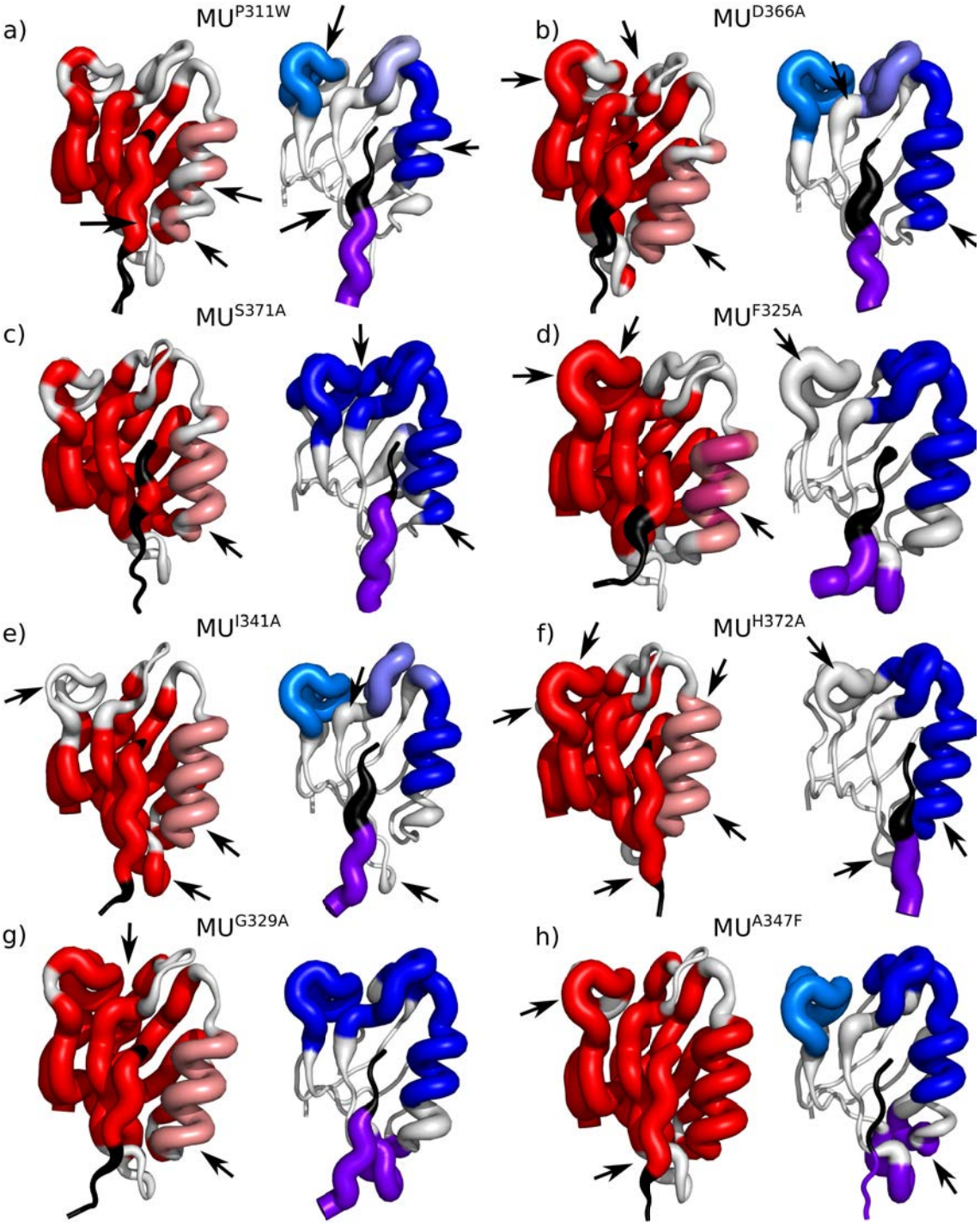

Figure 5.7: **Communication blocks identified by COMMA in PDZ mutants.** **(a)** MU^{P311W}, **(b)** MU^{D366A}, **(c)** MU^{S371A}, **(d)** MU^{F325A}, **(e)** MU^{I341A}, **(f)** MU^{H372A}, **(g)** MU^{G329A} and **(h)** MU^{A347F}. The differences with block detected in the wild-type form (**Figure 5.4**) are indicated by arrows.

sured over the substitutions that belong to other classes. Second, we defined $mean^{**}\Delta E$ that is similar to $mean^{*}\Delta E$ but the classes are different. Here we report the set of physico-chemical classes were considered: 1) hydrophobic: $\{V, I, L, M, F, W, A, P, G\}$, 2) negatively charged: $\{D, E\}$, 3) positively charged: $\{K, R\}$, 4) polar: $\{C, Y, H, N, S, T, Q\}$.

For every position on PDZ3 the following properties are measured from COMMA analysis of the WT and MUs and shown in form of a matrix (**Figure 5.8**):

- positions that belong to CBs^{path} of more than 3 residues and more than 6 residues
- position that belong to CBs^{clique}
- length of pathways
- number of pathways
- the number of residues that every position is connected to, through pathways
- number of direct contacts between loops
- number of direct contacts between loops and secondary structures (α helix or β strand)
- number of pathways connecting directly a protein residue to a residue on the the ligand
- the robustness of residues that belong to pathway-based and CBs^{clique}.
- Residues detected as sector (McLaughlin et al., 2012)
- the set of 4 clusters of coevolving residues in PDZ3 detected by MST (Baussand and Carbone, 2009)
- max $\Delta E(p)$ of (McLaughlin et al., 2012)
- max $\Delta E(p)$ of (McLaughlin et al., 2012), when ignoring the effect of substitutions to proline, because mutation to proline is the most unfavourable substitution
- the measured $mean\Delta E$ of (McLaughlin et al., 2012)
- the measured $mean^{*}\Delta E$
- the measured $mean^{**}\Delta E$
- the measured $mean\Delta E$ of (McLaughlin et al., 2012) when ignoring the effect of substitutions to proline
- $mean^{*}\Delta E$ when ignoring the effect of substitutions to proline
- $mean^{**}\Delta E$ when ignoring the effect of substitutions to proline
- differences of pathway lengths, number of pathways and number of contacts between WT and every MU.

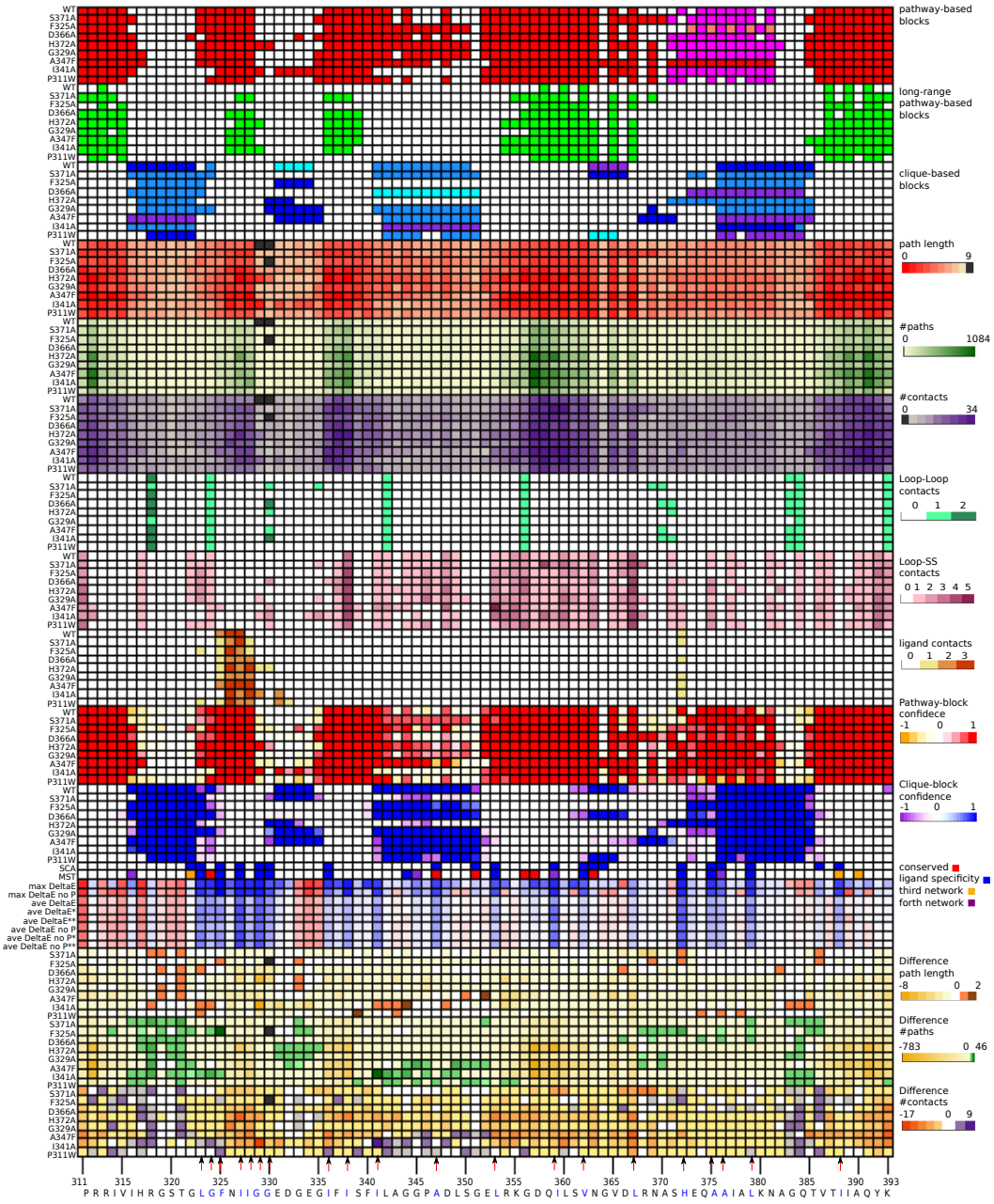

Figure 5.8: **PDZ matrix.** *mean*ΔE, *mean*$^{*}\Delta E$ and *mean*$^{**}\Delta E$ for every position is measured over the set of residues that do not belong to the same physico-chemical classes. The sum direct contacts between loops (white to green), loops and secondary structures (white to purple) and also contacts with the ligand (white to orange) are measures for every position in wild type and mutants. The newly defined *max*ΔE, *mean*ΔE, *mean*$^{*}\Delta E$ and *mean*$^{**}\Delta E$ while ignoring mutations to proline, are also shown here. The sum direct contacts between loops, loops and secondary structures and also contacts with the ligand are measures for every position in wild type and mutants. (The corresponding color code is from white to orange.)

5.5 Characterizing the effect of single mutations

The total number of communication pathways is increased in all mutants compared to the wild type (WT) complex (**Table 5.4**). The gain of pathways is higher for the deleterious mutations, A347F, H372A, G329A and I341A, than for the other ones. Where are located these additional paths? In all the mutants, the CBs^{path} cover a larger portion of the protein structure compared to WT (**Figure 5.9b**). This indicate that new residues are crossed by pathways in the mutants (**Figure 5.1b-c** and **5.1b-c**). Furthermore, the deleterious mutants display the largest numbers of residues (**Figure 5.9b**, in blue tones). This observation holds when considering subsets of pathways, defined by varying the minimum pathway length up to 6 residues.

The ensemble of CBs^{path} is also less fragmented in the mutant compared to the wild type (**Figure 5.7**). To quantify the differences between mutants and the wild type, first we measured the number of pathways (> 3 residues) that are crossing through every residue of the WT and MUs. **Figure 5.9a** represents number of pathways of the wild-type PDZ3 on the 3D structure.

Second we counted the number of residues comprised in CBs^{path} (**Figure 5.9b**). In all the mutants, the CBs^{path} cover a larger portion of the protein structure compared to WT. Furthermore, the deleterious mutants, A347F, H372A, G329A and I341A (in blue tones), display the largest numbers. This observation holds when considering subsets of pathways, defined by varying the minimum pathway length up to 6 residues.

Third, we defined a score that accounts for the total number of residues in CBs^{path} as well as the size of the largest CB^{path}:

$$S(MU) = \frac{\sum_i \#(X_i^{MU})}{\sum_i \#(X_i^{WT})} \frac{\max_i(\#(X_i^{MU}))}{\max_i(\#(X_i^{WT}))} \tag{5.1}$$

where $\#(X_i^{WT})$ (resp. $\#(X_i^{MU})$) is the number of residues comprised in the ith CB^{path} of the wild type (resp. the considered mutant). The largest and the less fragmented the CB^{path} ensemble in the mutant compared to the wild type, the higher the score. To compute the score, the CBs^{path} were obtained from inclusive subsets of pathways defined by using cutoff lengths ranging from 3 to 5 residues (**Figure 5.9d**). The scores are higher in the deleterious mutants (blue tones) compared to the other ones (gray tones and pink). The highest value (4.0) is reached by MU^{H372A} with pathways longer than 5 residues.

Last and forth, we counted the number of highly connected residues, *i.e.* residues crossed by >60, >80 or >100 pathways (**Figure 5.1c** and **5.9c**). There are more highly connected residues in all studied mutants compared to WT, with the deleterious mutants displaying the largest increases (**Figure 5.9c**). In MU^{A347F}, MU^{H372A} and MU^{G329A}, more than one third of the protein's residues are crossed by more than 100 pathways. This is more than twice as much as WT. The most highly connected residues in WT are also those displaying the highest gain of pathways upon mutations.

In all three plots reporting values for the three metrics considered here (**Figure 5.9**), the curves corresponding to deleterious mutations (in blue tones) are systematically above the other ones (in grey tones and in pink). Moreover, there is a clear distinction between the 2 groups: crossing between curves is observed within but not between groups. In addition, the variability between the neutral and gain-of-function mutations is smaller

than that observed between the deleterious mutations (pink and gray curves are closer to each other than the blue ones).

This analysis showed that although the single-point mutations did not drastically affect the structural dynamics of the PDZ-CRIPT peptide complex on a relatively short time scale, COMMA was able to highlight significant differences between the dynamical architectures of the mutants and that of the wild type. Specifically, the deleterious mutations induce a dramatic increase in the number and length of communication pathways, resulting in extended CBs^{path} and larger numbers of highly connected residues. This observation could be interpreted as an increased stiffness of the complex. Noticeably, the studied neutral and gain-of function mutants also show stiffened structures, although to a much smaller extent than the deleterious ones. This suggests that the structure of the complex has to adapt to any substitution introduced *in silico*, even if it has no experimentally measurable effect. The effects recorded on the neutral and gain-of function mutants can be considered as background noise, from which the signal corresponding to deleterious mutations can clearly be distinguished. Finally, one should notice that the metrics used here do not enable to single out the effet of the gain-of-function mutation P311W, whose magnitude is in any case twice as small as the least deleterious mutation.

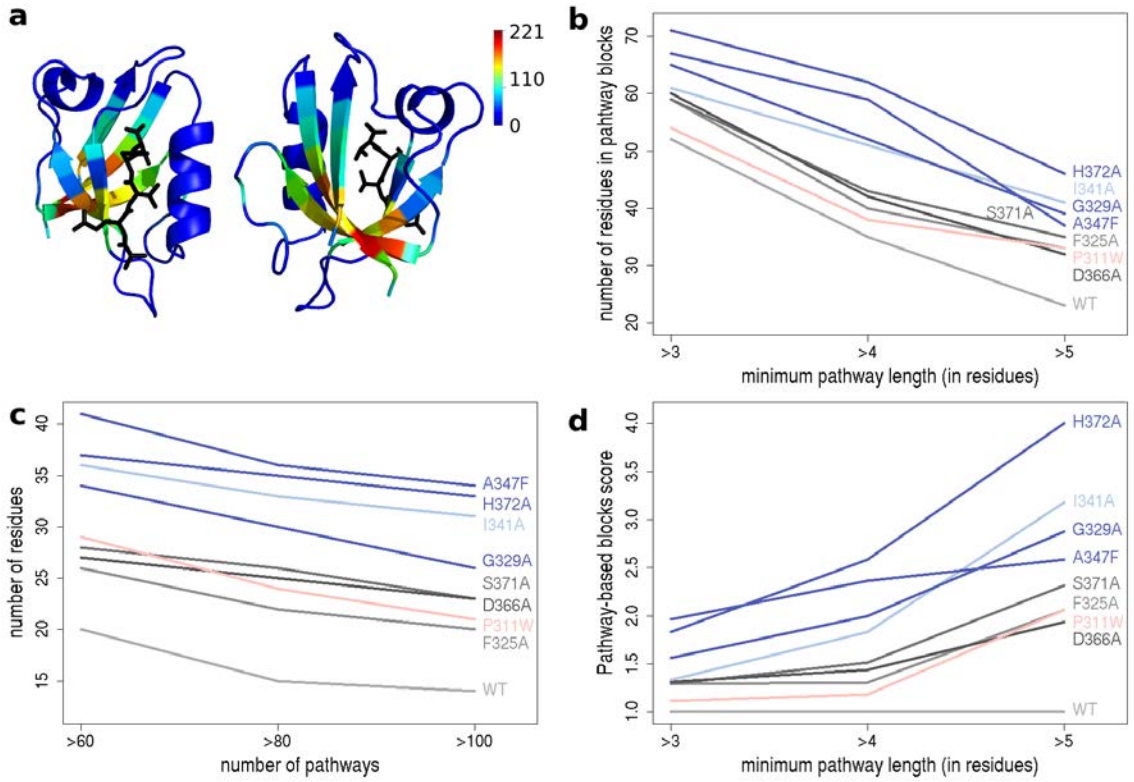

Figure 5.9: **Effect of single-point mutations on the conformational dynamics of PDZ. (a)** Mapping of the number of communication pathways (>3 residues) onto PDZ domain 3D structure. The ligand is colored in black and represented in sticks. **(b)** Number of residues in CB^{path}. The CBs are defined with pathways of varying length, from >3 to >5 residues. **(c)** Number of highly connected residues, *i.e.* crossed by a large number of pathways, from >60 to >100. **(d)** CB^{path} score computed for different pathway lengths, from >3 to >5 residues. The curves are colored according to the experimentally measured effects of the mutations: beneficial in pink, neutral in grey tones and deleterious in blue tones.

5.6 Predicting mutational hotspots

In this section we study the wild-type PDZ3 and its mutants, with the aim of predicting the mutational hotspots by applying COMMA analysis.

5.6.1 Structural dynamics-based prediction of deleterious hotspots using WT

Do the 20 deleterious positions identified in (McLaughlin et al., 2012) play a particular role in the structural dynamics of the wild-type complex and can we identify them? A first observation one can make is that all these residues belong to the interior of the complex (rsa<25%). In total, almost half of the protein residues are buried and display a distribution of experimental ΔE measurements shifted toward lower values compared to the solvent-exposed residues (**Figure 5.10**, compare boxplots in blue and orange). This finding agrees with previous studies showing that sites with lower solvent accessibility are typically less tolerant to mutations (Bustamante et al., 2000; Ramsey et al., 2011).

COMMA detected 7 residues belonging to both a CB^{path} and a CB^{clique} with high confidence (**Figure 5.4**). This dual character make them highly versatile as they transmit information in two different ways, via stable interactions and via high atomic fluctuations, to different regions of the protein. If we filter out surface residues (rsa>25%), we end up with 4 residues which were all identified in (McLaughlin et al., 2012) as deleterious hotspots: G324, I341, A376 and V379 (**Table 5.5**).

Another ensemble of residues potentially crucial for the stability of the complex are those comprising the ligand-binding pocket. One could expect that mutating them could lead to impairment of the protein-peptide association. We used COMMA to select residues forming *direct contacts* with the ligand. This means that they are immediately preceding or following a residue from the ligand in a communication pathway. This concept is more restrictive than the usual notion of physical contact, as residues adjacent in a path must: *(i)* form stable non-covalent interaction(s), present in more than 40 % of the simulation time, and *(ii)* have highly correlated movements, *i.e.* the variance of their inter-residue distance must be smaller than 0.08. We found 4 residues from PDZ being in *direct contact* with the ligand: F325, N326, I327 and H372 (**Figure 5.11a,c**). All of them are buried, and, except for N326, they are highly sensitive to mutation, representing 15% of the 20 deleterious hotspots (**Table 5.5**). Let us stress that the ensemble of residues forming stable non-covalent interactions along the simulations is 4 times bigger (16 residues) than that of residues in *direct contact* with the ligand. This illustrates the important contribution of condition *(ii)* in defining *direct contacts*.

