Prédiction des résidus impliqués dans le noyau du repliement et classification structurale de fragments protéiques en interaction.

Nicolas Prudhomme

Thèse

Prédiction des résidus clés du repliement et Classification structurale de fragments protéiques en interaction.

Nicolas Prudhomme

09 Novembre 2009

Sommaire

0. Introduction

0.1. Qu'est-ce qu'une protéine ?

Une protéine est une suite d'acides aminés reliés les uns aux autres. Dès lors qu'un acide aminé est inclus dans la suite on l'appelle résidu. Cette suite de résidus donne une chaîne linéaire qui va se replier sur elle-même afin de former la protéine dans sa forme native. Notre étude du repliement se fera sur les protéines globulaires compactes, néanmoins il existe plusieurs types de protéines et on peut citer en dehors des protéines globulaires ou hydrosolubles, les protéines fibrillaires et les protéines membranaires. Pour qu'une protéine soit active il se peut qu'elle ait besoin de petites molécules pour fonctionner, que l'on appelle des cofacteurs. Certaines protéines ne sont actives qu'en présence de, et complexées à d'autres protéines. Il existe des protéines structurales dont la fonction est de permettre l'intégrité de la cellule en participant au cytosquelette. D'autres sont enzymatiques et sont là pour assurer le bon fonctionnement de toutes les réactions chimiques nécessaires au métabolisme.

Le long du squelette protéique, commun aux vingt acides aminés naturels, on peut observer des structures régulières. Celles-ci résultent du fait que la conformation du squelette d'un acide aminé est parfaitement décrite par la connaissance de trois angles dièdres, et que ceux-ci ne peuvent pas prendre toutes les valeurs possibles, car ils sont limités par les contraintes stériques imposées par la présence des chaînes latérales. La suite d'acides aminés peut ainsi se structurer en hélice, se tendre en brin et former des feuillets en rapprochant ces brins. Ceci est la structuration secondaire.

Ces structures régulières se forment tout au long de la séquence (constituée par la liste des acides aminés successifs en allant du N terminal au C terminal) et sont espacées par des régions du squelette moins structurées. On appelle ces régions des boucles, et comme leur nom le décrit, elles permettent aux structures secondaires régulières (SSR) de se rapprocher dans l'espace et de s'associer entre elles. La figure 0.1 représente les deux structures secondaires régulières majoritairement représentées (Hélice α et Brin β).

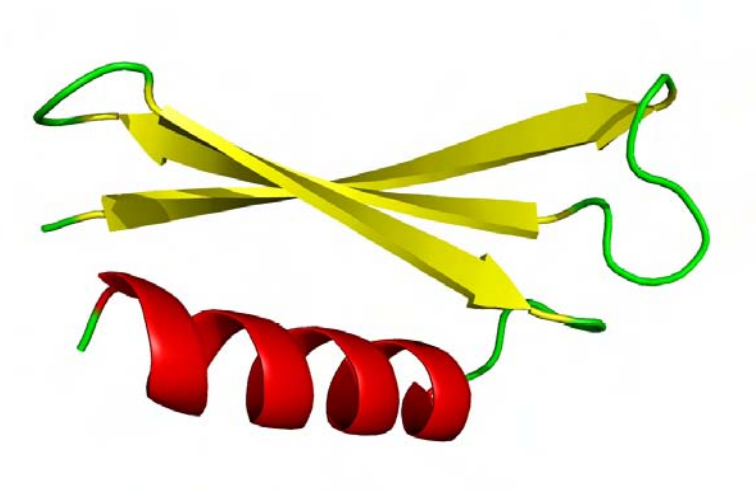


Figure 0.1 Exemple d'hélice alpha (rouge) et de feuillet bêta (jaune) à trois brins antiparallèles reliées par des boucles (vert)

Lorsque la totalité de la chaîne est structurée et que toutes les parties de structuration secondaire sont associées et en interaction, la chaîne polypeptidique est appelée globule. C'est au globule protéique compact que correspond la structure tertiaire.

C'est en se regroupant et en établissant des interactions entre eux que plusieurs globules forment la structure quaternaire, que ces globules soient constitués de plusieurs chaînes de séquences identiques ou non (respectivement, homopolymères et hétéro-polymères).

Il arrive que dans un globule il y ait plusieurs groupes de structures secondaires associées entre elles. Chaque groupe de structures secondaires regroupées et compactes forme un domaine. Le domaine est la plus grosse dimension à pouvoir être prédite avec une précision satisfaisante à l'heure actuelle. Afin de prédire des chaînes protéiques composées de multiple domaines, il faut d'abord prédire la structure tertiaire de chaque domaine et les assembler ensuite. Il existe des banques de données décrivant les limites de chaque domaine dans les protéines multi domaines. La banque de données CATH (Orengo et al., 1997; Orengo et al., 1998), en plus de nous donner les cartographies des protéines avec les limites des domaines qui les composent, nous apporte l'information du type de repliement (au sens de l'architecture structurale) que le domaine adopte.

Chaque domaine possède un enchaînement de structures secondaires et une topologie dans leur agencement qui lui est propre. C'est cet agencement que nous appelons repliement. Une estimation du nombre de repliements que l'on peut tirer des banques telles que CATH ou SCOP (Murzin et al., 1995) indique qu'il y aurait un peu plus d'un millier de repliements différents dans lesquels les séquences prennent leur conformation native.

Nous avons utilisé la banque CATH pour la première partie du travail. En effet cette banque est très facilement utilisable et mieux documentée que SCOP. Ce ne sont que sur des considérations pratiques que notre choix à été motivé, car il n'y a pas de différences entre les résultats des deux banques quant à l'attribution des limites des domaines.

Modularité des protéines

Au niveau du domaine, les extrémités ont tendance à se retrouver dans un espace confiné, avec les deux extrémités N et C proches dans l'espace (Fourty, 2006). Lorsque l'on regarde l'organisation du vivant, on s'aperçoit que la nature a l'air de conserver des analogies tout au long de la diminution des dimensions étudiées. Selon ces deux informations, il semblerait logique que les domaines doivent être construits par des modules régissant les mêmes propriétés que les parties qui les constituent. Ainsi, il est intéressant de regarder la conservation de fragments refermés du squelette protéique. C'est pourquoi nous allons nous intéresser à des fragments ayant certaines propriétés au cœur des domaines. La propriété la plus importante de ces fragments est que leurs extrémités doivent être rapprochées dans l'espace. On peut ensuite penser que les domaines des protéines ne sont qu'un assemblage de tels fragments. L'analyse structurale de ces fragments peut se révéler intéressante dans la compréhension du cœur protéique, et plus précisément du noyau du repliement.

0.2. Les acides aminés

Les acides aminés sont de petites molécules composées d'un groupement acide carboxylique, un groupement amine et une chaîne latérale variable. La partie commune est composée également de deux atomes de carbones. Sur le carbone dénommé alpha est branchée la chaîne latérale et le second porte le groupement acide.

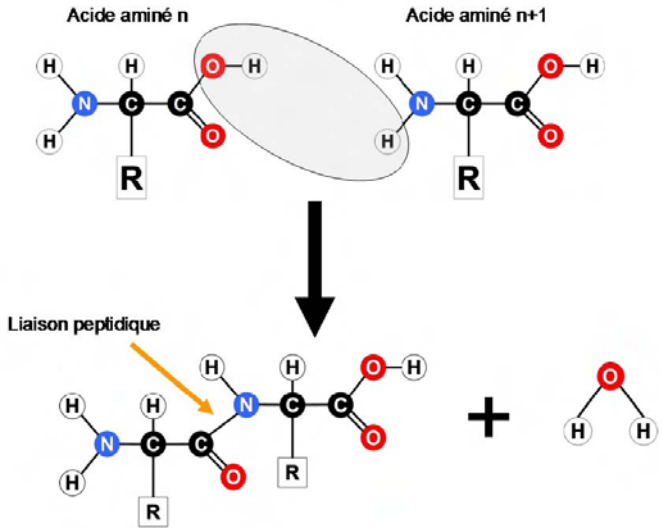


Figure 0.2 Liaison peptidique

Les acides aminés sont des molécules chirales, c'est-à-dire que leur conformation n'est pas superposable à l'image dans un miroir de cette dernière. Il y a donc deux conformations possibles d'acides aminés, appelées L et D. La conformation majoritairement représentée dans le monde du vivant, et donc dans les protéines, est la conformation L.

Les propriétés particulières des acides aminés, telles que le fait d'être polaires ou apolaires, aromatiques, aliphatiques sont conférées par leur chaîne latérale. La figure 0.1.1. illustre ces différentes propriétés. Les acides aminés sont représentés par leur code à une lettre selon la nomenclature internationale et sont regroupés par des cercles d'ensembles partageant une même propriété. Il existe dans la littérature plusieurs échelles d'hydrophobie des acides aminés (Koshi et Goldstein, 1997; Ladunga et Smith, 1997; Rose et al., 1985), établies en fonction de la propriété étudiée. Dans notre groupe, dont l'activité est essentiellement tournée vers la prédiction structurale, sont considérés comme hydrophobes les résidus suivants : Isoleucine, Leucine, Méthionine, Phénylalanine, Tryptophane, Tyrosine et Valine (Callebaut et al., 1997).

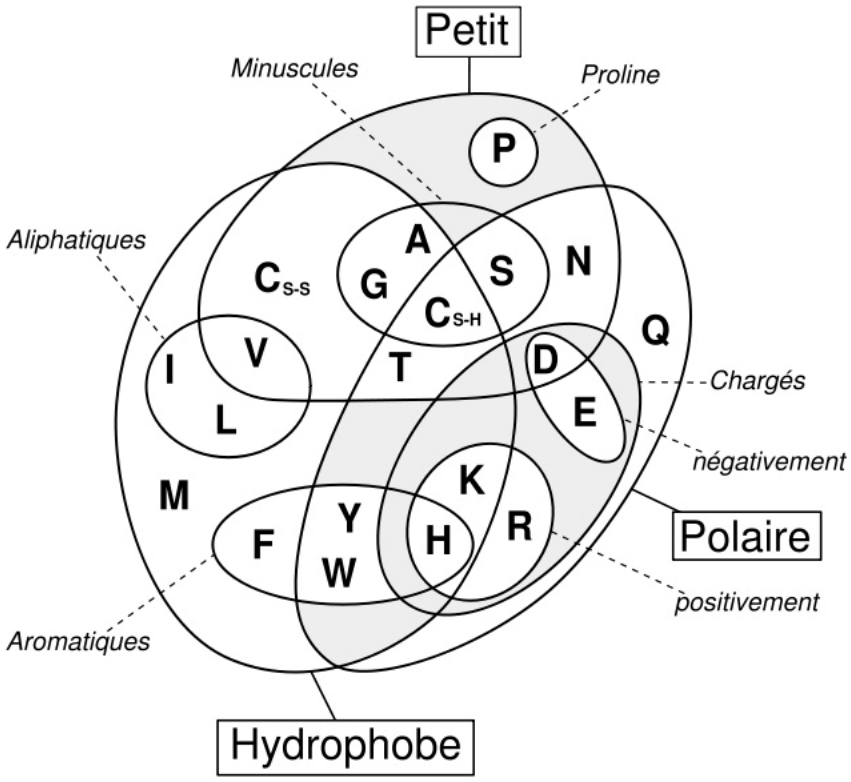


Figure 0.1.1. Diagramme de Venn des 20 acides aminés naturels selon le regroupement en classes, d'après (Taylor, 1986).

La polymérisation des acides aminés s'effectue toujours de la même façon. C'est l'attaque du groupe carboxyle sur le groupe amine. C'est parce que la synthèse est unidirectionnelle que les chaînes sont polarisées et possèdent un N- et un C-terminal.

La structure des acides aminés et la connexion entre eux par une liaison peptidique leur confèrent des propriétés géométriques particulières. En effet, pour cause d'encombrement stérique des chaînes latérales, certaines conformations ne sont pas autorisées. Ramachandran (Ramachandran et Sasisekharan, 1968; Ramakrishnan et Ramachandran, 1965) fut le premier à énoncer ces propriétés et à les mettre en rapport avec les angles de torsions le long des liaisons covalentes entre les différents atomes du squelette de l'acide aminé. Pour illustrer ce propos, la figure 0.1.2. montre comment deux acides aminés contigus peuvent être représentés par deux plans. A l'intérieur d'un plan, la position des atomes du squelette peut être décrite par les angles de torsions des liaisons N-CA, CA-C et de la liaison peptidique, dénommés respectivement PHI (φ), PSI (ψ) et OMEGA (ω). Il se trouve que la liaison peptidique perd en partie son caractère covalent par conjugaison avec la double liaison du groupe CO voisin. Ceci produit comme avantage que l'on peut complètement définir la conformation du squelette d'une protéine par la seule connaissance de la paire (φ, ψ) d'angles dièdres de chaque acide aminé.

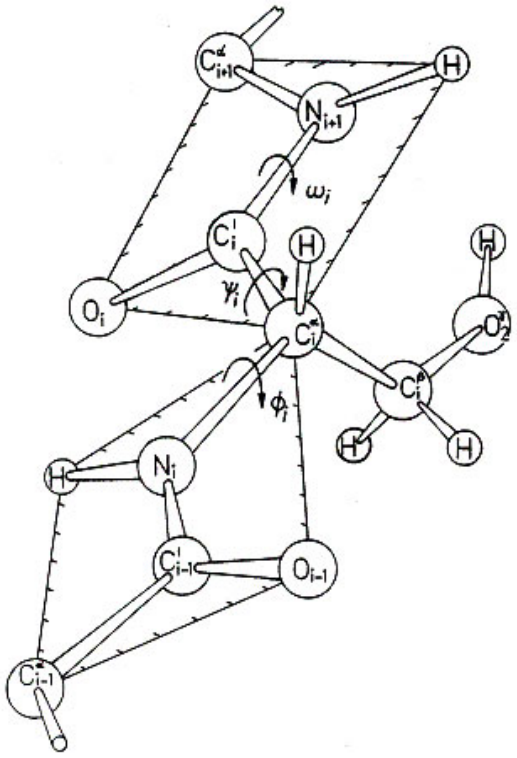


Figure 0.1.2. Conformation de deux acides aminés successifs, déterminée par les angles dièdres.

Sur un jeu constitué par les quelques dizaines de protéines dont la structure était disponible au milieu de années 1960, il a été montré que les valeurs prises par les paires (φ, ψ) sont limitées à certaines valeurs. Ces seules conformations autorisées sont dues à l'encombrement stérique provoqué par les chaînes latérales. Sur une carte dont les abscisses et les ordonnées correspondent aux angles dièdres φ et ψ, on peut positionner ces valeurs pour tous les acides aminés d'une protéine donnée. Cette carte est

plus connue sous le nom de diagramme de Ramachandran, comme la figure 0.1.3 en donne un exemple. On s'aperçoit que ce diagramme comporte trois zones favorables. Lorsqu'on analyse une structure protéique, on observe que tous les acides aminés ont des combinaisons d'angles (φ, ψ) qui s'inscrivent à l'intérieur de ces trois zones. Les deux principales régions correspondent aux structures secondaires régulières qui sont majoritairement observées dans les protéines : la région des hélices α et celle des feuillets β. La troisième région, qui est plus petite, correspond à une conformation en hélice gauche (φ>0). On y trouve principalement des glycines, qui ne comportent pas de chaîne latérale et ont une plus grande liberté conformationnelle.

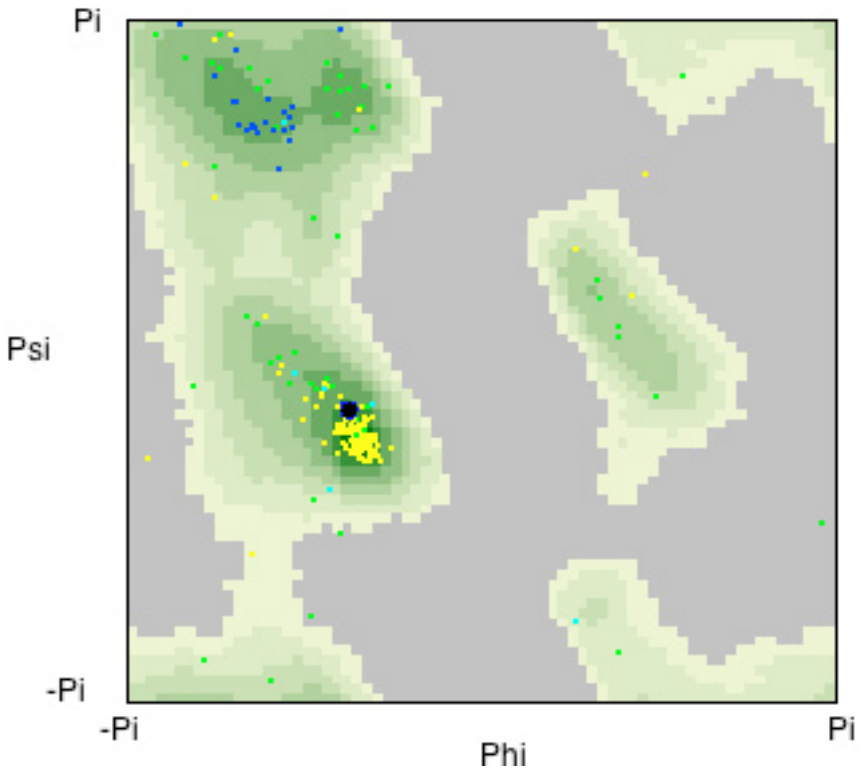


Figure 0.1.3. Diagramme de Ramachandran de la protéine de code PDB 1u3i. Les zones énergétiquement favorables sont représentées par des zones colorées allant du blanc au vert, et on les appelle les zones autorisées. Chaque acide aminé est représenté par un point.

La zone autorisée correspondant aux feuillets bêta est la zone en haut à gauche, celle correspondant aux hélices alpha est celle au milieu à gauche alors que la zone correspondant aux boucles et hélices alpha gauches est celle du milieu à droite.

0.3. Objectifs de la thèse

Le premier objectif de ce travail de thèse est de valider les résultats d'une technique de prédiction des résidus impliqués dans le noyau du repliement. Cette technique produit les positions appelées Most Interacting Residues (MIR) à partir de l'unique séquence. Afin de vérifier la précision de la méthode nous l'avons couplée à des méthodes d'analyse structurale (Tightened End Fragments ou TEF et alignements multiples) en employant une banque de protéine bien documentée (Immunoglobuline). De plus les résultats seront comparés à d'autres méthodes de prédiction.

Pour la deuxième partie de la thèse, l'objectif est de construire une base de données aidant à la prédiction des interactions protéines-protéines. Le premier objectif était de constituer une base de données de multimères « propre », c'est-à-dire dont l'existence à l'état physiologique n'est pas à remettre en cause. Ensuite, à partir de cette base construire une base de données de fragments représentatifs de ces interactions à des fins de prédiction.

1. Première Partie : Prédiction du noyau du repliement

1.1. Le repliement protéique

On dit que l'information génétique est contenue dans l'ADN, cependant les protéines portent l'information de leur structure dans la chaîne linéaire d'acides aminés. Selon Anfinsen (Anfinsen, 1973; Anfinsen et al., 1961)« toute l'information génétique requise pour le repliement correct est contenue dans la séquence d'acides aminés ». Or le dogme de la biologie structurale, posé dès la fin du XIXème siècle par Emil Fischer, énonce que la fonction biologique d'une protéine est due à sa structure tridimensionnelle, puisqu'une protéine dénaturée devient inactive. De plus, la mauvaise formation de la structure peut altérer les formations de complexes protéiques et les rendre ainsi inactifs.

Le repliement, au sens dynamique du terme, est ce qui assure le passage correct et unique de l'information génétique à une conformation du squelette protéique particulière pour qu'il acquière une énergie minimale. Ceci n'est plus possible du moment où l'information linéaire est altérée.

La première des propriétés des acides aminés responsables du repliement est attribuée à l'effet hydrophobe (Kauzmann, 1959). Les résidus hydrophobes ont tendance à se regrouper entre eux afin de contrer la dépense énergétique à rester entourés d'eau. Cette séparation des globules protéiques en un cœur hydrophobe entouré d'une couche majoritairement hydrophile (ou moins hydrophobe) est connue depuis le milieu du siècle précédent, à partir de considérations d'ordre géométrique (Bresler et Talmud, 1944a et 1944b). Les autres propriétés des acides aminés viennent ensuite finir le processus et maintenir la structure native de la protéine. Grâce à la possibilité d'engendrer des liaisons hydrogènes entre oxygène accepteur et azote donneur, les squelettes des résidus sont capables de former des structures élémentaires comme les hélices alpha et les brins bêta. Le repliement parait être guidé par les résidus internes, car par mutagénèse on voit que les substitutions sur les résidus les plus enfouis sont les plus délétères. Ce n'est que dans les dernières étapes du repliement que les chaînes latérales sont correctement positionnées.

1.1.1. Les différentes théories du repliement

Les fondements du repliement protéique découlent d'une expérience faite par Anfinsen (1973). Dans cette expérience il a observé la dénaturation de la structure d'une protéine sous l'effet de solvants dénaturants. Ceci a permis de confirmer que les protéines se replient d'elles mêmes dans la bonne conformation puisqu'après dialyse des agents dénaturants la protéine retrouve sa fonction, donc sa structure. Ceci a écarté les premières théories qui stipulaient que les protéines adoptaient leur

conformation finale grâce à une matrice. On peut imaginer qu'une protéine explore tous les possibles de son espace conformationnel. C'est sans compter sur la démonstration de Levinthal (Levinthal, 1968; Rooman et al., 2002; Zwanzig et al., 1992). Il a calculé le nombre total de conformations possibles que peut adopter une petite protéine de cent acides aminés. Il ne faudrait pas moins de 10^27 années pour explorer toutes ces conformations. Or on sait qu'une protéine se replie en quelques secondes au plus dans la plupart des cas. C'est donc que le repliement d'une chaîne protéique fait intervenir des mécanismes séquentiels au cours desquels l'évolution vers l'état natif s'accompagne d'une augmentation de la stabilité conformationnelle.

1.1.2. Le modèle hiérarchique

Certaines hypothèses tendent à démontrer que le processus du repliement est hiérarchique. C'est-à-dire que les structures secondaires se forment en premier et s'assemblent ensuite pour donner naissance à la structure tertiaire. Elles mêmes définissent des domaines qui s'associent à leur tour pour former la protéine dans sa conformation native. Une expérience de stopped flow (Jennings & Wright, 1993) a démontré que l'apparition des premières structures secondaires se fait dans la milliseconde ou moins après le début du repliement. Le début du repliement étant contrôlé par le passage de la chaîne peptidique des conditions dénaturantes à renaturantes. La formation des structures secondaires est étudiée par l'analyse du dichroïsme circulaire de la solution. Les structures secondaires ne sont pas nécessairement formées dans l'ordre qu'elles occupent dans la chaîne protéique. Ce modèle a été proposé par Karplus au début des années 70 sous l'appellation de diffusion-collision (Karplus & Weaver, 1996) : les structures secondaires sont formées avant la structure tertiaire.

1.1.3. Le modèle de nucléation condensation

Le modèle de la nucléation condensation stipule que le repliement débute par la formation d'interactions entre résidus hydrophobes qui forment un noyau autour duquel le reste de la structure va prendre sa forme native. Ce n'est qu'une fois que la structure tertiaire est grossièrement dessinée, que les structures secondaires se mettent en place (Kim & Baldwin, 1982). Le terme nucléation a été utilisé avec deux sens légèrement différents. Dans les deux cas, le noyau est la structure formée au commencement du repliement. Selon le premier sens de nucléation, celui de la chimie, le repliement implique une réaction qui s'initie rapidement à partir du moment où le noyau est formé, comme en cristallisation ou dans la formation d'une hélice alpha alors que la molécule nucléée n'est pas observable comme une espèce peuplée. En effet, le repliement protéique peut être décrit en accord avec ce premier sens, car il s'opère rapidement après la nucléation car le noyau est instable et il casse s'il n'est pas stabilisé ultérieurement. Le deuxième sens suppose qu'il doit y avoir des conformations intermédiaires. Cet usage s'est développé graduellement dans le domaine du repliement protéique car il est apparu aux chercheurs que des molécules nucléées sont néanmoins observables dans les

expériences sur le repliement. Ainsi le terme fut retenu par plusieurs équipes. Dans le modèle de nucléation condensation, le repliement ne peut pas commencer tant qu'une réaction initiale n'est pas accomplie : la nucléation. Et le repliement ultérieur prend place rapidement comparé à la réaction observée du repliement. Les états intermédiaires ne peuplent pas une population envisageable car une fois commencé le repliement est trop rapide. Cependant des conditions expérimentales particulières permettent de mettre en évidence ces états intermédiaires. Ce modèle s'applique plutôt aux petites protéines (de moins d'une centaine de résidus) mais ce critère n'est pas suffisant.

1.1.4. Le modèle de l'entonnoir

Le repliement se déroule sur une surface énergétique qui représente les différentes conformations auxquelles la protéine a accès. La protéine suivrait donc sa route énergétique vers l'état correspondant à l'énergie libre minimale (la conformation native), dont la forme ressemblerait à un entonnoir (voir la figure 1.1.1). La route énergétique de la chaîne protéique ne suit pas une piste particulière. La chaîne polypeptidique peut alors être piégée dans des minima locaux du paysage, qui rendent l'entonnoir « cabossé ». Ces minima correspondent à des états conformationnels intermédiaires de la chaîne polypeptidique, séparés du minimum global par des barrières d'énergie libre. Le repliement doit commencer quand le squelette peptidique adopte une conformation suffisamment proche de la protéine native. Cette hypothèse émane du modèle de recherche aléatoire biaisée. Selon Baldwin (Kim & Baldwin, 1982), le nombre de conformations de la chaîne peptidique est sérieusement restreint lorsqu'on accepte uniquement les conformations stériquement possibles. De plus, une fraction non négligeable de ces conformations sont compactes. Ces conformations sont appelées états intermédiaires (« transition state » dans la littérature anglo saxone).