The direct and indirect communications detected by COMMA within the PDZ domain are reported on **Figure 5.12**. Two residues are in indirect communication if they are linked by a communication path but they are not adjacent in the path (they do not form stable non-covalent interaction). We observed that most direct communications between residues far away in the sequence (black dots) are (1) grouped together in the dot plot and (2) surrounded by indirect ones (colored dots). (1) means that if two residues a_k and a_j, at positions k and j in the sequence with $|j-k|>4$, are adjacent in a pathway, then some of their immediate neighbours in the sequence are also likely to be directly connected (see lower left cartoon on **Figure 5.12**, black lines between I338 and I359,

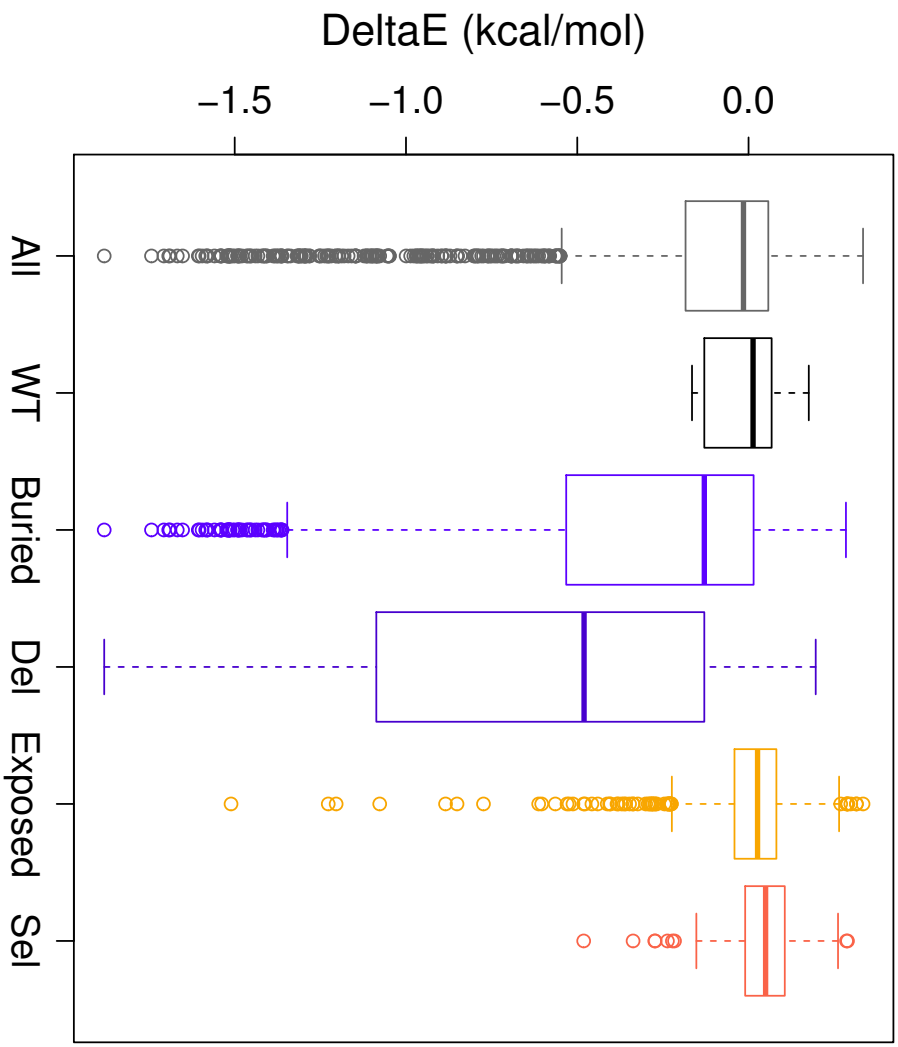

Figure 5.10: **Distributions of experimentally measured mutation-induced changes in PDZ3 affinity for its cognate ligand, the C-terminal peptide of CRIPT.** The values are taken from (McLaughlin et al., 2012). **All**: all 20 possible amino acid substitutions for all 83 positions (1660 values). **WT**: wild-type to wild-type substitutions (83 values). **Bur**: all substitutions for the 40 buried residues (solvent accessible surface area, *rsa* <25%). **Lig**: all substitutions for the 13 residues forming hydrogen-bond or hydrophobic contact with the ligand in the PDB structure (1BE9). **Del**: all substitutions for the 20 deleterious mutational hotspots identified in (McLaughlin et al., 2012). **Exp**: all substitutions for the 43 exposed residues (*rsa* ≥25%). **Sel**: all substitutions for the 9 residues exposed and located in CB^{clique} that lose communication in at least two mutants. **Benef**: all substitutions for the 20 beneficial positions.

F337 and D357, I336 and I359...). (2) means that if one or more pairs from two protein segments, $(a_{k-n}, ..., a_{k+n})$ and $(a_{j-m}, ..., a_{j+m})$, with $|j - k| > 4$, are connected by direct communication(s), they are likely part of longer pathways that extend along each segment (see lower left cartoon on **Figure 5.12**, red dotted lines between I336 and D357, F337 and I359...). Nevertheless, a few direct communications appear isolated in the plot (isolated black dots, encircled in blue, see *Materials and Methods*). They correspond to residue

Table 5.5: **Detection of deleterious mutational hotspots**

Strategy		Sens	PPV	Spe	Acc
Infostery Analysis of the Wild type	path- and clique-based CBs[1]	20	100	100	81
	direct communications with ligand only[2]	15	75	98	78
	isolated direct communications[3]	65	81	95	88
	all criteria	**80**	**80**	**94**	**90**
Coevolution Analysis	SCA	75	75	92	88
	MST	80	64	86	84
	DCA	70	70	94	86

The performance values, sensitivity (*Sens*), precision or positive predictive value (*PPV*), specificity (*Spe*) and accuracy (*Acc*), are given in percentages. [1] Buried residues detected in both a CB^{clique} and a CB^{path} with very high confidence. [2] Buried residues forming direct communications with the ligand. [3] Buried residues forming isolated direct communications between them (see *Materials and Methods*).

pairs that form communication bridges between two protein segments while the other residues from the two segments communicate with significantly poorer efficiency (**Figure 5.12**, upper left cartoon). We hypothesized that the residues involved in these bridges may be critically important in stabilizing the complex and that mutating them could result in highly deleterious outcome.

To test this hypothesis, we extracted the communication bridges (up to 5 black dots that are not surrounded by colored dots) from the dot plots obtained for different values of the communication propensity threshold (see *Materials and Methods*). In total, we identified 17 communication bridges involving 24 residues (**Figure 5.12** and **Figure ??**). When we filter out residues exposed to the solvent, we end up with a network of 16 residues, divided into 5 connected components (**Figure 5.11a,b**). Each component encompasses several secondary structure elements remote from each other in the primary sequence. Three components (in green, cyan and yellow) comprise 10 residues from the α-helix H2, the β-sheets S2-4 and the loops L2, L5 that enclose the ligand (**Figure 5.11b**). The second biggest component (in orange) is formed by 4 residues from the α-helix H1, the loop L1 and the β-sheet S5. The remaining one (in purple) comprises 2 residues from S3 and L4. 13 out of 16 (81%) residues are deleterious hotspots (**Table 5.5**). Noticeably, none of the isolated direct communications is located within the same secondary structure element (α-helix, β-strand or loop) or motif (β-sheet).

Our characterization of the wild-type complex infostery revealed that most of the

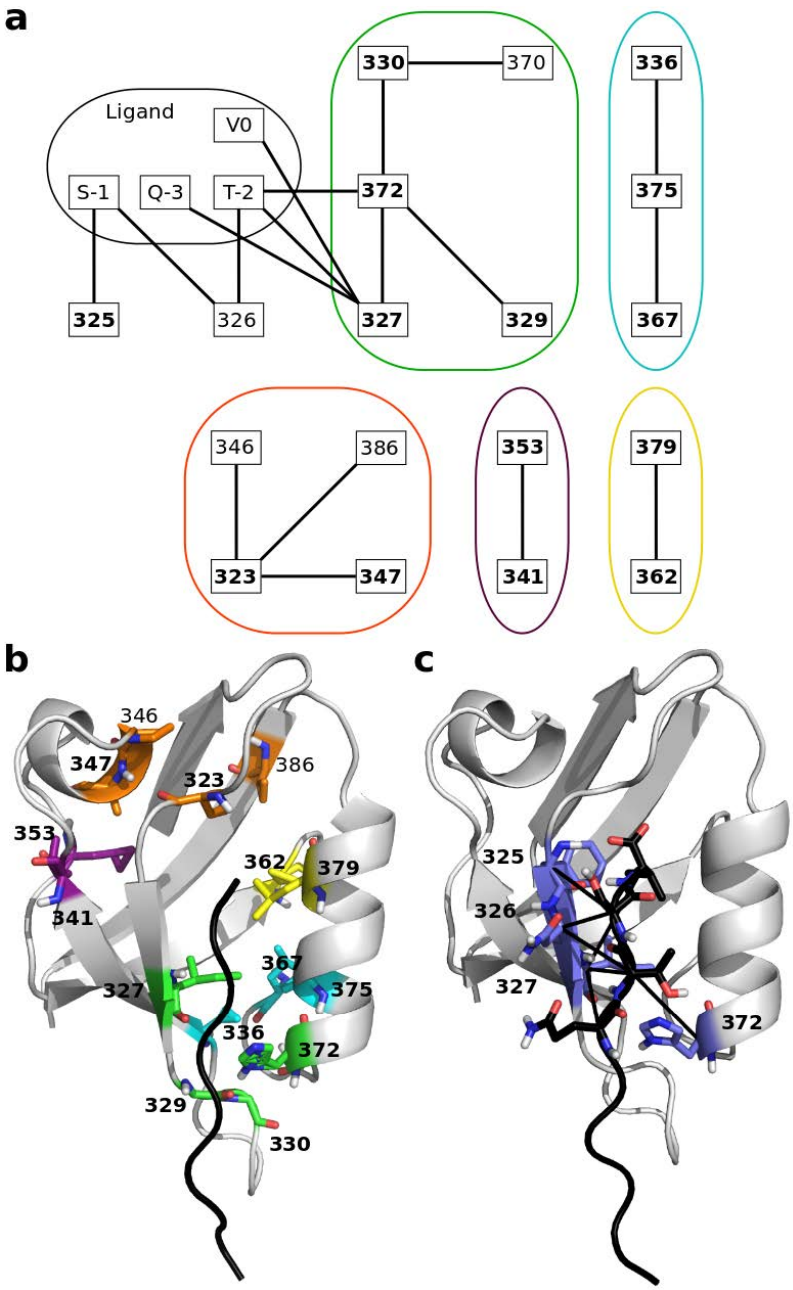

Figure 5.11: **Network of residues in direct communication in wild-type PDZ3-CRIPT peptide complex. (a)** Each node corresponds to a residue and each edge corresponds to a direct communication, detected either as isolated within the PDZ domain, or between PDZ and its ligand. Residues in bold are deleterious hotspots. The connected components extracted from the subnetwork where the nodes and edges associated to the ligand are removed are encircled in different colors. **(b)** The residues involved in communications within PDZ are shown as sticks and colored according to the connected component to which they belong. **(c)** The residues from the ligand (in black) and from PDZ (in slate) in direct communication are shown as sticks. The communications are displayed as black lines.

highly deleterious positions correspond to residues serving as critical bridges between either the protein and the peptide, or a CB^{path} and a CB^{clique}, or two distinct secondary

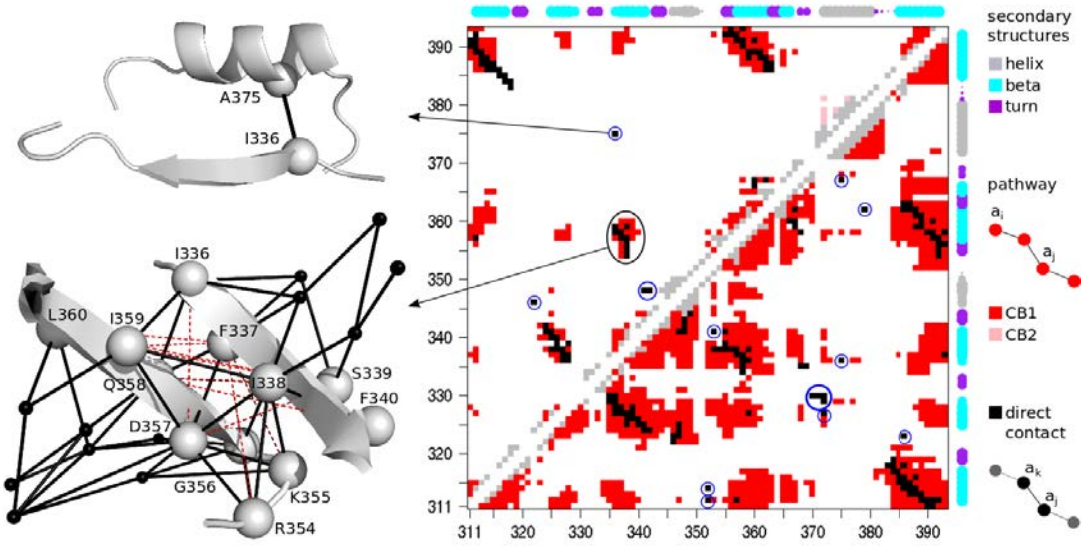

Figure 5.12: **Dotplot representing direct and indirect communication between PDZ residues.** Upper triangle: default communication propensity threshold. Lower triangle: threshold corresponding to 65% quantile of the communication propensity distribution. Each dot stands for the existence of a communication pathway linking the 2 residues indicated in x and y-axis. If the 2 residues are less than 4 residues away in the protein sequence, the dot is colored in grey. Otherwise, if the 2 residues are adjacent in a pathway (direct communication), the dot is in black. If they are not adjacent (indirect communication), the dot is colored according to the CB^{path} to which the residues belong (red or pink, same color-code as in **Figure 5.4a**). Isolated direct communications are encircled in blue. The secondary structures are also indicated. On the left, two communication motifs are mapped onto the 3D structure of PDZ, represented as a cartoon. The pathways (> 3 residues) linking the residues in the motifs are displayed as black solid lines. The C-α atoms of the residues belonging to the motif are represented as grey spheres (black smaller spheres outside the motif). Dashed red lines indicate indirect communications.

structure elements/motifs. These regions of the complex would otherwise behave independently, since they are not covalently bound, or they display different dynamical properties. Interestingly, some residues can play multiple roles. I327 and H372 form both direct communication with the ligand and isolated direct communications. I341 and L379 are detected in both types of CBs and form isolated direct communications. This analysis also demonstrated that by exploiting MD simulations of only one conformational state of the wild-type PDZ3-CRIPT peptide complex, without any insight into the conformational changes induced by any mutation, we could predict 80% of the deleterious hotspots with a precision of 80% (**Table 5.5**). The residues that are not detected, I328, I338, I359 and I388, all belong to a CB^{path} with propensity value of 1, are completely buried (rsa<5%) and highly connected (crossed by >70 paths).

5.6.2 Decreased communications for specific positions associated to beneficial mutations

Despite a global increase of pathway concentration in the mutants, some residues display a lower number of direct and/or indirect communications than in the wild type (**Figure**

5.1c and **5.6c**, red spheres). Between 8 and 24% of the residues, depending on the mutant considered, can lose up to 100% of their communications (**Figure 5.13**). Moreover, more than half of these residues lose communications in only one mutant. This indicates that specific residues in the different mutants measure loss of communication, contrary to what was observed above for pathway concentration (see on **Figure 5.1c** and **5.6c**, the red spheres are located in various places while the black spheres grow in the same regions). Moreover, the number of residues being affected and the magnitude of the loss are different depending on the experimentally measured mutational outcome from (McLaughlin et al., 2012). Noticeably, the beneficial mutation, P311W (**Table 5.4**), affects 1.4 to 3.1 times more residues than the other mutants (22 versus 7-16, see **Figure 5.13** and red spheres on **Figure 5.6c**). In three of the deleterious mutants, MU^{I341A}, MU^{G329A} and MU^{A347F}, the losses are more important (44-47% on average) than in the other mutants (23-30%).

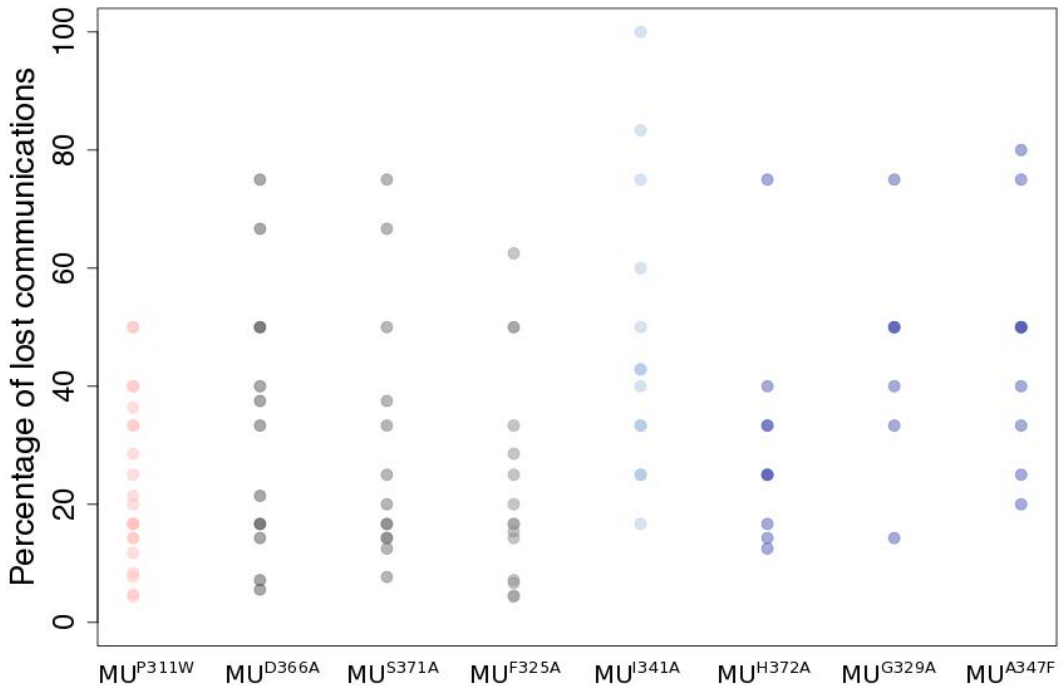

Figure 5.13: **Lost communications in mutated PDZ.** The 8 studied mutants are reported in x-axis and colored according to the experimentally measured effect of the mutation (beneficial in pink, neutral in gray tones, deleterious in blue tones). Each point represents a residue that loses at least one communication in at least one mutant and the percentage of lost communication for this residue is reported in y-axis. There are 22, 13, 13, 15, 16, 10, 7 and 10 residues that lose communication in MU^{P311W}, MU^{D366A}, MU^{S371A}, MU^{F325A}, MU^{I341A}, MU^{H372A}, MU^{G329A} and MU^{A347F}, respectively.

More than two thirds of the residues losing communications are linked by pathways to less than 8 other residues in the wild-type complex. Such residues are generally detected in CBs^{clique}. We considered the subset of 9 residues *(i)* detected in a CB^{clique} (with propensity value of 1) in WT, *(ii)* exposed to the solvent (rsa≥25%) in WT and *(iii)* losing communications in at least two mutants. These residues correspond to positions that could be considered as beneficial (**Figure 5.10**, *Sel*). Their average ΔE over all substitutions is

positive and for each position, between 12 and 20 substitutions (over 20 possible amino acid types) lead to positive ΔE values, up to 0.3 kcal/mol.

This analysis showed that the strain introduced by the mutations in the PDZ3-CRIPT peptide complex, manifested by a global increased pathway concentration, can be relaxed through different residues losing communications. It also suggested a set of criteria that could be applied to identify good candidates for beneficial mutations in a protein or protein complex.

5.7 Predicting mutatioanl hotspots using sequence analysis

We performed sequence analysis on PDZ3 to predict its mutational hotspots. We used the matrix of 20 (amino acid types) × 83 (positions) experimentally measured ΔE values reported in (McLaughlin et al., 2012) as our reference for defining beneficial, neutral and deleterious mutations. These values estimate the changes in binding affinity of the PDZ3 domain for its cognate ligand, the C-terminal CRIPT peptide, upon single-point mutations. Due to the availability of the experimental data which reveals the mutational landscape of PDZ3, we were able to examine the efficiency and accuracy of our results. Our analysis led to the detection of deleterious hotspots with high precision. Here we report the results

5.7.1 Homologous sequences

The set of sequences for PDZ3 were extracted following the method in (Baussand and Carbone, 2009). The PDZ3 was selected as the model (pdb 1BE9) and PSI-BLAST (Altschul et al., 1997) was applied using ClustalW (Larkin et al., 2007) with default parameters. Consequently, a set of 1384 sequences were extracted. Among the obtained sequences, 1263 represent 20%-40% sequence identity with the reference, 67 display 40%-60% and 53 represent more than 60% identity. Then multiple sequence alignment was applied on the sequence. The average sequence identity of the obtained sequences with PDZ3 is around 30%.

5.7.2 Evolutionary constraints

Can we relate our findings on the PDZ3-CRIPT peptide complex infostery to the evolutionary constraints exerted on the system. We extracted coevolution signals from a large set of PDZ homologous sequences. Such signals are indicative of functional dependencies between residues. We considered three different methods, namely Statistical Coupling Analysis (SCA) (Lockless and Ranganathan, 1999), Direct-Coupling Analysis (DCA) (Weigt et al., 2009) and Maximal SubTrees (MST) (Baussand and Carbone, 2009) (see *Materials and Methods*). SCA and DCA are statistical methods that infer couplings between residues from the alignment and require a large set of sequences. By contrast, MST relies on a combinatorial approach based on the analysis of the alignment, on the associated distance tree and on the combinatorics of its subtrees. The three methods display comparable accuracies (**Table 5.5**). SCA detected a physically contiguous network

	324	325	328	338	341	359	367	376	388
SCA	✓		✓	✓	✓		✓		
DCA		✓	✓		✓		✓	✓	✓
MST			✓	✓	✓		✓		
COMMA			✓	✓		✓			✓

Table 5.6: **False negatives given by coevolution-based (SCA, DCA and MST) and infostery-based (COMMA) analyses.**

	316	322	326	344	345	346	351	356	357	360	363	364	370	386	390
SCA		✓					✓				✓	✓		✓	
DCA					✓			✓	✓		✓		✓	✓	
MST	✓	✓		✓	✓		✓	✓	✓		✓				✓
COMMA			✓			✓							✓	✓	

Table 5.7: **False positives given by coevolution-based (SCA, DCA and MST) and infostery-based (COMMA) analyses.**

(sector) of 20 coevolving amino acids (McLaughlin et al., 2012), containing 15 (75%) of the 20 deleterious hotspots. Among the 20 best-ranked positions identified by DCA, 14 (70%) are deleterious positions. MST detected 25 coevolving positions, among which 16 are deleterious. Consequently, most of the positions highly sensitive to mutations (between 70 and 80%) are detected as coevolved.