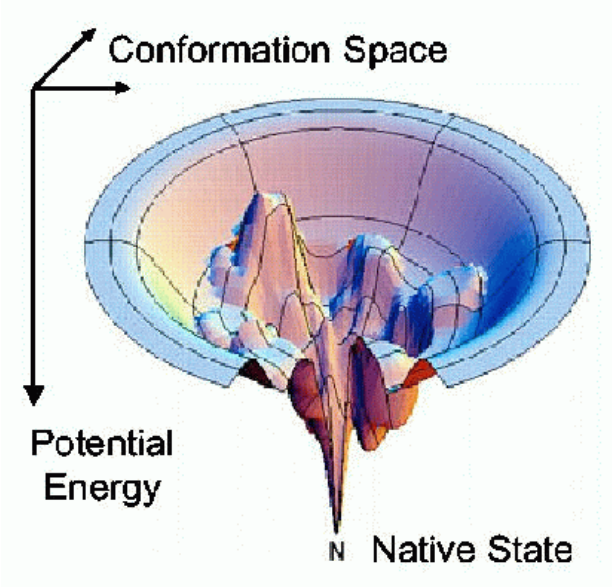


Figure 1.1.1. Représentation de l'entonnoir du payage du repliement. D'après (Dill et Chan, 1997)

Les modèles qui donnent des résultats en accord avec les expériences, laissent deux choix aux structuralistes. D'un côté prendre un modèle existant et le tester, de l'autre extraire un modèle des données expérimentales. Le premier correspond à un modèle structuré impliquant des états intermédiaires (nucléation condensation) et le second fondé sur l'assemblage modulaire dans lequel la quasi-totalité du repliement de chaque partie de la protéine s'opère au même moment, les modules s'assemblant à des temps différents (hiérarchique). Notons que le mécanisme du repliement peut combiner les deux modèles, ainsi la formation d'états intermédiaires avec des structures secondaires peut précéder la formation d'interactions tertiaires comme dans le modèle structuré alors que certains sous-domaines peuvent se replier à des moments différents, en accord avec le modèle assemblage modulaire.

1.1.5. Lien entre états intermédiaires et noyau du repliement

Il existe dans le mécanisme du repliement des intermédiaires, qui correspondent à des minima locaux d'énergie, contrairement à un état de transition, maximum local d'énergie (qui correspond en fait à un point selle), et donc de ce fait non observable expérimentalement. De nombreuses équipes ont mis en évidence par des expériences de stopped flow, de cinétique enzymatique ou encore d'échange de proton, des états multiples de la chaîne protéique qui ne sont pas bien repliés. Ces états intermédiaires peuvent se trouver soit sur le chemin qui mène à la forme native, soit en dehors. Dans ce dernier cas, cela veut dire que pour passer d'un tel état intermédiaire à la forme stable, il faut d'abord opérer un dépliement qui conduit à une conformation dénaturée.

Les expériences cinétiques et énergétiques montrent clairement l'existence d'états intermédiaires parfois instables mais qui mènent néanmoins la chaîne peptidique vers sa conformation native. Bien que ces états soient instables, ils sont quand même formés au tout début du repliement par quelques interactions de certains résidus appartement au noyau du repliement.

Comment des états de transition avec des structures différentes peuvent mener à l'identification de quelques résidus responsables du noyau du repliement ? On peut prendre l'exemple du repliement de l'ACBP (acyl-coenezyme binding protein) étudié par Thomsen et al. (Thomsen et al., 2002) qui apparait être de cinétique à deux états (sans intermédiaire stable). Or, des expériences de dichroïsme circulaire et de RMN ont montré clairement des structures régulières formées avant que la protéine atteigne sa forme native, remettant en cause cette cinétique apparente. Ainsi par la nature transitoire des structures formées (états intermédiaires) dans la population des molécules non repliées, l'étude suggère qu'une fraction de ces molécules non repliées forme une ou plusieurs sous-populations qui ont une forte propension à se replier. Seules les molécules ayant un ou plusieurs sous-groupes d'interactions spécifiques formées vont se replier. Ainsi plusieurs sous-structures ou états intermédiaires peuvent donner plusieurs « sous-noyaux » et ce sont celles là qui pourront se replier

correctement. Ceci est en accord avec le fait qu'une chaîne polypeptidique peut adopter des chemins énergétiques multiples vers sa structure native comme le montre le modèle du paysage.

Ces interactions spécifiques formant le noyau pourraient être décrites par la localisation des positions « serrées » de boucles fermées.

1.2. Concept de TEF

1.2.1. Vision modulaire des protéines : TEF, foldons et building blocks

Beaucoup de protéines sont à l'évidence composées d'une concaténation de plusieurs domaines, qui peuvent se replier indépendamment dans leur conformation native. Des unités plus petites que le domaine ont été envisagées. Le fait que différentes parties de la protéine se replient indépendamment entraîne que des interactions guidant la protéine vers sa structure native sont contenues dans des parties discrètes et contiguës de la séquence (Kippen et al., 1994). De plus un argument entropique et évolutionniste vient démontrer la cohérence de la construction modulaire des protéines. Selon le facteur entropique, les parties de séquences correspondant à des structures repliées (donc ordonnées) auraient tendance au cours de l'évolution à se raréfier en parallèle de l'augmentation en taille des séquences. Panchenko et al. ont une vision modulaire des protéines qui les ont mené à réfléchir sur les unités structurales fondamentales constituant les protéines (Panchenko et al., 1996). Il est vrai que les structures des grandes protéines montrent plusieurs domaines, chacun comparable en taille aux plus petites protéines correctement repliées, donc d'une taille moyenne d'environ 150 acides aminés. Dans de nombreux cas, des domaines ayant des fonctions physiologiques bien différentes ont été mélangés ou dupliqués au cours de l'évolution. Si des modules protéiques peuvent se replier indépendamment, le temps de repliement de grandes molécules devrait être comparable aux plus petites, plutôt que d'être fortement proportionnel à la longueur de la chaîne (ce que l'on ne constate pas expérimentalement). Des modules isolés ont été observés (Wetlaufer, 1981 ; Ikura et al., 1993) et montrent que certains modules de protéines peuvent se replier en structures natives. De plus les expérimentations sur les intermédiaires du repliement indiquent que des unités structurales bien définies de la protéine native peuvent être présentes dans les intermédiaires pas complètement repliés (Kippen et al., 1994 ; Jennings & Wright, 1993). Ceci suggère que des parties distinctes d'une protéine peuvent se replier quasi-indépendamment à des temps différents.

Beaucoup d'équipes de chercheurs aiment voir les domaines comme un assemblage de structures élémentaires. La plupart des structures natives peuvent être reproduites en assemblant un nombre restreint de motifs simples composés de quelques structures secondaires en interaction. Ces motifs

simples, qui pourraient être des super structures secondaires, peuvent être vus comme les briques élémentaires qui participeront à la construction de l'édifice. Ces briques élémentaires ont été appelées Foldons ou Building Blocks selon les équipes. Cette approche repose sur l'hypothèse d'une modularité des protéines. En effet, la recherche de la plus petite unité structurale d'une protéine se révèle intéressante afin de faciliter la compréhension de la complexité de la séquence et de la structure des protéines (Panchenko et al., 1996).

L'approche fondée sur les fragments a été initiée par Lesk et Rose (Lesk et Rose, 1981) et a été suivie par bon nombre de chercheurs. Par exemple, le groupe de Nussinov a proposé une dissection hiérarchique des domaines grâce à un algorithme fondé sur trois paramètres : la compacité, le degrè d'isolation et l'hydrophobie des fragments (Kifer et al., 2008; Tsai et al., 2000; Tsai et Nussinov, 2001). Le niveau le plus bas dans la décomposition des structures des chaînes est celui des Building Blocks, dont la longueur moyenne est de 29 acides aminés (Haspel et al., 2003a; Haspel et al., 2003b).

Ainsi ces unités structurales ont une stabilité intrinsèque, ils seront dénommés par Panchenko et Al. par le terme Foldon. L'équipe de Panchenko a donc développé un algorithme afin de définir des foldons tout au long de la chaîne peptidique. L'analyse du paysage énergétique du repliement suggère qu'une fonction énergétique réaliste facilite le repliement rapide en stabilisant l'état correctement replié par rapport à toutes les autres structures mal repliées, dont les énergies peuvent être estimées selon la théorie du « spin glass » (qui ne sera pas décrite ici). La température de repliement T_f est directement reliée à la différence en énergie entre les structures correctement repliées et l'ensemble des structures compactes mal repliées, qui sera appelé l'écart de stabilité ΔE. La température de « transition vitreuse » T_g est reliée à l'écart quadratique moyen des énergies des structures mal repliées δE. La configuration entropique des globules fondus rend difficile l'attribution d'une amplitude à T_f/T_g. Si l'entropie est extensive, T_f/T_g est une fonction croissante monotone de la quantité $\Theta = \Delta E/(\delta E\sqrt{N})$ où N est le nombre de résidus de la protéine. Les limites des foldons sont définies selon la procédure suivante : la chaîne polypeptidique est coupée à un résidu j et la valeur moyenne de Θ du N terminal (du premier résidu au résidu j) et C terminal (du résidu j au dernier résidu) est calculée ainsi : $\Theta = (\Theta_{N,j} + \Theta_{C,j})/2$. Le point de coupe est déplacé ensuite le long de la chaîne et la position du premier maximum local de Θ définit la limite du premier foldon. La procédure de coupe est répétée afin que chaque fois qu'un foldon se finit, un autre commence. Les résultats émanant de cette méthode indiquent que les frontières entre foldons correspondent étroitement à celles des modules structuraux, ce qui laisse penser que de telles unités compactes ont tendance à se replier indépendamment. Cette équipe a aussi comparé le découpage en foldons de quelques protéines avec les résultats expérimentaux de cinétique et thermodynamique. Les limites de foldons semblent être

sensibles à un seuil de taille minimale (15 résidus) et à la séquence de la chaîne polypeptidique. D'après l'étude de Panchenko, la longueur moyenne d'un foldon est de 38 résidus.

Les calculs sur réseaux de points, qui permettent de discrétiser l'espace, suggèrent que la nature des unités du repliement dépend aussi bien de la séquence que de la structure. Un foldon se replie indépendamment si la structure l'englobant génère un piège énergétique prenant en charge le processus du repliement. Dans ce cas, bon nombre de foldons « transitoires » peuvent être éliminés du cheminement cinétique par quelques mutations. Des expériences *in vitro* et *in vivo* démontrent l'existence d'un assemblage modulaire des protéines mono domaine. Quant aux très grosses protéines, il est certain qu'elles sont composées de plusieurs domaines, la limite supérieure de taille d'un domaine étant de l'ordre de 500 résidus. Il est néanmoins clair que différentes parties de la protéine se replient quasi indépendamment à différents moments du repliement. On peut dire que les foldons présentent une stabilité intrinsèque.

1.2.2. Description du concept

Des études physiques sur les homo-polymères en globules compacts mettent en exergue un facteur de boucle. Le facteur de boucle se comporte de la manière suivante : pour des tailles très petites de chaîne le facteur est très faible, pour des tailles très grandes le facteur est aussi faible car les longues boucles ne sont pas favorisées par le facteur entropique. Il en résulte que le facteur de boucle est optimal à une taille préférentielle du polymère. Berozovsky (Berezovsky et al., 2000) et son équipe ont analysé un ensemble de 300 protéines cristallisées avec un faible taux d'identité en séquence (inférieure à 25%). Il sera discuté plus tard de l'intérêt de choisir des séquences hautement divergentes en séquence. La distribution des longueurs des boucles avec des carbones alpha en contacts, c'est-à-dire à moins de 10 Angstrom de distance, montre un maximum bien défini autour de 25 résidus de longueur. Il faut ici entendre le terme boucle comme fragment polypeptidique sans nœud dont l'origine du terme (closed loop) est due à Haas (Ittah et Haas, 1995) dans un article où ce mot figure mais a été refusé par l'éditeur dans le titre. Ce terme n'est pas du tout ici en rapport avec la classification des structures secondaires en hélice alpha, brin beta et boucle. Les boucles fermées sont définies comme les sous-trajectoires contiguës de chaînes repliées avec une faible distance entre leurs extrémités. Une cartographie des boucles fermées (closed loops dans la littérature) a été entreprise sur les grandes familles de repliements. Lorsque une boucle est identifiée, la partie correspondant à cette dernière est exclue des calculs ultérieurs. Cependant un chevauchement de 1 à 5 résidus est admis. Si deux boucles sont imbriquées l'une dans l'autre, c'est celle avec les extrémités les plus resserrées qui est choisie. L'algorithme commence avec une distance entre extrémités inférieure à 4 Angstrom et est épuisé à une distance de 10 Angstrom des fins de boucles. L'expérience a été réalisée avec différents seuils de distance des extrémités des boucles. Il en résulte que le maximum du nombre de boucles observé pour une longueur d'environ 25 résidus n'est pas dépendant du seuil. Cette étude a donné une nouvelle

propriété aux structures protéiques : la taille préférentielle des boucles fermées. Ces boucles couvrent entre 40 et 80 pour-cent de la séquence. Il a été conclu que les protéines globulaires et membranaires sont largement faites de boucles fermées indépendamment du type de repliement. Les boucles analysées se sont avérées hétérogènes en forme et en composition de structures secondaires. La question principale de cette étude était de savoir si la longueur des boucles fermées dans les chaînes protéiques correspondait au facteur statistique de fermeture des boucles. Ce facteur est en rapport direct avec la longueur de persistance de la chaîne d'un homopolymère. Un homopolymère constitué de différents acides aminés montre une longueur de persistance d'à peu près 4 à 5 résidus. La longueur de boucles statistiquement optimale est de 10-25 résidus et de 20-50 résidus après correction due à l'insertion de structures secondaires. Cette longueur statistique est proche de la valeur de 27 ±5 observée pour les structures natives de protéines. La ponctuation de la chaîne polypeptidique par les boucles fermées suggère un schéma franc du repliement des protéines. Les 'points de coutures' par les boucles fermées pourraient nucléer le repliement.

Berezovsky et al. se sont ensuite intéressés à la trajectoire de la chaîne peptidique des protéines. Ils se sont aperçus que la trajectoire du squelette effectuait de nombreux retours sur elle-même en décrivant des boucles fermées. Cette étude est fondée sur d'autres contemporaines démontrant la construction des protéines globulaires avec des unités « loop-n-lock » d'à peu près 30 résidus, proches de la taille moyenne d'un foldon. Trifonov et al (Trifonov et al., 2001) suggèrent que les boucles fermées sont des descendants évolutifs de prototypes structuraux ou séquentiels. Les versions actuelles de boucles fermées peuvent apparaître divergentes en séquence par rapport aux prototypes hypothétiques. De même on attend une certaine conservation des structures secondaires des boucles fermées. Afin de vérifier cette hypothèse, cette équipe a réalisé une recherche sur 23 protéomes bactériens de motifs de 30 résidus consécutifs de long. Tous les motifs de 30 de long sont ensuite classés selon la fréquence avec laquelle ils sont observés dans ces 23 protéomes. Afin de laisser une certaine tolérance et ne pas prendre en compte uniquement les identités strictes, il est laissé une certaine relaxation en autorisant un taux de similitude d'un tiers (donc 20 identités conservées sur les 30 de long). Les fragments choisis sont ceux montrant le plus haut score après plusieurs tours d'itérations avec le taux d'identité comme paramètre. La distribution des fréquences des résidus « consensus » montre des limites précises et clairement identifiables. Beaucoup de motifs trouvés par cette procédure correspondent aux boucles fermées dans la structure des protéines. De nombreux éléments d'identité en séquence même marginaux avec les prototypes montrent eux aussi les caractéristiques des boucles fermées, bien que les structures secondaires varient (Berezovsky et Trifonov, 2002). Ainsi le fait qu'il y ait une conservation de la « fermeture » de la chaîne aux extrémités de ces motifs permet de penser qu'il s'agit d'une propriété de retour de boucle. Et cette propriété doit ainsi être primordiale pour le repliement protéique.

L'étude des boucles fermées a été reprise par Lamarine et al. (Lamarine et al, 2001). L'équipe a changé le terme en Tightened End Fragments (TEF) ou fins de fragment resserrées. Désormais les boucles fermées seront appelées TEF. L'équipe s'est intéressée à la correspondance entre les extrémités des TEF et la présence de résidus hydrophobes particuliers, dénommés topo-hydrophobes, car ils sont hautement conservés au sein de familles fonctionnelles très divergentes en séquence (Poupon et Mornon, 1998; Poupon et Mornon, 1999). Ils sont la preuve que les familles protéiques gardent un patron de positions topologiques. On décèle ces positions conservées au sein de familles structurales de protéines en alignant les séquence des protéines dans chaque famille (les alignements seront discutés plus loin). Si au moins trois-quarts des résidus occupant une position dans l'alignement sont hydrophobes, la position est alors considérée comme topo-hydrophobe. Ainsi sur un ensemble de 372 protéines, dans le cœur interne des protéines il existe deux populations d'acides aminés hydrophobes. La première peut être remplacée par des acides aminés hydrophiles ou neutres. La seconde est toujours occupée par un hydrophobe (VIMWYLF), ce sont ses positions qui sont des topo-hydrophobes. Si la question de leur responsabilité dans la nucléation du repliement n'est pas totalement résolue, il est par contre d'évidence qu'elles en sont les principaux composants. Les topo-hydrophobes possèdent aussi la propriété d'être profondément enfouis dans la protéine et d'avoir de nombreuses interactions entre chaînes latérales. Si les topo-hydrophobes étaient corrélés aux limites des TEF, ce serait un argument en faveur des deux concepts. La première chose que l'équipe a regardé pour cette étude est la distribution des distances entre carbone alpha des résidus topo-hydrophobes et des hydrophobes occupant des positions non topo-hydrophobes. Les courbes montrent bien une différence entre ces deux populations. Comme illustre la figure 1.1.2. tirée de l'article de Lamarine et al.

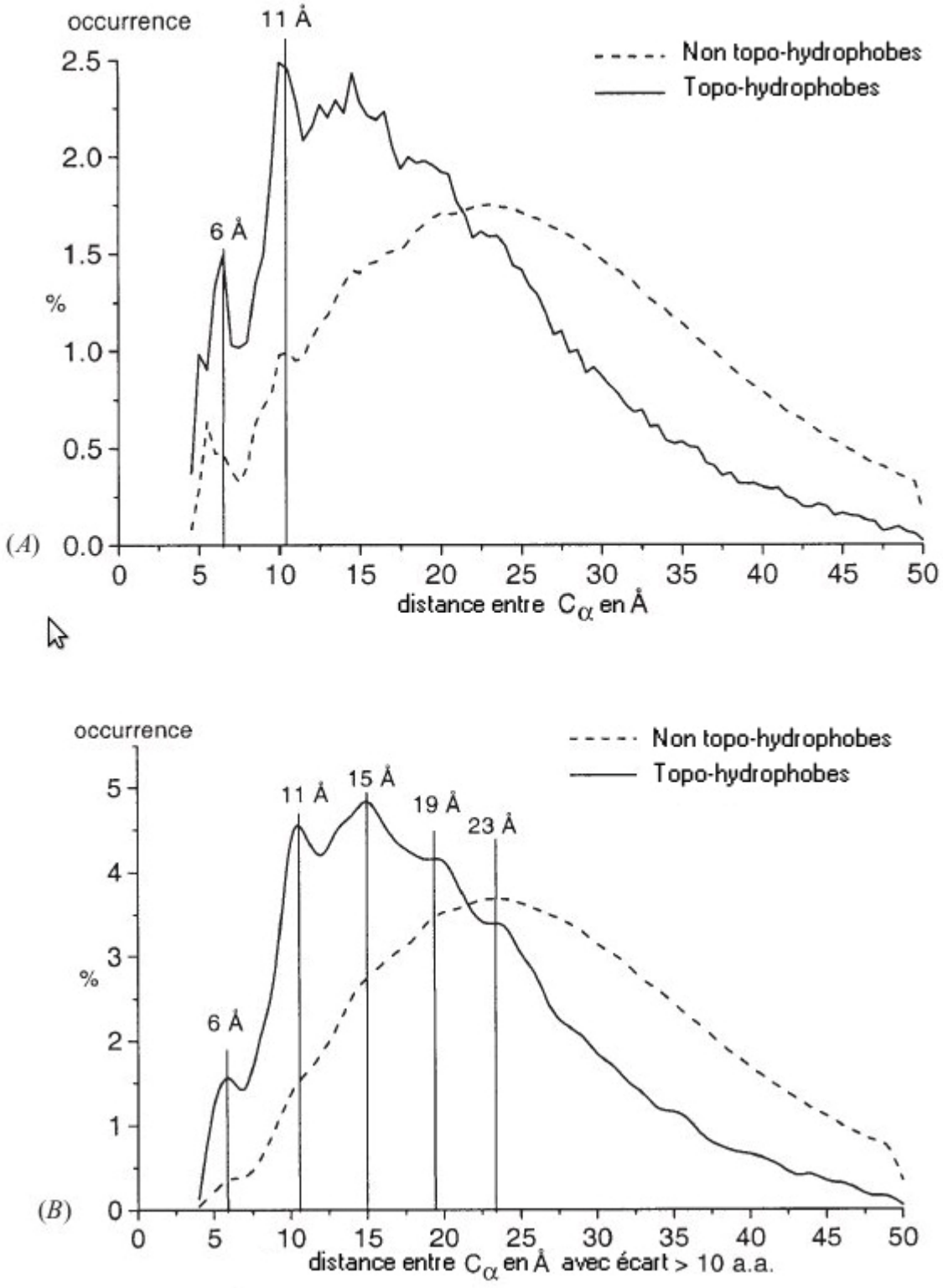


Figure 1.1.2. (A) histogramme, échantillonné tous les 0,5 Angstrom, des distances entre paires de résidus sans restriction de séparation en séquence entre les deux membres de la paire, pour les paires de topo-hydrophobes (courbe pleine) et pour les paires de non topo-hydrophobes (courbe en pointillés). (B) Histogramme, échantilloné tous les 1 Angstrom, des distances entre paires avec une

séparation en séquence (écart) supérieure à 10 résidus (a.a.) pour les paires topo-hydrophobes (courbe pleine) et pour les paires de non topo-hydrophobes (courbe en pointillés). D'après (Lamarine et al., 2001).

Alors que la distribution des distances entre résidus non topo-hydrophobes est assez lisse et montre un maximum à 25 Angstrom, la distribution des distances entre topo-hydrophobes montre deux pics à 6 et 11 Angstrom avec le dernier comme maximum. Ces résultats indiquent clairement une interaction à courte distance dans l'espace tridimensionnel correspondant à de grandes distances au niveau de la séquence. On constate sur la figure 1.2.3. une correspondance statistique entre les limites des TEF et la localisation des positions topohydrophobes.

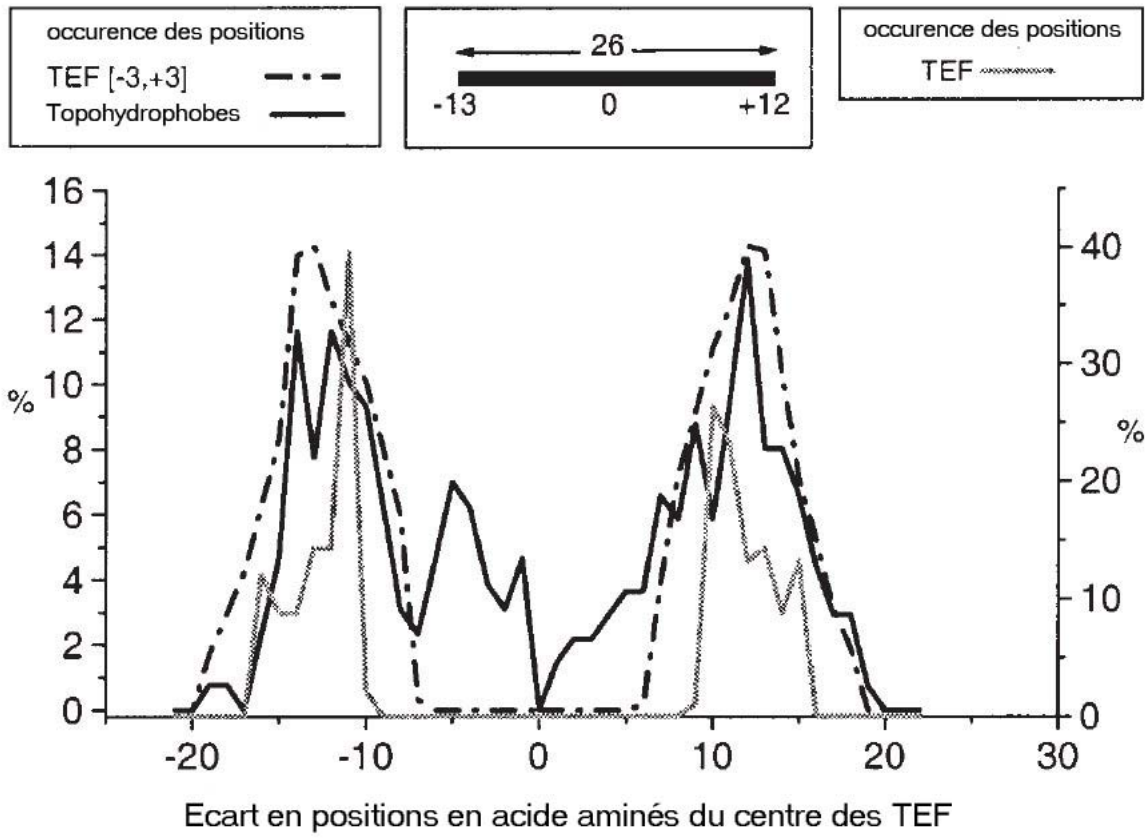


Figure 1.1.3. L'origine de l'abscisse est située au milieu des TEF sur une banque de 372 protéines. En ordonnée figure la localisation des topohydrophobes et des limites de TEF, strictes ou avec une fenêtre de ±3 résidus. D'après (Lamarine et al., 2001).

La distribution de la taille des TEF a ensuite été comparée à la distribution de la séparation en séquence des topo-hydrophobes comme cela est indiqué sur la figure 1.1.4. La distribution des distances entre C alpha inférieures à 7 Angstrom montre une série de pics réguliers distants entre eux

de 13 résidus environ. Lorsque cette distribution est étudiée pour les topo-hydrophobes et non topo-hydrophobes, on retrouve un pic pour une séparation de 27 résidus.