Using COMMA, we were able to detect 80% of the deleterious positions with higher accuracy (*Acc*=90%) than the three sequence-based methods (**Table 5.5**). Noticeably, the deleterious position I328 is missed by all methods (**Table 5.6**). I341 and L367 are not detected as co-evolved but they are identified by the infostery analysis. Some non-deleterious positions are detected as coevolved but not as important for the complex infostery (**Table 5.7**).

The fact that both infostery-based and sequence-based analyses retrieve most of the deleterious positions clearly indicates a link between the evolutionary constraints and the structural constraints that apply to the PDZ domain to ensure/adapt its function. Furthermore, our characterization of PDZ3-CRIPT peptide complex infostery provides a physical interpretation of the functional importance of the coevolved residues. For instance, the sector residues identified in (McLaughlin et al., 2012) from sequence analysis play different roles in the complex infostery: 55% belong to a CB^{path}, 15% to a CB^{clique}, 15% to both and 15% to none (**Table 5.8**).

5.7.3 Predicting mutational effects

A mutational landscape represents the genotype-to-phenotype mapping, with a quantitative phenotype that is assigned to each sequence. We followed a similar protocol to the independent model (*IND*) in (Figliuzzi et al., 2016), to measure the phenotypic effect of mutations. Subsequently we generated a matrix of data, from the set of multiple sequence

sector residue	CB^{path} propensity	CB^{clique} propensity	direct communication with ligand	Isolated direct communication
322	0	1	0	0
323	0.8	0	0	1
325	1	0	1	0
327	1	0	1	1
329	0	0	0	1
330	0	0	0	1
336	1	0	0	1
347	0	1	0	1
351	0	0	0	0
353	1	0	0	1
359	1	0	0	0
362	1	0	0	1
363	1	1	0	0
364	0	1	0	0
372	1	0	1	1
375	1	0	0	1
376	1	1	0	0
379	1	1	0	1
386	1	0	0	1
388	1	0	0	0

Table 5.8: **Role of the residues detected in a sector by coevolution analysis (McLaughlin et al., 2012) in the infostery of the wild-type PDZ3-CRIPT peptide complex.**

alignments (**Figure 5.14**c). In this table, every column i corresponds to one position on the wild-type sequence and every row j to one of the 20 amino acid types. Then, we counted the number of sequences that represent amino acid j at position i and put this number in the table. Then we applied the *IND* model to obtain the predicted values from sequence analysis for every substitution, where the effect of substituting amino acid a^0 by a at position i is estimated as:

$$\Delta E^{IND}_{i(a^0 \to a)} = log\left(\frac{\#(a_i)}{\#(a_i^0)}\right) \tag{5.2}$$

where $\#(a_i)$ (resp. $\#(a_i^0)$)) is the number of sequences where a (resp. a^0) occurs at position i. The Pearson correlation R and R^2 between the experimental and predicted values of *IND* model equal 44% and 19% (**Figure 5.14a**).

On the other hand, we proposed to used the structural data in order to improve the accuracy of the results. Hence, we measured the C_α distance between all pairs of residues, considering the crystal 3D structure (pdb 1BE9). For each amino acid a^0 occupying position i, we considered the structural neighbours b^0 found at a distance <10 Å in the tertiary structure and occupying positions j in the alignment:

$$\Delta E^{pred}_{i(a^0 \to a)} = log\left(\frac{\#(a_i)}{\#(a_i^0)}\right) + log\left(\Pi_{b_j^0 \in neighb(a_i^0)} \frac{\#(a_i, b_j^0)}{\#(a_i^0, b_j^0)}\right) \tag{5.3}$$

where $\#(a_i, b_j^0)$ is the number of sequences containing at positions i, j the amino acids a, b^0 respectively and $\#(a_i^0, b_j^0)$ is the number of sequences containing the amino acids a^0 and b^0 respectively. Intuitively, if a^0 is found co-occurring with wild-type residues b^0 more often than a is, then it means that the latter is less fitted for its structural neighbourhood.

We compared the performance of our sequence-based approaches to five state-of-the-art methods to predict mutational outcome: PopMusic (Dehouck et al., 2011), Polyphen-2 (Adzhubei et al., 2010), MUpro (Cheng et al., 2006), SIFT (Ng and Henikoff, 2003) and I-mutant 2.0 (Capriotti et al., 2005). All these methods displayed lower predictive power than what we obtained with Equations 2.1 and 5.3 (**Figure 5.14c**). We also compared our results to a previous study (Figliuzzi et al., 2016), where the authors used Direct-Coupling Analysis (Weigt et al., 2009) to estimate the pairwise interaction terms between positions i and j. They obtained an R^2 value around 27%, very similar to what we obtained (26%) with simple sequence counts.

This analysis showed that most of the residues in PDZ3 whose substitutions result, on average, in strong impairment of CRIPT peptide binding can be detected in a very straightforward way from the alignment of many PDZ homologous sequences. 80% of these residues correspond to positions that are highly conserved in the alignment. By using an additional filter based on residue burial we could achieve an accuracy (87%) similar to that obtained from a more sophisticated analysis, called SCA, detecting co-evolution signals (88%) (McLaughlin et al., 2012). Noticeably, 3 mutational hotspots identified experimentally, I328, I338 and L367, were not detected by us from conservation signals, nor by McLaughlin and coauthors who used SCA (McLaughlin et al., 2012). In addition, our analysis led to the identification of 12 highly conserved positions that do not correspond to mutational hotspots. One could hypothesize that these positions are

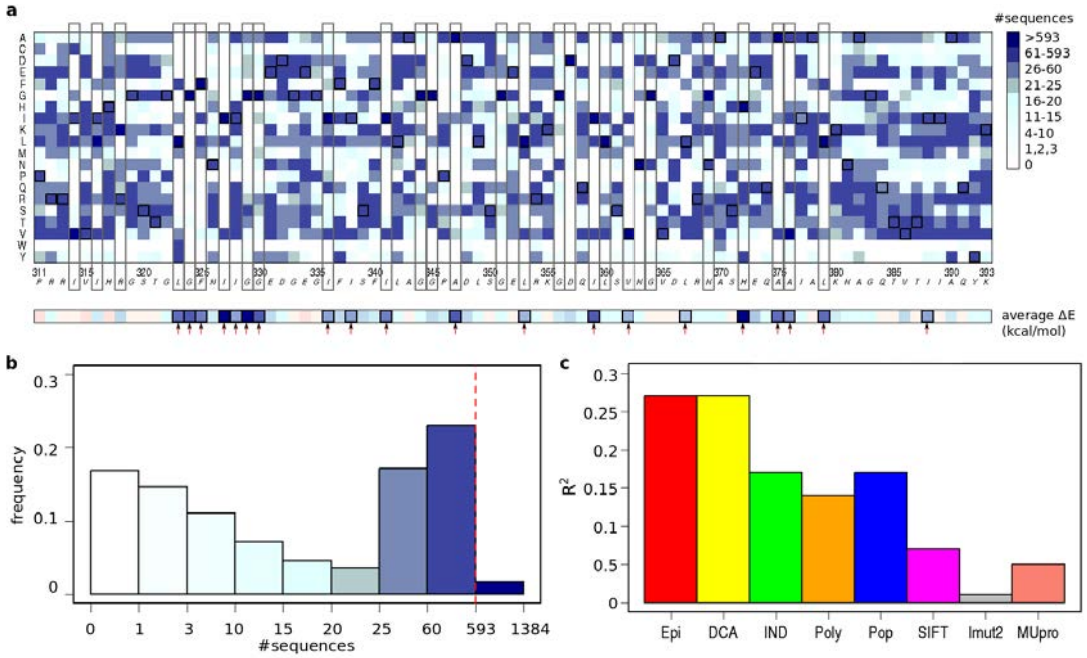

Figure 5.14: **Sequence analysis for PDZ. (a)** Matrix showing the number of sequences containing each amino acid at each position of the protein. At each position, the amino acid present in the wild-type PDZ3 sequence is highlighted by a black square. Highly conserved positions, displaying one dominant amino acid (> 593 sequences) are highlighted by gray rectangles. The strip below the matrix reports the experimental PDZ3-CRIPT ligand binding affinity changes (ΔE) averaged over the 20 possible amino acid substitutions for each position (values taken from (McLaughlin et al., 2012)). Deleterious mutational hotspots are indicated by red arrows. **(b)** Distribution of the number of sequences. **(c)** Correlation R^2 between experimental and predicted mutational effects. Tested methods are Epistatic Sequence Analysis (EPI), Direct Coupling Analysis (DCA), Independent Sequence Analysis (IND), PopMuSiC (Pop), Polyphen-2 (Poly), MUpro (MUpro), SIFT (SIFT) and I-Mutant2.0 (Imut2).

important for other aspects of PDZ function, for instance homo-dimerization (some PDZ domains are known to form dimers (Lee and Zheng, 2010)).

Beyond hotspot detection, single-site conservation signals, combined with pairwise interaction terms between each position and its structural neighbours, could be used to predict the phenotypic outcomes of all 20 × 83 possible substitutions. A correlation of 0.51 was obtained with the whole matrix of experimental ΔE values. This correlation is very good, owing to the noise contained in the experimental data (values reported for wild-type to wild-type substitutions range from -0.17 to 0.18 kcal/mol, **Figure 5.10**, *WT*). Moreover, it is significantly better than what is obtained from more sophisticated methods. Yet, it means that sequence analysis only partially explain the experimental values, which can be explained by discrepancies between the measured phenotype and that under evolutionary selection (fitness). The experimental measurements do not reflect mutation-induced changes in PDZ domain function, but a much more specific outcome, that is the binding affinity between one particular PDZ domain, $PSD95^{pdz3}$, and its cognate ligand, the CRIPT peptide. Different PDZ domains specifically recognize different target peptides and some of them even form homo-dimers (Lee and Zheng, 2010). These aspects of the functional variability of PDZ domains are not captured by the experiment designed

by McLaughlin and co-authors (McLaughlin et al., 2012) but could influence the correct interpretation of the results.

5.8 Conclusions

In this work, we have proposed new measures to study the dynamical behavior of proteins and protein complexes in solution and to probe their mutational landscape. We have introduced the concept of infostery, from 'info' - information - and 'steric' - arrangement of residues in space. We have characterized the infostery of the wild-type PDZ3-CRIPT peptide complex and single-point mutants with COMMA, a method to describe how residues communicate with each other across a protein structure. We have applied different criteria based on the geometry and dynamics of the complex to capture pertinent information.

First, we demonstrated that the wild-type complex contain all information necessary to identify almost all ($Sens$=80%) the positions significantly sensitive to mutations (deleterious hotspots) with very high precision (PPV=80%). We found that the residues at these positions were either crucial for stabilizing the binding of the ligand by contacting it directly, critical for the structural stability of the protein by connecting segments remote in the primary sequence, or versatile in the dynamical architecture of the protein by being involved in two different types of communication. Our approach does not require any *a priori* knowledge about the effect of any substitution nor about the system (*e.g.* residues or regions known to be important).

Second, we assessed the effects of chosen amino acid substitutions at 8 particular positions. The phenotypic outcomes (beneficial, neutral or deleterious) of these mutations were suggested by experiments (McLaughlin et al., 2012). We showed that the mutations did not drastically change the shape or motions of the complex on the time scale of a few tens of nanoseconds. Yet, we could exploit the data by characterizing the mutants infostery to distinguish the different types of mutations: deleterious, neutral and beneficial. Our results revealed a stiffening of the PDZ3-CRIPT peptide complex induced by the deleterious mutations and manifested as a largely increased concentration of communication pathways. This global increase was accompanied by communication losses affecting different residues, specific to each mutant. The beneficial mutant could be singled out as it displayed the highest number of such residues.

From the experimental data, deleterious mutations are much easier to identify than beneficial ones. Indeed, the magnitude of the beneficial effects are rather small and the data are noisy, as exemplified by the fact that values reported for wild-type to wild-type amino acid substitutions are not zero. Our analysis suggested that residues exposed to the solvent and located in clique-based communication blocks are good candidates for beneficial mutations. This finding shall be confirmed by future studies on systems where beneficial mutational effects are more clearly assessed.

We also put in evidence a link between the evolutionary constraints and the structural constraints that apply to the PDZ domain. Most of the highly deleterious positions have coevolved along evolution and they play particular roles in the complex communication. It was suggested that evolutionarily coupled residues in dopamine D2 receptor are links in a chain of allosteric interactions (Sung et al., 2016). Such results let envisage the possibility of reconstructing communication networks across protein structures based on conservation and coevolution signals. This would require further developments aimed at

deciphering the physical basis for such signals and for the functional dependencies they underlie.

Our work contributes to better understand the sequence-structure-dynamics relationship as it provides means to predict the phenotypic outcomes of mutations in a systematic way. It can be applied to any pair of mutations, or triplets, not just point-wise mutations, for the analysis of combined mutational effects, that might be deleterious but also compensatory (for the re-establishment of the function). It opens new avenues for developing efficient strategies to describe the mutational landscape of a protein in a computationally tractable way. We addressed the study of a difficult case, where the effects of the mutations are not obvious from classical analysis of the simulations. We were able to extract pertinent information from relatively short MD simulations and we demonstrated that the wild-type complex contained all information to identify most of the positions that 'matter'. This is very encouraging and let envisage large-scale applications of our approach.

Our results open new avenues for the prediction of mutational effects and let envisage the possibility of developing efficient strategies to characterize/explore the conformational dynamics of a protein in a computationally tractable way. (1) Simple sequence analysis and structural information from X-ray crystallography already furnish a lot of information about the function of the protein. (2) Relatively short MD simulations of the wild-type complex are sufficient to identify most of the positions that 'matter' for ligand binding. One could think of collecting information from (1) and (2) to define a set of potentially important positions. Systematic substitutions applied to a set of about 20 pre-selected positions could be applied to further investigate them. The total amount of required computing time for simulating 20×20 substitutions with 5 replicates of 20 ns would be 1 280 000 CPU hours.

Chapter 6

Disorder in coiled-coils

Contents

6.1 Introduction

Coiled-coils are ubiquitous oligomerisation motifs in proteins, where up to 7 amphiphatic α-helices (the most common number of helices are 2 and 3) intertwine together similar to the strings of a rope. The most common coiled-coils, are left-handed. They feature a specific sequence motif called heptad repeat, comprised of seven residues *abcdefg* where *a* and *d* are hydrophobic and the other residues are apolar. The number of residues per turn in a regular α-helix is 3.6. In the case of heptad repeats, it reduces to 3.5 residues per turn which leads to the left-handing (Stetefeld et al., 2000). A few cases of naturally occurring

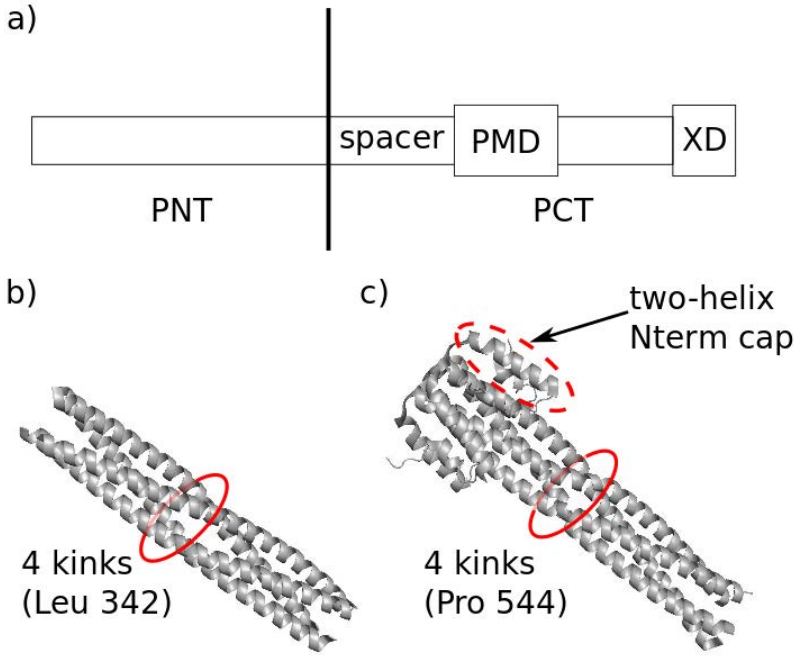

Figure 6.1: **Domain organization of MeV and NiV Phosphoprotein.** The organization of the MeV and NiV phosphoprotein is shown (**a**). The protein is comprised of two domains, PNT and PCT. PCT is composed of three parts: the spacer that is a disordered region, the Poly Multimerization Domain (PMD) that is a structured region, a disordered linker and the X domain (XD) that is a globular region. MeV PMD and NiV PMD are shown here (**b)** and **c)**, respectively). Each monomer of NiV PMD, has a two-helix Nterm cap.

right-handed coiled coils were also identified, characterized by 11-residue repeats, where the periodicity of residues per helices increases up to 3.67. In this chapter we will focus our study on phosphoprotein multimerisation domains (PMD) of two viruses, namely Measles virus (MeV) and Nipah virus (NiV) that adopt a left-handed coiled-coil structure in solution.

Measles virus (MeV) is a negative single stranded, nonsegmented virus that belongs to the family of paramyxoviridae and it is encapsidated by monomers of nucleoprotein (N) (Blocquel et al., 2014). MeV is the template for both transcription and replication by the viral polymerase complex. Polymerase complex consists of RNA-dependent RNA polymerase (L) and phosphoprotein (P) (**Figure 6.1**). P plays role to tether the L onto the nucleocapsid template. P comprises a N-terminal domain (PNT, res 1-230) which is a disordered region and the C-terminal (PCT, res 231-507) is composed of the following parts: a disordered region (res 231-303), the P multimerization domain (PMD) (res 304-375), a disordered linker (res 377-458) and a globular region (res 459-507) known as X domain (XD) (Karlin et al., 2003) (**Figure 6.1**).

Different crystal structures were solved for this protein. One tetrameric coiled-coil structure of MeV PMD (PDB code 3ZDO) was solved by (Communie et al., 2013). Our collaborators, Sonia Longhi (Université d'Aix-Marseille) and Denis Gerlier (Ecole Normale Supérieure de Lyon), first solved the crystal structure of the shortened form of MeV PMD, residues 304-360, called PMD-Ctrunc (PDB code 4BHV) (Blocquel et al., 2014). Then they employed MeV PMD-Ctrunc as a model to generate the long form crystal

PDB code	method	resolution (Å)	chains	res. present in the construct	res. resolved in the structure
3ZDO (Communie et al., 2013)	X-ray	2.07	A/B/C/D/ E/F/G/H	304-377	308-371
4BHV (Blocquel et al., 2014)	X-ray	2.10	A/B/C/D/ E/F/G/H	304-360	307-360
4C5Q (Blocquel et al., 2014)	X-ray	2.20	A/B/C/D	304-375	307-357
4N5B (Blocquel et al., 2014)	X-ray	2.2	A/B/C/D/ E/F/G/H	470-578	477-576

Table 6.1: **MD preparation and equilibration details.** The counter-ions employed to neutralize the systems are Na^+ and Cl^-. Root mean square deviations were computed on the backbone atoms of the equilibrated conformations versus the initial template.

structure of the MeV PMD (PDB code 4C5Q) and showed that only residues 308 to 357 are structured while the rest of the polypeptide chain (res 358-375) is disordered. The details of the crystal structures are described in **Table 6.1**.

In the case of MeV PMD, a and d are always leucine (L), isoleucine (I) and valine (V), except for N329 and Q356 (**figure 6.2**). The side chains adopt a "knobs into holes" packing, where a residue from one helix inside the space is encompassed by four side-chains of another helix in front of it (a and d registers in **figure 6.2**). The stability of coiled-coils is closely related to the geometry of knobs into holes. In MeV MPD, there is a breakage in the repetition of abcdefg motif around L342 which leads to the appearance of a kink at this position. Therefore positions L339 to L342 form a 3_10 helix which leads K343 outward.

Our collaborators performed a structural comparison of three MeV PMD structures (4C5Q, 4BHV and 3ZDO) and discovered some differences (Blocquel et al., 2014). The kink occurs at L342 in all chains of 4C5Q, 4BHV, whereas it is missing in the chains C and F of 3ZDO. The association of the tetramer is less tight and helices are significantly less twisted in PMD-Ctrunc (4BHV) compared to the other two structures. Differences were also observed in the geometry of knobs in those three structures: the structure of 3ZDO, represents fewer knobs on each protomer. In addition, Blocquel et al. showed that all tetramers are energetically stable and the crystal structure of PMD-Ctrunc has the highest stability. According to these findings, the same protein sequence may lead to different coiled-coil structures (ie different content of disorder and different packing of the tetramer). Where do the differences come from? Are they due to different crystal packings, or to different states of the complex? These differences can have an impact on the conclusions about the function and mechanisms of the same protein. Consequently, it is important to investigate how to resolve the discrepancies between the different structural data.