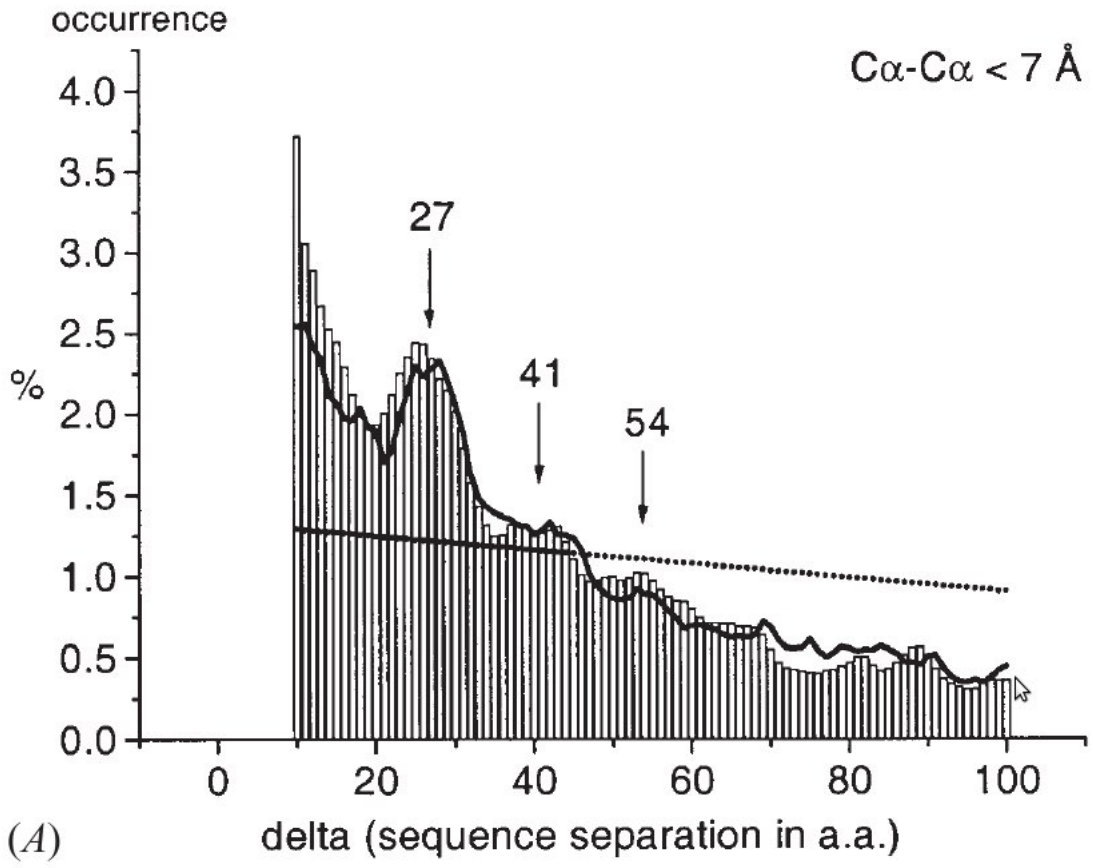


Figure 1.1.4. Séparation en séquence pour les acides aminés distants de moins de 7 Angstroms sur deux jeux de protéines différents. D'après (Lamarine et al., 2001).

Il est néanmoins clair que les topo-hydrophobes sont plus concernés par une séparation en séquence préférentielle. Dans la majorité des cas, les TEF commencent et finissent à une position occupée par un topo-hydrophobe (figure 1.1.3.). Cette étude confirme en partie les fondements du repliement par fragments, et étaie la théorie de coopérativité du repliement protéique avec des positions topo-hydrophobes agissant comme les clés de la nucléation protéique.

Yew et al. (Yew et al., 2007) ont récemment testé la conservation des TEF. Ils proposent que la distribution de type géométrique puisse arriver simplement parce que les protéines contiennent des boucles fermées. Cependant la distribution de type Gaussien, centrée à environ 25 résidus, ne peut pas être triviale et doit avoir une origine fonctionnelle importante. Le type Gaussien de la distribution est à mettre en rapport au pic de longueurs préférentielles des TEF. D'autres études ont montré que le nombre de voisins exprimé en fonction de la distance en séquence atteint un maximum à la longueur

de 27 résidus. Dans cette étude de Yew, une approche un peu différente de celle de Berezovsky et Trifonov a été employée. Elle est fondée sur des alignements multiples de séquences de neuf super repliements. La conservation est déterminée sur ces alignements, l'entropie, la variabilité et la proportion maximale. Les conclusions de cette étude montrent bien une corrélation entre l'observation des résidus conservés et la proportion d'acides aminés hydrophobes aux extrémités des TEF. Taylor et Waissman (Taylor et Vaisman, 2006) ont effectué un histogramme de la distance entre C alpha sur une banque de structures, en employant une méthode de tesselation de Delaunay. Ceci présente l'avantage de ne pas introduire de distance seuil au delà de laquelle on admet que deux résidus ne sont plus voisins. Deux acides aminés sont considérés comme premiers voisins s'ils partagent une face de leurs cellules de Delaunay. La distance de séparation en séquence entre premiers voisins spatiaux est entre 20 et 25 résidus, donc comparable avec les résultats sur les TEF, et elle dépend de la classe structurale à laquelle appartient la protéine, la séparation la plus courte ayant lieu pour la classe principalement beta, et la plus longue pour la classe mixte alpha bêta.

1.3. Concept de MIR

1.3.1. Les réseaux cubiques

Plusieurs méthodes sont employées pour représenter les acides aminés de la chaîne peptidique afin de simuler le processus du repliement. Cela va du modèle gros grains où l'acide aminé entier est modélisé par une boule dure, jusqu'aux modèles tout atome. D'autres perspectives parallèles à cet axe des dimensions décrites, sont de modéliser l'acide aminé entier par un point à un nœud d'un réseau. Le réseau choisi est celui développé par Skolnick et Kolinski (Skolnick & Kolinski, 1991) et il correspond à un déplacement de type (2, 1, 0) par rapport à un réseau cubique sous jacent. Ce réseau montre une certaine flexibilité et les structures secondaires peuvent être représentées avec un RMSD de 0,7 Angstrom pour les motifs hélicaux et 0,6 Angstrom pour les brins bêta car il autorise une gamme d'angles entre carbones alpha successifs allant de 64° à 143°.

D'autres types de réseau ont été testés, tel le réseau en « diamant ». Cependant ces réseaux ne se sont pas montrés cohérents avec les conformations observées au sein des protéines. Ils montrent de grandes déficiences quant à l'entassement d'hélices alpha et de brins bêta dans une orientation parallèle, quant au pré-requis de 3,6 résidus par tour dans les hélices et l'inaptitude des feuillets bêta à adopter un vrillage (Skolnick & Kolinski, 1991).

Le groupe de Shakhnovich (Abkevich et al., 1994) a été parmi les premiers à étudier le mécanisme du repliement sur un réseau cubique. Les résultats montrent selon eux une croissance par nucléation,

c'est-à-dire la formation d'un noyau essentiellement hydrophobe avec un patron spécifique de contacts. La recherche de ce noyau est le facteur limitant du repliement et correspond au passage de la barrière énergétique la plus importante. « Le noyau est une structure spatialement localisée d'une sous-structure de la protéine native ».

L'environnement, le comportement d'un état natif de la chaîne protéique et sa dynamique peuvent être étudiés par des modèles avec une représentation tout atome de la protéine. Cependant l'investigation de l'espace conformationnel par de tels modèles implique des temps de calculs prodigieusement longs. Il est donc nécessaire de simplifier les modèles utilisés pour étudier le repliement protéique. De tels modèles perdent en détails, mais l'un des termes du paradoxe de Levinthal est conservé : en effet sur un réseau cubique la chaîne protéique modélisée par un chapelet de perles conserve un éventail de conformations initiales possibles. Shakhnovich et al soulèvent néanmoins une condition pour l'utilisation de ces modèles. En effet, un pré-requis est que la simulation doit aboutir au même résultat quelle que soit la conformation initiale de la chaîne, comme le font les protéines dans la nature.

Un autre critère nécessaire au modèle est que des interactions entre paires de chaînes latérales doivent être indépendantes de la conformation du squelette protéique. Ce critère doit être respecté afin d'éliminer la construction d'un motif particulier. En d'autre termes, la structure native ne peut pas être spécifiée par un potentiel cible de l'algorithme (Skolnick & Kolinski, 1991). La représentation la plus employée sur de tels réseaux consiste en une permutation cyclique de vecteurs de type (±2, ±1, 0) joignant les carbones alphas contigus en séquence. Ces vecteurs correspondent sur le réseau à un « pas de cavalier » dans un jeu d'échecs, mais ici dans les trois dimensions. Ce type de réseau sera nommé par la suite réseau (2,1,0) pour plus de simplicité.

1.3.2. Description du modèle utilisé

La protéine est constituée d'une représentation de chaque résidu par son seul carbone alpha. Le système entier est inclus dans un réseau cubique construit par des vecteurs de type (±1, 0, 0) qui joignent n'importe quels nœuds adjacents du réseau. La longueur des vecteurs constituant ce réseau cubique est de 1,7 Angstrom. Chaque carbone alpha occupe un nœud du super réseau (2,1,0) (Skolnick & Kolinski, 1991). La chaîne latérale n'est pas représentée dans ce modèle qualifié de « toy model » (jouet) dans la littérature.

Les interactions entre paires de résidus sont autorisées si les sites d'interaction se trouvent à une distance inférieure à 3,8 Angstrom, ce qui correspond à la distance entre carbones alpha premiers voisins. Chaque nœud est entouré de 24 premiers voisins sur le réseau (2, 1, 0). Le potentiel d'interaction entre les paires de résidus est représenté par un potentiel de forces moyennes. Les paires hydrophiles et hydrophobes interagissent avec un potentiel attractif et les paires mixtes avec un potentiel répulsif. Les valeurs de ces interactions sont tabulées dans une matrice et nous avons utilisé

le potentiel statistique de Miyazawa et Jernigan (Miyazawa et Jernigan, 1996) qui sera décrit ultérieurement. Les conformations initiales sont choisies en plaçant au hasard les résidus sur les nœuds du réseau en respectant les contraintes imposées par l'enchaînement des résidus et par le fait qu'un nœud ne peut pas être occupé par plus d'un résidu. La conformation initiale n'est donc pas totalement étendue et elle reste largement non compacte.

1.3.3. Monte-Carlo

Bien que les simulations sur réseau cubique soient moins gourmandes en temps de calcul que les méthodes tout atome, le temps pour que la chaîne modélisée se replie reste très long. Une méthode a été appliquée afin d'optimiser les simulations dynamiques. C'est la méthode dite de Monte-Carlo inventée en 1947 par Nicholas Metropolis. Cette méthode fait appel à des procédés aléatoires d'où son nom en rapport aux jeux pratiqués dans la principauté. L'application de cette méthode aux simulations de repliement sur réseau prend place dans les mouvements de la chaîne à chaque itération. Il existe un critère dit de Metropolis afin d'accepter ou de rejeter le mouvement qui positionne le résidu tiré au sort dans une nouvelle position voisine libre. En effet la dynamique du modèle est simulée par un processus stochastique de petits mouvements aléatoires dans la conformation de la chaîne. La dynamique de Monte-Carlo a un sens physique pour ces propriétés dynamiques dont les échelles de temps caractéristiques sont considérablement plus grandes que celles correspondant aux petites modifications implémentées dans l'algorithme. Les réseaux cubiques avec dynamique de Monte-Carlo ont prouvé être des méthodes efficaces pour l'étude de dynamique longue de systèmes de polymères (Kolinski & Skolnick, 1994).

La dynamique de la chaîne est simulée par une séquence pseudo-aléatoire de réarrangements conformationnels pratiqués à chaque itération. Les mouvements effectués regroupent les mouvements de pointe (en vert sur la figure 1.3.1), de manivelle (en rouge), et les mouvements de coin (en bleu) de chaîne où un résidu seul en milieu de chaîne peut bouger.

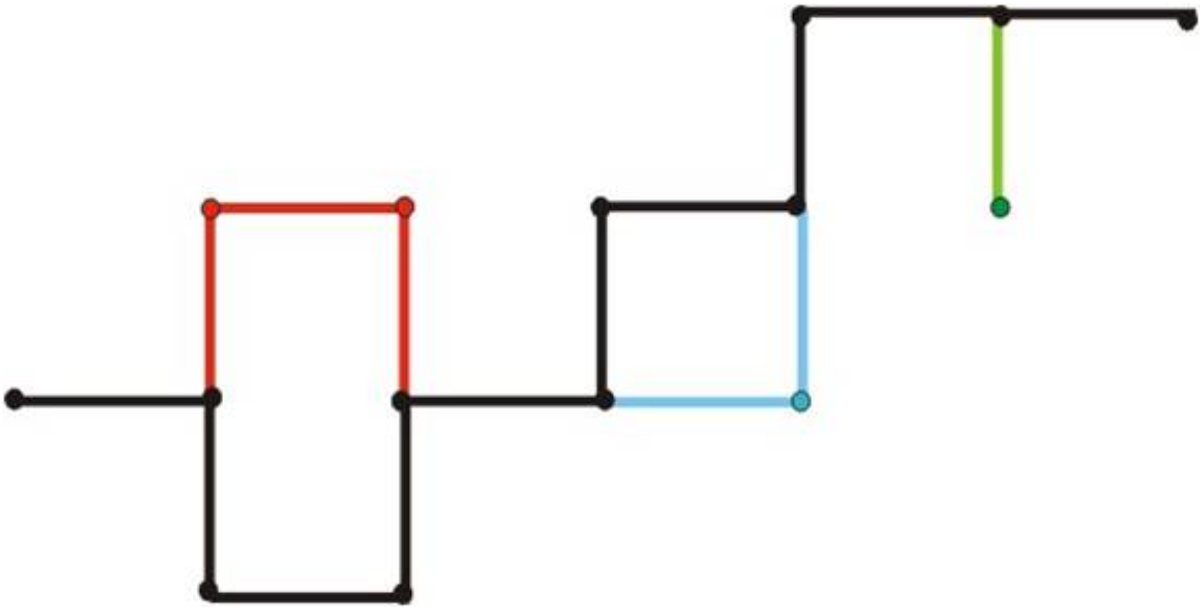


Figure 1.3.1. Mouvements autorisés dans un réseau cubique.

1.3.4. Champ de force

La chaîne peptidique simulée doit subir une contrainte pour acquérir sa conformation finale. Ainsi un ensemble de forces est appliqué au squelette peptidique. Cette somme de forces régissant la physique de toutes molécules dérive d'une énergie potentielle appelée champ de force. Le champ de force employé dans la méthode qui nous intéresse est celui développé par S. Miyazawa et R. L. Jernigan (Miyazawa et Jernigan, 1996). En règle générale, le champ de force utilisé en dynamique moléculaire évalue la force appliquée à un atome en fonction de sa distance par rapport aux autres. De tels types de champs de force ne sont pas applicables aux simulations du repliement car l'échelle de temps à observer est bien trop grande, de l'ordre de la milliseconde alors qu'une trajectoire typique de dynamique moléculaire est largement inférieure à la microseconde. C'est pourquoi Miyazawa et Jernigan ont choisi un champ simplifié, afin que les calculs soient plus rapides, en le décomposant en une composante simulant les interactions longue distance et une autre pour les courtes distances. Néanmoins il reste réaliste en terme d'énergie potentielle et de plus il inclut les effets du solvant. Une table d'énergies de contact inter-résidus par paire pour les 20 acides aminés a été indexée en les moyennant sur de nombreuses structures cristallines. Pour déterminer l'énergie potentielle appliquée à un résidu donné, on détermine dans un premier temps l'ensemble des autres résidus situés à une distance suffisamment faible pour contribuer de façon significative à cette énergie. Puis on somme les valeurs tabulées pour toutes les paires dénombrées précédemment et qui comportent le résidu pour lequel on est en train de calculer son énergie. Cette étape est donc rapide, et explique en partie le succès des potentiels statistiques. La seconde raison est que les énergies de contact utilisées dans le

potentiel de Miyazawa et Jernigan différencient avec succès les structures natives des structures mal repliées, ce qui est important pour la validation des modèles.

1.3.5. Description de la méthode

Il existe au sein des protéines des positions qui ont été conservées au cours de l'évolution. Les positions conservées hydrophobes le sont d'autant plus qu'elles sont en général plus enfouies et que leur rôle structural est fort. Il a été démontré que ces positions sont corrélées aux extrémités des TEF (Papandreou & Chomilier, 2004). Le défi est maintenant de prédire ces positions à partir de l'unique séquence. Les TEF sont universellement représentés dans toutes les sortes de protéines. Les extrémités des TEF correspondent à des groupes de résidus hydrophobes hautement conservés, les topo-hydrophobes. Les positions topo-hydrophobes sont d'une grande importance pour la formation et la stabilité du cœur protéique, ainsi que cela a été démontré sur un petit nombre d'exemples (Poupon et Mornon, 1999a). La formation des interactions entre topo-hydrophobes favoriserait la capacité des premiers fragments (correspondant aux TEF) à se replier et permettrait d'un point de vue dynamique d'accélérer le repliement et ainsi de se défaire du paradoxe de Levinthal. Le fait que des positions soient particulièrement conservées comme les topo-hydrophobes nous indique que le facteur principal qui mène à un cœur protéique stable et compact ne repose pas sur les détails des chaînes latérales des résidus. Les topo-hydrophobes sont disséminés tout au long de la séquence selon un patron spécifique de l'architecture qu'adoptera la structure native. C'est cette notion qui nous suggère que le facteur entrainant la nucléation du noyau est la répartition entre les résidus hydrophiles et hydrophobes tout au long de la séquence. Ainsi les modèles simplifiés comme ceux des réseaux cubiques se trouvent être tout à fait adéquats pour modéliser ce processus. Chomilier et Papandreou (Chomilier et al., 2004) ont développé et testé cette méthode de simulation du repliement sur un set d'une centaine de protéines représentant des repliements différents. Les premiers pas du repliement ont été simulés en utilisant un modèle simplifié ne représentant que le squelette de carbones alpha de la chaîne peptidique.

Sur le réseau cubique de géométrie identique à celui de Kolinski et Skolnick, les carbones alpha contigus sont reliés entre eux par un vecteur (±2, ±1,0). La longueur de ce vecteur est de √5 fois l'unité du réseau et vaut ainsi 3,8 Angstrom comme la longueur typique entre deux carbones alpha dans une protéine (voir la figure 1.3.2.). Sur ce réseau, chaque résidu en un nœud peut avoir 24 premiers voisins.

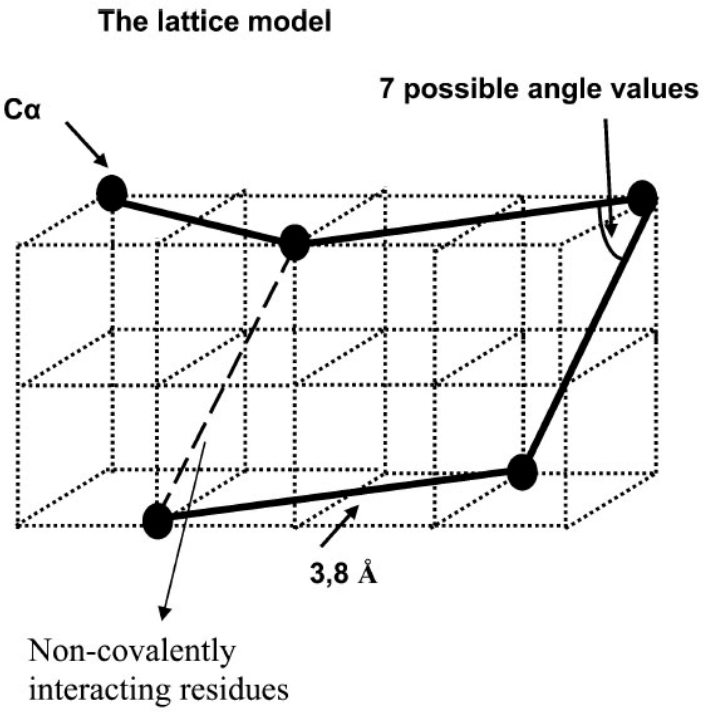


Figure 1.3.2. Construction du réseau (2, 1, 0) à partir d'un réseau cubique sous-jacent.

Le champ de force utilisé prend en compte la nature de chaque acide aminé. La matrice 20 par 20 d'énergie par paire de résidus de Miyazawa et Jernigan a été utilisée. Si deux acides aminés non contigus se retrouvent à moins de 5,88 Angstrom un terme énergétique correspondant au couple de résidus est ajouté à l'énergie totale du système. Pour chaque protéine, cent différentes conformations initiales sont générées. Toutes ces conformations initiales sont essentielles afin que la simulation soit indépendante des conditions initiales. La seule contrainte imposée à la conformation initiale est que la chaîne ne doit pas être compacte. Cette contrainte a été établie afin d'éviter que des résidus éloignés en séquence ne se retrouvent dans un groupe de résidus à cause de la conformation initiale. Les mouvements de résidus seuls sont de deux types : le mouvement du résidu de fin et un mouvement de coin pour les autres, le réseau (2, 1, 0) ne permettant pas le mouvement de manivelle. Après chaque mouvement, l'énergie conformationnelle est soumise au critère standard de Metropolis à température constante. Ainsi le mouvement est gardé ou rejeté. Le but de la méthode étant de déterminer la propension de chaque résidu à être enfoui dans le cœur protéique, la simulation doit aller jusqu'à la formation des premiers fragments compacts de la chaîne. L'algorithme étant en série, le temps de calcul est proportionnel à la longueur en séquence. Il a été démontré empiriquement que pour une chaîne de 50 résidus il fallait 10^6 pas afin d'observer les premier fragments compacts. Ainsi le nombre de pas de Monte Carlo est proportionnel à la longueur en séquence L. L'équation suivante régit le nombre de pas à appliquer : $Npas = 10^6 \left(\frac{L}{50}\right)^2$

Pour chaque simulation, 10^4 enregistrements des conformations intermédiaires sont effectués à intervalles réguliers. Pour cent conformations initiales, nous avons donc 10^6 enregistrements. Pour chaque enregistrement le nombre de voisins non covalents est comptabilisé pour chaque résidu. Une moyenne est ensuite opérée de la sorte : pour une protéine donnée et un résidu i à l'enregistrement r, le nombre moyen de voisins non covalents NC est : $NC(i) = \frac{1}{10^6}\sum_{r=1}^{10^6} nc(i,r)$.

De la distribution des occurrences selon la valeur NC, on observe que 13% des résidus ont un NC supérieur à six. Cette proportion est en accord avec différentes théories sur la proportion de résidus impliqués dans la nucléation du repliement (Papandreou et al., 2004). Ainsi les résidus dont le NC est supérieur ou égal à six seront définis comme des résidus interagissant le plus, ou Most Interacting Residues (MIR) (Papandreou et al., 2004). Les résidus qui comportent le plus de voisins aux premiers stades du repliement sont les résidus qui constituent le patron spécifique de contacts.

Pour valider ce modèle les positions des MIR ont été comparées avec celles des topo-hydrophobes et des extrémités de TEF. L'accord entre les MIR et les topo hydrophobes est clair, en effet 63% des positions de MIR sont dans un intervalle de ±5 résidus sur la banque d'une centaine de repliements déjà présentée. Quant aux extrémités de TEF, les positions de MIR se trouvent à 57 % dans le même intervalle de ±5.

Les concepts couplés des topo-hydrophobes et des TEF offrent un scénario simple et général au mécanisme de repliement des protéines globulaires et produisent un set de positions du cœur protéique.

1.3.6 Lissage des MIRs

Un inconvénient de la méthode précédemment décrite est le fait que les MIR sont sur-prédits par rapport aux extrémités des TEF. En effet, la moyenne de résidus prédits comme MIR est de 15% or le noyau du repliement est estimé à une moindre proportion (Shakhnovich et al. 1996). De plus, la méthode produit des positions de MIR qui apparaissent pour certaines regroupées dans la séquence. Dans ces regroupements, en accord avec les théories du repliement explicités plus tôt, un seul de ces résidus doit contribuer réellement au noyau du repliement.

La méthode employée ici est de lisser la courbe des voisins non covalents par la méthode du triangle de Pascal. Cette méthode permet de lisser une courbe en prenant en compte les voisins situés de part et d'autre d'un point considéré. Pour calculer une valeur on va considérer l'ensemble des points P tel que $P = \{p_{i-n} + \cdots + p_i + \cdots + p_{i+n}\}$ avec i le point étudié et n la puissance du lissage. Pour une puissance de lissage donnée, les valeurs de l'ensemble de points sont multipliées par les valeurs correspondantes dans la ligne égale à la puissance du triangle de Pascal. On utilise les valeurs définies

par le triangle de Pascal (représentées dans le tableau 1.3.1) pour calculer la nouvelle coordonnée du point p_i. La puissance du lissage est proportionnelle au nombre de voisins choisis. Il est à noter que dans le cas des extrémités de la séquence, le nombre de voisins diminue au fur et à mesure qu'on s'en rapproche, de même la puissance du lissage décroit jusqu'à être nulle.

n	Valeurs des points P
1	1 1
2	1 2 1
3	1 3 3 1
4	1 4 6 4 1
5	1 5 10 10 5 1
6	1 6 15 20 15 6 1
7	1 7 21 35 35 21 7 1
8	1 8 28 56 70 56 28 8 1
9	1 9 36 84 126 126 84 36 9 1
10	1 10 45 120 210 252 210 120 45 10 1

Tableau 1.3.1 Valeurs du triangle de Pascal. La première colonne n correspond à la puissance du lissage tandis que les colonnes suivantes correspondent aux coefficients dont sont affectés les points voisins.

Ci-dessous l'exemple de lissage des MIRs prédits pour la myohemerythrine de *Themiste zostericola* (code PDB : 2mhr) avec une puissance de lissage de ±5. La figure 1.3.3 représente la courbe des valeurs de « Non Covalent Neighbors » (NCN) en fonction de la position dans la séquence. La figure 1.3.4 représente la courbe lissée avec les valeurs du triangle du Pascal.