Nipah virus (NiV) is a newly emerged human pathogen in the family of paramyxoviridae (Eaton et al., 2007) and no vaccine or antiviral therapeutics is detected yet for

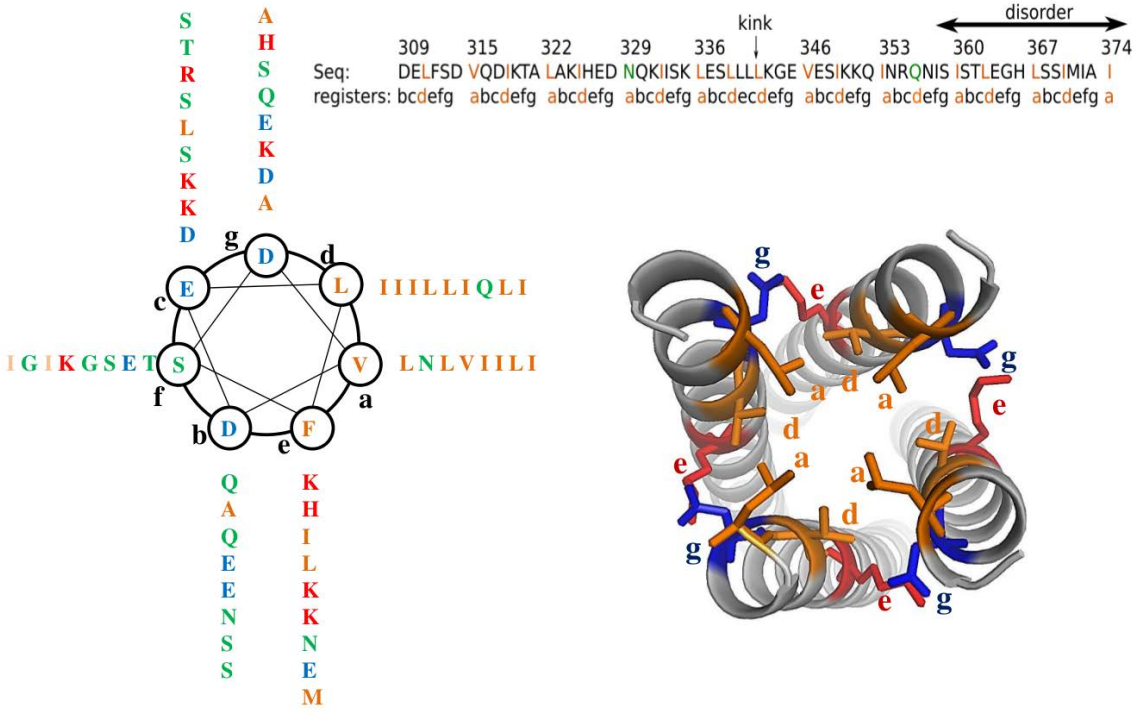

Figure 6.2: **Sequence of registers for MeV PMD.** The list of MeV PMD amino acids that appear at every position of the heptad repeat are shown. Colors represent physico-chemical properties of amino acids: orange for hydrophobic, red for negatively charged, blue for positively charged and green for polar.

human use (Broder, 2012). The N-terminal region of Nipah virus phosphoprotein (P) is intrinsically disordered (residues 1-50). The C-terminal consists of a well-ordered region, P multimerization domain (PMD) that spans over residues 470-578, a flexible linker and the X domain (residues 660-709) (**Figure 6.1**). Kink position is at Pro 544 in Nipah virus. Each monomer has a two-helix N-terminal cap in Nipah virus whereas in measles virus there is no cap at the N-terminal.

The crystal structure of NiV PMD was solved as a long parallel tetrameric coiled-coil (Bruhn et al., 2014), where in the N-terminal each monomer forms a two-helix cap. There is a kink formed at position Pro 544 in the middle of each helix which corresponds to a coil frameshift, braking from ideal Crick parameter for coiled-coils. The stabilization of NiV PMD is primarily due to the hydrophobic interactions (isoleucines, leucines and valines), a typical characteristic of coiled-coils. Although the protein forms a tetramer in the PDB structure, our collaborators found evidence that the NiV PMD exists as a trimer in solution through different experiments (the elution profile, cross-linking experiments with SAB, experiment of sedimentation velocity, far-UV spectroscopy and obtaining molecular mass from SAXS studies) (Blocquel et al., 2013).

There are two main biological questions that we are interested to investigate. First we would like to predict the disordered region of PMD of the two mentioned viruses. Three X-ray structures were solved experimentally for the PMD of MeV, however the body of structural data available are conflicting. The full-length domain (residues 304-375) is well-ordered (Communie et al., 2013) in one structure (PDB code: 3ZDO), whereas for the other structure (PDB code: 4BHV) the C-terminal part (residues 360-375) is missing,

indicating that the region is intrinsically disordered (Blocquel et al., 2014). Second, the paramyxoviridae viruses were shown to form trimers or tetramers. In particular the PMDs of the Measle (MeV) and Nipah (NiV) viruses were crystallized as tetramers. There is strong experimental evidence obtained from SAXS studies and far-UV spectroscopy for the existence of a trimeric form of NiV PMD in solution (Blocquel et al., 2013). Therefore, there is a strong interest to know which one is the most stable form for NiV PMD and MeV PMD, the trimeric or tetrameric.

In the present work, we show that COMMA can detect protein regions that are prone to disorder or substantial conformational rearrangements, without requiring the input MD trajectory to actually sample the unfolded states of these regions. Furthermore the analysis of results obtained from COMMA enabled us to propose hypothesis for mutations on MeV PMD, in order to control the stability of the coiled-coil structure.

6.2 Methods

6.2.1 Studied systems

The following homo-tetramer coiled-coils were studied (6.1):

- MeV PMD (PDB code: 3ZDO, chains A: 309-371, B: 309-370, C: 311-373 and D: 308-371, 2.07Å)
- NiV PMD (PDB code: 4N5B, chain A: 475-578, B: 476-575, C: 477-576 and D: 476-576, 2.2Å)
- NiV PMD (PDB code: 4GJW, chains A: 476-571, B: 476-571, D: 471-571 and H: 476-571, 3.0 Å)
- RhcC (PDB code: 1YBK, chains A: 1-52, B: 1-52, C: 4-52 and D: 1-52, 1.45 Å)

Among the three mentioned structures for MeV PMD, we chose 3ZDO, because it has a full length structure and the C-term is resolved. Moreover the right-handed coiled-coil homo-tetramer of the RhcC protein (*Staphylothermus marinus*) was studied as a control for our analysis on left-handed coiled-coils. Furthermore, these proteins were also simulated as monomers. In addition, we studied 4 different mutations of MeV PDM. In all these mutants the wild-type amino acid was mutated to an aspartic acid (D): V315D, L322D, V346D and I353D.

6.2.2 Molecular dynamics simulations

Set up of the systems The 3D coordinates for the studied proteins were retrieved from the Protein Data Bank (PDB) (Berman et al., 2000). All crystallographic water molecules and other non-protein molecules were removed. All models were prepared using the LEAP module of AMBER 12 (Case et al., 2012), with the ff12SB forcefield parameter set: (*i*) hydrogen atoms were added, (*ii*) Na^+ or Cl^- counter-ions were added to neutralise the systems charge, (*iii*) the solute was hydrated with a cuboid box of explicit TIP3P water molecules with a buffering distance up to 10Å. The environment of the histidines was manually checked and they were consequently protonated with a hydrogen at the

ϵ nitrogen. The mutated forms were generated by *in silico* substitutions using Rosetta Backrub (Smith and Kortemme, 2008).

Minimisation, heating and equilibration The systems were minimised, thermalised and equilibrated using the SANDER module of AMBER 12. The following minimisation procedure was applied: (*i*) 10,000 steps of minimisation of the water molecules keeping protein atoms fixed, (*ii*) 10,000 steps of minimisation keeping only protein backbone fixed to allow protein side chains to relax, (*iii*) 10,000 steps of minimisation without any constraint on the system. Heating of the system to the target temperature of 310 K was performed at constant volume using the Berendsen thermostat (Berendsen et al., 1984) and while restraining the solute C_α atoms with a force constant of 10 kcal/mol/Å^2. Thereafter, the system was equilibrated for 100 ps at constant volume (NVT) and for further 100 ps using a Langevin piston (NPT) (Loncharich et al., 1992) to maintain the pressure. Finally the restraints were removed and the system was equilibrated for a final 100-ps run.

Production of the trajectories For every protein, 2 replicates of 50 ns, with different initial velocities, were performed in the NPT ensemble using the PMEMD module of AMBER 12. The temperature was kept at 310 K and pressure at 1 bar using the Langevin piston coupling algorithm. The SHAKE algorithm was used to freeze bonds involving hydrogen atoms, allowing for an integration time step of 2.0 fs. The Particle Mesh Ewald method (PME) (Darden et al., 1993) was employed to treat long-range electrostatics. The coordinates of the system were written every ps. Standard analyses of the MD trajectories were performed with the *ptraj* module of AMBER 12.

Stability of the trajectories The RMSD of the studied coiled-coils (MeV PMD, NiV PMD and RhcC) and the mutants of MeV OMD are measured along simulation time for all the replicates of wild type and mutants (figures 6.3 and 6.4). All systems are fully relaxed after 10 ns. Consequently, the last 40 ns of each replicate were retained for subsequent analyses.

6.2.3 COMMA analysis

COMMA was applied to each system, over the 2 replicates of 50-ns MD simulations to extract communication blocks. COMMA identified pathway-based communication blocks, *i.e.* groups of residues that move together and are linked by non-covalent interactions, and clique-based communication blocks, *i.e.* groups of residues that display high concerted atomic fluctuations and that are close in 3D space. Pathways are chains of residues linked by non-covalent interactions that move together. We should emphasize that all the backbone-backbone non-covalent interactions are ignored for the analysis of COMMA. We define to set of pathway-based communication blocks, namely short-range and long-range blocks. Short-range block are consist of pathways of at least 4 residues, whereas long-range blocks are detected from the set of long-range pathways. The length of such pathways are system-depended, therefore we may have pathways of at least 7 or 8 residues.

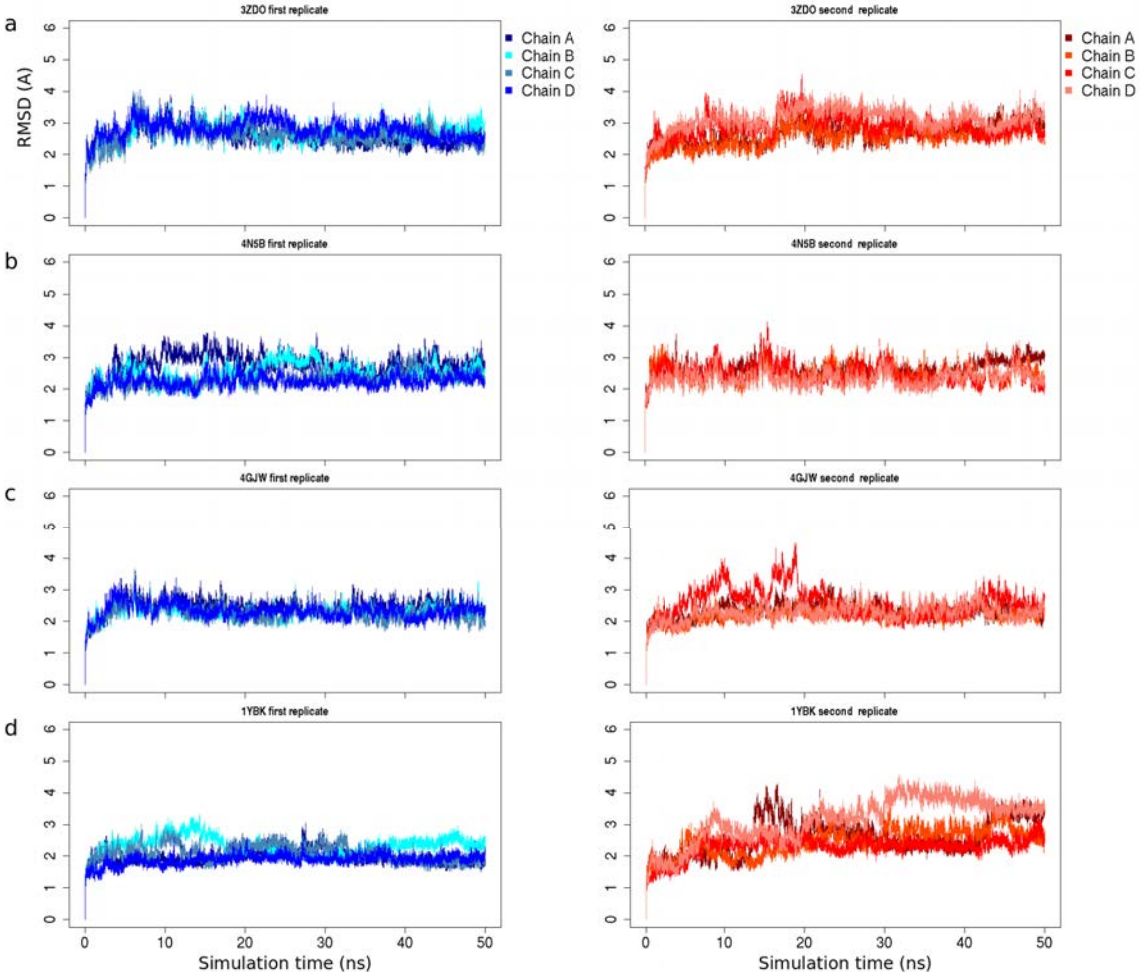

Figure 6.3: **RMSD plot for the studied coiled-coils.** The root mean square deviation is measured for every chain along the MD simulation time for the two replicates of **a)** MeV PMD (3ZDO), **b)** NiV PMD (4N5B), **c)** NiV PMD (4GJW) and **a)** RhcC (1YKB).

6.3 Results

We performed MD simulations on MeV PMD (PDB code: 3ZDO) and NiV PMD (PDB code: 4N5B) and applied COMMA to extract the communication blocks for each system. CBs^{path} define rigid bodies that move together, whereas CBs^{clique} define flexible regions that represent concerted atomic fluctuations.

6.3.1 Non-symmetric organization of the rigid communication in MeV and NiV PMD

The four helices of MeV PMD tetramer, although they share exactly the same sequence and that sequence follows the same pattern of 7 registers (abcdefg), helices do not play equivalent nor symmetrical roles in the communication of the protein. The largest CB^{path} identified by COMMA (**Figure 6.5a**, in red) comprises most of the residues from helices B, C and D and only half of the residues from helix A (77% of the total number of residues in the protein). The remaining half of helix A is detected as an independent

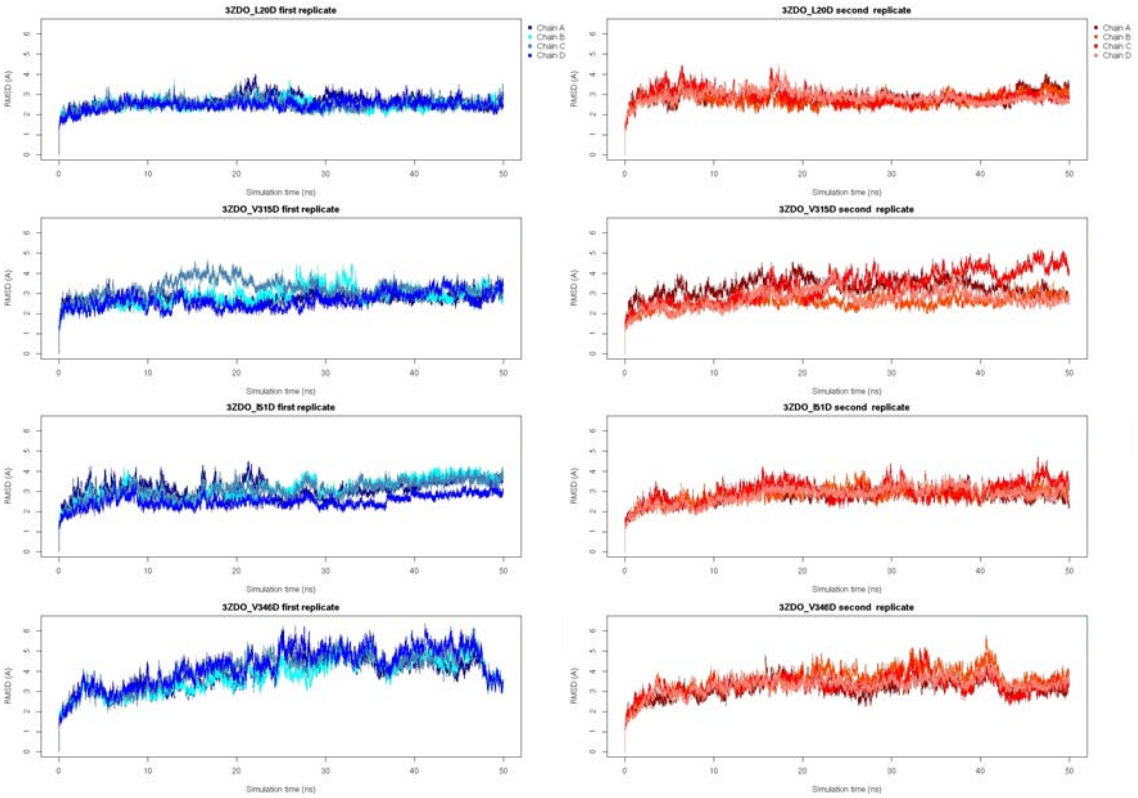

Figure 6.4: **RMSD plot for the wild type and mutants of MeV PMD.** The root mean square deviation is measured for every residue along the MD simulation time for the wild type and mutants.

CB^{path} (in pink), which partly overlaps with a CB^{clique} (**Figure 6.5e**, in purple). These observations show that almost half of helix A is not integrated in the communication of the complex. Residues that belong to long-range CB^{path}, represent pertinent behaviour in communication with other residues. They highlight regions with strong interactions and high communication propensities, that are expected to form the communication core of the protein. On the other hand, the communication hierarchy of a protein can be inferred by changing the length of pathways considered to define CBs^{path}. While considering all the pathways, almost all the residues are involved in communication pathways, the increase of pathway length lead us toward communication core of the protein. The hierarchy of communication pathways between the helices in MeV and NiV PMD, can be further refined by considering only pathways of at least 8 residues (**Figure 6.5c**), where the cores of communication in MEV PMD, highlight only the N-term halves. Furthermore, helices A and D are coupled together, while helices B and C are coupled together. The two blocks, which represent the communication core of the protein, comprise 34% of the protein residues.

In the case of NiV PMD, COMMA identified a large block (**Figure 6.5b**, in red) comprises most of the residues from helices A, C and D and only half of the residues from helix B (58% of the total number of residues in the protein). The N-term of helix B is not integrated in the larger block and about 26% of its residues (6% of the total number of residues in the protein) are comprised in an independent CB^{path} (**Figure 6.5b**, in pink). The pairing of the helices in long-CBs^{path} is different from that observed in MeV PMD,

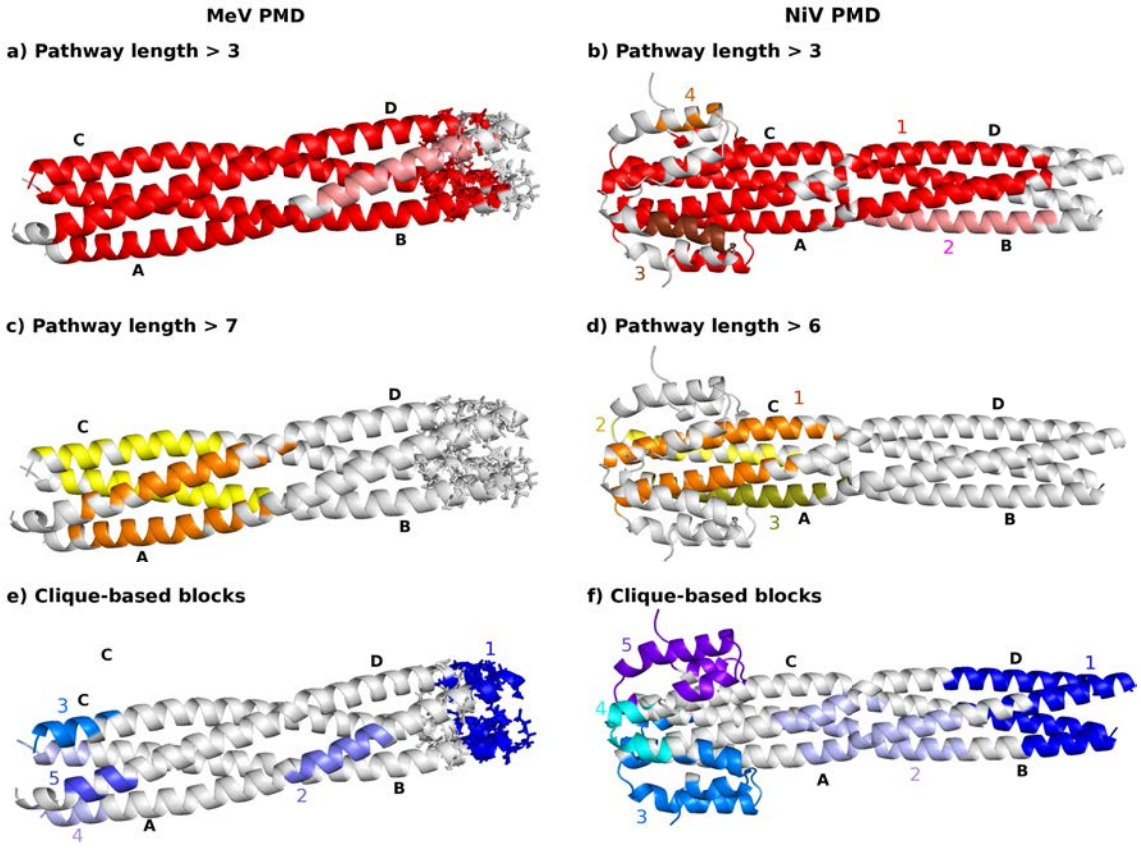

Figure 6.5: **Communication Blocks of MeV PMD and NiV PMD identified by COMMA.** Pathway-based and clique-based communication blocks identified by COMMA are mapped on the average MD conformation. Those blocks are obtained by considering pathways with length equal or greater than 4 (a and b), 8 (c) and 7 (d) residues. The clique-based communication blocks identified by COMMA are colored in blue tones (e and f). Known residues involved in disorder region of MeV PMD are shown by sticks.

with the C- D in the same blocks and A and B in two different blocks (**Figure 6.5d**). It should be also stressed out that the minimum path length considered for the long-range CB^{path} is smaller in NiV PMD (at least 7 residues) compared to MeV PMD (at least 8 residues).