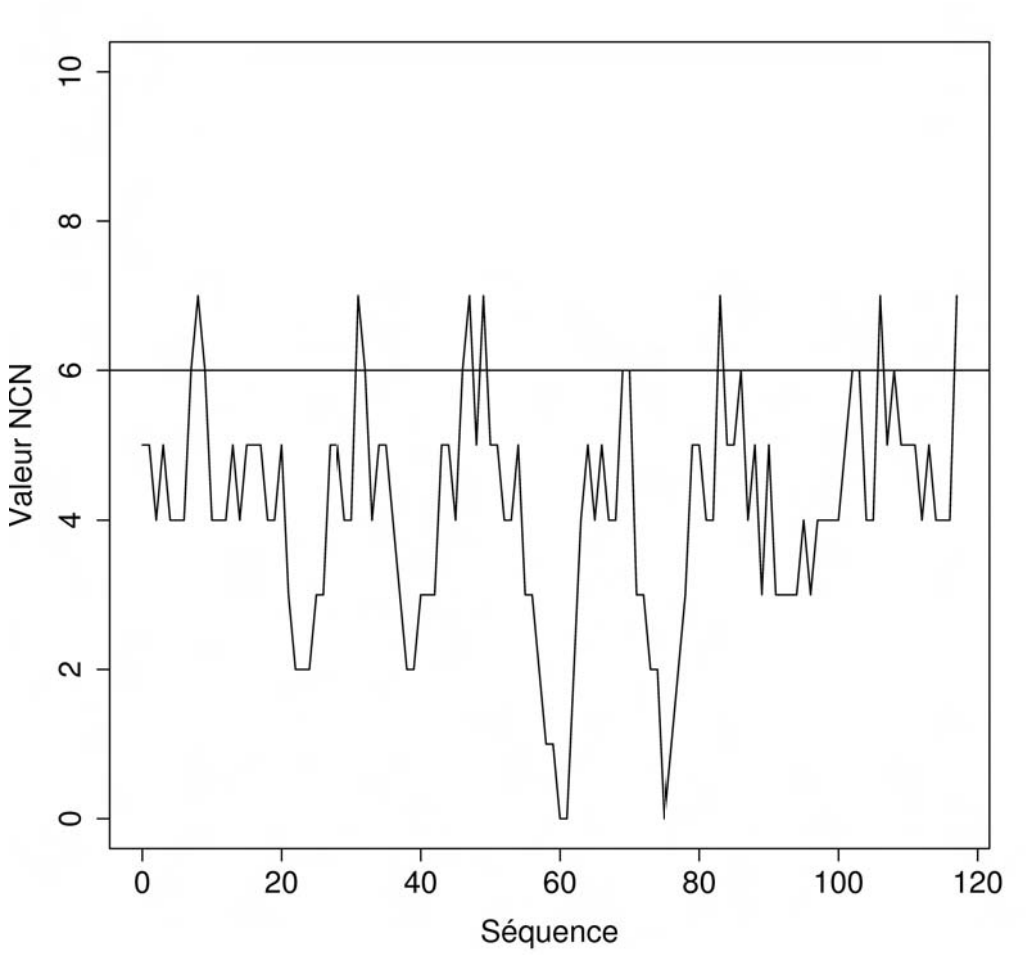


Figure 1.3.3 Profil NCN de la myohemerythrine de *Themiste zostericola* (code PDB : 2mhr). La ligne horizontale de valeur y = 6 représente le seuil de NCN choisi pour déterminer si un acide aminé est un MIR ou non.

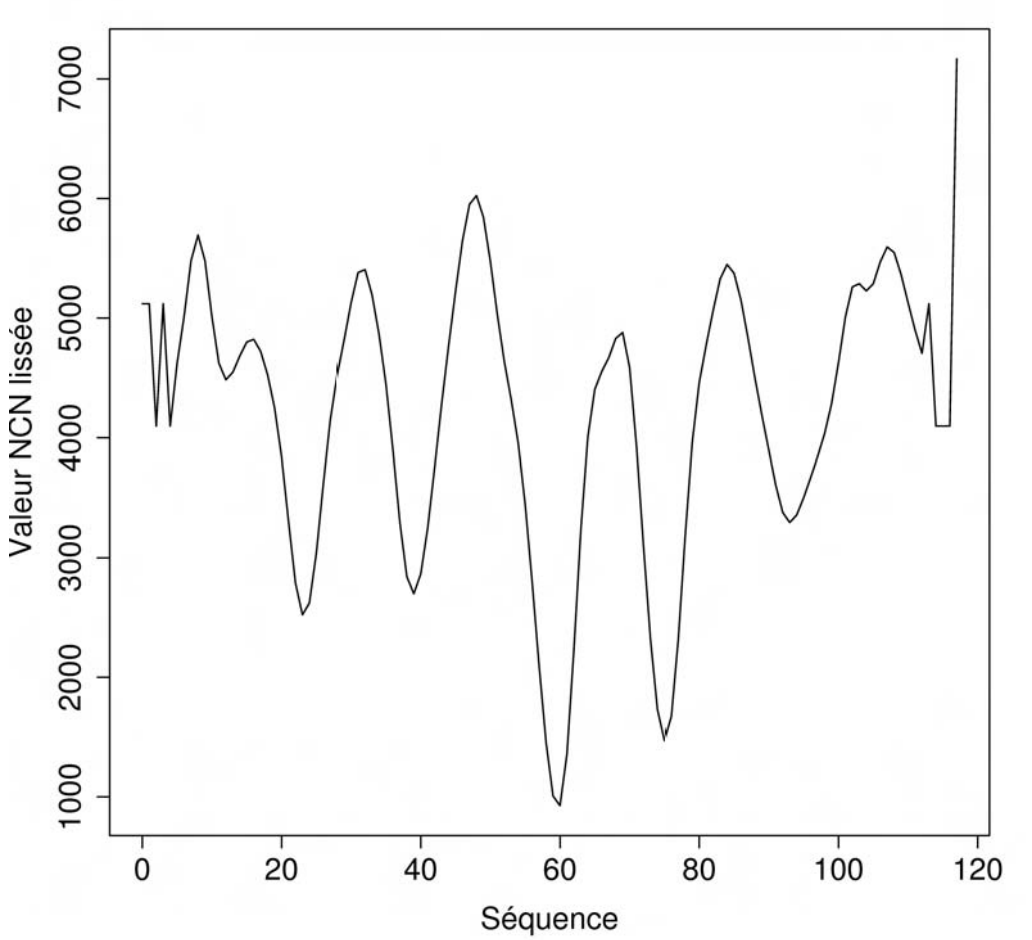


Figure 1.3.4 Profil NCN lissé de la myohermerytrine de *Themiste Zostericola* (code PDB : 2mhr); la méthode du triangle de Pascal est utilisée avec une fenêtre de ± 5 voisins autour de l'acide aminé considéré.

Les valeurs représentées sur la figure 1.3.3 sont multipliées par les valeurs du triangle de Pascal et produisent les valeurs de la courbe lissée représentée sur la figure 1.3.4. Cette opération n'a aucune incidence sur la définition qualitative des maxima en tant que MIR. Ainsi les maxima locaux de la courbes seront attribués à une position de SMIR (smoothed MIR). La prise en compte des maximas locaux permet de réduire le nombre total de résidus prédits et de s'affranchir d'un seuil de voisins non-covalents qui peut être discutable.

Les résultats utilisant les SMIR dans les travaux décrits plus loin, on été obtenue par le serveur RPBS où la puissance de lissage est fixée à ±5.

1.4. Conservatism of conservatism

Mirny et Shakhnovich (Mirny et Shakhnovich, 1999) ont étudié cinq des repliements les plus peuplés : l'immunoglobine, l'oligonucléotide–binding, le repliement Rossman, le plateau alpha/bêta et le tonneau TIM. Afin de distinguer la conservation en acides aminés « historiques », fonctionnelle et structurale, les séquences étudiées n'ont pas d'homologies évidentes. Pour chaque famille de repliement, les conservations en acides aminés qui coïncident après alignement structural ont été étudiées. Cette étude a été motivée par les découvertes du concept de nucléation et de la cinétique de repliement à deux états. Le scénario du repliement grâce à la formation d'un noyau spécifique suggère qu'un nombre de contacts obligatoires doit être formé afin que la chaîne protéique atteigne sa structure d'état de transition. Le noyau spécifique constitue un regroupement spatial de résidus qui sont cependant dispersés le long de la structure primaire. Le but de cette étude est de mettre en accord la compréhension théorique du repliement avec l'analyse de l'information évolutionnaire. Comme dans la plupart des cas, les analogues montrent un repliement commun mais une fonction différente. La comparaison en séquence devrait mettre en exergue des positions où le conservatisme est relié à la stabilité structurale et à la cinétique de repliement plutôt qu'à la fonction. Cependant l'analyse en séquence doit se faire de manière très prudente et ne fonctionne pas toujours à cause des possibilités de mutations d'acides aminés reliés. Pour illustrer ce propos on peut citer l'exemple d'interactions stabilisantes telles qu'une interaction hydrophobe ou un pont disulfure à des positions identiques dans deux analogues. Cependant l'interaction stabilisante n'étant pas de la même nature, les acides aminés en question seront différents et la comparaison en séquence ne pourra pas reconnaitre l'analogie. En d'autres termes, dans une famille de protéines homologues on peut attendre que les résidus conservés vont former une sous–structure : le changement de résidus en ces positions requiert qu'il y ait aussi un changement aux positions reliées, ce qui est extrêmement rare. Cette analyse suggère qu'un facteur qui peut pointer une propriété de structure commune reliée avec tous les analogues peut être la conservation intra–famille elle-même plutôt que l'actuel résidu en cette position. Ceci mène au concept de « conservatism of conservatism » ou CoC.

La détermination des Coc commence par celle de l'entropie de séquence. Pour ce faire il faut disposer d'un set d'analogues partageant un repliement commun (les protéines représentatives) et pour chaque protéine représentative un set de protéines homologues (la famille protéique). Ainsi un alignement multiple en séquence est réalisé sur les protéines homologues à chaque protéine représentative. Ensuite, les positions conservées sont identifiées, puis les familles de protéines sont structuralement alignées les unes aux autres et enfin les sites où les positions sont conservées entre les différentes familles de protéines sont identifiés. Le degré de conservation évolutionnaire pour chaque famille de séquences homologues est calculé par l'entropie en séquence :

$$s(l) = -\sum_{i=1}^{6} p_i(l) log p_i(l)$$

Où $p_i(l)$ est la fréquence de chacune des six classes de résidus pour la position l dans l'alignement multiple en séquence. Les six classes de résidus sont les résidus aliphatiques {AVLIMC}, aromatiques {FWYH}, polaires {STNQ}, chargés positifs {KR} ou négatifs {DE}, et spéciaux (vu leurs conformations particulières) {GP}. Une faible valeur du conservatisme intra–famille $s(l)$ indique que la position a été sous pression évolutionnaire pour garder un type particulier de résidu.

Ensuite les protéines représentatives et leurs familles respectives sont superposées structuralement et le Conservatism of Conservatism (CoC) est calculé de la façon suivante :

$$S(l) = \sum_{m=1}^{M} s^m(l)/M$$

Où l est maintenant la position dans l'alignement multiple et $s^m(l)$ le conservatisme intra–famille pour la famille m. Une faible valeur de $S(l)$ indique que la position l est conservée dans la plupart des familles protéiques possédant un repliement commun.

1.5. Alignements structuraux

1.5.1. Généralités sur les alignements

Un alignement est l'assignation des correspondances résidus-résidus pour deux séquences protéiques apparentées d'un point de vue de l'évolution, c'est à dire possédant un ancêtre commun. Les événements élémentaires qui peuvent se produire au cours de l'évolution sont la mutation, la disparition (délétion ou gap) ou l'apparition (insertion) d'un ou de plusieurs résidus (on parle d'indels dans ce dernier cas). Ce sont les indels qui posent problème, puisque l'on est contraint de comparer deux séquences dont les longeurs sont différentes, ce qui rend difficile la mise en évidence des mutations.

Les alignements en séquences sont nécessaires à beaucoup de sujets en biologie, tels que l'étude de l'évolution au sein des familles protéiques, l'identification de patrons conservés qui se replient en structures semblables, la modélisation par homologie ou encore la résolution de structures cristallines par replacement moléculaire. Les alignements structuraux ont été créés à la base pour remédier aux familles à faible taux d'identité qui ne donnaient pas de résultats avec les alignements de séquences

traditionnels. Ils nécessitent de réaliser dans une première étape une superposition structurale des deux protéines, dont on déduit ensuite l'alignement de séquences. La comparaison systématique de structures de protéines de familles très différentes a mené à la reconnaissance de sous-structures et l'identification de repliements (Schuler et al., 1991) Les informations de séquence et de structure peuvent être parfois contradictoires, comme le montrent les exemples suivants.

Les serpines constituent une famille de protéines pouvant adopter plusieurs états conformationnels. En effet la même structure primaire peut produire deux topologies différentes. Puisque les séquences des deux états sont identiques, l'alignement est trivial. Cependant, des parties de la protéine conservent des similitudes locales en structure. Ainsi une superposition structurale peut ajouter une information en indiquant quelles sont les zones qui se correspondent au niveau tertiaire. Par opposition, elle permet de déterminer les régions charnières, de flexibilité, que l'alignement des séquences ne permet pas de mettre en évidence.

Le second exemple est le « domain swapping » (Liu & Eisenberg, 2002). Ce processus peut faire partie du mécanisme d'assemblage des complexes oligomériques. Ainsi un (ou plusieurs) élément de structure secondaire d'une protéine monomérique est remplacé, dans un homodimère, par le même élément mais provenant de l'autre chaîne. Le « domain swapping » peut aller de la structure secondaire jusqu'au domaine structural. Il arrive qu'une protéine puisse être homologue avec un dimère à domaine « swappé ». Dans un tel cas un alignement en séquence alignera une protéine avec un des domaines du dimère. Alors qu'un alignement en structure alignera la protéine d'intérêt avec des fragments des deux domaines du dimère. La figure 1.5.1 illustre ce phénomène par l'exemple du dimère GB1 (domaine B1 de la protéine G du streptocoque)

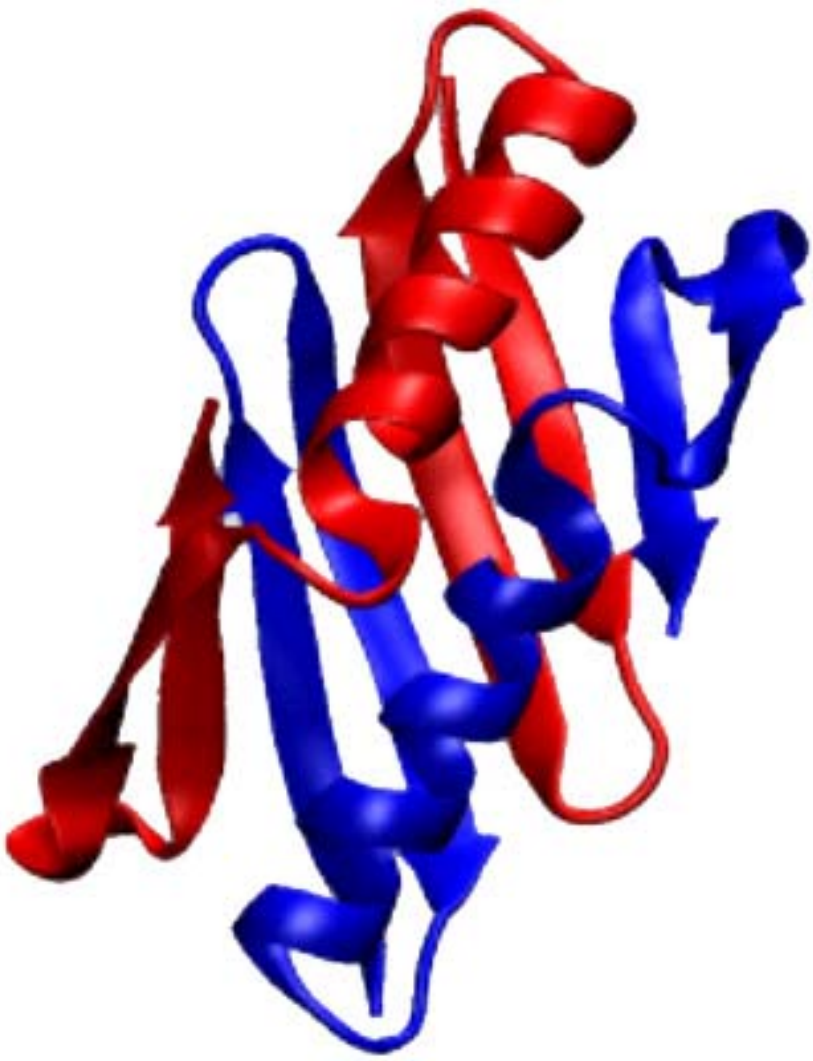


Figure 1.5.1 Dimère du domaine B1 de la protéine G du streptocoque, chaque chaîne peptidique est colorée d'une couleur différente.

De plus lors de l'évolution il est classique que les séquences divergent plus vite que les conformations. C'est pourquoi on valide souvent la qualité d'un alignement en séquence par une comparaison des structures.

1.5.2. Alignement structural par paire

L'alignement structural résulte de la superposition de deux structures. Le critère pour quantifier l'alignement est la déviation quadratique moyenne ou RMSD. Entre une chaîne composées de *n* carbones alpha dénommés $x_{1,i}$ et une autre chaîne composée de *n* résidus $x_{2,i}$ le RMSD est alors :

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n}(x_{1,i} - x_{2,i})^2}{n}}$$

Les données fournies par les programmes d'alignements structuraux sont au minimum de deux sortes : premièrement les coordonnées tridimensionnelles de chaque structure après superposition et le calcul du RMSD mutuel ; la seconde est une correspondance à une dimension des résidus de l'alignement sous forme de séquence.

La superposition de structures rigides marche bien pour les protéines raisonnablement reliées, mais pour des protéines plus divergentes le cœur conservé se retrouve être de petite taille. La plupart des algorithmes dynamiques, qui sont une optimisation visant à éviter de faire plusieurs fois le même calcul, soulagent ce genre de problème mais recèlent quelques difficultés avec les mouvements internes aux domaines. Certaines méthodes fondées sur la comparaison des environnements structuraux qui peuvent être définis en terme de structure secondaire, d'accessibilité au solvant ou encore du statut des liaisons hydrogène, ont la capacité d'outrepasser ces problèmes (Shi et al., 2001). Pour produire des superpositions structurales de plusieurs protéines, le fonctionnement le plus employé à l'heure actuelle consiste à effectuer des superpositions par paires, en commençant par celles qui sont les plus proches en terme de séquence, et en procédant de proche en proche sur toute la famille. De nouvelles techniques ont été développées pour produire un alignement multiple de manière concertée, cependant on ne dispose pas encore d'assez de tests sur ces méthodes (Carpentier, 2005).

Ces méthodes de superposition par paires que nous allons décrire cherchent à superposer les carbones alpha correspondant aux coordonnées tridimensionnelles du squelette protéique tout en diminuant la distance géométrique entre ces derniers.

Ci-dessous figure une liste des différents programmes d'alignement structuraux les plus utilisés.

1.5.2.1. MAMMOTH

MAtching **M**olecular **M**odels **O**btained from **TH**eory. Comme son nom le suggère, MAMMOTH a été développé à l'origine pour comparer des modèles venant de la prédiction protéique (THeory). Ainsi le programme accepte de larges régions divergentes, mais a démontré qu'il fonctionnait bien avec des modèles expérimentaux, plus particulièrement quand l'homologie est éloignée. Les tests ont été effectués sur des structures prédites non validées du concours CASP (Critical Assignemnt of Structural Predictions) (Moult, 2005) et les annotations automatisées ont montré une corrélation serrée avec les annotations faites à la main (Ortiz et al., 2002 ; Qiu et al., 2007). Une base de données a ainsi été créée à partir de l'annotation fondée sur les structures prédites de protéines inconnues par MAMMOTH. Le programme décompose la structure protéique en peptides cours (heptapeptides) qu'il compare à ceux d'une autre protéine. La similitude est quantifiée par un score de similitude de matrices et l'alignement optimal des résidus est calculé avec un programme dynamique hybride (global – local). Une évolution de ce programme a donné MAMMOTH-mult pour aligner des familles

entières de protéines homologues. Cet algorithme est assez rapide et produit un alignement structural de qualité.

MAMOTH est accessible à l'adresse suivante : http://ub.cbm.uam.es/mammoth/mult

1.5.2.2. RAPIDO

Rapid **A**lignements of **P**roteins **I**n terms of **DO**mains est un serveur web pour l'alignement de structures tridimensionnelles issues de la cristallographie. RAPIDO identifie les fragments qui sont similaires structuralement dans deux protéines par une approche fondée sur la différence en matrice de distances. Les fragments superposables ou Matching Fragments Pairs sont ensuite représentés comme les nœuds d'un graphe et sont concaténés pour former un alignement moyen. Le traitement final du serveur applique un algorithme génétique pour l'identification de régions invariables en conformation (Schneider, 2002). En plus des fonctionnalités existant dans la plupart des programmes d'alignements structural, RAPIDO permet d'identifier des régions de structures équivalentes même quand il est question de fragments relativement éloignés en terme de séquence, voir séparés par d'autres domaines.

RAPIDO est accessible à l'adresse suivante : http://webapps.embl-hambourg.de/

1.5.2.3. SSAP

La méthode **S**equential **S**tructure **A**lignment **P**rogram ou SSAP produit un alignement structural basé sur des vecteurs inter atomiques dans la structure spatiale. Ce ne sont pas les carbones alpha qui sont utilisés comme à l'accoutumée, mais les carbones bêta, excepté pour la glycine. Ainsi l'algorithme prend en compte l'état rotamérique d'une partie des chaînes latérales de chaque résidu. SSAP fonctionne d'abord en construisant pour chaque protéine des séries de vecteurs de distance inter-résidus entre chaque résidu et ses voisins les plus proches. Une série de matrices est alors générée contenant les vecteurs différences entre chaque paire de résidus voisins. La programmation dynamique appliquée à chaque matrice résultante détermine une série d'alignements locaux optimaux qui sont ensuite sommés dans une matrice « sommaire » à laquelle est encore appliquée une programmation dynamique pour déterminer l'alignement structural dans son entier. SSAP ne produit à l'origine que des alignements par paires, mais il possède une extension récente pour les alignements multiples (Taylor et al., 1994). Il a été utilisé d'une façon "tout pour tout" pour produire une classification hiérarchique de schémas de repliements utilisée dans la banque de structures CATH (Class, Architecture, Topology, Homology) (Orengo et al., 1997; Orengo et al., 1998) .

SSAP est disponible à l'adresse suivante : http://www.cathdb.info/cgi-bin/SsapServer.pl

1.5.2.4. MATRAS

Matras est un serveur web comparant les structures tridimensionnelles des protéines (Kawabata, 2003). Un avantage de MATRAS est son score de similitude de structures, qui est défini comme les log-odds des probabilités évolutives, c'est-à-dire la probabilité qu'une structure change en une autre (Kawabata, 2003). Les log-odds sont des mesures d'effets de taille décrivant l'association ou l'indépendance entre deux groupes de données binaires. Ce score est dédié à la reconnaissance de la similitude structurale de deux protéines reliées en termes d'évolution (homologues). Le serveur propose des superpositions par paires, des superpositions d'une chaîne sur elle-même pour la recherche de domaines dupliqués, mais aussi des alignements multiples jusqu'à 10 structures. Le programme emploie un algorithme d'alignement progressif par lequel les alignements par paires sont assemblés dans le bon ordre. Le dernier service du serveur est de comparer une structure à un grand nombre de structures de la PDB (Berman et al., 2000).

MATRAS est disponible à l'adresse suivante : http://biunit.aist-nara.ac.jp/matras/

1.5.2.5. MUSTANG

Konagarthu et Lesk ont développé un algorithme robuste d'alignement multiple reposant sur les structures protéiques, qu'ils ont nommé **MU**ltiple **ST**ructural **A**lig**N**ment Al**G**orithm (Konagurthu et al., 2006).

Un alignement structural peut faire la différence entre les régions alignables et les régions non alignables des protéines, ce qui n'est pas possible par un alignement en séquences par paires. Ainsi les alignements structuraux ne doivent pas être vus seulement comme des extensions aidant à la superposition de séquences de faible taux d'identité mais aussi comme un outil fournissant des informations sur les similitudes et divergences de conformations.

MUSTANG aligne les résidus sur la base de la similitude de patrons de contacts inter résidus et de la topologie structurale. L'algorithme utilise un cadrage progressif afin de construire l'alignement final à partir des alignements par paires. Au centre de la méthode existe une fonction de score robuste pour les alignements par paires, fonction qui permet d'utiliser un algorithme de programmation dynamique donnant un alignement par paires précis sans avoir besoin d'introduire des pénalités de gap. Avant la construction de l'alignement final, les scores par paires sont recalculés par rapport aux structures restantes. Ceci réduit considérablement le problème de faire les mauvais choix pour reconstruire l'alignement final. La construction de l'alignement final est réalisée le long d'un arbre guide (qui sera expliqué avec plus de précision plus loin).

Phase 1 : Calcul de scores de correspondance résidu – résidu par paire

MUSTANG extrait les carbones alpha des n protéines qu'il doit aligner et ignore complètement la séquence des protéines. Dans un premier temps MUSTANG calcule la matrice complète des distances inter carbones alpha pour chaque structure. La fonction de score des paires initiales se sert de superpositions optimales et de valeurs de RMSD pour détecter des sous-structures contigües à l'intérieure des deux structures de la paire. Ceci afin d'obtenir un état numériquement appréciable de chaque correspondance résidu – résidu possible. Les résultats de cette fonction de score sont déduits indépendamment pour chaque paire, afin d'aligner les paires de structures par un algorithme de programmation dynamique simple. Les pénalités de gap ne sont pas exigées par cet algorithme car la fonction de score de la paire de résidus est influencée par l'appartenance à des sous structures similaires, d'abord localement puis globalement. La fonction de score est décomposée comme suit :

- Une liste des sous-structures similaires contigües maximales est générée.
- Un score brut est ensuite calculé pour chaque paire de résidus correspondants.
- Un essai d'alignement par paire est ensuite généré.
- Ensuite la liste de sous-structures est élaguée, car la liste contient de nombreux fragments alignés par paires redondants. C'est cette redondance qui est éliminée afin d'accélérer l'étape suivante.
- Les scores de similitude de la correspondance résidu – résidu sont enfin recalculés.

Phase 2 : Réunion des équivalences entre résidus de toutes les paires de structures de l'ensemble.

Les équivalences produites à partir de chaque alignement par paire sont stockées dans une structure de données.