Although in each system, helices have identical sequence of residues between the chains, the analysis of COMMA reveals different behaviour for each chain. In addition, the second half of chain A in MeV PMD and the second half of chain B in NiV PMD, are detected as separate CBs^{path}. In addition the grouping of the helices is different between the PMD of the two viruses. The same analysis were applied to another structure of the coiled-coil tetramer of NiV PMD (PDB code: 4GJW) and we obtain very similar results.

6.3.2 Flexible regions mediating communication in MeV and NiV PMD

CBs^{clique} detected by COMMA, are groups of residues that display concerted atomic fluctuations along the MD simulations. These residue are the most striking regions for the

flexibility of the protein and fluctuate independently from the rest of the protein. Three types of cliques are detected for the MeV and NiV PMDs: at N-term, kink-region (around the kink) and C-term.

CBs^{clique} are detected at N-term and C-term, which is expected at extremities. They tend to be highly flexible in solution. Interestingly, in the case of MeV PMD 3 CBs are detected at the N-term that comprise residues from individual (C, D) or a pair of (A and B) helices, while only 1 CB encompassing residues from all four helices (7-9 residues from each helix) is detected at the C-term. 65% of the residues known to be ambiguous (disordered in PDB structure 3ZDO, shown as sticks on **Figure 6.5e**) are included in this block. For the NiV PMD, COMMA identified 3 CBs^{clique} in the N-terminal part of the tetramer, representing 81% of the residues from the two-helix caps (**Figure 6.5f**, in purple, cyan and marine). As in MeV PMD, 1 CB (in dark blue) encompassing all four helices was detected in the C-terminal part of the tetramer. This block is noticeably larger that that detected in MeV PMD, as it comprises between 13 and 24 residues from each helix (73 residues). We can hypothesize that those residues are, to some extent, intrinsically disordered.

Another CB^{clique} (in light purple) is detected around the kinks in both PMDs. However, in MeV PMD the kink-region clique contains 11 residues from one helix, while in NiV PMD, it contains 57 residues from all 4 helices (**Figure 6.5e**). The presence of this four-helix CB in NiV PMD may be indicative of a weaker stability of the tetrameric arrangement for this protein compared to MeV PMD, in agreement with experiments suggesting that NiV PMD may be more stable as a trimer than as a tetramer (Blocquel et al., 2013).

It should be also stressed that although the sequence is identical between the chains in each system, there is no symmetrical or identical behaviour detected by COMMA. Even along the helices, different behaviours are observed. For example, the second half of chain A in MeV PMD is detected as a separate CB^{path}, while the single chain kink-region CB^{clique} is detected at this very same region (they have an overlap of 6 residues). The aforementioned identical behaviour of chain A in MeV PMD, is not observed for NiV PMD.

6.3.3 Study of a right-handed coiled-coil as a control

We investigated whether these observations, in particular the detection of a C-terminal disordered part, could be reproduced for any coiled-coil structure. As a control, we analysed the RhcC protein from *Staphylothermus marinus* which was solved as a right-handed coiled-coil tetramer. We expected to observe a different behaviour due to the lack of a kink in this protein. COMMA was applied to the ensemble of trajectories produced by MD simulations. Here we report the results:

When all paths are considered, 4 different pathway blocks are detected, each of which covers almost one helix out of four. Through these four blocks about 72% of residues in the protein are communicating (**Figure 6.6a**). Considering only long pathways, the number of residues in every block is decreased. Also blocks are reduced toward the center of the protein (**Figure 6.6b**). Three different CBs^{clique} are identified by COMMA (they are colored by different blue tones): 2 cliques that cover different N terminals and one clique that is positioned on the C terminal region and contains three helices, all together

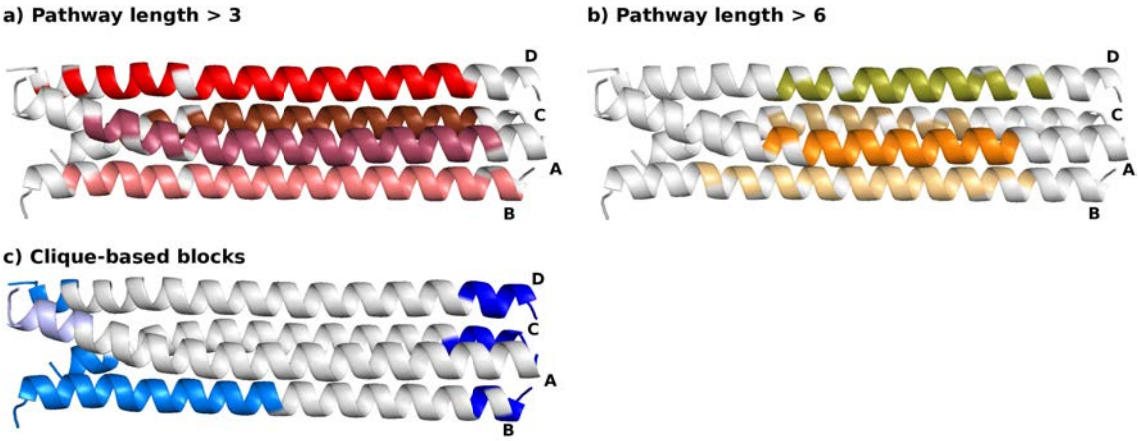

Figure 6.6: **Communication Blocks of RhcC identified by COMMA.** Pathway-based communication blocks identified by COMMA are mapped on the average MD conformation. Those pathways are obtained by considering lengths equal or greater than (a) 4 and (b) 7 residues. The clique-based communication blocks identified by COMMA are colored in blue tones (c).

in one clique. This largest clique involves about 10% of the residues (**Figure 6.6c**).

Those results suggest a strong different behaviour from what we observed for the left-handed coiled-coils. Every chain of the protein is detected as a separate CB^{path}, when we consider all pathways and when we consider only long pathways. What we observed for MeV and NiV PMDs, is not reproduced as a standard right handed coiled-coil. The lack of a unique communication core across the helices is visible here, where every helix displays an independent role.

The CBs^{clique} at the N-term and C-term represent a roughly identical behaviour in terms of the pairing of the helices. The C-term of chains B, C and D is detected as a single C-term clique, while the N-term of the same chains is identified as a separate clique. On the other hand there is a small clique on chain A at the N-term. None of the identified CBs^{clique} at the N- and C-terminal extremities of the tetramer encompasses the four helices, which suggests the lack of a disordered region on the structure of RhcC.

COMMA analysis revealed that the four helices of MeV and NiV PMD tetramers do not play equivalent nor symmetrical nor independent roles in the communication of the protein. This is very different from what can be observed for the right-handed coiled-coil structure of the RhcC protein, where the residues from each individual helix belong to only one CB^{path} and the block roughly contain the same number of residues. One reason for the observed contrast between the two coiled-coils, could be the lack of the kink in the middle of helices for the right-handed coiled-coils.

6.3.4 Comparison with single chains (for monomers)

Two replicates of 50ns MD trajectories were also produced for each of the MeV PMD, NiV PMD and RhcC monomers. The stability of the systems were reached after 25ns, therefore the last 25ns of each replicate were considered to be analysed using COMMA. In the simulations of both MeV and NiV PMDs, the unfolding of the C-term was observed. The single helices strongly bent during the simulation with the kink serving as a hinge point. The average structure obtained form MD simulations reveals the appearance of

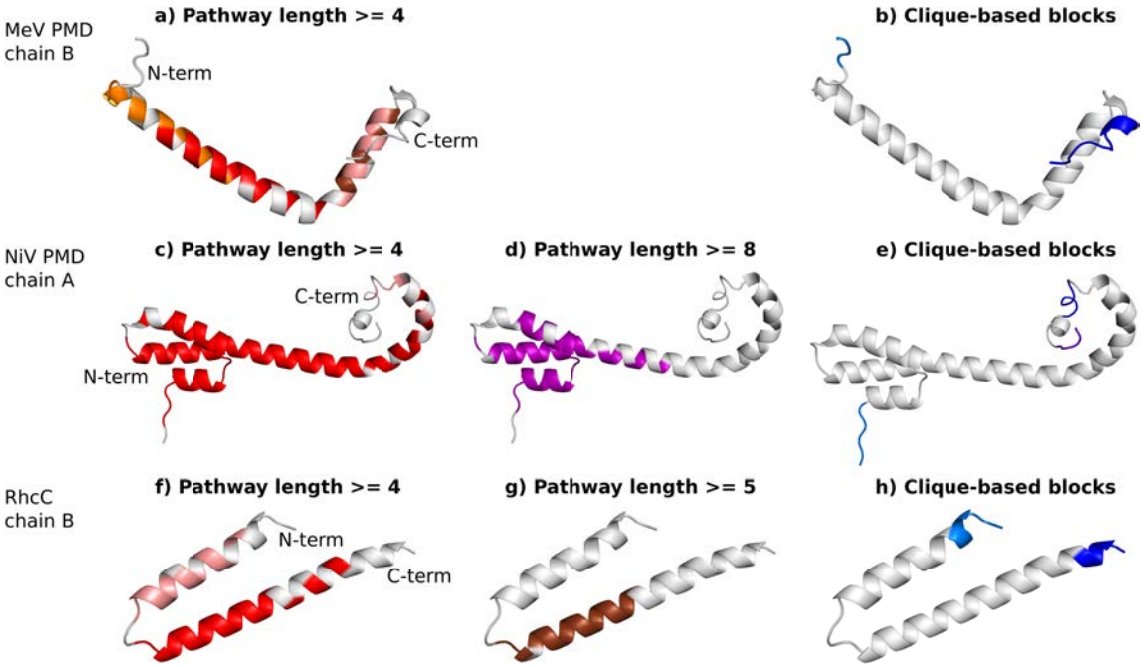

Figure 6.7: **Communication blocks identified by COMMA from single chain simulations.** Communication blocks are defined from applying COMMA to single chains of MeV PMD, NiV PMD and RhcC. Pathway and CBs^{clique} identified by COMMA are mapped on the average MD conformation. Those pathways are obtained by considering interactions involving side chains, short- and long-range pathways. The clique-based communication blocks identified by COMMA are colored in blue tones.

such a strong kink with high degree of bending for all the three systems (**Figure 6.7**). Here we report COMMA results:

- **Chain B of MeV PMD (3ZDO)**: CBs^{path} defined from short-range pathways are extracted (**Figure 6.7a**). There is no pathway block constructed from long-range pathways, the maximum number of residues involved in pathways is 4. The region known as disordered is 77% covered by the CB^{clique} (res 361-370) (**Figure 6.7b**).

- **Chain A of NiV PMD (4N5B)**: CBs^{path} defined from short (more than 3 res.) and long-range (at least 8 res.) pathways are extracted (**Figure 6.7c and d**). A portion of the residues that were previously predicted as disordered, is covered by two CBs^{clique} (res 563-568 and res 575-578).

- **Chain B of RhcC (1YBK)**: CBs^{path} defined from short and long-range (at least 5 residues) pathways are extracted (**Figure 6.7f and g**). The C-terminal clique covers 5 residues (res 48-52).

Several CBs^{path} were observed for all the three monomers. The residues from 360 to 370, known to be ambiguous in MeV PMD, were completely unfolded and detected as a CB^{clique} by COMMA (**Figure 6.7b**, in blue).

The number of residues that are detected as CB^{clique}, in addition to the number of unfolded residues in the N-term and C-term of the monomers are reported in **Table 6.2**.

monomer protein	N-term		C-term	
	CB^{clique}	unfolded	CB^{clique}	unfolded
MeV PMD	4	7	10	15
NiV PMD	5	0	10	19
RhcC	6	2	5	3

Table 6.2: CB^{clique} **of the monomers.** The number of residues detected as CB^{clique} in the N-term and C-term of each monomer, MeV PMD, NiV PMD and RhcC are shown here.

These results indicate that MeV and NiV PMD do not adopt stable monomeric conformations in solution and that their C-terminal parts are intrinsically disordered. Whereas, in the case of RhcC monomer the C-term is folded and we cannot suggest a disorder behaviour. In addition the monomeric form of RhcC, is bended in the middle but there is no kink. But from the MD simulations we can say that the monomeric form is not stable. At last, we can interpret the results as a transition from unfolded state to "not-so-folded" state upon binding for the two PMDs.

6.3.5 Comparison between COMMA and other tools to predict disorder

In order to evaluate the power of COMMA to predict the disordered region in coiled-coil structures, we compared our results with three sequence-based programs: Coils server, IUPred and ANCHOR, using their default parameters. The results obtained from these methods are shown in Figure 6.8.

From the analysis of COMMA, the presence of the clique is averaged over the four chains to predict the region of disordered residues, where the predicted disordered regions for MeV PMD starts from Gly 365 and for NiV PMD starts from Leu 563. The unfolding of the helices (loops detected by DSSP (Kabsch and Sander, 1983)), is reported over MD simulations and averaged over four chains (shown in red on the Figure 6.8). Coils server measures the probability to form a coiled-coil structure, as mentioned in the method section. But here we present the probability not to form a coiled-coil structure (which is $1 - Prob(CoilsServer)$, in order to better compare the results (Figure 6.8 green bars).

Residues for which we obtain the maximum number of helices (that is 4 in the case of studied coiled-coils) in CB^{clique} and maximum confidence (which is 1), are detected as predictions of disordered by COMMA. In the case of MeV PMD, the comparison of the predictions with the experimental data, reveals the power of COMMA to predict the disordered residues with strikingly better precision, compared to the other methods (**Figure 6.8**). The set of disordered residues in NiV PMD are not known, experimentally, However, the results obtained from different sequence-based predictors, suggest a similar region on the C-term to be disordered. In that case, the predictions of COMMA and Coils server suggest roughly the same set of C-term residues as disordered.

The three mentioned sequence-based methods (Coils server, IUPred and ANCHOR) and also the unfolding of the helices are good and fast measures to predict disorder, however they are not able to detect flexible and unstable regions of coiled-coils. Although

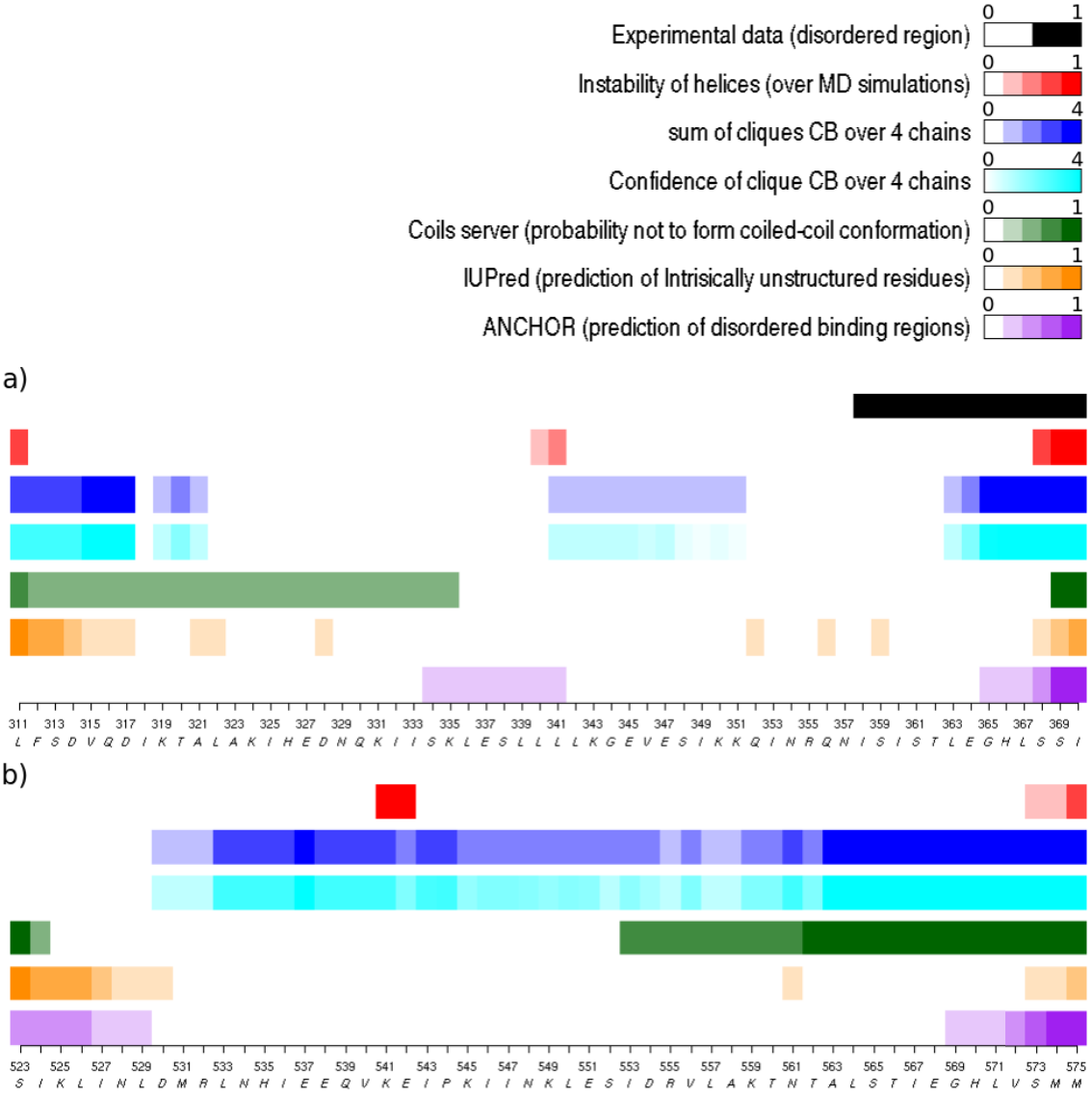

Figure 6.8: **Comparison between different methods to predict disorder residues.** For the **a)** MeV PMD and **b)** NiV PMD the comparison is done between different methods to predict disorder residues. The unfolding of helices along MD simulations, number of chains detected as cliques and the confidence of CB^{path} residues are shown in red, blue and cyan, respectively. The predictions of Coils server to form coiled-coil structures are shown in green. The predictions of disorder residues obtained from IUPred and ANCHOR are shown in orange and purple, respectively.

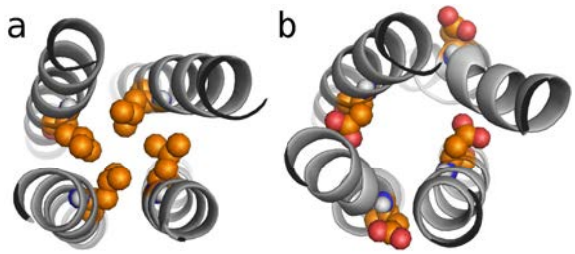

Figure 6.9: **Mutation from a hydrophobic amino acid to a negatively charged one.** The hydrophobic a registers are mutated to negatively charged residues. Changes of the side chains are shown in the figure.

COMMA is computationally expensive due to the fact that it depends on the results of time consuming MD simulations, it is able to predict both types of regions. Interestingly, COMMA also provides a mean to distinguish the two different behaviours. The presence of a single CB^{clique} over the four chains represent the probability of a disordered region, while the existence of other type of CBs^{clique} provide hint on flexible region for the dynamics of the coiled-coils.

While, COMMA is capable of identifying disordered residues through a CB^{clique} that spans C-term of the four chains, it also highlights the differences between MeV PMD and NiV PMD. One can observe that NiV PMD is less stable, where a larger number of residues are detected as CB^{clique}. This observation suggests a larger disordered region for NiV PMD, which may lead to the conclusion of instability of the tetrameric form for NiV PMD.

6.4 Controlling MeV PMD flexibility/communication through mutations

COMMA enabled us to predict disordered residues with high accuracy compared to the other predictors. Consequently our collaborators and us, we were interested to design a mutation which brings more stability to the structure. For that reason, our collaborators suggested two mutations of MeV PMD, L322D and I353D to be tested by COMMA analysis. These two mutations are *a* positions on the sequence of MeV PMD, where hydrophobic amino acids (leucine and isoleucine) are mutated to a negatively charged amino acid (aspartic acid). The hydrophobic amino acids in position *a* have their side chains oriented toward the interior of the coiled-coils (inward) and mutating them to Asp leads to outward orientation of the side chains because of its negative charges (**Figure 6.9**).