Phase 3 : Nouveau calcul des scores de correspondance résidu-résidu dans le contexte de l'alignement multiple (phase d'extension)

L'approche de l'alignement progressif par paires pour construire un alignement multiple le long d'un arbre guide implique la fusion de série d'alignements par paire. Dans de telles procédures, de nombreux choix « gourmands » en temps de calcul incorrects sont faits. L'ordre dans lequel les alignements par paires sont forcés à être reconstruits affecte la précision et la qualité de l'alignement multiple final. La phase d'augmentation du nombre de structures incorporées dans l'arbre guide de MUSTANG permet de réduire le nombre de ces choix « gourmands » incorrects. Lors de cette phase,

une nouvelle matrice est générée, contenant les scores de toutes les correspondances possibles résidu - résidu entre deux structures. Ces scores sont initialisés à zéro dans la première itération. Des correspondances transitives sont ensuite établies entre chaque paire de structures par toutes les structures intermédiaires possibles. Ces correspondances sont évaluées avec les intermédiaires que l'on peut introduire entre deux structures d'une paire. Plus il existe d'intermédiaires supportant un alignement entre deux structures, plus le score de la paire est augmenté et tend à subsister dans l'alignement final. Pour une meilleure compréhension les figures 1.5.2. à 1.5.4. illustrent un exemple de la phase d'extension.

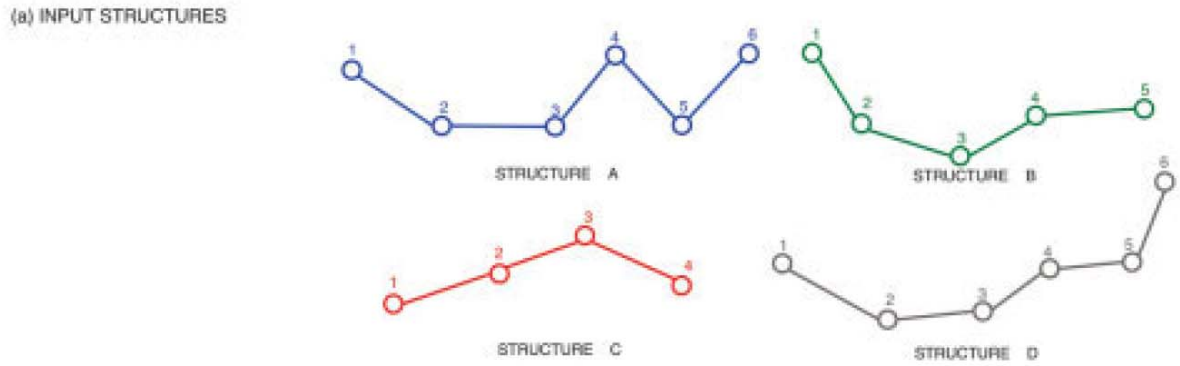


Figure 1.5.2. Structures en entrée dans l'algorithme de MUSTANG.

Soit quatre structures dont les résidus représentés par de petits cercles sont notés de un à six. Les structures A, B, C et D sont respectivement de couleur bleue, verte, rouge et grise.

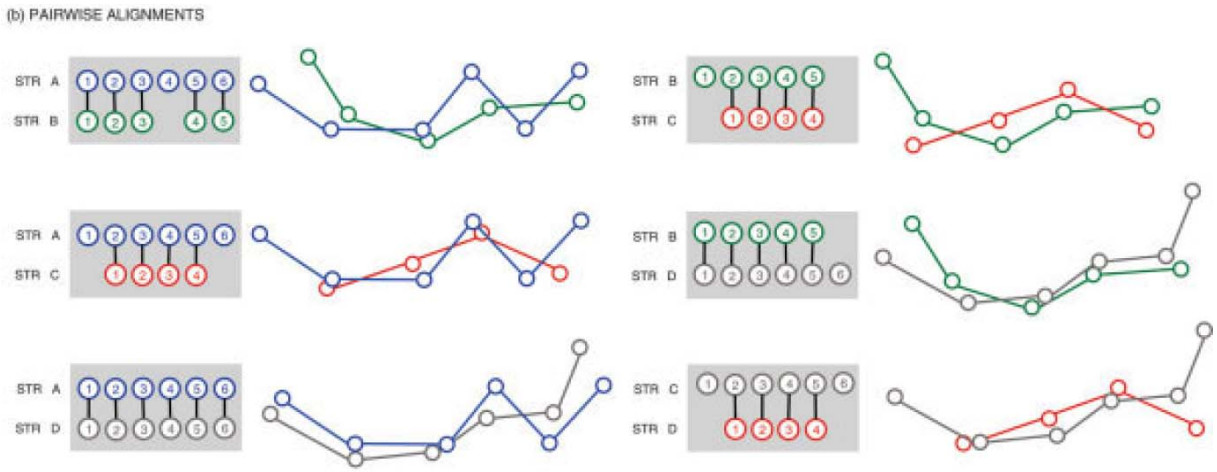


Figure 1.5.3. Alignements par paires des structures entrées

Le figure 1.5.3. montre tous les alignements par paires possibles que MUSTANG a généré dans la phase 2.

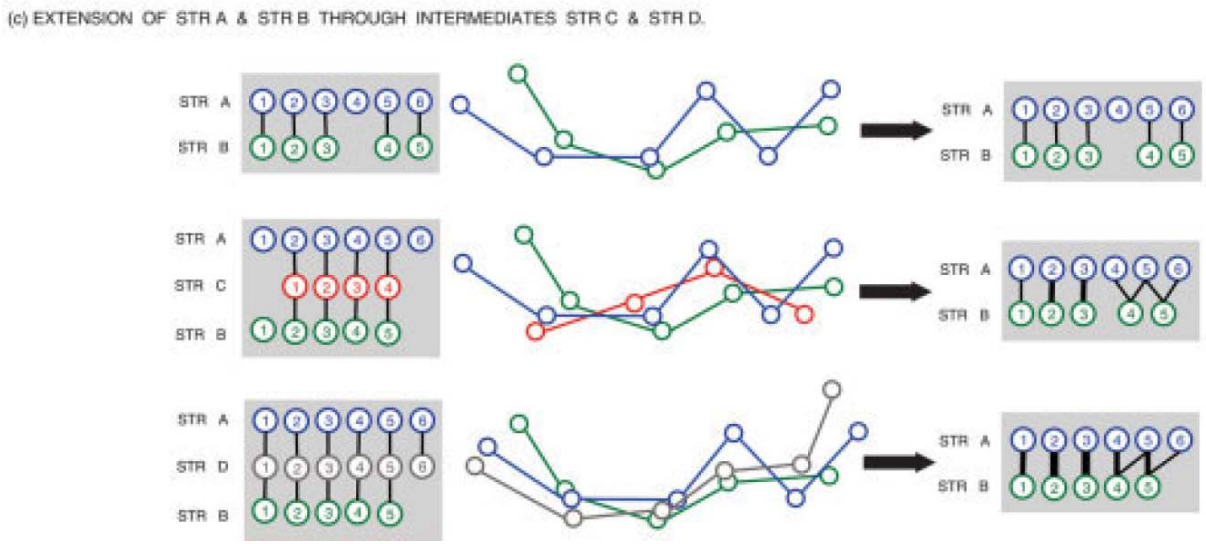


Figure 1.5.4. Extension des correspondances par paires entre les structures A et B utilisant C et D comme intermédiaires.

La figure 1.5.4. montre la phase d'extension dans laquelle la paire A/B est pondérée par le recalcul dans le contexte des structures restantes C et D utilisées comme intermédiaires. Cette phase commence par l'attribution de poids étendus correspondant aux alignements dans la paire A /B. La finesse des lignes reliant chaque résidu représente le poids attribué à chaque paire de résidus. Les poids par paires de A/B sont premièrement étendus par l'utilisation de C. Nous avons A2 (le deuxième résidu de la structure A) équivalent à C1 et C1 équivalent à B2 dans l'alignement de la paire B/C. Ainsi la correspondance transitive est établie entre A2 et B2. Dans ce cas, la correspondance transitive renforce l'équivalence originelle entre A2 et B2 et le poids de la correspondance est augmenté de 1, comme l'illustre l'épaisseur du trait entre A2 et B2 sur la figure 1.5.4. De la même manière A3 et B3 voient leur poids de correspondance augmenté par l'intermédiaire de C2. En revanche la correspondance transitive entre A4 et B4 par C4 est incohérente avec l'équivalence entre A4 et le gap dans l'alignement entre A et B. Ainsi le poids de la correspondance entre A4 et B4 est établi comme celui de la paire originelle A5/B4. Cette équivalence incohérente entre A5 et B5 par l'intermédiaire de C4 est assignée au maximum des poids de correspondance A5/B4 et A6/B5. L'extension est ensuite appliquée par l'intermédiaire de D. Les correspondances A1/B1, A2/B2 et A3/B3 sont toutes cohérentes avec l'alignement originel et les poids sont augmentés comme auparavant. Les correspondances A4/B4 et A5/B5 sont aussi incohérentes dans ce cas. A la fin de toutes les extensions nous avons une nouvelle matrice pour la paire A B avec des poids reflétant les informations d'équivalence avec les autres structures C et D.

Ces extensions sont réalisées pour toutes les paires avant d'utiliser les nouvelles matrices pondérées pour la phase de l'alignement progressif.

Phase 4 : Alignement progressif

Enfin, la dernière phase de l'algorithme est la phase progressive de sélection de structures pour former l'arbre guide. Le score d'alignement par paire, recalculé lors de la précédente phase, est transformé pour chaque paire de structures en une matrice de divergence en distance. L'arbre guide est construit en utilisant ces matrices de distance – divergence par une technique de « neighbour joining » (McLachlan, 1972). Les alignements multiples sont évalués entre deux groupes de sous – alignements. Une fonction de profil calcule la somme des correspondances par paires entre chaque colonne des sous – alignements à être joints. Il en résulte un alignement multiple où chaque structure est superposée de façon optimale à chaque autre.

MUSTANG possède quelques inconvénients, néanmoins partagés par la majorité des programmes existants. Le principal est que chaque structure à superposer doit correspondre à un mono – domaine. Ceci demande un traitement des structures afin de les « découper » en domaines. L'alignement multiple reste une concaténation de plusieurs alignements par paires, ce qui est le cas de tous les programmes existants, à l'exception de celui de Carpentier malheureusement pas encore disponible.

Les avantages de MUSTANG sont nombreux. Premièrement, MUSTANG est facilement exploitable en local. Ensuite, le nombre de structures à superposer n'est pas limité, ce qui a été le facteur déterminant de notre choix pour ce programme vu que nous avons plus d'une cinquantaine de structures à aligner. De plus la fonction de reconstruction et d'affinement de l'alignement multiple confère à Mustang une robustesse appréciable.

MUSTANG est téléchargeable à l'adresse qui suit : http://www.bx.psu.edu/arun/research/mustang. Un taux élevé de conservation dans un alignement multiple veut-il dire l'appartenance au cœur protéique ? Ceci sera discuté plus loin dans le chapitre prédiction des résidus clé du noyau du repliement.

1.6. Concept HTOO

Gerstein et Altman (Gerstein & Altman, 1995) ont développé une méthode pour calculer un noyau structural moyen et décrire la variabilité d'une famille de protéines, les immunoglobulines. L'algorithme de recherche du noyau ou « core-finding algorithm » peut être décomposé en 5 étapes :

- Premièrement, il faut aligner les structures afin de décrire un "noyau putatif".
- Deuxièmement, la construction d'une structure non-biaisée à partir de tous les noyaux putatifs est effectuée.
- La détermination de la variation spatiale de chaque groupe d'atomes alignés est ensuite calculée. Le calcul de la variation pour une position donnée est effectué comme la somme des variabilités de chaque position alignée sur l'ensemble des structures en utilisant une matrice de variance/covariance. Chaque élément de cette matrice trois par trois est représenté par la covariance entre deux positions représentées par leurs coordonnées en x, y ou z. La matrice est translatée en ellipsoïde d'erreur centrée sur la position moyenne de chaque atome.
- On retire ensuite du noyau putatif la position ayant la plus grande variation spatiale. L'algorithme retourne ensuite à la deuxième étape où le noyau putatif a été construit. Et ainsi de suite jusqu'a ce que toutes les positions alignées soient retirées.
- Finalement on obtient une liste de positions atomiques ordonnées selon un facteur de rejet (throw-out order), pour représenter la séparation entre les atomes de cœur et les autres.

Cette méthode peut être utilisée afin de calculer un RMSD plus fin, contenant plus d'informations que celui que l'on emploie habituellement. L'idée de base est qu'on utilise la variation observée à chaque position dans le noyau pour échelonner la distance interatomique entre deux structures. Si un atome a une faible variation structurale dans le cœur, mais que sa variation de position dans l'alignement est grande alors cette différence devrait plus contribuer dans la différence totale. Inversement, si la différence en position dans l'alignement est faible par rapport à la variation structurale, c'est que la contribution est moins grande vis-à-vis de la valeur totale de la déviation entre les structures. Un poids relatif à la position est ainsi introduit, contrairement au RMS classique, en utilisant les ellipsoïdes d'erreur. Gerstein et Altman ont appelé cette mesure SD-RMS, exprimée en unité de déviation standard. Il est ensuite introduit un facteur de conversion entre la moyenne de l'unité de déviation et l'Angstrom. Ainsi la distance calibrée en Angstrom est différente de la distance réelle en fonction de la variabilité au sein de la famille étudiée.

L'amélioration du RMSD par la méthode décrite précédemment ne sera pas utilisée ici, nous nous servirons uniquement des résultats produits par le « core-finding algorithm ». Ces résultats représentent les positions du cœur et seront dénommées « High Throw Out Order (HTOO) residues.

1.7. Concept des valeurs de Phi

Il existe une technique expérimentale qui analyse la différence énergétique entre les conformations native et dépliée lors de la réaction de repliement d'une protéine de type sauvage et pour la même protéine dont l'un des résidus, que l'on présuppose appartenir au noyau du repliement, est muté. Soit un résidu i dans une protéine à l'état natif noté N, l'état de transition noté TS, et l'état dénaturé noté D. La valeur de Phi est le rapport de la différence d'énergie libre observée entre état de transition et état dénaturé (qui sert de référence) pour la protéine sauvage (W) ou mutée (M) sur la différence d'énergie libre entre états natifs et dénaturé pour les mêmes deux conformations :

$$\Phi = \frac{(\Delta G_W^{TS-D} - \Delta G_M^{TS-D})}{(\Delta G_W^{N-D} - \Delta G_M^{N-D})}$$

En d'autres termes, la valeur de Phi représente le ratio de la déstabilisation introduite par une mutation sur l'état de transition contre celle introduite sur l'état natif (Fersht et Sato, 2004; Fersht, 1997; Fersht, 2000; Itzhaki et al., 1995) .

1.8. Prédiction des résidus clés du noyau du repliement

1.8.1. Pourquoi la théorie de la nucléation condensation ?

Les premières étapes de la simulation du repliement d'une chaîne peptidique sur un réseau de points font apparaître des agrégats autour de résidus hydrophobes comme cela se constate sur la figure 1.8.1.

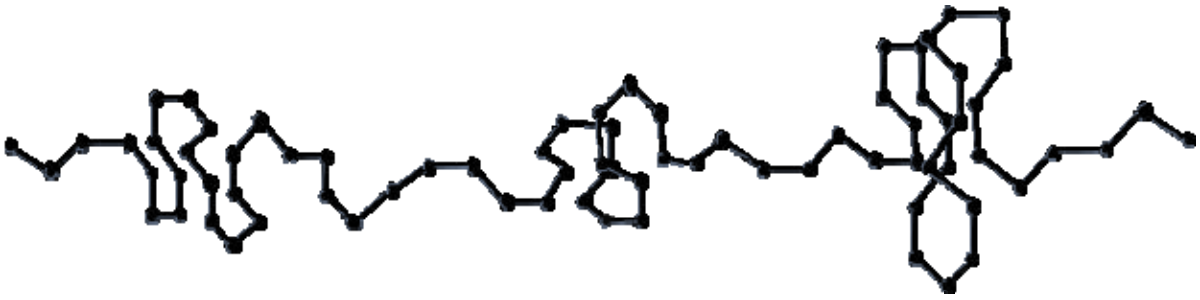


Figure 1.8.1. Exemple d'une simulation du repliement sur un réseau de points après la formation de quelques agrégats autour de résidus majoritairement hydrophobes. Les régions peu compactes correspondent à des fragments qui sont des boucles dans la structure native.

Cette observation peut être rapprochée de l'analyse fragmentaire et modulaire des protéines. C'est pourquoi nous nous fonderons sur les concepts de MIR qui sont en accords avec les théories de nucléation condensation, et le concept de TEF qui permet l'analyse fragmentaire des domaines protéiques, pour la prédiction des résidus impliqués dans le processus du repliement.

1.8.2. Pourquoi le repliement immunoglobuline ?

Cette famille de protéines montrant le même repliement offre une multitude de fonctions différentes, qui ont été depuis longtemps étudiées en détail et dont les résultats sont bien référencés dans la littérature d'un point de vue théorique et expérimental. Ainsi, cette grande divergence en fonctions permet d'effectuer une comparaison sur des bases structurales. Le repliement immunoglobuline inclut les super familles immunoglobuline (Ig) et fibronectine de type III (FnIII) qui comptent respectivement 5771 et 2468 membres (Fowler et Clarke, 2001). Les domaines Ig sont retrouvés dans une large variété de protéines comme les récepteurs de surface, les molécules d'adhésion cellulaire, des protéines musculaires et d'autres. Dans des études précédentes, des corrélations évidentes entre la stabilité protéique et le taux de repliement, ou encore entre la stabilité et la population d'intermédiaires, ont été mis en exergue.

Nous avons utilisé une particularité du repliement immunoglobuline, qui est de type bêta sandwich et par ce fait nous pouvons affirmer que les brins se retrouvent au cœur du domaine. Ainsi nous avons pris cette information de structuration secondaire en brin bêta comme un élément d'appartenance au coeur protéique. La conservation au sein de l'alignement structural permet-elle de décrire le cœur protéique ? Les parties conservées structuralement sont les parties les moins flexibles de la protéine, donc les brins. Le cœur protéique est la partie de la chaîne protéique qui est au centre du globule et qui montre une grande compacité. Le cœur protéique doit montrer une flexibilité faible, ainsi les parties alignées d'une famille de protéines de même repliement doivent correspondre au cœur protéique.

Le cœur protéique n'est pas à confondre avec le noyau de repliement. Le cœur englobe tous les résidus au centre du globule qui sont en compaction maximale. Le noyau du repliement est un réseau de quelques résidus qui initient la nucléation de l'édifice protéique. Cependant on peut penser que le noyau du repliement est compris dans le cœur protéique.

1.8.3. Pourquoi un alignement multiple ?

Les MIR sont sur-prédits, car nous avons déjà vu que 63% des MIR sont des positions topo-hydrophobes à ±5 résidus près, sur une centaine de repliements. Or les topo-hydrophobes correspondent raisonnablement bien aux résidus du noyau du repliement. Il en résulte que notre simulation prédit trop de résidus dans le noyau, à peu près d'un facteur deux. Des tentatives afin de discriminer les « bons » MIR, ceux qui sont aux extrémités des TEF et donc correspondent statistiquement aux topo hydrophobes, des « mauvais » MIR (les autres), ont été conduites et une corrélation avec les extrémités de TEF a été étudiée. La corrélation des MIR et des extrémités de TEF a été réalisée au cours de la thèse de Mathieu Lonquety mais elle n'apporte pas de solution totalement satisfaisante. Il a fallu introduire un nouveau procédé afin de discriminer les MIR, qui consiste à inclure la notion d'alignement structural. Les régions conservées le long de cet alignement doivent définir les cœurs protéiques et pondérer favorablement les MIR qui s'y trouvent.

Le concept de nucléation émerge tel un paradigme pour décrire le repliement protéique, spécifiquement par une cinétique en deux étapes. Le scénario suggère qu'un certain nombre de contacts obligatoires doivent être formés afin que la chaîne prenne la conformation de l'état intermédiaire. Des séquences différentes mais homologues d'un point de vue évolutif se repliant en une même structure devraient avoir un noyau de nucléation voisin. Par conséquent, la localisation du noyau du repliement d'une structure peut servir d'empreinte d'un repliement protéique.

Nous commencerons cette étude sur des bases structurales. Comme on l'a vu précédemment, les protéines peuvent être décrites comme un assemblage de sous–structures tels que les TEF. L'analyse des TEF renseigne sur le cœur protéique. Un inconvénient de cette méthode est la nécessité d'avoir une structure. Nous partons néanmoins de cette analyse afin de valider ultérieurement la prédiction des MIR. Pour des domaines différents d'immunoglobulines, le cœur commun a été établi à 76 résidus (Chothia et al. 1998) ce qui est une proportion assez large par rapport a l'évaluation du cœur des protéines globulaires prévu entre 15 et 30 résidus selon les études de (Chen et al., 2008). Ainsi le cœur protéique peut montrer une fluctuation conséquente et nous espérons contribuer à l'étude du noyau par sa prédiction et l'analyse de structures divergentes. Un inconvénient potentiel de cette stratégie concerne les protéines multi–domaines, car avec plusieurs cœurs il va être difficile de les différencier expérimentalement et de les prédire. C'est pourquoi nous nous attachons à la sélection rigoureuse des domaines de manière manuelle et que nous travaillerons uniquement sur ces domaines. Le choix de

domaines relativement petits devrait réduire les risques de mauvaises prédictions. Montrant des histoires évolutionnaires et des fonctions très différentes des analogues, le seul trait que partagent les immunoglobulines est leur structure native, c'est-à-dire leur repliement.

Nous avons comparé notre approche avec d'autres méthodes telles que les HTOO qui définissent les résidus clés à partir de la structure, et avec l'approche de Conservatism of Conservatism qui détermine ces positions à partir d'alignements multiples. Nous allons vérifier aussi que notre prédiction ne peut pas être remplacée par la détermination des résidus les plus enfouis. De plus pour vérifier la justesse de la méthode, nous l'avons appliquée à un autre repliement : le repliement de type flavodoxine qui appartient à la classe mixte alpha–bêta. Nous avons employé ce repliement pour une extension de la méthode car ce repliement a la particularité d'avoir un noyau du repliement qui pourrait être en partie exposé au solvant.

1.9. Matériels et méthodes

1.9.1. Récupération des données de séquences et de structures

Mon premier travail a été de constituer une banque de domaines. Une étude précédente a été réalisée sur le repliement immunoglobuline par Hallaby et al. (Halaby et al., 1999). Elle s'appuyait sur une banque de domaines immunoglobulines, c'est naturellement que nous avons repris ces entrées. A cette banque, nous avons ajouté les entrées de la banque utilisée dans l'étude de Gerstein et Altman (Gerstein & Altman, 1995). Les structures des protéines ont été obtenues par la base de données Protein Data Bank (Berman et al., 2000), ainsi que les séquences associées. Les structures des protéines sont encodées au format pdb et les séquences sont au format fasta.

Une fois les structures et séquences rapatriées sur un ordinateur local, il a fallu analyser chaque entrée PDB et y reconnaître les limites des domaines. Ces informations ont été obtenues sur le serveur de la banque de domaines CATH. A chaque domaine correspondra alors un fichier fasta comprenant sa séquence propre. Les fichiers seront nommés comme suit : code PDB-code chaîne-numéro du domaine. Par exemple : 1ACXA01.fasta. Grace à cette information en séquence on récupère les coordonnées tridimensionnelles associées. A chaque domaine correspondra un fichier contenant les coordonnées atomiques du domaine. Ces fichiers seront nommés sur le même schéma que ceux de séquences, mais l'extension changera en « pdb », comme par exemple : 1ACXA01.pdb. Au total, nous avons formé un ensemble comprenant 56 domaines Ig.

1.9.2. Formatage de données de structures

Une fois les domaines découpés, j'ai utilisé l'algorithme d'alignement MUSTANG v3 afin de superposer les différentes structures. Le logiciel produit en sortie l'alignement des séquences correspondant à la superposition des structures. Cet alignement est fourni au format multi-fasta, qui présente les séquences en « paragraphes ». C'est-à-dire que l'alignement est disséqué dans sa longueur et présenté en plusieurs paragraphes pour correspondre à un fichier texte communément lisible. Or pour être traité, il est plus facile que l'alignement soit sur une ligne pour chaque protéine. Le reformatage de l'alignement a été un de mes premiers travaux effectués au cours de ma thèse. Ainsi les corrélations ultérieures seront beaucoup plus facilement analysées.

1.9.3. Analyse des identités en séquence et vérification de la non redondance

Lors des études structurales il est important que les identités soient faibles. En effet si les acides aminés sont conservés en séquence afin de préserver la fonction, les positions particulières sont conservées afin de préserver la structure. Ainsi il est important de ne « récolter » que les conservations porteuses d'information structurale. L'analyse de l'identité s'est faite par couple. J'ai programmé un script domestique afin de générer tous les couples possibles de la banque de domaines. Tous ces couples sont ensuite alignés en séquence par le programme T-coffee (Notredame et al., 2000). L'alignement produit par T-coffee est ensuite analysé. Le taux d'identité (*TI*) est calculé de la façon suivante : le rapport de la somme des résidus alignés strictement identiques (x_{sa}) sur la somme des résidus alignés (x_a).

$$TI = \frac{\sum_{i=1}^{n} x_{sa,i}}{\sum_{i=1}^{n} x_{a,i}}$$

Parmi les couples de séquences montrant un taux d'identité supérieur à 70%, l'une des deux séquences est retirée de la banque de domaines. Le choix du partenaire à retirer a été guidé en faveur des structures présentant les meilleures informations, telles que la résolution, la méthode expérimentale (la cristallographie étant préférée à la RMN), l'absence de ligand ou de trous dans la structure, afin que l'alignement structural soit le plus précis possible. L'identité de la banque est appréciée par la distribution des différents pourcentages d'identité observés entre paires (figure 1.9.1.).