From previous COMMA analysis of the wild-type MeV PMD, one may infer that the first half of the structure is more rigid and represents the communication core of the system, whereas the second half is more flexible and contains more disordered residues. For any mutation from a hydrophobic amino acid to a negatively charged one in coiled-coils, we expect to observe some flexibility, because the side chains of the mutated residues are exposed to the solvent and not packed inside the coiled-coils. Consequently, we expected the mutation in the first half (L322D) to bring some flexibility to the N-term, which results in a bridge between the two halves. On the other hand, the mutation on the second half

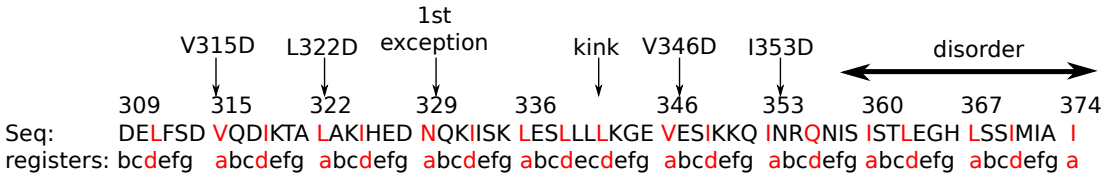

Figure 6.10: **The sequence of MeV PMD with the corresponding registers.** The sequence of MeV PMD along with the registers assigned to it are listed here. Arrows are pointed to the a positioned that are mutated in this study.

(I353D), was expected to bring even more flexibility to the C-term which may cause the increase of CBs^{clique}.

In order to test the mentioned hypothesis, we decided to study to study the mutation of all hydrophobic *a* positions to the same negatively charged amino acid, *Asp*. **Figure 6.10** represents the sequence of a single chain for the PMD of MeV in addition to the registered assigned to every position. Here we report four different mutations, each at different *a* of the heptad repeat. Two of them are selected before the kink and two others are after the kink.

MD simulations were performed for every mutation, then we applied COMMA analysis to the ensemble of conformations obtained for the each mutation and here we report the obtained results:

6.4.1 Mutations before the kink

V315D This mutation is positioned at the N-terminal, the first *a* register on every chain. Two CBs^{clique} are detected at N-term, one spans three chains (A, B and C and the other one contains residues from chain D (**Figure 6.11b**). On the other hand, the C-term CB^{clique} spans all four chains and contains 23 residues of the disorder region, 9 residue less than wild type (32 res.) (**Figure 6.12**). Almost all residues (83%) are in a single CB^{path} (in red), when considering all pathways (> 3 res), whereas in wild type the second half of chain A is detected as a second block (**Figure 6.11b**). Similarly, a single CB^{path} is detected when considering long-range pathways (at least 8 residues). Whereas in WT, this block contains mainly residues on the first, but also contains few residues after the kink. In addition the pairing of helices is not observed upon mutation and all the helices are placed in the same block.

L322D 2 CBs^{clique} are detected, one at N-term (25 res.) and the other at C-term (32 res) (**Figure 6.12**). Both of these CBs spans residues from the four chains. On the other hand, one large CB^{path} is defined when considering pathways of at least 4 residues (79%), similar to V315D (**Figure 6.11c**). A single long-range CB^{path} (>=8 res.) is detected that contains residues from both halves of the helices. It extends toward the C-term significantly more than in V315D.

Segment pairs We applied COMMA analysis to define the pairs of secondary structure elements (*SSEs*) and to monitor the direct communications between them (see *Methods*) (**Figure 6.12**). Applying the mutations on *a* positions at the first half, leads to the increase in number of direct communications at the second half, in total 11 communications in WT,

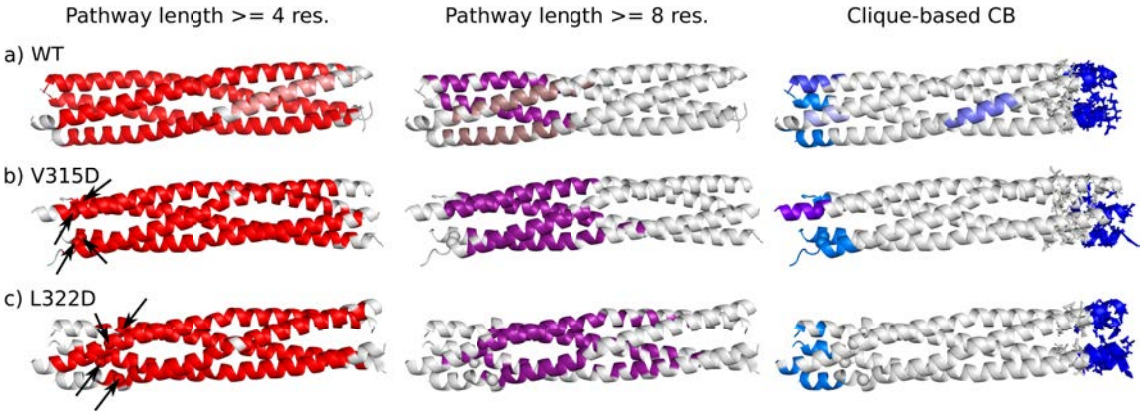

Figure 6.11: **Study of mutants (V315D and L322D) of MeV PMD communication.** Communication blocks identified by COMMA for the **a)** wild type and mutations at the first half of the helices, **b)** V315D and **c)** L322D), are mapped on the average MD conformation. CBs^{path} are obtained by considering interactions involving side chains and pathways with length equal or greater than 4 and 8 residues. Known residues involved in disorder region are shown as sticks. The clique-based communication blocks identified by COMMA are colored in blue tones.

	path CB	CB^{path}			direct pathways	
		N-term	2nd half	C-term CB	N-term	C-term
wild type	195, 12	19, 8, 6	11	32	35	11
V315D	209	28, 9	-	23	65	17
L322D	198	25	-	32	58	31
V346D	123, 20, 19, 19, 19	15, 9	19	30	63	4
I353D	170, 27, 20	17	20	36	55	12

Table 6.3: **Summary of the number of residues involved in path-based and CBs^{clique} and direct contacts between the chains.** For the wild type and four mutants the detail of the path-based and CB^{clique} are reported here. In addition number of direct communications between the chains in the two halves are recorded.

17 in V315D and 31 in L322D (**Table 6.3**). In the case of L322D, this is reflected in the shift of the CB^{path} toward the C-term (**Figure 6.11**). Consequently, the mutant (L322D) induce new communications between chains A and C (1st half) and between chains B and D (2nd half). Also the CB^{clique} on second half of the wild type disappears. In addition the results suggest a symmetrical or almost equivalent detection of CBs^{clique} at N-term and C-term.

We can conclude that upon mutation of the hydrophobic *a* positions on the first half to a negatively charged amino acid, establish/reinforce communication between the two halves. The roles of the two halves in the communication of the complex become almost symmetrical/equivalent.

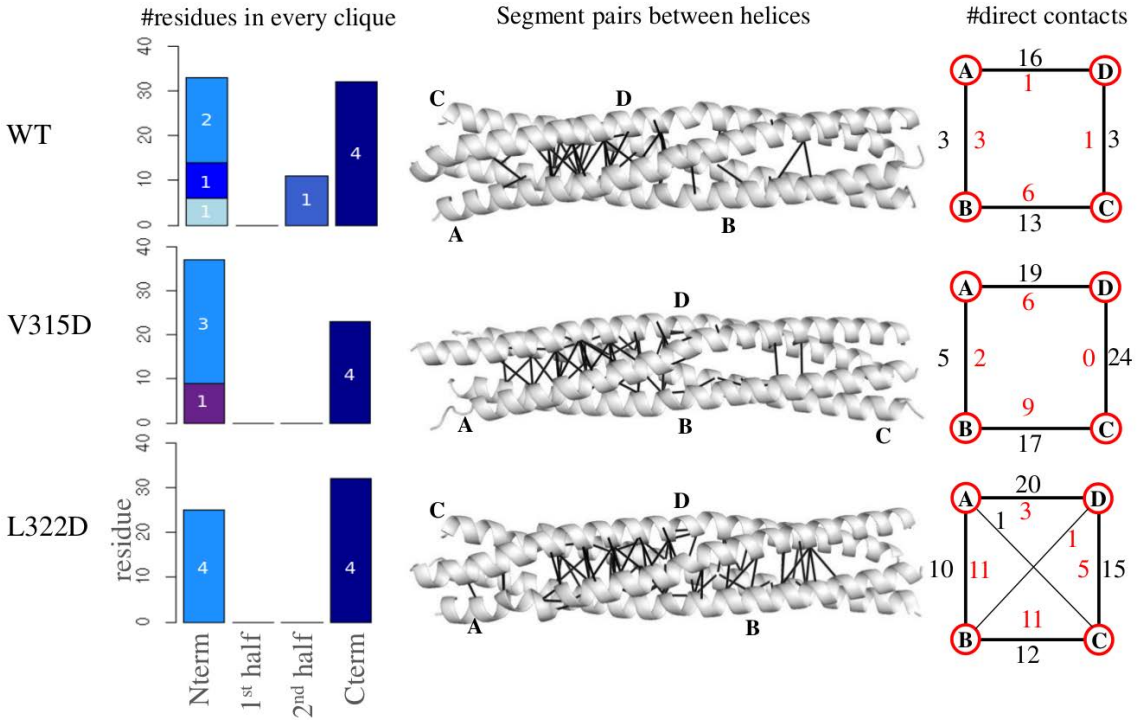

Figure 6.12: **Study of mutations of MeV PMD at the first half of helices by COMMA segment pairs.** Number of residues detected as *CBclique* at the terminals, before and after the kind, is reported for the WT and mutants on the left column. All communications between helices extracted from the analysis of segment pairs in COMMA are show on the average structure, in the middle column of the figure. The schematic representation on the right (wild type and mutants), reports the number of direct pathways between helices in the first and second halves (numbers colored in black and red, respectively).

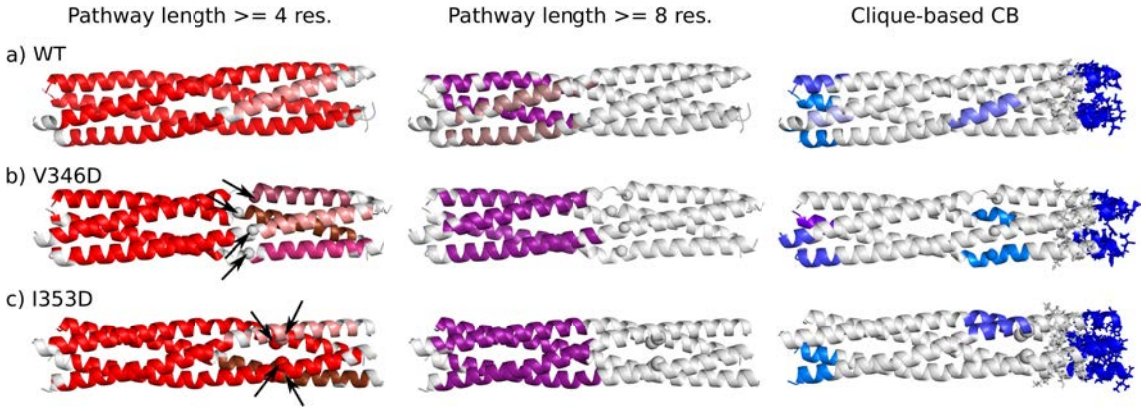

Figure 6.13: **Study of mutants (V346D and I353D) of MeV PMD communication.** Communication blocks identified by COMMA for the **a)** wild type and mutations at the second half of the helices, **b)** I353D and **c)** V346D, are mapped on the average MD conformation. Those pathways are obtained by considering interactions involving side chains and pathways with length equal or greater than 4 and 8 residues. Known residues involved in disorder region are shown by sticks. The clique-based communication blocks identified by COMMA are colored in blue tones.

6.4.2 Mutations after the kink

V346D This mutation is placed very close to the kink. 4 CBs^{clique} are detected (**Figure 6.13b** and **6.14**). The C-term clique spans the four chains but contains roughly the same number of residues as the wild type. Two other N-term cliques are detected, where two of them cover residues from chains A and B, and the third clique is positioned on chain C. The last clique (in dark blue), is positioned around second halves of chains B and C, in the region between the kink and disordered residues. CBs^{path} obtained from this mutation are significantly different from the wild type. When considering all pathways (>=4res.), the second half of chains A and C are detected as two separate blocks. Also considering long-range pathways (at least 8 residues) the pairing of the chains is not present any more and only one block is detected by COMMA.

I353D 3 CBs^{clique} are detected (**Figure 6.13c**). The area detected as disordered at C-terminal contains a larger number of residues (36 aa) compared to the wild type (32 aa) (**Figure 6.14**. The other N-term CB^{path} covers residues of chains A and B, very similar to the clique detected at the same region in wild type. The third clique is positioned around second half of chains A and D, in the region between the kink and disorder residues. When considering all pathways (>=4res.), the larger block (in red) is smaller than the same block in WT and new blocks (in pink, magenta, violet and brown) appear in the 2nd half, each of the them containing residues from only one chain. Long-range pathway block are similar to the wild type in terms of number of residues in the block, although the pairing of the chains is not present any more and only one block is detected by COMMA.

Segment pairs The total number of residues that belong to N-term clique, is significantly decreased, whereas a sharp increase is observed in the clique that appear at the second half. In the case of V 346D, a significant decrease is observed, there is only one contact between chains C and D and 3 contacts between chains A and D. Comparison of

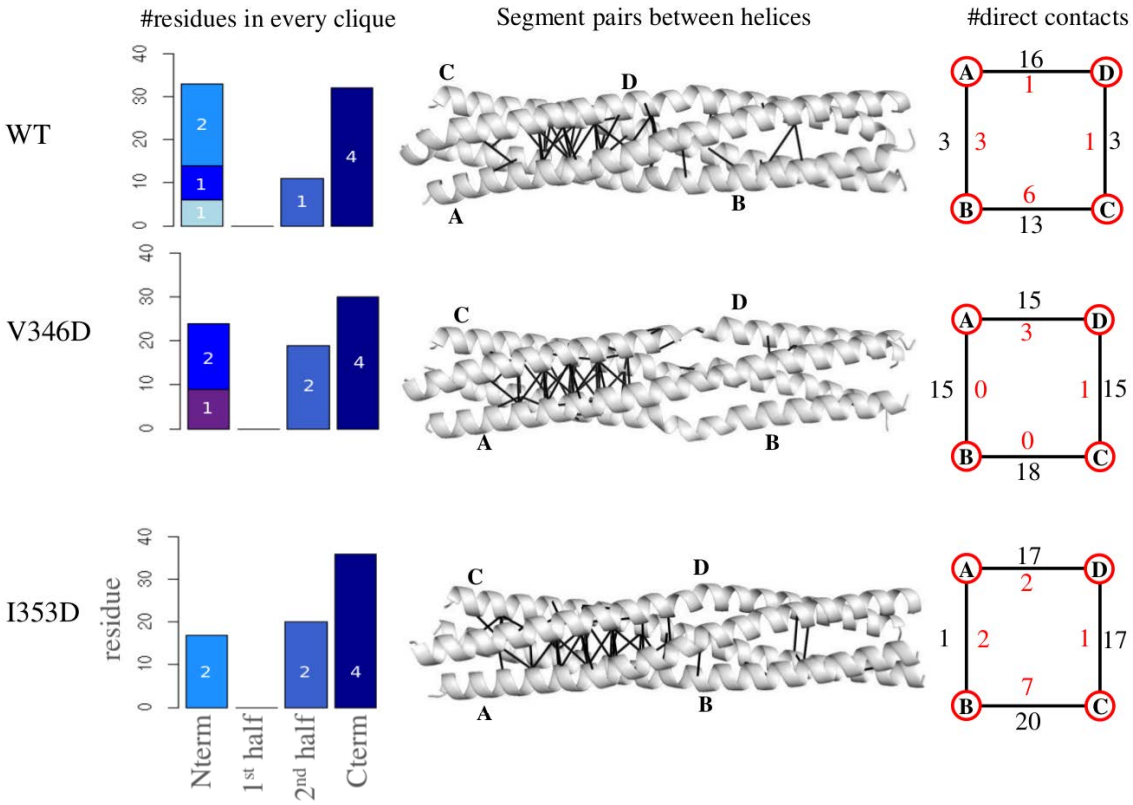

Figure 6.14: **Study of mutations of MeV PMD at the second half of helices by COMMA segment pairs.** Number of residues detected as *CBclique* at the terminals, before and after the kink, is reported for the WT and mutants on the left column. All communications between helices extracted from the analysis of segment pairs in COMMA are show on the average structure, in the middle column of the figure. The schematic representation on the right (wild type and mutants), reports the number of direct pathways between helices in the first and second halves (numbers colored in black and red, respectively).

the results represent a separation between the two halves of the helices, before and after the kink. The first part is more rigid, while the second part is more flexible.

CBs^{path} (of at least 4 res.) of the mutants are more fragmented compared to the wild type, at the C-term. The separation between such CBs happens around the mutation position. This observation suggests stiffening of the N-term and reduction of communication in C-term. The analysis of segment pair suggests the lack or small number of communications between the chains on the second half (**Table 6.3**). Furthermore, the presence of a second half CB^{clique} that spans two chains and the total increase in the number of residues detected as CB^{path}, suggest the increase of flexibility on the second half due to the mutation. We can conclude that upon mutation of the hydrophobic *a* positions on the second half to a negatively charged amino acid, breakage of the communication between 1st and 2nd halves happens, due to which the communication cannot propagate across the structure. The structure is more rigid on the first half while it is even more flexible on the second half. Chains of the second half represent a more independent behavior, that is similar to the right-handed coiled-coil (RhcC) (**Figure 6.6**).

6.5 Conclusions

This analysis revealed that COMMA is able to detect a region known to be ambiguous, *i.e.* well ordered in a PDB structure and unresolved in another one, in the coiled-coil tetramer of the MeV PMD. The region is detected as a clique-based communication block that spans all four chains. The application of COMMA to the coiled-coil tetramer of the NiV PMD yielded similar results, suggesting that the C-terminal part of the NiV PMD coiled-coil tetramer also has substantial disorder content. This disorder seems to be an intrinsic feature of the monomer. This property is not shared by the RhcC protein, which forms right-handed tetrameric coiled coils. Interestingly, COMMA also predicts disorder around the kink in NiV PMD, which may be indicative of a weaker stability of the tetrameric form of this protein compared to MeV PMD. Furthermore, comparisons with the existing tools to predict disordered residues, revealed the power of COMMA to predict disordered residues.

Surprisingly the sequence of amino acids for each studied system, MeV PMD, NiV PMD and RhcC, is identical between their chains, whereas the behaviour of the chains are different. In the case of wild-type MeV PMD, residues are grouped in the same CB^{path}, whereas another CB is detected over the second half of chain A and a similar behaviour was observed for NiV PMD. Such separation may suggest the existence of the trimeric form. On the other hand chains are fragmented in four different path CBs for RhcC, representing independent communications. Also the organization of CBs^{path} are different between chains and between the systems. Consequently, chains have different roles, although they possess the same sequence. In addition, by using classical MD analysis, it is not possible to differentiate the role of helices.

Pathways extracted from COMMA contain important information about the communication between residues and specifically between chains. Based on our analysis, there are many such interactions between all the chains in the coiled-coil structures studied in this work. This analysis along with the study of mutations suggested by our collaborators, enabled us to propose mutations that modulate the stability of disordered coiled-coils.

Applying the mutations at N-term, led to the increase in number of direct communications at the second half and establishing communication between the two halves, which in turn enabled the two halves to communicate across the structure. On the other hand, the mutation of *a* positions at C-term, led to the fragmentation of path CBs and decrease of communication at C-term and significant increase of communication at the N-term at the same time. Comparison of the results represent a separation between the two halves of the helices, before and after the kink. The N-term became more stable and rigid, while the C-term turned to be more flexible. Finally, experimental characterization of the mutants could help validate these hypotheses.

Chapter 7

Conclusion

In this thesis we presented COMMA, a method to describe and compare the dynamical architectures of different proteins or different variants of the same protein. COMMA extracts dynamical properties from conformational ensembles to identify *communication pathways*, chains of residues linked by stable interactions that move together, and *independent cliques*, clusters of residues that fluctuate in a concerted way. Pathways and cliques are used to define *communication blocks*. The term 'communication' refers to the way information is transmitted across the protein structure. The originality of the method lies in the fact that it accounts for two different modes of communication, through the use of pathways and cliques. Consequently, it enables to contrast the different types of communication occurring between residues and to hierarchise the different regions of a protein depending on their communication efficiency. COMMA provides a description of the infostery of a protein or protein complex that goes beyond the notions of chain, domain and secondary structure element/motif, and beyond classical measures of how a protein moves and/or changes its shape.

We showed the efficiency of our approach in providing mechanistic insights on the effects of deleterious mutations by pinpointing residues playing key roles in the propagation of these effects, through different case studies. The discussion of examples, revealed physical interpretation on how the study of conservation brings significant insights on the sensitivity of conserved positions to mutations. Moreover, our work contributed to better understanding the sequence-structure-dynamics relationship as it provides means to predict the phenotypic outcomes of mutations in a systematic way. It has to be emphasized that in the case of PDZ domain, we were able to extract pertinent information from relatively short MD simulations and we demonstrated that the wild-type complex contained all information to identify most of the positions that 'matter'. Our proposed method to study the dynamics of proteins, can detect protein regions that are prone to disorder or substantial conformational rearrangements, without requiring the input MD trajectory to actually sample the unfolded states of these regions. Moreover, COMMA analysis of disordered coiled-coils, enabled us to suggest mutations that regulate the stability of the coiled-coils.

Further investigation can be applied to improve the proposed methods, some of which are: **(1)** The automatic set-up of the thresholds used in COMMA need to be modified for the study of large complexes. **(2)** The link between clusters of coevolving residues and networks of dynamically correlated positions has yet to be further studied. We observed

some interesting overlaps or complementarity between coevolution signals and communication pathways that should be further investigated. **(3)** We proposed a hypothesis for mutations to regulate the stability of disordered coiled-coils, experimental investigation of those mutations can add more evidence to our results.

Bibliography

I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev. A method and server for predicting damaging missense mutations. *Nat. Methods*, 7(4):248–249, Apr 2010.

Hiroshi Akashi. Within-and between-species dna sequence variation and the 'footprint'of natural selection. *Gene*, 238(1):39–51, 1999.

Berni J Alder and TE Wainwright. Studies in molecular dynamics. i. general method. *The Journal of Chemical Physics*, 31(2):459–466, 1959.

BJ Alder and TEf Wainwright. Phase transition for a hard sphere system. *The Journal of chemical physics*, 27(5):1208, 1957.

Ariane Allain, Isaure Chauvot de Beauchêne, Florent Langenfeld, Yann Guarracino, Elodie Laine, and Luba Tchertanov. Allosteric pathway identification through network analysis: from molecular dynamics simulations to interactive 2d and 3d graphs. *Faraday Discussions*, 2014.

S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25(17):3389–3402, Sep 1997.

William R Atchley, Kurt R Wollenberg, Walter M Fitch, Werner Terhalle, and Andreas W Dress. Correlations among amino acid sites in bhlh protein domains: an information theoretic analysis. *Molecular biology and evolution*, 17(1):164–178, 2000.

RP Bahadur and M Zacharias. The interface of protein-protein complexes: analysis of contacts and prediction of interactions. *Cellular and Molecular Life Sciences*, 65(7-8): 1059–1072, 2008.

Y. Bai, A. Karimi, H. J. Dyson, and P. E. Wright. Absence of a stable intermediate on the folding pathway of protein A. *Protein Sci.*, 6(7):1449–1457, Jul 1997.

J. Baussand and A. Carbone. A combinatorial approach to detect coevolved amino acid networks in protein families of variable divergence. *PLoS Comput. Biol.*, 5(9): e1000488, Sep 2009.

Herman JC Berendsen, J Pl M Postma, Wilfred F van Gunsteren, ARHJ DiNola, and JR Haak. Molecular dynamics with coupling to an external bath. *The Journal of chemical physics*, 81(8):3684–3690, 1984.

A Berk, Sl Zipursky, and H Lodish. Molecular cell biology 4th edition. 2000.

H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Res.*, 28(1): 235–242, Jan 2000.

Moitrayee Bhattacharyya, Chanda R Bhat, and Saraswathi Vishveshwara. An automated approach to network features of protein structure ensembles. *Protein Science : A Publication of the Protein Society*, 22(10):1399–1416, 10 2013.

Kristin Blacklock and Gennady M Verkhivker. Differential modulation of functional dynamics and allosteric interactions in the hsp90-cochaperone complexes with p23 and aha1: A computational study. *PLoS ONE*, 8(8):e71936, 2013.

D. Blocquel, M. Beltrandi, J. Erales, P. Barbier, and S. Longhi. Biochemical and structural studies of the oligomerization domain of the Nipah virus phosphoprotein: evidence for an elongated coiled-coil homotrimer. *Virology*, 446(1-2):162–172, Nov 2013.

David Blocquel, Johnny Habchi, Eric Durand, Marion Sevajol, François Ferron, Jenny Erales, Nicolas Papageorgiou, and Sonia Longhi. Coiled-coil deformations in crystal structures: the measles virus phosphoprotein multimerization domain as an illustrative example. *Acta Crystallographica Section D: Biological Crystallography*, 70(6):1589–1603, 2014.

AC Bloomer, JN Champness, G Bricogne, R Staden, and A Klug. Protein disk of tobacco mosaic virus at 2.8 a resolution showing the interactions within and between subunits. *Nature*, 276(5686):362–368, 1978.

Wolfram Bode, Peter Schwager, and Robert Huber. The transition of bovine trypsinogen to a trypsin-like state upon strong ligand binding: the refined crystal structures of the bovine trypsinogen-pancreatic trypsin inhibitor complex and of its ternary complex with ile-val at 1.9 å resolution. *Journal of molecular biology*, 118(1):99–112, 1978.

C Boede, IA Kovacs, MS Szalay, R Palotai, T Korcsmaros, and P Csermely. Network analysis of protein dynamics. {*FEBS*} *Letters*, 581(15):2776 – 2782, 2007.

K. V. Brinda and S. Vishveshwara. A network representation of protein structures: implications for protein stability. *Biophys. J.*, 89(6):4159–4170, Dec 2005.

C. C. Broder. Henipavirus outbreaks to antivirals: the current status of potential therapeutics. *Curr Opin Virol*, 2(2):176–187, Apr 2012.

Celeste J Brown, Audra K Johnson, A Keith Dunker, and Gary W Daughdrill. Evolution and disorder. *Current opinion in structural biology*, 21(3):441–446, 2011.

Jessica F Bruhn, Katherine C Barnett, Jaclyn Bibby, Jens MH Thomas, Ronan M Keegan, Daniel J Rigden, Zachary A Bornholdt, and Erica Ollmann Saphire. Crystal structure of the nipah virus phosphoprotein tetramerization domain. *Journal of virology*, 88(1): 758–762, 2014.

A. N. Bullock, J. Henckel, B. S. DeDecker, C. M. Johnson, P. V. Nikolova, M. R. Proctor, D. P. Lane, and A. R. Fersht. Thermodynamic stability of wild-type and mutant p53 core domain. *Proc. Natl. Acad. Sci. U.S.A.*, 94(26):14338–14342, Dec 1997.

C. D. Bustamante, J. P. Townsend, and D. L. Hartl. Solvent accessibility and purifying selection within proteins of Escherichia coli and Salmonella enterica. *Mol. Biol. Evol.*, 17(2):301–308, 2000.

J. S. Butler and S. N. Loh. Structure, function, and aggregation of the zinc-free form of the p53 DNA binding domain. *Biochemistry*, 42:2396–2403, 2003.

J. S. Butler and S. N. Loh. Zn(2+)-dependent misfolding of the p53 DNA binding domain. *Biochemistry*, 46:2630–2639, 2007.

S. Calhoun and V. Daggett. Structural effects of the L145Q, V157F, and R282W cancer-associated mutations in the p53 DNA-binding core domain. *Biochemistry*, 50(23): 5345–5353, Jun 2011.

J. M. Canadillas, H. Tidow, S. M. Freund, T. J. Rutherford, H. C. Ang, and A. R. Fersht. Solution structure of p53 core domain: structural basis for its instability. *Proc. Natl. Acad. Sci. U.S.A.*, 103(7):2109–2114, Feb 2006.

E. Capriotti, P. Fariselli, and R. Casadio. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.*, 33(Web Server issue):W306–310, Jul 2005.

Alessandra Carbone. Extracting coevolving characters from a tree of species. In *Discrete and Topological Models in Molecular Biology*, pages 45–65. Springer, 2014.

Alessandra Carbone and Linda Dib. Co-evolution and information signals in biological sequences. *Theoretical Computer Science*, 412(23):2486–2495, 2011.

DA Case, TA Darden, TE Cheatham III, CL Simmerling, J Wang, RE Duke, R Luo, RC Walker, W Zhang, KM Merz, et al. Amber 12. *University of California, San Francisco*, 1(2):3, 2012.

R. Champeimont, E. Laine, S. W. Hu, F. Penin, and A. Carbone. Coevolution analysis of Hepatitis C virus genome to identify the structural and functional dependency network of viral proteins. *Sci Rep*, 6:26401, 2016.

Jean-Pierre Changeux. The feedback control mechanism of biosynthetic l-threonine deaminase by l-isoleucine. In *Cold Spring Harbor symposia on quantitative biology*, volume 26, pages 313–318. Cold Spring Harbor Laboratory Press, 1961.

Rajagopal Chattopadhyaya, William E Meador, Anthony R Means, and Florante A Quiocho. Calmodulin structure refined at 1.7 å resolution. *Journal of molecular biology*, 228(4):1177–1192, 1992.

I. Chauvot de Beauchene, A. Allain, N. Panel, E. Laine, A. Trouve, P. Dubreuil, and L. Tchertanov. Hotspot mutations in KIT receptor differentially modulate its allosterically coupled conformational dynamics: impact on activation and drug sensitivity. *PLoS Comput. Biol.*, 10:e1003749, 2014.

H. Chen, J. Ma, W. Li, A. V. Eliseenkova, C. Xu, T. A. Neubert, W. T. Miller, and M. Mohammadi. A molecular brake in the kinase hinge region regulates the activity of receptor tyrosine kinases. *Mol. Cell*, 27(5):717–730, Sep 2007.

G. Cheng, B. Qian, R. Samudrala, and D. Baker. Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. *Nucleic Acids Res.*, 33(18):5861–5867, 2005.

J. Cheng, A. Randall, and P. Baldi. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins*, 62(4):1125–1132, Mar 2006.

Chakra Chennubhotla and Ivet Bahar. Signal propagation in proteins and relation to equilibrium fluctuations. *PLoS Computational Biology*, 3(9):e172, 2007.

C. N. Chi, L. Elfstrom, Y. Shi, T. Snall, A. Engstrom, and P. Jemth. Reassessing a sparse energetic network within a single protein domain. *Proc. Natl. Acad. Sci. U.S.A.*, 105 (12):4679–4684, Mar 2008.

Federica Chiappori, Ivan Merelli, Giorgio Colombo, Luciano Milanesi, and Giulia Morra. Molecular mechanism of allosteric communication in hsp70 revealed by molecular dynamics simulations. *PLoS Computational Biology*, 8(12):e1002844, 12 2012.

Marcio F Colombo, Donald C Rau, and V Adrian Parsegian. Protein solvation in allosteric regulation: a water effect on hemoglobin. *Science*, 256(5057):655, 1992.

Guillaume Communie, Thibaut Crépin, Damien Maurin, Malene Ringkjøbing Jensen, Martin Blackledge, and Rob WH Ruigrok. Structure of the tetramerization domain of measles virus phosphoprotein. *Journal of virology*, 87(12):7166–7169, 2013.

S. Couve, C. Ladroue, E. Laine, K. Mahtouk, J. Guegan, S. Gad, H. Le Jeune, M. Le Gentil, G. Nuel, W. Y. Kim, B. Lecomte, J. C. Pages, C. Collin, F. Lasne, P. R. Benusiglio, B. Bressac-de Paillerets, J. Feunteun, V. Lazar, A. P. Gimenez-Roqueplo, N. M. Mazure, P. Dessen, L. Tchertanov, D. R. Mole, W. Kaelin, P. Ratcliffe, S. Richard, and B. Gardie. Genetic evidence of a precisely tuned dysregulation in the hypoxia signaling pathway during oncogenesis. *Cancer Res.*, 74(22):6554–6564, Nov 2014.

P. Da Silva Figueiredo Celestino Gomes, N. Panel, E. Laine, P. G. Pascutti, E. Solary, and L. Tchertanov. Differential effects of CSF-1R D802V and KIT D816V homologous mutations on receptor tertiary structure and allosteric communication. *PLoS ONE*, 9 (5):e97519, 2014.

Tom Darden, Darrin York, and Lee Pedersen. Particle mesh ewald: An nlog(n) method for ewald sums in large systems. *The Journal of Chemical Physics*, 98:10089–10092, 1993.

T. A. de Beer, K. Berka, J. M. Thornton, and R. A. Laskowski. PDBsum additions. *Nucleic Acids Res.*, 42(Database issue):D292–296, Jan 2014.

David de Juan, Florencio Pazos, and Alfonso Valencia. Emerging methods in protein co-evolution. *Nature Reviews Genetics*, 14(4):249–261, 2013.

A. M. de Vos, M. Ultsch, and A. A. Kossiakoff. Human growth hormone and extracellular domain of its receptor: crystal structure of the complex. *Science*, 255(5042):306–312, Jan 1992.

Y. Dehouck, J. M. Kwasigroch, D. Gilis, and M. Rooman. PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinformatics*, 12:151, 2011.

Yves Dehouck, Dimitri Gilis, and Marianne Rooman. A new generation of statistical potentials for proteins. *Biophysical journal*, 90(11):4010–4017, 2006.

Antonio del Sol, Hirotomo Fujihashi, Dolors Amoros, and Ruth Nussinov. Residues crucial for maintaining short paths in network communication mediate signaling in proteins. *Molecular systems biology*, 2(1), 2006.

Antonio del Sol, Marcos J Araúzo-Bravo, Dolors Amoros, and Ruth Nussinov. Modular architecture of protein structures and allosteric communications: potential implications for signaling proteins and regulatory linkages. *Genome biology*, 8(5):1, 2007.

Warren L DeLano. The pymol molecular graphics system. 2002.

Xin Deng, Jesse Eickholt, and Jianlin Cheng. A comprehensive overview of computational protein disorder prediction methods. *Molecular BioSystems*, 8(1):114–121, 2012.

Linda Dib and Alessandra Carbone. CLAG: an unsupervised non hierarchical clustering algorithm handling biological data. *BMC bioinformatics*, 13(1):194, 2012a.

Linda Dib and Alessandra Carbone. Protein fragments: functional and structural roles of their coevolution networks. *PloS ONE*, 7(11):e48124, 2012b.

Ruxandra I Dima and D Thirumalai. Determination of network of residues that regulate allostery in protein families using sequence analysis. *Protein Science*, 15(2):258–268, 2006.

A. Dixit and G. M. Verkhivker. Hierarchical modeling of activation mechanisms in the ABL and EGFR kinase domains: thermodynamic and mechanistic catalysts of kinase activation by cancer mutations. *PLoS Comput. Biol.*, 5(8):e1000487, Aug 2009.

Anshuman Dixit and Gennady M Verkhivker. Computational modeling of allosteric communication reveals organizing principles of mutation-induced signaling in abl and egfr kinases. *PLoS Computational Biology*, 7(10):e1002179, 2011.

Z. Dosztanyi, V. Csizmok, P. Tompa, and I. Simon. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, 21(16):3433–3434, Aug 2005.

Z. Dosztanyi, B. Meszaros, and I. Simon. ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics*, 25(20):2745–2746, Oct 2009.

D. A. Doyle, A. Lee, J. Lewis, E. Kim, M. Sheng, and R. MacKinnon. Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ. *Cell*, 85(7):1067–1076, Jun 1996.

Ron O. Dror, Robert M. Dirks, J.P. Grossman, Huafeng Xu, and David E. Shaw. Biomolecular simulation: A computational microscope for molecular biology. *Annual Review of Biophysics*, 41(1):429–452, 2012.

Chester L Drum, Shui-Zhong Yan, Joel Bard, Yue-Quan Shen, Dan Lu, Sandriyana Soelaiman, Zenon Grabarek, Andrew Bohm, and Wei-Jen Tang. Structural basis for the activation of anthrax adenylyl cyclase exotoxin by calmodulin. *Nature*, 415(6870): 396–402, 2002.

A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner, and Z. Obradovic. Intrinsically disordered protein. *J. Mol. Graph. Model.*, 19(1):26–59, 2001.

Laurent Duret, Saïd Abdeddaim, et al. Multiple alignment for structural, functional, or phylogenetic analyses of homologous sequences. *Bioinformatics: Sequence, Structure, and Databanks*, pages 51–76, 2000.

B. T. Eaton, J. S. Mackenzie, and L. F. Wang. Henipaviruses. In B. N. Fields, D.M. Knipe, and P. M. Howley, editors, *Fields Virology*, page 1587–1600. Philadelphia: Lippincott-Raven, 5 edition, 2007.

Charlotte E. Edling and Bengt Hallberg. c-kit – a hematopoietic cell essential receptor tyrosine kinase. *The International Journal of Biochemistry and Cell Biology*, 39(11): 1995–1998, 2007.

S. Engelen, L. A. Trojan, S. Sacquin-Mora, R. Lavery, and A. Carbone. Joint evolutionary trees: a large-scale method to predict protein interfaces based on sequence sampling. *PLoS Comput. Biol.*, 5(1):e1000267, Jan 2009.

Daniel P Faith. Conservation evaluation and phylogenetic diversity. *Biological conservation*, 61(1):1–10, 1992.

Mario A Fares and Simon AA Travers. A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses. *Genetics*, 173(1):9–23, 2006.

Matteo Figliuzzi, Hervé Jacquier, Alexander Schug, Oliver Tenaillon, and Martin Weigt. Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase tem-1. *Molecular biology and evolution*, 33(1):268–280, 2016.

András Fiser, Richard Kinh Gian Do, and Andrej Šali. Modeling of loops in protein structures. *Protein science*, 9(9):1753–1773, 2000.

Z. H. Foda, Y. Shan, E. T. Kim, D. E. Shaw, and M. A. Seeliger. A dynamically coupled allosteric network underlies binding cooperativity in Src kinase. *Nat Commun*, 6:5939, 2015.

D. M. Fowler and S. Fields. Deep mutational scanning: a new style of protein science. *Nat. Methods*, 11(8):801–807, 2014.

H. Frauenfelder, S. G. Sligar, and P. G. Wolynes. The energy landscapes and motions of proteins. *Science*, 254(5038):1598–1603, 1991.

Agya K Frimpong, Rinat R Abzalimov, Vladimir N Uversky, and Igor A Kaltashov. Characterization of intrinsically disordered proteins with electrospray ionization mass spectrometry: Conformational heterogeneity of α-synuclein. *Proteins: Structure, Function, and Bioinformatics*, 78(3):714–722, 2010.

Karl J Fryxell. The coevolution of gene family trees. *Trends in Genetics*, 12(9):364–369, 1996.

K. S. Gajiwala, J. C. Wu, J. Christensen, G. D. Deshmukh, W. Diehl, J. P. DiNitto, J. M. English, M. J. Greig, Y. A. He, S. L. Jacques, E. A. Lunney, M. McTigue, D. Molina, T. Quenzer, P. A. Wells, X. Yu, Y. Zhang, A. Zou, M. R. Emmett, A. G. Marshall, H. M. Zhang, and G. D. Demetri. KIT kinase mutants show unique mechanisms of drug resistance to imatinib and sunitinib in gastrointestinal stromal tumor patients. *Proc. Natl. Acad. Sci. U.S.A.*, 106:1542–1547, 2009.

S. Gianni, S. R. Haq, L. C. Montemiglio, M. C. Jurgens, A. Engstrom, C. N. Chi, M. Brunori, and P. Jemth. Sequence-specific long range networks in PSD-95/discs large/ZO-1 (PDZ) domains tune their binding selectivity. *J. Biol. Chem.*, 286(31): 27167–27175, Aug 2011.

F. Glaser, T. Pupko, I. Paz, R. E. Bell, D. Bechor-Shental, E. Martz, and N. Ben-Tal. ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, 19(1):163–164, Jan 2003.

Gregory B Gloor, Louise C Martin, Lindi M Wahl, and Stanley D Dunn. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry*, 44(19):7156–7165, 2005.

Chern-Sing Goh, Andrew A Bogan, Marcin Joachimiak, Dirk Walther, and Fred E Cohen. Co-evolution of proteins with their interaction partners. *Journal of molecular biology*, 299(2):283–293, 2000.

James Griffith, James Black, Carlos Faerman, Lora Swenson, Michael Wynn, Fan Lu, Judith Lippke, and Kumkum Saxena. The structural basis for autoinhibition of {FLT3} by the juxtamembrane domain. *Molecular Cell*, 13(2):169 – 178, 2004.

George Harauz, Vladimir Ladizhansky, and Joan M Boggs. Structural polymorphism and multifunctionality of myelin basic protein. *Biochemistry*, 48(34):8094–8104, 2009.

Bo He, Kejun Wang, Yunlong Liu, Bin Xue, Vladimir N Uversky, and A Keith Dunker. Predicting intrinsic disorder in proteins: an overview. *Cell research*, 19(8):929–949, 2009.

Katherine Henzler-Wildman and Dorothee Kern. Dynamic personalities of proteins. *Nature*, 450(7172):964–972, 2007.

Nico Tjandra Hitoshi Kuboniwa, Hao Ren Stephan Grzesiek, B Claude, and Ad Bax. Solution structure of calcium-free calmodulin. *Nature structural biology*, 2(9), 1995.

S. Hubbard and J. Thornton. http://www.bioinf.manchester.ac.uk/naccess/., 1992-6.

Morgan Huse and John Kuriyan. The conformational plasticity of protein kinases. *Cell*, 109(3):275 – 282, 2002.

Mitsuhiko Ikura, Marius Clore, et al. Solution structure of a calmodulin-target peptide complex by multidimensional nmr. *Science*, 256(5057):632, 1992.

C. A. Innis. siteFiNDER—3D: a web-based tool for predicting the location of functional sites in proteins. *Nucleic Acids Res.*, 35(Web Server issue):W489–494, Jul 2007.

Gaetano Invernizzi, Matteo Tiberti, Matteo Lambrughi, Kresten Lindorff-Larsen, and Elena Papaleo. Communication routes in arid domains between distal residues in helix 5 and the dna-binding loops. *PLoS Computational Biology*, 10(9):e1003744, 09 2014.

Takashi Ishida and Kengo Kinoshita. Prdos: prediction of disordered protein regions from amino acid sequence. *Nucleic acids research*, 35(suppl 2):W460–W464, 2007.

David T Jones, Daniel WA Buchan, Domenico Cozzetto, and Massimiliano Pontil. Psicov: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, 28(2):184–190, 2012.

Robert Roskoski Jr. Structure and regulation of kit protein-tyrosine kinase – the stem cell factor receptor. *Biochemical and Biophysical Research Communications*, 338(3):1307 – 1315, 2005.

W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, Dec 1983.

Robert Kalescky, Jin Liu, and Peng Tao. Identifying key residues for protein allostery through rigid residue scan. *The Journal of Physical Chemistry A*, 119(9):1689–1700, 2014.