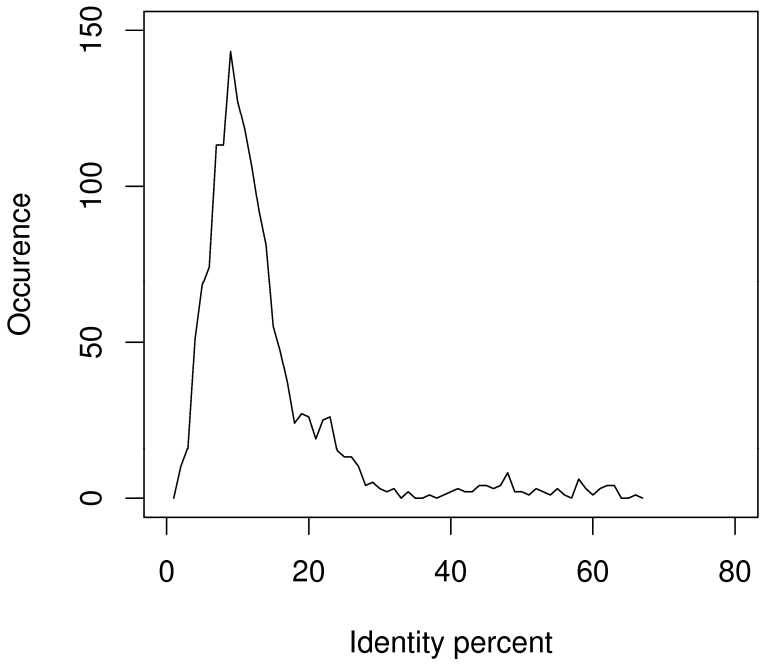
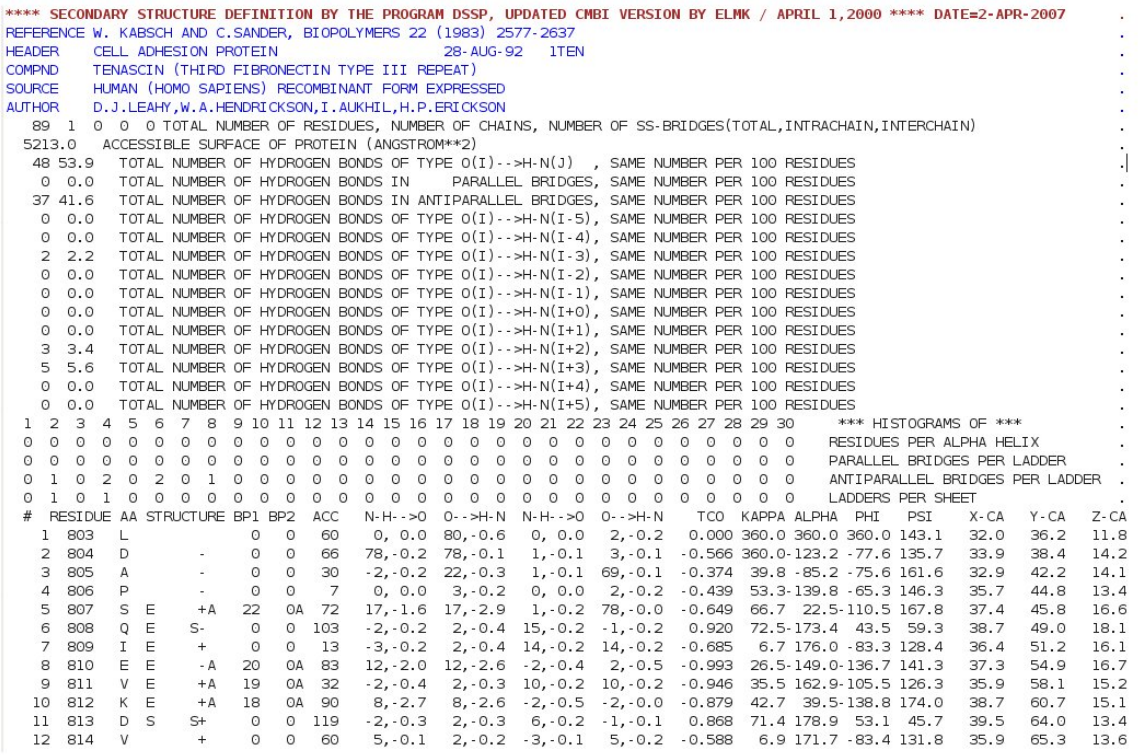


Figure 1.9.1. Occurrence de paires en fonction du pourcentage pour la banque de 56 structures de domaines immunoglobulines.

Le taux d'identité à été calculé après l'alignement structural avec la même méthode, seulement au lieu que l'alignement des paires de séquence soit fait par T-coffee, on récupère l'alignement en séquence correspondant à l'alignement structural.

1.9.4. Données de structuration secondaire

Les informations de l'appartenance de chaque résidu à une structure secondaire sont disponibles au format DSSP (Kabsch & Sander, 1983). Le format DSSP présente les résidus par ligne. A chaque résidu sont attribuées plusieurs informations telles que l'accessibilité au solvant, les angles de torsion, les liaisons hydrogènes et la structuration secondaire, entre autres. La figure 1.9.2 représente une capture d'écran du fichier résultat de 1TEN au format DSSP.



```
**** SECONDARY STRUCTURE DEFINITION BY THE PROGRAM DSSP, UPDATED CMBI VERSION BY ELMK / APRIL 1,2000 **** DATE=2-APR-2007
REFERENCE W. KABSCH AND C.SANDER, BIOPOLYMERS 22 (1983) 2577-2637
HEADER    CELL ADHESION PROTEIN                   28-AUG-92   1TEN
COMPND    TENASCIN (THIRD FIBRONECTIN TYPE III REPEAT)
SOURCE    HUMAN (HOMO SAPIENS) RECOMBINANT FORM EXPRESSED
AUTHOR    D.J.LEAHY,W.A.HENDRICKSON,I.AUKHIL,H.P.ERICKSON
   89  1  0  0  0 TOTAL NUMBER OF RESIDUES, NUMBER OF CHAINS, NUMBER OF SS-BRIDGES(TOTAL,INTRACHAIN,INTERCHAIN)
  5213.0   ACCESSIBLE SURFACE OF PROTEIN (ANGSTROM**2)
   48 53.9   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(J)  , SAME NUMBER PER 100 RESIDUES
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS IN     PARALLEL BRIDGES, SAME NUMBER PER 100 RESIDUES
   37 41.6   TOTAL NUMBER OF HYDROGEN BONDS IN ANTIPARALLEL BRIDGES, SAME NUMBER PER 100 RESIDUES
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-5), SAME NUMBER PER 100 RESIDUES
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-4), SAME NUMBER PER 100 RESIDUES
    2  2.2   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-3), SAME NUMBER PER 100 RESIDUES
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-2), SAME NUMBER PER 100 RESIDUES
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-1), SAME NUMBER PER 100 RESIDUES
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+0), SAME NUMBER PER 100 RESIDUES
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+1), SAME NUMBER PER 100 RESIDUES
    3  3.4   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+2), SAME NUMBER PER 100 RESIDUES
    5  5.6   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+3), SAME NUMBER PER 100 RESIDUES
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+4), SAME NUMBER PER 100 RESIDUES
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+5), SAME NUMBER PER 100 RESIDUES
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30     *** HISTOGRAMS OF ***
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    RESIDUES PER ALPHA HELIX
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    PARALLEL BRIDGES PER LADDER
 0  1  0  2  0  2  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    ANTIPARALLEL BRIDGES PER LADDER
 0  1  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    LADDERS PER SHEET
  #  RESIDUE AA STRUCTURE BP1 BP2  ACC    N-H-->O  O-->H-N  N-H-->O  O-->H-N    TCO  KAPPA ALPHA  PHI   PSI    X-CA   Y-CA   Z-CA
    1  803   L              0   0   60    0, 0.0  80,-0.6   0, 0.0   2,-0.2   0.000 360.0 360.0 360.0 143.1   32.0   36.2   11.8
    2  804   D        -     0   0   66   78,-0.2  78,-0.1   1,-0.1   3,-0.1  -0.566 360.0-123.2 -77.6 135.7   33.9   38.4   14.2
    3  805   A        -     0   0   30   -2,-0.2  22,-0.3   1,-0.1  69,-0.1  -0.374  39.8 -85.2 -75.6 161.6   32.9   42.2   14.1
    4  806   P        -     0   0    7    0, 0.0   3,-0.2   0, 0.0   2,-0.2  -0.439  53.3-139.8 -65.3 146.3   35.7   44.8   13.4
    5  807   S  E     +A   22   0A  72   17,-1.6  17,-2.9   1,-0.2  78,-0.0  -0.649  66.7  22.5-110.5 167.8   37.4   45.8   16.6
    6  808   Q  E     S-    0   0  103   -2,-0.2   2,-0.4  15,-0.2  -1,-0.2   0.920  72.5-173.4  43.5  59.3   38.7   49.0   18.1
    7  809   I  E     +     0   0   13   -3,-0.2   2,-0.4  14,-0.2  14,-0.2  -0.685   6.7 176.0 -83.3 128.4   36.4   51.2   16.1
    8  810   E  E     -A   20   0A  83   12,-2.0  12,-2.6   -2,-0.4   2,-0.5  -0.993  26.5-149.0-136.7 141.3   37.3   54.9   16.7
    9  811   V  E     +A   19   0A  32   -2,-0.4   2,-0.3  10,-0.2  10,-0.2  -0.946  35.5 162.9-105.5 126.3   35.9   58.1   15.2
   10  812   K  E     +A   18   0A  90    8,-2.7   8,-2.6   -2,-0.5  -2,-0.0  -0.879  42.7  39.5-138.8 174.0   38.7   60.7   15.1
   11  813   D  S     S+    0   0  119   -2,-0.3   2,-0.3   6,-0.2   -1,-0.1   0.868  71.4 178.9  53.1  45.7   39.5   64.0   13.4
   12  814   V        +     0   0   60    5,-0.1   2,-0.2   -3,-0.1   5,-0.2  -0.588   6.9 171.7 -83.4 131.8   35.9   65.3   13.6
```

Figure 1.9.2. Début du fichier de sortie de DSSP pour la protéine de code 1ten.

Il s'agit de récupérer dans les fichiers DSSP, exécutés sur une machine locale, la structuration secondaire de chaque résidu afin d'établir le fichier correspondant à l'alignement multiple.

A chaque domaine, le code de la structuration secondaire est reformaté séquentiellement, c'est-à-dire que le formatage en colonne est transformé en ligne pour chaque domaine en correspondance avec la séquence en résidu. Ensuite un alignement contenant l'information de la structuration secondaire est généré en correspondance avec l'alignement multiple structural donné par MUSTANG. Ainsi les gaps sont introduits fidèlement à l'alignement produit dans les séquences assignées en structures secondaires. DSSP fournit 7 classes de structures secondaires :

H : hélice alpha

G : hélice 3/10

I : hélice Pi

E : brin bêta

S : « bend »

B : résidu isolé dans un pont bêta

T : tour de liaison hydrogène

L'algorithme de recherche de conservation des structures secondaires traite ces données de la sorte : H, G et I seront traités comme hélice, E et S comme brins bêta si un S est contigu à un E, sinon S sera considéré avec B et T comme de la boucle ou « coil ».

1.9.5. Données de MIR

Les informations correspondant aux MIR sont récupérées sur le serveur RPBS (Alland et al., 2005). Les fichiers MIR contiennent les informations de séquences, du nombre de voisins non covalents (Non Covalent Neighbour ou NCN) associés et de la présence ou non d'un MIR en une position, notés respectivement par un tiret et un « M ». Les gaps sont introduits conformément à l'alignement produit.

1.9.6. Données de TEF

Les TEF ont été calculés par un script domestique codé par Mathieu Lonquety strictement identique à celui disponible sur le serveur RPBS. Le programme produit un fichier avec toutes les entrées, leurs séquences, et l'appartenance à un TEF ou non de chaque résidu de chaque séquence. Vu qu'il arrive que des TEF se chevauchent, les résultats sont présentés sur deux lignes.

Comme pour les MIR et les structures secondaires, l'appartenance ou non à un TEF est introduite conformément à l'alignement multiple. Ainsi on obtient deux lignes par structure contenant les informations d'alignement et d'appartenance aux TEF.

1.9.7. Données de HTOO

Les HTOO sont définis sur des alignements selon la numérotation de Kabbat (Deret et al., 1999) des domaines VH et VL (variables lourd et léger). Il y a un set d'HTOO pour chaque type de domaine VH et VL. Il faut ensuite faire la concordance avec les séquences et définir quel résidu est un HTOO ou non.

Les HTOO sont ensuite répartis tout au long de l'alignement en concordance avec leurs positions dans les séquences.

1.10. Analyse des corrélations

Les corrélations entre différentes caractéristiques (MIR, HTOO…) sont calculées à partir des fichiers correspondants. L'alignement est représenté comme suit : verticalement les occurrences, et horizontalement les positions dans l'alignement. On parcourt l'alignement dans le sens horizontal et à chaque pas du parcours l'alignement est parcouru verticalement afin de compter les occurrences. Cette méthode sera utilisée pour définir si une position est alignée, si une position est de structuration secondaire conservée et pour compter les occurrences en MIR, TEF et HTOO à chaque position de l'alignement. Voici un exemple de la représentation des données pour leur analyse :

Code du domaine ***---**---X----*-*---**---*----X--***

Où les positions en astérisques (*) sont les gaps introduits par l'alignement structural, et où les tirets (-) sont des résidus ne portant pas d'information d'appartenance à une position de MIR, TEF ou HTOO, contrairement aux positions représentées sur le schéma par un X.

1.10.1. Détermination des positions du cœur et de structure secondaire conservée

Les positions du cœur sont déterminées à partir de l'alignement en séquence correspondant à l'alignement structural. Chaque position de l'alignement est parcourue le long des 56 structures. Une position est attribuée à du cœur quand le taux d'occupation de celle-ci dans l'alignement est supérieur à 80%. Pourquoi 80% ? C'est sur ces positions que seront comptabilisés les MIR et les extrémités de TEF et il est donc important de choisir une valeur élevée de similitude.

Chaque position de l'alignement est parcourue le long des 56 structures, et celles composées à plus de 60% de structure secondaire identique sont considérées comme positions de structure secondaire conservée. Il est reconnu que les programmes d'attribution de structures secondaires ne sont pas encore assez robustes, et l'on observe des divergences dans les résultats de différents programmes. C'est pour tenir compte de cette imprécision dans les limites des brins et hélices que le seuil d'assignation a été descendu de 80% à 60%.

Sur toutes les positions de cœur, sont comptabilisés le nombre d'extrémités de TEF et les MIR. Cette comptabilisation est effectuée en considérant strictement chaque position, puis en l'encadrant par trois

fenêtres dont les tailles varient de ±1 à ±3 positions. C'est-à-dire qu'on regarde pour une position de cœur les positions voisines à plus ou moins un écart fixé.

1.10.2. TEF et cœur

La corrélation entre les extrémités de TEF et les MIR a été étudiée sur l'alignement structural. Il en résulte que les MIR correspondant à une extrémité de TEF représentent un pourcentage de 6,6 % des MIR pour une comparaison stricte en position (on n'accepte pas de fenêtre). Si l'on diminue la précision en prenant une fenêtre de ±5, on atteint un pourcentage de 57,6%. Le tableau 1.10 liste les correspondances aux différentes fenêtres de précision regardées.

Ecart de la fenêtre de précision	Pourcentage de MIR localisés aux extrémités de TEF(%)
0	6,56
±1	12,5
±2	26,8
±3	36,6
±4	49
±5	57,6

Tableau 1.10 Correspondances des différentes fenêtres de précision aux pourcentages de MIR localisés aux extrémités de TEF.

1.11. Présentation des résultats

La corrélation des extrémités de TEF avec les positions de cœur et des structures secondaires conservées est représentée par des courbes dont l'axe des abscisses représente la position dans l'alignement et celui des ordonnées les occurrences des extrémités de TEF. Les occurrences en extrémités de TEF sont comptabilisées sur les seules positions de cœur, la courbe de structuration secondaire est arbitrairement assignée à une valeur fixe quand la position est conservée en structuration secondaire et à zéro lorsqu'elle ne l'est pas. En effet, l'information binaire d'appartenance à une position de structuration secondaire conservée doit apparaître comme une courbe dont on fixe arbitrairement l'ordonnée pour aider à la lecture des graphiques. Ainsi, lorsque la courbe a une valleur nulle la position n'est pas conservée en SSR, quand la courbe montre la valeur arbitrairement fixée, la position est de SSR conservée. Pour des raisons de lisibilité, les courbes ont été lissées. Pour ce faire, nous avons sommé toutes les trois positions dans l'alignement. C'est pourquoi l'alignement apparait sur les courbes d'une longueur divisée par un facteur 3, comme on peut le voir dans la figure 1.12.4.

La corrélation des MIR avec les positions du cœur et les positions de structuration secondaire conservées est effectuée avec la même méthode que les TEF.

1.12. Résultats

1.12.1. Banque

Ainsi 52 domaines de la banque d'Halaby et Mornon et 25 domaines de Gerstein et Altman ont été analysés. 21 montraient une identité de séquence supérieure à 70% avec une autre et ont donc été retirées de la banque. Le tableau suivant récapitule les différents domaines utilisés.

Protéine	**Espèce**	**Code PDB**	**Résolution**	**Chaîne**	**Domaine**
Actinoxantin	Actinomyces Globisporus	1ACX	2.0	_	_
Endoglucanase	Clostridium Thermocellum	1CLC	1.9	_	01(35-136)
Neocarzinostatin	Streptomices Carzinostaticus	1NCO	1.8	A	_
Immunoglobin FAB	Homo sapiens	2FB4	1.9	H	01(1-118) 02(119-218)
				L	02(110-210)
Human Growth Hormone	Homo sapiens	3HHR	2.8	B	02(132-234)
T cell Antigen Receptor	Mus Musculus	1BEC	1.7	_	01(3-117) 02(118-246)
Cytochrome F	Brassica Rapa	1CTM	2.3	_	01(1-170
Synaptotagmin I	Escherichia Coli	1RSY	1.9	_	_
Chitinase A	Escherichia Coli	1CTN	2.3	_	01(24-130)
Human Class I Histocompatibility antigen	Homo sapiens	1HLA	3.5	A	02(183-270)
				M	_
Transcription factor NFKB	Homo sapiens	1SVC	2.6	P	01(43-244)
				P	02(249-353)
Immunoglobin FAB	Homo sapiens	7FAB	2.0	H	01(1-117)
				L	01(1-103)

Protéine	**Espèce**	**Code PDB**	**Résolution**	**Chaîne**	**Domaine**
Beta-Galactosidase	Escherichia Coli	1BGL	2.5	A	02(219-333)
Cyclodextrin Glucanotransferase		1CYG	2.5	_	02(397-491)
T Cell Receptor	Homo sapiens	1CD8	2.6	_	_
FC Fragment	Homo sapiens	1FC1	2.9	A	01(238-337) 02(338-443)
CD2	Homo sapiens	1HNF	2.5	_	01(4-103) 02(104-182)
Tenascin	Escherichia Coli	1TEN	1.8	_	_
Drosophilia Neuroglian	Drosophilia Melanogaster	1CFB	2.0	_	01(610-708) 02(709-814)
Telokin	Meleagris Gallopavo	1TLK	2.8	_	_
Macromomycin	Streptomyces Macromomycet icus	2MCM	1.5	_	_
CD2	Rattus Rattus	1HNG	2.8	A	01(2-98) 02(99-176)
Cell adhesion Protein	Homo sapiens	1VCA	1.8	A	01(1-89) 02(90-199)
Superoxide Dismutase	Bos Taurus	2SODO	2.0	O	_
CD4	Rattus Rattus	1CID	2.8	_	01(1-106) 02(107-177)
Coagulation Factor	Homo sapiens	1GGT	2.65	A	03(516-630) 04(631-729)
CD4	Homo sapiens	3CD4	2.2	_	01(1-98) 02(99-173)
N-Cadherin	Mus Musculus	1NCI	1.9	A	_
Immunoglobin FAB	Mus Musculus	3HFM	3.0	H	01(1-113)
				L	01(1-108)
FAB Fragment	Mus Musculus	4FAB	2.7	H	01(1-118)
				L	01(1-113)
FAB Fragment (Galctin binding)	Mus Musculus	2FBJ	1.95	H	01(1-118)
				L	01(1-107)
Antigene binding fragment	Mus Musculus	6FAB	1.9	H	01(1-121)
				L	01(1-107)

Protéine	Espèce	Code PDB	Résolution	Chaîne	Domaine
FAB Fragment (Anti-lyzozyme antibody)	Mus Musculus	1FDL	2.5	H	01(1-116)
				L	01(1-108)
FAB Fragment	Mus Musculus	1IGF	2.8	H	01(1-113)
FAB Fragment	Mus Musculus	1MCP	2.7	H	01(1-123)
N9 Neuraminidase-FAB complex	Anous Minutus	1NCA	2.5	H	01(1-114)
				L	01(1-108)

1.12.2. Identité

La distribution des différents pourcentages d'identité observés montre que pour les pourcentages allant de 30 à 70%, il n'y a qu'au maximum trois couples affichant ce pourcentage.

La figure 1.12.1. représente le TI après alignement structural :

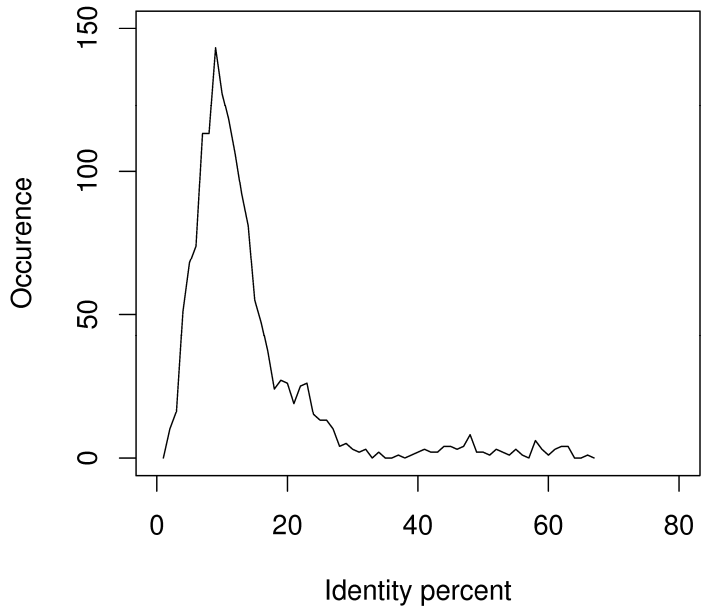


Figure 1.12.1. Distribution de l'identité de séquences entre paires de la banque de domaines immunoglobulines, après superposition des structures.

La valeur moyenne du taux d'identité entre toutes les paires après alignement structural est de 12,6%. Le Taux d'identité comprenant le maximum de couples est de 8%, ainsi 951 paires de séquences sur 1540 sont au-dessus de cette valeur de TI.

1.12.3. Alignement

L'alignement des domaines a été réalisé par le programme MUSTANG. Il présente des blocs de 4 à 10 résidus successifs (ce chiffre correspond aux longueurs habituellement décrites pour les brins β) avec une forte densité de résidus alignés, mais aussi des régions de grandes insertions. On peut penser que, comme le seuil de structuration secondaire (60%) est inférieur à celui définissant le cœur (80%) en terme de taux d'occupation d'une position de l'alignement, il arriverait qu'on ait des structures secondaires en dehors du coeur. Ce point sera discuté plus loin. La figure 1.12.2. représente l'alignement en abscisse ; la courbe noire représente les positions du cœur (et est fixée à 2) et la courbe bleue les positions de structures secondaires conservées (fixée à 4).

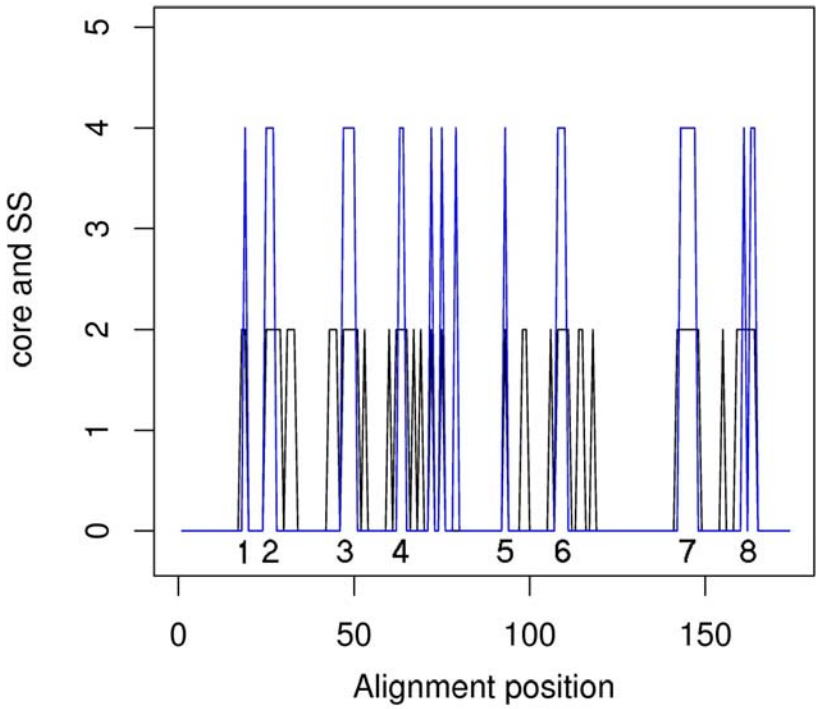


Figure 1.12.2. Correspondance entre cœur et structures secondaires conservées dans l'alignement de la famille de domaines immunoglobulines.

Les régions conservées en structures secondaires (la courbe bleue) montrent les différents brins du modèle canonique des domaines immunoglobulines, marqués par des nombre (le brin numéro 8 étant divisé en deux pics), et les régions des boucles hyper variables conservées (les trois pics entre les brins numéro 4 et 5). Les deux premiers pics correspondent au premier brin, ce qui provient d'une insertion originale pour la structure 1clc01 dans un brin alors que toutes les autres structures montrent un premier brin sans rupture. Il se trouve que cette insertion est en accord avec la structure de la titine (structure ne faisant pas partie de la banque alignée mais utilisée pour la corrélation avec les valeurs de Phi) qui montre aussi un premier brin divisé en deux. Les pics plus larges correspondent aux brins suivants. Entre les pics 4 et 5 (entre les positions 70 et 90) on observe trois petits pics qui correspondent aux boucles hypervariables en séquence mais conservées structuralement et assignées en brin par DSSP. Ces régions selon la littérature seraient responsables du couplage antigène/anticorps (Chothia et al., 1989).