B. Kamaraj and A. Bogaerts. Structure and Function of p53-DNA Complexes with Inactivation and Rescue Mutations: A Molecular Dynamics Simulation Study. *PLoS ONE*, 10(8):e0134638, 2015.

Y. Karami, E. Laine, and A. Carbone. Dissecting protein architecture with communication blocks and communicating segment pairs. *BMC Bioinformatics*, 17 Suppl 2:13, 2016.

D. Karlin, F. Ferron, B. Canard, and S. Longhi. Structural disorder and modular organization in Paramyxovirinae N and P. *J. Gen. Virol.*, 84(Pt 12):3239–3252, Dec 2003.

M. B. Kennedy. Origin of PDZ (DHR, GLGF) domains. *Trends Biochem. Sci.*, 20(9):350, Sep 1995.

Dorothee Kern and Erik RP Zuiderweg. The role of dynamics in allosteric regulation. *Current Opinion in Structural Biology*, 13(6):748 – 757, 2003.

Bostjan Kobe, Gregor Guncar, Rebecca Buchholz, Thomas Huber, Bohumil Maco, Nathan Cowieson, Jennifer L Martin, Mary Marfori, and Jade K Forwood. Crystallography and protein–protein interactions: biological interfaces and crystal contacts. *Biochemical Society Transactions*, 36(6):1438–1441, 2008.

John Kuriyan. Allostery and coupled sequence variation in nuclear hormone receptors. *Cell*, 116(3):354–356, 2004.

E. Laine, I. Chauvot de Beauchene, D. Perahia, C. Auclair, and L. Tchertanov. Mutation D816V alters the internal structure and dynamics of c-KIT receptor cytoplasmic region: implications for dimerization and activation mechanisms. *PLoS Comput. Biol.*, 7(6): e1002068, Jun 2011a.

E. Laine, I. Chauvot de Beauchene, D. Perahia, C. Auclair, and L. Tchertanov. Mutation D816V alters the internal structure and dynamics of c-KIT receptor cytoplasmic region: implications for dimerization and activation mechanisms. *PLoS Comput. Biol.*, 7(6): e1002068, Jun 2011b.

Elodie Laine, Christian Auclair, and Luba Tchertanov. Allosteric communication across the native and mutated kit receptor tyrosine kinase. *PLoS Computational Biology*, 8(8), 2012.

Meytal Landau, Itay Mayrose, Yossi Rosenberg, Fabian Glaser, Eric Martz, Tal Pupko, and Nir Ben-Tal. Consurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic acids research*, 33(suppl 2):W299–W302, 2005.

Mark A Larkin, Gordon Blackshields, NP Brown, R Chenna, Paul A McGettigan, Hamish McWilliam, Franck Valentin, Iain M Wallace, Andreas Wilm, Rodrigo Lopez, et al. Clustal w and clustal x version 2.0. *bioinformatics*, 23(21):2947–2948, 2007.

Odile Lecompte, Julie D Thompson, Frédéric Plewniak, Jean-Claude Thierry, and Olivier Poch. Multiple alignment of complete sequences (macs) in the post-genomic era. *Gene*, 270(1):17–30, 2001.

H. J. Lee and J. J. Zheng. PDZ domains and their binding partners: structure, specificity, and modification. *Cell Commun. Signal*, 8:8, 2010.

H. Lei, C. Wu, Z. X. Wang, Y. Zhou, and Y. Duan. Folding processes of the B domain of protein A to the native state observed in all-atom ab initio folding simulations. *J Chem Phys*, 128(23):235105, Jun 2008.

Mark A Lemmon and Joseph Schlessinger. Cell signaling by receptor-tyrosine kinases. *Cell*, 141(7):1117–1134, 06 2010.

T. Li, N. Kon, L. Jiang, M. Tan, T. Ludwig, Y. Zhao, R. Baer, and W. Gu. Tumor suppression in the absence of p53-mediated cell-cycle arrest, apoptosis, and senescence. *Cell*, 149(6):1269–1283, Jun 2012.

L-Y Lian. Nmr structural studies of glutathione s-transferase. *Cellular and Molecular Life Sciences CMLS*, 54(4):359–362, 1998.

O. Lichtarge and A. Wilkins. Evolution: a guide to perturb protein function and networks. *Curr. Opin. Struct. Biol.*, 20(3):351–359, Jun 2010.

O. Lichtarge, H. R. Bourne, and F. E. Cohen. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, 257(2):342–358, Mar 1996.

Olivier Lichtarge and Mathew E Sowa. Evolutionary predictions of binding surfaces and interactions. *Current opinion in structural biology*, 12(1):21–27, 2002.

R. C. Liddington. Anthrax: a molecular full nelson. *Nature*, 415(6870):373–374, Jan 2002.

J. Liu and R. Nussinov. Allosteric effects in the marginally stable von Hippel-Lindau tumor suppressor protein and allostery-based rescue mutant design. *Proc. Natl. Acad. Sci. U.S.A.*, 105(3):901–906, Jan 2008.

Steve W Lockless and Rama Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438):295–299, 1999.

Richard J Loncharich, Bernard R Brooks, and Richard W Pastor. Langevin dynamics of peptides: The frictional dependence of isomerization rates of n-acetylalanyl-N'-methylamide. *Biopolymers*, 32(5):523–535, 1992.

S. Lu, H. Jang, R. Nussinov, and J. Zhang. The Structural Basis of Oncogenic Mutations G12, G13 and Q61 in Small GTPase K-Ras4B. *Sci Rep*, 6:21949, 2016.

W. J. Lu, J. F. Amatruda, and J. M. Abrams. p53 ancestry: gazing through an evolutionary lens. *Nat. Rev. Cancer*, 9(10):758–762, Oct 2009.

S. Lukman, D. P. Lane, and C. S. Verma. Mapping the structural and dynamical features of multiple p53 DNA binding domains: insights into loop 1 intrinsic dynamics. *PLoS ONE*, 8(11):e80221, 2013.

A. Lupas, M. Van Dyke, and J. Stock. Predicting coiled coils from protein sequences. *Science*, 252(5009):1162–1164, May 1991.

E. Lyman and D. M. Zuckerman. Ensemble-based convergence analysis of biomolecular trajectories. *Biophys. J.*, 91:164–172, 2006.

M. A. Marti-Renom, A. C. Stuart, A. Fiser, R. Sanchez, F. Melo, and A. Sali. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct*, 29:291–325, 2000.

J Andrew McCammon, Bruce R Gelin, and Martin Karplus. Dynamics of folded proteins. *Nature*, 267(5612):585–590, 1977.

Christopher L. McClendon, Alexandr P. Kornev, Michael K. Gilson, and Susan S. Taylor. Dynamic architecture of a protein kinase. *Proceedings of the National Academy of Sciences*, 111(43):E4623–E4631, 2014.

I. K. McDonald and J. M. Thornton. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.*, 238(5):777–793, May 1994.

Robert T. McGibbon, Kyle A. Beauchamp, Christian R. Schwantes, Lee-Ping Wang, Carlos X. Hernández, Matthew P. Harrigan, Thomas J. Lane, Jason M. Swails, and Vijay S. Pande. Mdtraj: a modern, open library for the analysis of molecular dynamics trajectories. *bioRxiv*, 2014. doi: 10.1101/008896.

R. N. McLaughlin, F. J. Poelwijk, A. Raman, W. S. Gosal, and R. Ranganathan. The spatial architecture of protein function and adaptation. *Nature*, 491(7422):138–142, Nov 2012.

Richard N McLaughlin Jr, Frank J Poelwijk, Arjun Raman, Walraj S Gosal, and Rama Ranganathan. The spatial architecture of protein function and adaptation. *Nature*, 491 (7422):138–142, 2012.

Yang Mei, Minfei Su, Gaurav Soni, Saeed Salem, Christopher L Colbert, and Sangita C Sinha. Intrinsically disordered regions in autophagy proteins. *Proteins: Structure, Function, and Bioinformatics*, 82(4):565–578, 2014.

Lasota J. Miettinen M, Majidi M. Pathology and diagnostic criteria of gastrointestinal stromal tumors (gists): a review. *Eur J Cancer.*, Suppl 5(Suppl 5):S39–51, 2002.

I. Mihalek, I. Res, and O. Lichtarge. A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J. Mol. Biol.*, 336(5):1265–1282, Mar 2004.

Hisashi Mizutani, K Saraboji, SM Malathy Sony, MN Ponnuswamy, T Kumarevel, Krishna Swamy, DK Simanshu, MRN Murthy, and Naoki Kunishima. Systematic study on crystal-contact engineering of diphthine synthase: influence of mutations at crystal-packing regions on x-ray diffraction quality. *Acta Crystallographica Section D: Biological Crystallography*, 64(10):1020–1033, 2008.

J. Monod, J. Wyman, and J. P. Changeux. On the Nature of Allosteric Transitions: A Plausible Model. *J Mol Biol*, 12:88–118, 1965.

Jacques Monod and François Jacob. General conclusions: teleonomic mechanisms in cellular metabolism, growth, and differentiation. In *Cold Spring Harbor symposia on quantitative biology*, volume 26, pages 389–401. Cold Spring Harbor Laboratory Press, 1961.

F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U.S.A.*, 108(49):E1293–1301, Dec 2011.

Faruck Morcos, Biman Jana, Terence Hwa, and José N Onuchic. Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proceedings of the National Academy of Sciences*, 110(51):20533–20538, 2013.

G. Morra, A. Genoni, and G. Colombo. Mechanisms of Differential Allosteric Modulation in Homologous Proteins: Insights from the Analysis of Internal Dynamics and Energetics of PDZ Domains. *J Chem Theory Comput*, 10(12):5677–5689, Dec 2014.

Hesam N Motlagh, James O Wrabl, Jing Li, and Vincent J Hilser. The ensemble nature of allostery. *Nature*, 508(7496):331–339, 2014.

Javier Murciano-Calles, Carles Corbi-Verge, Adela M Candel, Irene Luque, and Jose C Martinez. Post-translational modifications modulate ligand recognition by the third pdz domain of the maguk protein psd-95. *PloS one*, 9(2), 2014.

P. C. Ng and S. Henikoff. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, 31(13):3812–3814, Jul 2003.

Brad Nolen, Susan Taylor, and Gourisankar Ghosh. Regulation of protein kinases: Controlling activity through activation segment conformation. *Molecular Cell*, 15(5):661 – 675, 2004.

Cédric Notredame. Recent progress in multiple sequence alignment: a survey. *Pharmacogenomics*, 3(1):131–144, 2002.

Cédric Notredame. Recent evolutions of multiple sequence alignment algorithms. *PLoS Comput Biol*, 3(8):e123, 2007.

A. L. Okorokov and E. V. Orlova. Structural biology of the p53 tumour suppressor. *Curr. Opin. Struct. Biol.*, 19(2):197–202, Apr 2009.

A Orfao, AC Garcia-Montero, L Sanchez, and L Escribano. Recent advances in the understanding of mastocytosis: the role of kit mutations*. *British Journal of Haematology*, 138(1):12–30, 2007.

Alessandro Pandini, Arianna Fornili, Franca Fraternali, and Jens Kleinjung. Gsatools: analysis of allosteric communication and functional local motions using a structural alphabet. *Bioinformatics*, 29(16):2053–2055, 2013.

Elena Papaleo, Giulia Renzetti, and Matteo Tiberti. Mechanisms of intramolecular communication in a hyperthermophilic acylaminoacyl peptidase: A molecular dynamics investigation. *PLoS ONE*, 7(4):e35686, 2012.

Florencio Pazos and Alfonso Valencia. Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein engineering*, 14(9):609–614, 2001.

Zhen-Ling Peng and Lukasz Kurgan. Comprehensive comparative assessment of in-silico predictors of disordered regions. *Current Protein and Peptide Science*, 13(1):6–18, 2012.

Max F Perutz and LF TenEyck. Stereochemistry of cooperative effects in hemoglobin. In *Cold Spring Harbor symposia on quantitative biology*, volume 36, pages 295–310. Cold Spring Harbor Laboratory Press, 1972.

Max F Perutz, AJ Wilkinson, M Paoli, and GG Dodson. The stereochemical mechanism of the cooperative effects in hemoglobin revisited. *Annual review of biophysics and biomolecular structure*, 27(1):1–34, 1998.

V. Petkovic, M. Godi, A. V. Pandey, D. Lochmatter, C. R. Buchanan, M. T. Dattani, A. Eble, C. E. Fluck, and P. E. Mullis. Growth hormone (GH) deficiency type II: a novel GH-1 gene mutation (GH-R178H) affecting secretion and action. *J. Clin. Endocrinol. Metab.*, 95(2):731–739, Feb 2010.

Stefano Piana, John L Klepeis, and David E Shaw. Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. *Current Opinion in Structural Biology*, 24(0):98 – 105, 2014.

DD Pollock and WR Taylor. Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. *Protein Engineering*, 10(6):647–657, 1997.

T. Pupko, R. E. Bell, I. Mayrose, F. Glaser, and N. Ben-Tal. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, 18 Suppl 1:S71–77, 2002.

F H Qiu, P Ray, K Brown, P E Barker, S Jhanwar, F H Ruddle, and P Besmer. Primary structure of c-kit: relationship with the csf-1/pdgf receptor kinase family–oncogenic activation of v-kit involves deletion of extracellular domain and c terminus. *The EMBO Journal*, 7(4):1003–1011, 04 1988.

A Rahman. Correlations in the motion of atoms in liquid argon. *Physical Review*, 136 (2A):A405, 1964.

Francesco Raimondi, Angelo Felline, Michele Seeber, Simona Mariani, and Francesca Fanelli. A mixed protein structure network and elastic network model approach to predict the structural communication in biomolecular systems: The pdz2 domain from tyrosine phosphatase 1e as a case study. *Journal of Chemical Theory and Computation*, 9(5):2504–2518, 2013.

Arun K Ramani and Edward M Marcotte. Exploiting the co-evolution of interacting proteins to discover interaction specificity. *Journal of molecular biology*, 327(1):273–284, 2003.

D. C. Ramsey, M. P. Scherrer, T. Zhou, and C. O. Wilke. The relationship between relative solvent accessibility and evolutionary rate in protein evolution. *Genetics*, 188(2):479–488, 2011.

Michael I Sadowski, Katarzyna Maksimiak, and William R Taylor. Direct correlation analysis improves fold recognition. *Computational Biology and Chemistry*, 35(5):323–332, 2011.

G. Saladino and F. L. Gervasio. Modeling the effect of pathogenic mutations on the conformational landscape of protein kinases. *Curr. Opin. Struct. Biol.*, 37:108–114, Apr 2016.

S. Sankararaman, B. Kolaczkowski, and K. Sjolander. INTREPID: a web server for prediction of functionally important residues by evolutionary analysis. *Nucleic Acids Res.*, 37(Web Server issue):W390–395, Jul 2009.

S. Sato, T. L. Religa, V. Daggett, and A. R. Fersht. Testing protein-folding simulations by experiment: B domain of protein A. *Proc. Natl. Acad. Sci. U.S.A.*, 101(18):6952–6956, May 2004.

S. Sato, T. L. Religa, and A. R. Fersht. Phi-analysis of the folding of the B domain of protein A using multiple optical probes. *J. Mol. Biol.*, 360(4):850–864, 2006.

Travis P Schrank, D Wayne Bolen, and Vincent J Hilser. Rational modulation of conformational fluctuations in adenylate kinase reveals a local unfolding mechanism for allostery and functional adaptation in proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 106(40):16984–16989, 10 2009.

Michele Seeber, Angelo Felline, Francesco Raimondi, Stefanie Muff, Ran Friedman, Francesco Rao, Amedeo Caflisch, and Francesca Fanelli. Wordom: A user-friendly program for the analysis of molecular structures, trajectories, and free energy surfaces. *Journal of Computational Chemistry*, 32(6):1183–1194, 04 2011.

David J Sidote and David W Hoffman. Nmr structure of an archaeal homologue of ribonuclease p protein rpp29. *Biochemistry*, 42(46):13541–13550, 2003.

Lars Skjærven, Xin-Qiu Yao, Guido Scarabelli, and Barry J Grant. Integrating protein structural dynamics and evolutionary analysis with bio3d. *BMC Bioinformatics*, 15(1): 399, 2014.

C. A. Smith and T. Kortemme. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *J. Mol. Biol.*, 380(4):742–756, Jul 2008.

J. Stetefeld, M. Jenny, T. Schulthess, R. Landwehr, J. Engel, and R. A. Kammerer. Crystal structure of a naturally occurring parallel right-handed coiled coil tetramer. *Nat. Struct. Biol.*, 7(9):772–776, Sep 2000.

Frank H Stillinger and Aneesur Rahman. Improved simulation of liquid water by molecular dynamics. *The Journal of Chemical Physics*, 60(4):1545–1557, 1974.

Read AP. Strachan T. *Human Molecular Genetics. 2nd edition.* New York: Wiley-Liss, 1999.

Gürol M Süel, Steve W Lockless, Mark A Wall, and Rama Ranganathan. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural & Molecular Biology*, 10(1):59–69, 2003a.

Gürol M Süel, Steve W Lockless, Mark A Wall, and Rama Ranganathan. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural & Molecular Biology*, 10(1):59–69, 2003b.

M. Sundstrom, T. Lundqvist, J. Rodin, L. B. Giebel, D. Milligan, and G. Norstedt. Crystal structure of an antagonist mutant of human growth hormone, G120R, in complex with its receptor at 2.9 A resolution. *J. Biol. Chem.*, 271(50):32197–32203, Dec 1996.

Y. M. Sung, A. D. Wilkins, G. J. Rodriguez, T. G. Wensel, and O. Lichtarge. Intramolecular allosteric communication in dopamine D2 receptor revealed by evolutionary amino acid covariation. *Proc. Natl. Acad. Sci. U.S.A.*, 113(13):3539–3544, Mar 2016.

Julie D Thompson, Frédéric Plewniak, and Olivier Poch. A comprehensive comparison of multiple sequence alignment programs. *Nucleic acids research*, 27(13):2682–2690, 1999.

Matteo Tiberti, Gaetano Invernizzi, Matteo Lambrughi, Yuval Inbar, Gideon Schreiber, and Elena Papaleo. Pyinteraph: A framework for the analysis of interaction networks in structural ensembles of proteins. *Journal of Chemical Information and Modeling*, 54(5):1537–1551, 2014.

Garima Tiwari and Debasisa Mohanty. An in silico analysis of the binding modes and binding affinities of small molecule modulators of pdz-peptide interactions. *PloS one*, 8(8), 2013.

Chung-Jung Tsai, Antonio del Sol, and Ruth Nussinov. Allostery: Absence of a change in shape does not imply that allostery is not at play. *Journal of Molecular Biology*, 378 (1):1 – 11, 2008.

Vladimir N Uversky and A Keith Dunker. Understanding protein non-folding. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1804(6):1231–1264, 2010.

S. W. Vetter and E. Leclerc. Novel aspects of calmodulin target recognition and activation. *Eur. J. Biochem.*, 270(3):404–414, Feb 2003.

Hansia P. Vishveshwara S, Ghosh A. Intra and inter-molecular communications through protein structure network. *Curr Protein Pept Sci*, 10(2):146–60, 2009.

B. Vogelstein, D. Lane, and A. J. Levine. Surfing the p53 network. *Nature*, 408(6810): 307–310, Nov 2000.

K. H. Vousden and C. Prives. Blinded by the Light: The Growing Complexity of p53. *Cell*, 137(3):413–431, May 2009.

D. M. Vu, J. K. Myers, T. G. Oas, and R. B. Dyer. Probing the folding and unfolding dynamics of secondary and tertiary structures in a three-helix bundle protein. *Biochemistry*, 43(12):3582–3589, Mar 2004.

Michael E Wall, James B Clarage, and George N Phillips. Motions of calmodulin characterized using both bragg and diffuse x-ray scattering. *Structure*, 5(12):1599–1612, 1997.

Iain M Wallace, Gordon Blackshields, and Desmond G Higgins. Multiple sequence alignments. *Current opinion in structural biology*, 15(3):261–266, 2005.

S. T. Walsh, J. E. Sylvester, and A. A. Kossiakoff. The high- and low-affinity receptor binding sites of growth hormone are allosterically coupled. *Proc. Natl. Acad. Sci. U.S.A.*, 101(49):17078–17083, Dec 2004.

James D Watson, Roman A Laskowski, and Janet M Thornton. Predicting protein function from sequence and structural data. *Current opinion in structural biology*, 15(3):275–284, 2005.

Gregorio Weber. Ligand binding and internal equilibiums in proteins. *Biochemistry*, 11 (5):864–878, 1972.

M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. U.S.A.*, 106(1):67–72, Jan 2009.

Willy Wriggers, Kate A. Stafford, Yibing Shan, Stefano Piana, Paul Maragakis, Kresten Lindorff-Larsen, Patrick J. Miller, Justin Gullingsrud, Charles A. Rendleman, Michael P. Eastwood, Ron O. Dror, and David E. Shaw. Automated event detection and activity monitoring in long molecular dynamics simulations. *Journal of Chemical Theory and Computation*, 5(10):2595–2605, 2009.

Peter E Wright and H Jane Dyson. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of molecular biology*, 293(2):321–331, 1999.

Chen-Hsiang Yeang and David Haussler. Detecting coevolution in and among protein domains. *PLoS Comput Biol*, 3(11):e211, 2007.

Z. Zhang and W. Wriggers. Local feature analysis: a statistical theory for reproducible essential dynamics of large macromolecules. *Proteins*, 64(2):391–403, Aug 2006.