Le modèle canonique comprend huit brins en sandwich de feuillet bêta (Chothia et al., 1998), or l'alignement n'en montre que sept. Après analyse des structures, on s'est aperçu que les huit brins ne sont jamais représentés. Les deux brins, C' et D selon la notation du modèle canonique, à la charnière des deux feuillets du modèle dont chaque structure n'ont que l'un ou l'autre se trouvent être à la charnière du sandwich de feuillets. C'est pourquoi MUSTANG les a alignés ensemble et que le brin D n'apparaît pas (voir figure 1.12.13.).

Une figure fondée sur les coordonnées spatiales des domaines alignés a été produite. On a coloré de couleur différente chaque brin correspondant au modèle canonique de chaque structure. Les régions représentées sur la figure 1.12.2 sont un peu élargies par rapport aux attributions en structure secondaire pour plus de lisibilité.

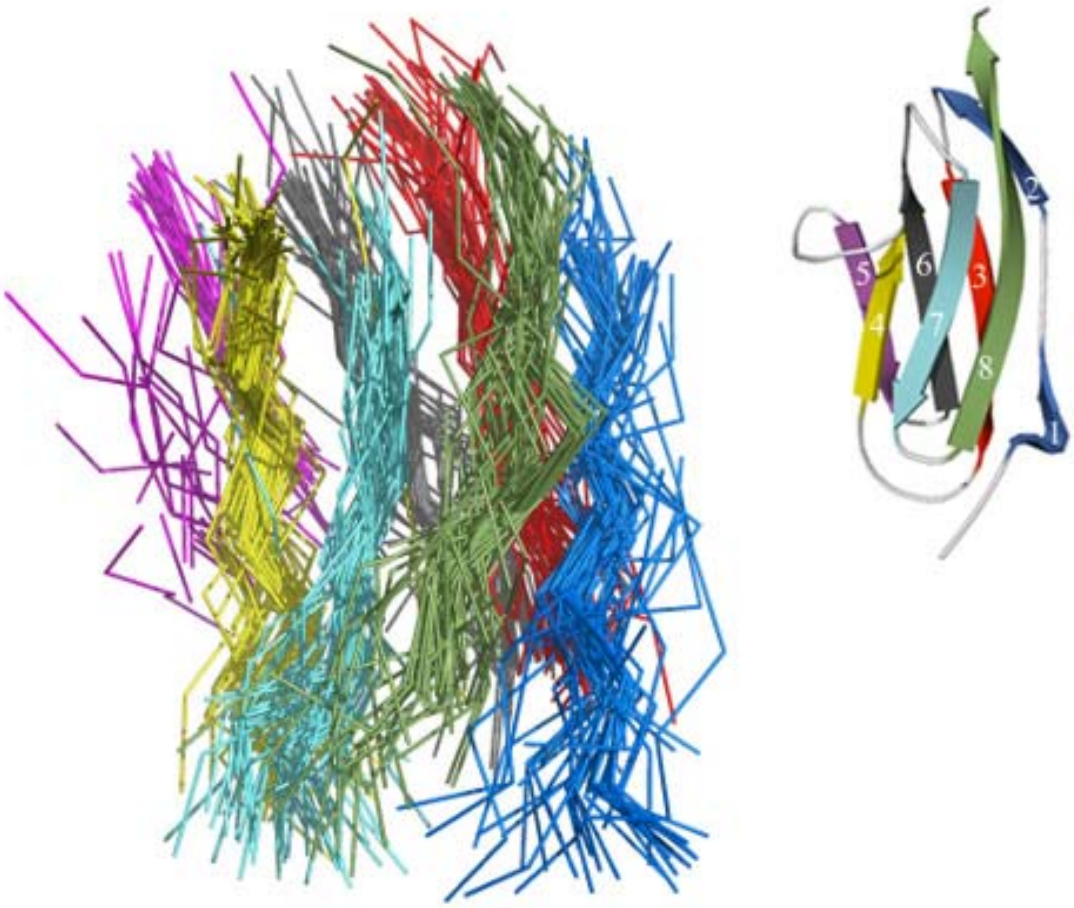


Figure 1.12.3. Conservation structurale des 56 domaines immunos avec une couleur par brin. L'encart permet de voir la notation des brins dans le cas de la titine.

On observe aussi pour l'une des structures (de code PDB 1cyg02) une variante dans la topologie, où deux brins sont inversés. On peut apercevoir un brin jaune au milieu des brins bleus clairs (numéro 7 sur l'encart) et vice versa.

1.12.4. Le cœur protéique

Le cœur parait étendu (68%) car le repliement immunoglobuline est très conservé et compact. Le calcul de taux moyen par domaine de résidus correspondant au cœur (soit les positions ayant un taux d'occupation dans l'alignement supérieur à 80%) a été effectué. Il en résulte un nombre moyen de 74 résidus appartenant au cœur protéique. Ceci est en accord avec les études de Chothia et al. (Chothia et al., 1998) et semble provenir d'une particularité du repliement immunoglobine.

1.12.5. Corrélation cœur/extrémité de TEF/structuration secondaire

Le nombre total de résidus appartenant à un TEF est de 4425, parmi lesquels 3067, c'est-à-dire 68%, sont sur des positions de cœur. En ce qui concerne la corrélation cœur et structuration secondaire conservée, il convient de retenir que 50% des positions de cœur sont corrélées à des structures secondaires conservées.

Pour une corrélation stricte, donc sans latitude autour des positions analysées, sur les 193 TEF, 156 débuts de TEF (soit 81% du total des TEF), et 152 fins de TEF (soit 78,8%) sont sur des positions de cœur. Sont situés sur des positions de structures secondaires conservées 144 débuts de TEF, ce qui correspond à 74,6% de tous les débuts de TEF et 136 fins de TEF, soit 70,5% de l'ensemble des fins de TEF. La figure 1.12.4. représente les résultats de corrélation d'extrémités de TEF avec les positions de cœur. La courbe rouge correspond aux départs de TEF et la courbe verte aux fins de TEF. La courbe bleue correspond aux positions de structures secondaires conservées et elle est bornée à 10 en ordonnée, pour améliorer la lisibilité du graphique.

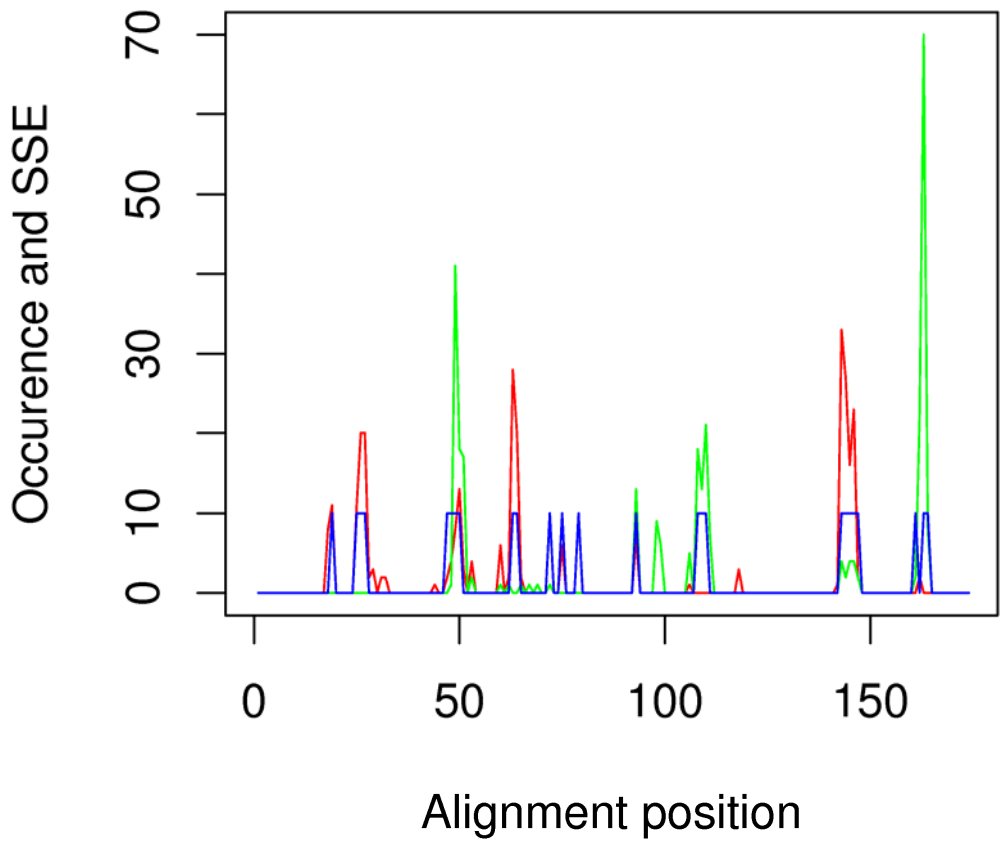


Figure 1.12.4. Occurrences des limites de TEF (débuts en rouge et fins en vert) comparées aux structures secondaires conservées parmi les 56 domaines immunoglobulines. Correspondance stricte entre limites de TEF et structures secondaires.

En acceptant une précision de ±1 autour des positions considérées comme du cœur, on trouve dans cet intervalle 374 TEF pour lesquels 294 départs (78,6% de tous les TEF) et 295 fins de TEF (78,9%) sont sur des positions de cœur (figure 1.12.5). En prenant la même fenêtre de ±1, les positions d'extrémités de TEF corrélées à des structures secondaires conservées sont de 275 pour les débuts de TEF (soit 73,5% du nombre des débuts) et 270 pour les fins de TEF (soit 72,2% du nombre des fins). Il ressort que les limites de TEF sont plus fortement corrélées avec les positions de cœur, pour lesquelles les contraintes de conservation sont les plus fortes, que pour les structures secondaires conservées. Etant donné que nous avons pris deux seuils différents pour définir le cœur et les structures secondaires conservées, si on applique deux seuils différents, 60% et 80%, à la conservation de la même propriété,

la corrélation avec les TEF devrait être plus importante pour le seuil le plus bas. Comme ce n'est pas le cas, cela ne peut être attribué qu'au fait que l'on considère deux grandeurs différentes : le cœur et les structures secondaires. On peut donc en déduire que le cœur est un concept pertinent, qui n'est pas limité aux seules structures secondaires conservées, bien qu'elles en soient la composante principale.

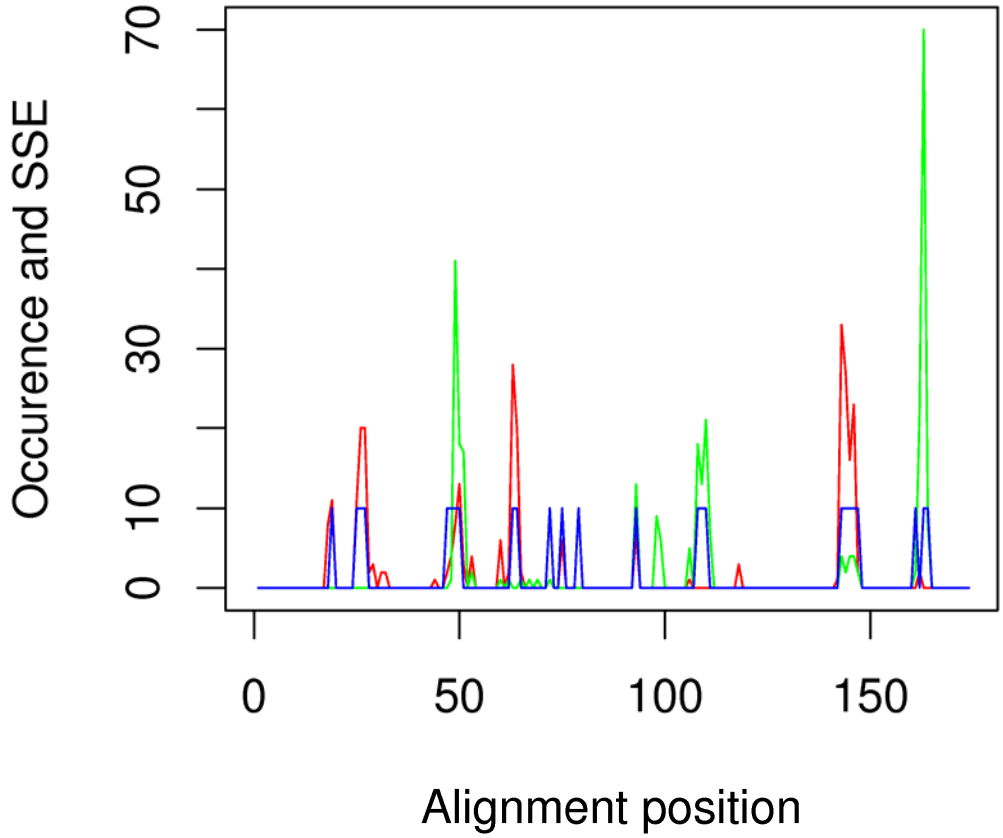


Figure 1.12.5. Occurrences des limites de TEF (débuts en rouge et fins en vert) comparées aux structures secondaires conservées parmi les 56 domaines immunoglobulines. Correspondance entre limites et structures secondaires prise en compte dans une fenêtre de ±1 position.

En analysant la corrélation des limites de TEF avec les positions de coeur sur une fenêtre de ±2 (figure 1.12.6), on trouve 570 TEF produisant 424 départs (74.4% du nombre des débuts) et 456 fins (80% du nombre de fins). Pour la même fenêtre, les positions d'extrémités de TEF corrélées à des structures secondaires conservées sont de 391 pour les débuts de TEF (soit 68,6% des débuts) et 418 pour les fins de TEF (soit 73,3% des fins).

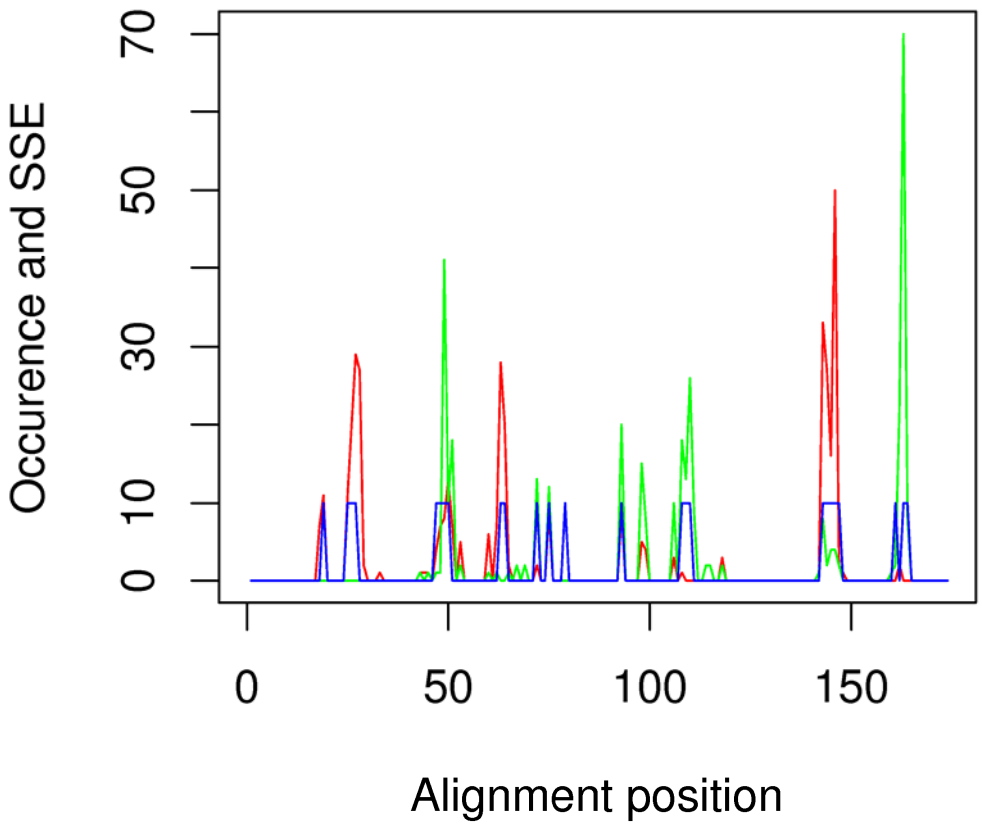


Figure 1.12.6. Occurrences des limites de TEF (débuts en rouge et fins en vert) comparées aux structures secondaires conservées parmi les 56 domaines immunoglobulines. Correspondance entre limites et structures secondaires prise en compte dans une fenêtre de ±2 positions.

Si on considère une fenêtre de ±3 autour des positions de cœur (figure 1.12.7), on comptabilise 640 TEF, avec 460 départs de TEF (soit 71,8% des débuts) et 546 fins de TEF (soit 85,3% des fins). Avec la même précision, les positions d'extrémités de TEF corrélés à des structures secondaires conservées sont de 474 pour les débuts (soit 74,1% des débuts) et 490 pour les fins de TEF (soit 76,6% des fins).

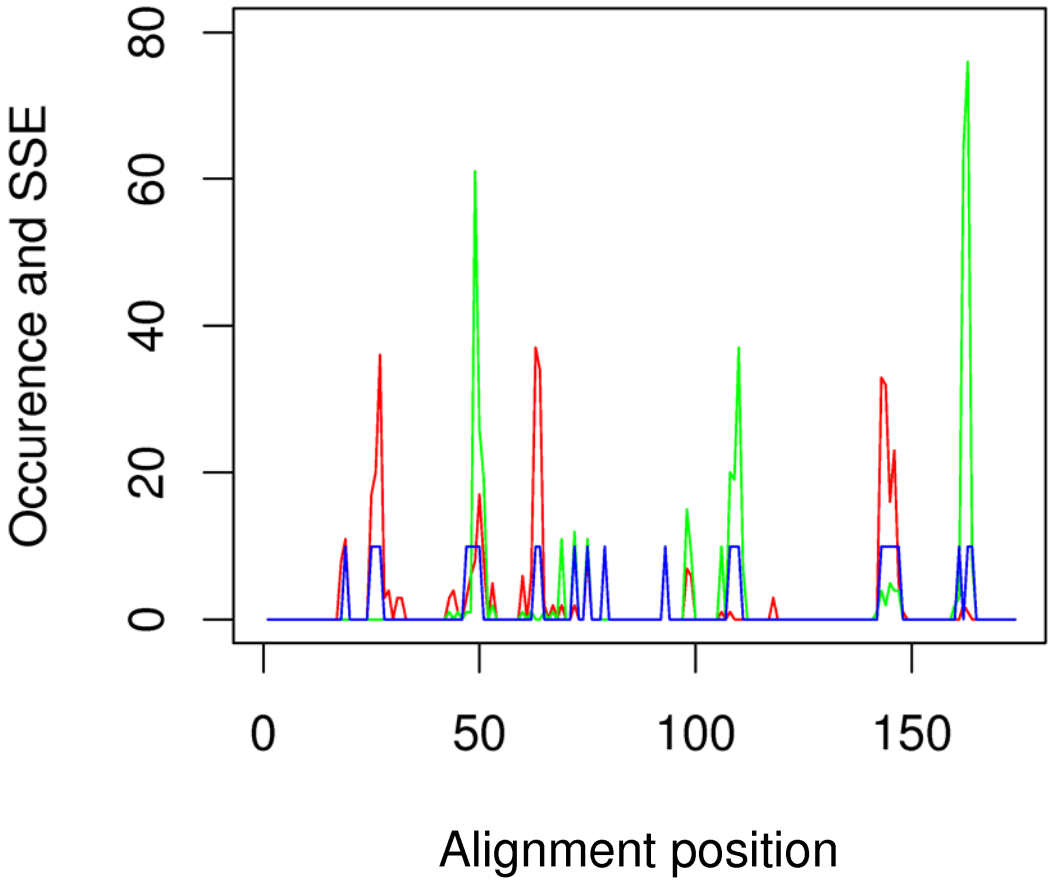


Figure 1.12.7. Occurrences des limites de TEF (débuts en rouge et fins en vert) comparées aux structures secondaires conservées parmi les 56 domaines immunoglobulines. Correspondance entre limites et structures secondaires prise en compte dans une fenêtre de ±3 positions.

Plus on augmente la tolérance pour définir une position, plus celle-ci sera comptabilisée plusieurs fois lorsque l'on parcourt l'alignement multiple. Ceci n'est pas la seule raison de l'augmentation des occurrences d'extrémités de TEF. En effet plus la fenêtre augmente plus des positions d'extrémités sont comptabilisées comme appartenant au cœur. Cependant bien que ce nombre croisse avec la fenêtre, les pourcentages de corrélation avec les structures secondaires n'ont pas le même comportement et on observe le pourcentage maximum lorsque la fenêtre de tolérance est fixée à 0.

Toutes ces données sont rassemblées dans le tableau récapitulatif 1.12.5 :

Fenêtre	TEF total	débuts TEF & cœur (%)	fins TEF & cœur (%)	début TEF & SSR (%)	fin TEF & SSR (%)
0	193	156 (81)	152 (78,8)	144 (74,6)	136 (70,5)
±1	374	294 (78,6)	295 (78,9)	275 (73,5)	270 (72,2)
±2	570	424 (74,4)	456 (80)	391 (68,8)	418 (73,3)
±3	640	460 (71,8)	546 (85,3)	474 (74,1)	490 (76,6)

Tableau 1.12.5 Récapitulatif des corrélations de TEF, cœur et SSR.

On observe une grande corrélation entre les extrémités de TEF et les structures secondaires. Ainsi, quand on filtre l'alignement à un taux d'occupation de 80% des positions, les extrémités des TEF se retrouvent principalement dans des régions de structures secondaires conservées, ce qui représente la principale contribution au cœur protéique, comme nous l'avons déjà indiqué. Pour mieux percevoir la corrélation entre les extrémités de TEF et les structures secondaires, le tableau 1.12.6 présente la distribution des limites de TEF dans les parties de structures secondaires conservées pour une correspondance stricte (une fenêtre de un, donc sans aucune latitude).

N° Brin	Longueur alignée (> 60%)	Nombre de débuts de TEF	Nombre de fins de TEF
1	1	11	0
2	3	20	0
3	6	14	21
4	6	25	0
5	3	3	5
6	6	0	30
7	8	21	7
8	5	0	38
Total	38	94	101

Tableau 1.12.6 Distribution des limites de TEF dans les parties de structures secondaires conservées.

Le fait que le cœur ne soit pas restreint aux parties de structures secondaires conservées est une indication que l'inclusion de résidus dans une structure secondaire régulière conservée n'est pas cruciale pour déterminer le cœur. Néanmoins la forte correspondance entre les limites de TEF et les régions de structures secondaires conservées indique que les fragments protéiques tels que les TEF commencent et finissent très majoritairement au cœur des domaines protéiques globulaires, bien qu'ils ne soient pas strictement assortis aux cœurs. Comme nous l'avons montré plus tôt, les limites de TEF peuvent être prédites par les MIR, ainsi nous allons étudier la correspondance entre les MIR, le cœur protéique et les structures secondaires conservées.

1.12.6. Corrélation cœur/MIR/structuration secondaire

Pour une correspondance stricte (fenêtre de 1, donc sans latitude) dans l'analyse de corrélation, 666 MIR sont corrélés aux positions de cœur parmi lesquelles 580 MIR sont sur des positions strictes de structuration secondaire conservée (soit 87.1% du total). Ceci est représenté sur la figure 1.12.8.

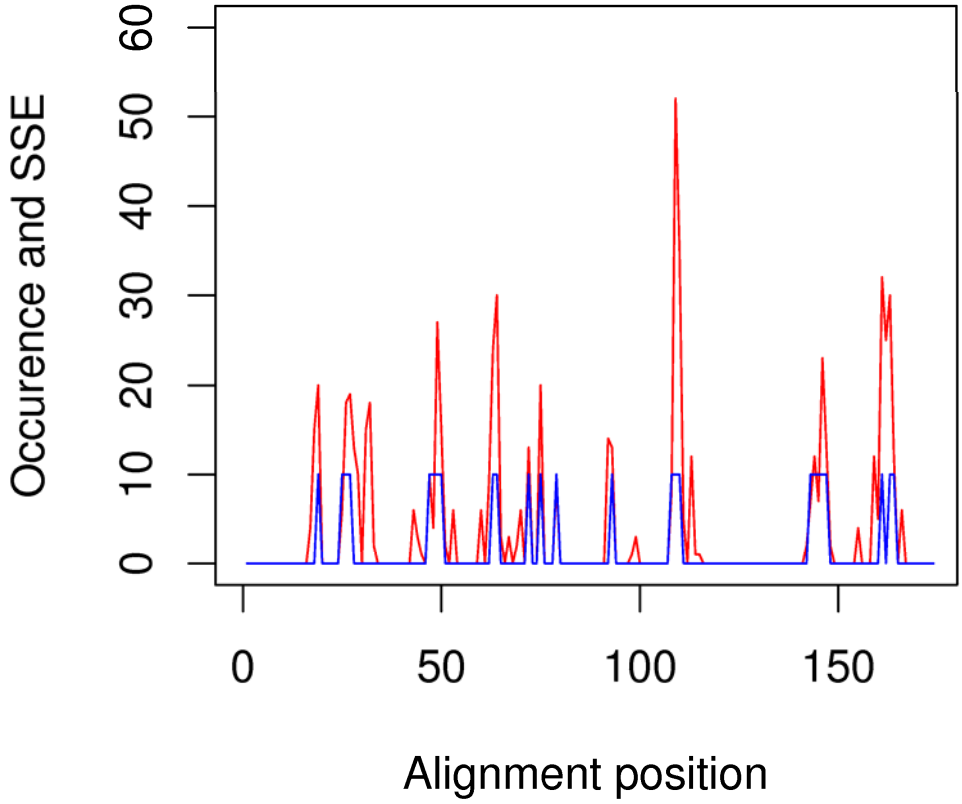


Figure 1.12.8. Occurrences des MIR (en rouge) par rapport aux brins conservés (en bleu) pour les 56 domaines immunoglobulines. La correspondance entre MIR et structures secondaires est prise en compte strictement.

Pour une précision de ± 1 dans l'analyse de corrélation, 4282 MIR sont corrélés aux positions de cœur, parmi lesquels 3267 MIR sont sur des positions de structuration secondaires conservées, soit 76,3%. Ceci est représenté sur la figure 1.12.9.

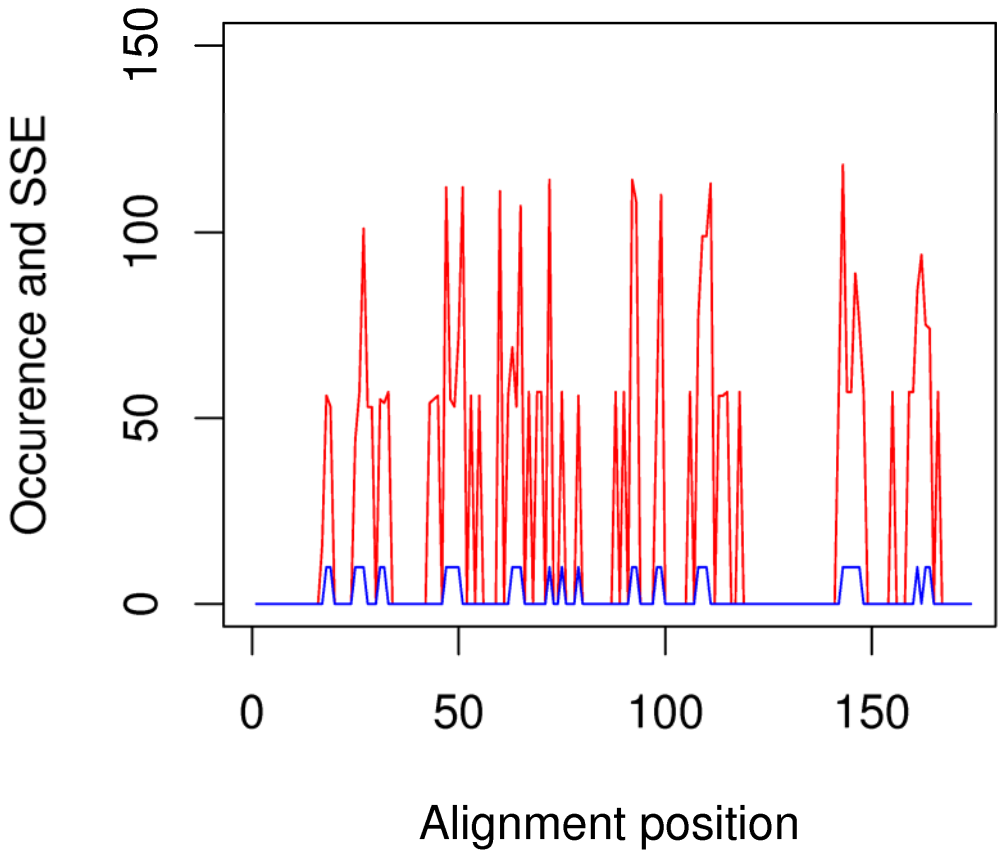


Figure 1.12.9. Occurrences des MIR (en rouge) par rapport aux brins conservés (en bleu) pour les 56 domaines immunoglobulines. La correspondance entre MIR et structures secondaires est prise en compte dans une fenêtre de ±1 position.

Pour une fenêtre de ±2 dans l'analyse de corrélation, 4671 MIR sont corrélés aux positions de cœur et 3767 MIR sont sur des positions de structuration secondaire conservée, soit 80,1%. Ceci est représenté sur la figure 1.12.10.

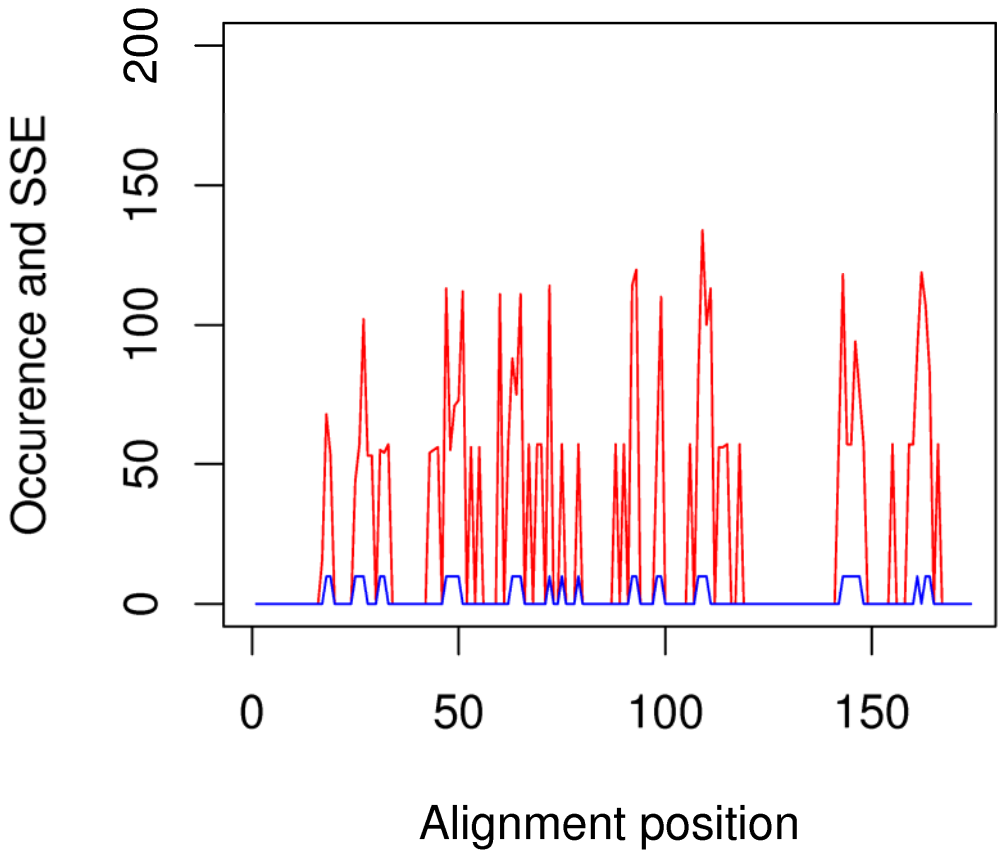


Figure 1.12.10. Occurrences des MIR (en rouge) par rapport aux brins conservés (en bleu) pour les 56 domaines immunoglobulines. La correspondance entre MIR et structures secondaires est prise en compte dans une fenêtre de ±2 positions.

Pour une précision de ±3 dans l'analyse de corrélation, 6244 MIR sont corrélés aux positions de cœur et 4471 sont sur des positions de structuration secondaire conservée, soit 71,6%. Ceci est représenté sur la figure 1.12.11.

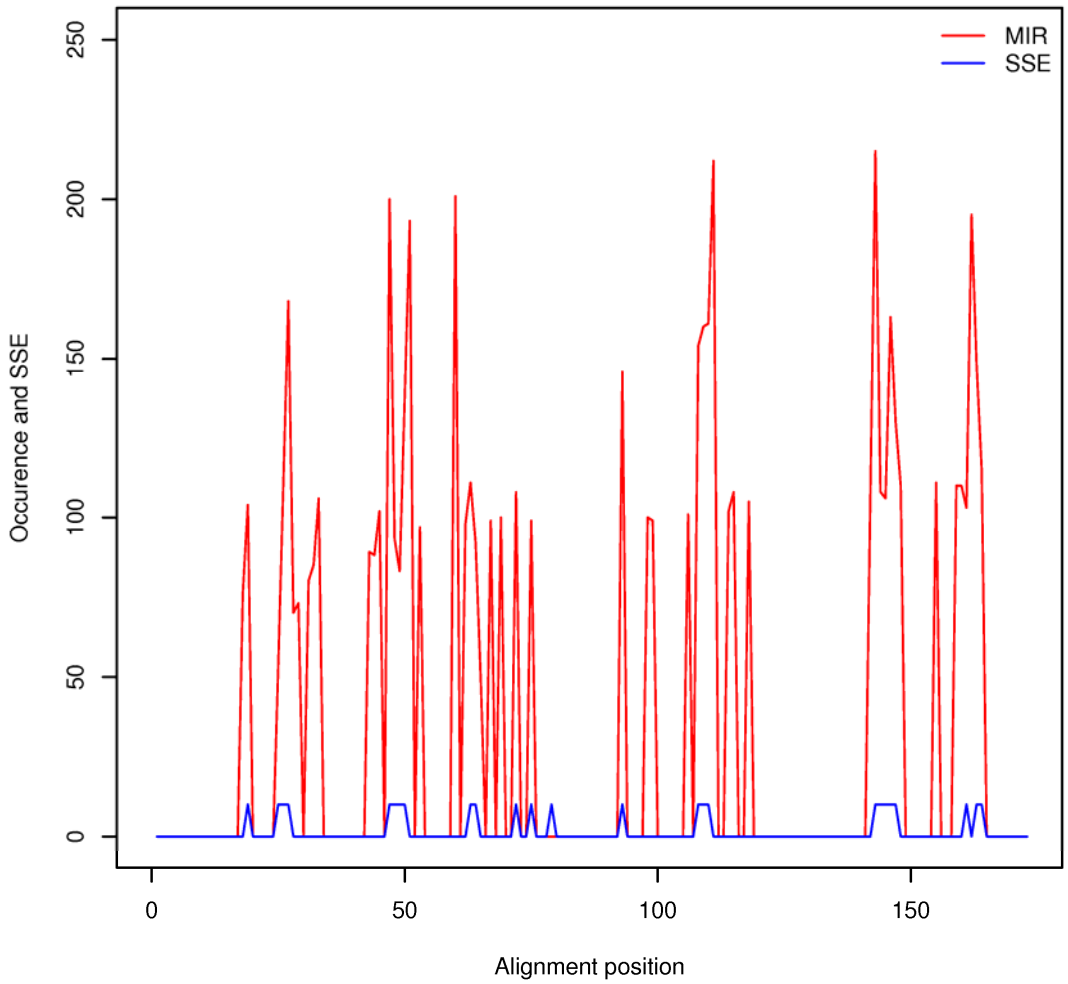


Figure 1.12.11. Occurrences des MIR (en rouge) par rapport aux brins conservés (en bleu) pour les 56 domaines immunoglobulines. La correspondance entre MIR et structures secondaires est prise en compte dans une fenêtre de ±3 positions.

Le tableau 1.12.7 récapitule les résultats concernant la corrélation MIR et cœur.

Fenêtre	MIR corrélés au cœur	MIR corrélés aux SSR conservées
0	666	580 (87,1%)
±1	4282	3267 (76,3%)
±2	4671	3767 (80,1%)
±3	6244	4471 (71,6%)

Tableau 1.12.7 Récapitulatif des corrélations de MIR, cœur et SSR.

1.12.7. Accord TEF/MIR

Le nombre total de MIR est de 899 dans notre banque de 56 domaines immunoglobulines.

Pour une localisation stricte, on trouve 32 MIR à des positions de départs de TEF et 27 à des positions de fins de TEF. Soit 59 MIR sont à des positions d'extrémités de TEF. A cet écart on peut dire que 6,56% des MIR sont corrélés de manière stricte aux extrémités de TEF. Cette valeur, très faible, nécessite d'augmenter la fenêtre sur laquelle nous allons chercher la correspondance entre MIR et extémités de TEF. Ceci prend en compte d'une part le fait que la simulation des MIR ne prétend pas à une précision à un résidu près, et que l'algorithme d'attribution des TEF a également des limites un peu floues dans la détermination des extrémités. Ceci est vrai car l'algorithme d'attribution des TEF disponible sur le serveur RPBS (http://bioserv.rpbs.jussieu.fr/) est légèrement différent de celui mis en ligne par Igor Berezovsky (Domain Hierarchy and Closed Loops) à Bergen en Norvège (http://ssitron.bccs.uib.no/dhcl) et les résultats peuvent varier à la marge entre ces deux serveurs.

Le tableau 1.12.8 ci-dessous reprend la correspondance entre les MIR et les différentes extrémités des TEF, selon la précision que l'on tolère sur la position des MIR (ou des limites de TEF).

Fenêtre	MIR = départs de TEF	MIR= fins de TEF	MIR à une limite de TEF	MIR à une limite de TEF (%)
0	32	27	59	6,6
1	58	54	112	12,5
2	109	132	241	26,8
3	149	180	329	36,6
4	196	244	440	48,9
5	227	291	518	57,6

Tableau 1.12.8 Récapitulatif des correspondances MIR et extrémités de TEF.

La corrélation entre extrémités de TEF et MIR n'est pas très encourageante, puisque pour une fenêtre de ±5, on a un peu plus de la moitié des MIR qui se trouvent aux limites d'un TEF. Ceci indique que notre algorithme de prédiction surévalue d'environ un facteur deux la prédiction des résidus les plus au cœur. C'est pourquoi l'information de structuration secondaire est nécessaire pour compléter l'analyse du noyau. On peut aussi dans l'avenir envisager d'ajouter d'autres critères, tels que la prédiction de l'accessibilité au solvant, la prédiction de la stabilité aux mutations…

1.12.8. Corrélation MIR/HTOO

Les résidus de grand ordre de rejet (High Throw Out Order ou HTOO) de Gerstein et Altman ont été analysés en fonction de l'écart en séquence qu'ils arboraient par rapport aux MIR et aux MIR lissés (SMIR pour Smoothed MIR). L'algorithme de Gerstein et Altman a tendance à produire des HTOO groupés. Or ceci ne correspond pas avec notre vision de résidus dispersés dans la séquence protéique qui provoquent la nucléation du repliement. Ainsi nous avons fait la comparaison des centres de gravité des groupes de HTOO avec les SMIR, qui est représentée sur la figure 1.12.12 pour l'ensemble des 25 domaines immunoglobulines de cette étude. On s'aperçoit avec satisfaction que le pic est centré sur l'origine, autrement dit que statistiquement les MIR lissés coïncident avec les centres de gravités des résidus HTOO. Notre outil prédictif est donc en accord qualitatif avec une détermination sur une base structurale des résidus les plus au cœur du globule protéique.

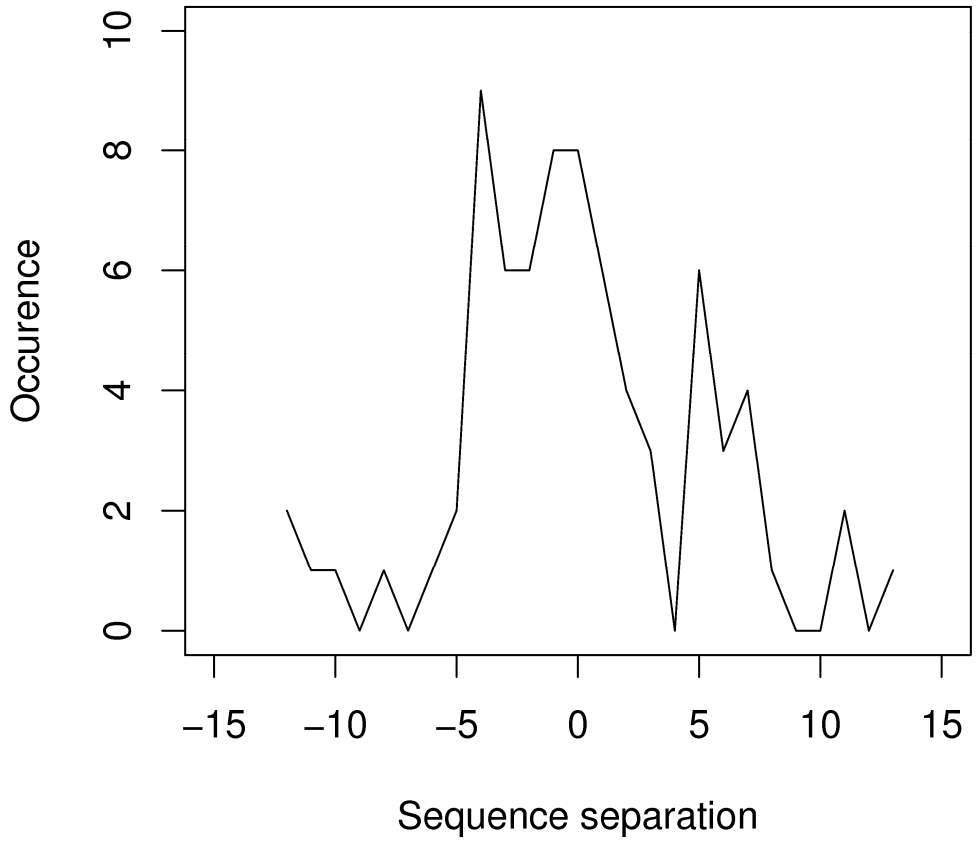


Figure 1.12.12. Séparation en séquence entre les centres de gravité des HTOO et les MIR lissés les plus proches sur les 25 domaines immunoglobulines étudiés par Gerstein et Altman.

1.12.9. Corrélation MIR / hautes valeurs de PHI

La détermination expérimentale du noyau protéique est une tâche difficile, rendue cependant envisageable par les valeurs de PHI. La valeur de PHI au-dessus de laquelle on peut déterminer de façon sure l'appartenance au noyau protéique est soumise à un débat controversé. « Il n'est pas possible de simplement prendre les résidus avec la plus haute valeur de PHI comme ceux qui vont former le noyau protéique : il faut aussi considérer l'entassement des résidus » selon Lappalainen et al. (Lappalainen et al., 2008). Le nombre d'interactions augmente avec les limites du noyau, donc avec sa taille, mais on peut imaginer que le cœur protéique est constitué de sphères concentriques avec un nombre décroissant d'interactions quand on s'éloigne du centre (Fowler & Clarke, 2001). L'équipe de Jane Clarke a étudié de façon approfondie le repliement de la super famille immunoglobuline et a déterminé les valeurs de PHI pour quatre protéines, dont les codes PDB sont les suivants : 1tit (Fowler & Clarke, 2001), 1ten (Hamill et al., 2000), 1k85 (Lappalainen et al., 2008) et 1ttf (Cota et al., 2001). Les séquences correspondantes ont été alignées avec celles de notre banque, et les localisations des valeurs de PHI ont été comparées aux positions de MIR. Un seuil de 0,3 pour les valeurs de PHI a été admis comme suffisant pour refléter une structuration significative dans l'état de transition, autrement dit pour considérer que le résidu appartient au noyau (Fowler et Clarke, 2001). La figure 1.12.13. représente les quatre protéines de repliement immunoglobuline pour lesquels les valeurs de PHI sont disponibles, alignées entre elles et avec les MIR lissés. Les étoiles indiquent les 10 valeurs de PHI les plus élevées.

```
1titA   -----------L----IEVEKPLYGVEVFVG--E-TAHFEIELSEPDVHGQWKLKGQPLTASPDCEIIEDGKKHILILHNC-QL----GMTGEV-SFQAANAKSAANLKVKEL
PHI     ------------------------------------*-*----------*---------------------*-*-----------------*--*----------------
SMIR    ----------------------S----------------S-----------S-----S-------S---------S------S----------------------------
1ten    R---LDAPSQIEVKDVTDTTALITWFKPLAE--IDGIELTYGIKDVPGDRTTIDLTEDENQYSIGNLKPDTEYEVSLISRR-GD----MSSNPAKETFTT-------------
PHI     ----------------------*----------------*-----------*----------*----------*-----------------------------------
SMIR    ----------S-------------S---------------S-----------S----------S------------S-----------S-------------------------
1ttfA   ---VSDVPRDLEVVAATPTSLLISWDAPAVT-VRYYRITYGETGGNSPVQEFTVPGSKSTATISGLKPGVDYTITVYAVT-GRGDSPASSKPI-SINYRT-------------
PHI     ----------*------------------------*-*--------------*------------------*----------------------*---------------
SMIR    -----------S----------S-------------S---------------S--------------------S-----------------------------------
1k85A   -HMAPTAPTNLASTAQTTSSITLSWTASTDNVGVTGYDVYNG------TALATTVT--GTTATISGLAADTSYTFTVKAKDAAG----NVSAASNAVSVKT------------
PHI     ----------------------*---------------*-*----------------------*----------*-*-----------------------------------
SMIR    ----------S----------S----------------S-----------S----------------S-----S-----------------S---------------------
SSt     ---------1-22------33333----------44444-------------5------66666-------7777777-------------888888----------------
SSA     --------AAAAAAA-----BBBBB--------CCCCCC-----------CCCCC----EEEEE-------FFFFFFFFF----------------GGGG-------------
```

Figure 1.12.13. Alignement des séquences de domaines immunos pour lesquels les valeurs de PHI ont été publiées. Les dix valeurs les plus importantes de PHI sont représentées par des étoiles. Les SMIR sont les MIR lissés.

Sur un domaine immuno de la titine cardiaque humaine, TNfn3 (code PDB : 1tit), Fowler et Clarke ont étudié vingt-neuf mutations. Les valeurs de PHI ont été produites à partir de cinétiques de dénaturation en présence de chlorure de guanidinium forçant une cinétique à deux états. La justification de l'emploi de cette méthode est que le mécanisme de repliement protéique à deux états est principalement la nucléation (Ding et al., 2002). Dix–neufs positions mutées donnent des valeurs de PHI entre 0,3 et 0,8 dont 11 sont plus grandes que 0,5. Quatre résidus au centre du cœur hydrophobe sont affectés au site de nucléation du repliement : I23 (Phi=0,82), W34 (une valeur de PHI

de 0,82 a été prédite depuis toutes les structures, mais il n'a pas été muté pendant cette expérience), L58 (Phi=0,79) et F73 (Phi=0,72).

Le domaine fibronectine de type III de la tenascine humaine (code PDB : 1ten) a été aussi étudié et produit les résidus suivants comme inclus dans le noyau du repliement : I20 (Phi=0.39), Y36 (0,53), I48 (0,67), I59 (0,64) et V70 (0,54). La titine et TNfn3 ont été comparées structuralement (Geierhaas et al., 2004). Le coefficient de corrélation entre les valeurs expérimentales de PHI et celles predites est de 0,4, indiquant la difficulté d'une prédiction précise. Cependant, la prédiction peut être améliorée si l'on s'attache uniquement aux positions incluses dans les brins superposés B, C, D, E et F, s'affranchissant ainsi des valeurs aux positions périphériques de la structure. Ainsi le repliement de la titine part de W34, et inclut F21 (Phi=0 ,5), I23 (0,82), H56 (0,52), L58 (0,79), V71 (0,63) et F73 (0,72). La figure 1.12.13 montre cette séquence avec la position des SMIR. La plupart des positions expérimentales du noyau sont prédites par l'algorithme de SMIR si l'on admet une fenêtre de précision de ±3, excepté les deux dernières. Le noyau du repliement de TNfn3 contient les résidus clés I20, Y36, I59 et V70. Ces positions, comme I48 qui est la plus haute valeur de PHI sont prédites par les SMIR. En effet, le noyau est un ensemble de résidus en interaction au milieu de chaque brin B, C, E et F en accord avec les résultats de la comparaison avec les résidus HTOO. On peut néanmoins noter que les brins C' (le second brin noté C sur la figure 1.12.13) ou D peuvent avoir des résidus aux hautes valeurs de PHI, indiquant que la valeur de PHI n'est pas une garantie stricte d'appartenance au noyau du repliement.

Le dixième domaine fnIII de la fibronectine humaine (code PDB : 1ttf) a été muté quarante–deux fois en vingt–neuf positions (Cota et al., 2001). Les valeurs de PHI plus grandes que 0,5 sont L8 (0,66), F32 or Y32 (0,52), I34 (0,67), V50 (0,58), I70 and A74 (0,85), Y92 (0,9). Les auteurs admettent qu'une valeur de PHI de plus de 0,35 est une preuve significative de la formation de la structure mais A74 est surprenant car il montre très peu d'interactions. Y36, F48, I59, V72 et A74 sont aussi inclus dans le noyau bien qu'ayant des plus petites valeurs de PHI. L18, I20 et W22 font de nombreux contacts avec le cœur protéique mais ont des valeurs de PHI faibles, ainsi la structuration définitive du brin B dans l'état de transition reste incertaine. La plupart des grandes valeurs de PHI se retrouvent dans les brins C et F. L'état de transition compact est plus étendu dans FNfn10 que dans TNfn3, qui présentent une faible identité en séquence mais des structures proches. Les prédictions de SMIR sont localisées aux positions de grande valeur de PHI avec une fenêtre de précision de ±4 résidus, excepté celle en position C terminale.

Le second domaine fnIII dans la chitine A1, CAfn2 (code PDB : 1k85), a un noyau composé des résidus suivants (Lappalainen et al., 2008) : L22(0,24), V38 (0,40), N40 (0), I55 (0,40), F66 (0,07) et V68 (0,25). Il est étonnant de trouver des résidus attribués au noyau avec une valeur de PHI nulle (N40),

mais pour cette protéine toutes les valeurs sont très petites et étonnement la valeur maximale V48(0,45) ne fait pas partie du noyau selon les auteurs. On touche ici du doigt la difficulté, du point de vue expérimental, d'accéder de manière certaine à la connaissance du noyau protéique. Néanmoins les positions considérées du noyau sont dans le top 10 des valeurs de PHI et sont relativement bien prédites par les SMIR.

Pour le moment les SMIR prédisent grossièrement le noyau du repliement, tant que son extension est concernée et à condition d'autoriser un peu de lattitude sur la position. Sur 26 positions de SMIR, 24 sont des positions du noyau du repliement, ce qui correspond à 7% de la séquence. Pour augmenter la précision de la prédiction, on a déjà vu que l'on peut envisager d'introduire une nouvelle information telle que la prédiction de l'accessibilité au solvant ou celle de la stabilité. Avec les cas 1tit et 1ttf, les MIR n'arrivent pas à prédire les positions du noyau proches du C terminal. Pour 1ttf, on peut noter que cette extrémité comprend un schéma de 10 résidus hydrophiles impliqués dans une boucle. Ainsi il est difficile de prédire un MIR car le potentiel utilisé par l'algorithme désavantage les interactions hydrophiles. Une option afin de détourner cet effet serait d'opérer une permutation circulaire de la séquence. Ceci placerait le C terminal au milieu de la séquence, mais les résultats doivent être interprétés avec prudence car les permutation circulaires ont été démontrées comme perturbatrices du repliement (Li et Shakhnovich, 2001) .

La figure 1.12.14. montre la séparation en séquence entre les positions du noyau du repliement définies par les plus hautes valeurs de PHI et les positions de SMIR. Le maximum correspond à une correspondance exacte, ce qui est encourageant car indicateur que statistiquement les MIR lissés correspondent aux résidus impliqués dans le noyau. En effet, dans une fenêtre raisonnable de ±3, qui tient compte à la fois du lissage appliqué et de la précision sur la prédiction des MIR, on dénombre 55% des LIR lissés à l'intérieur de cet intervalle.

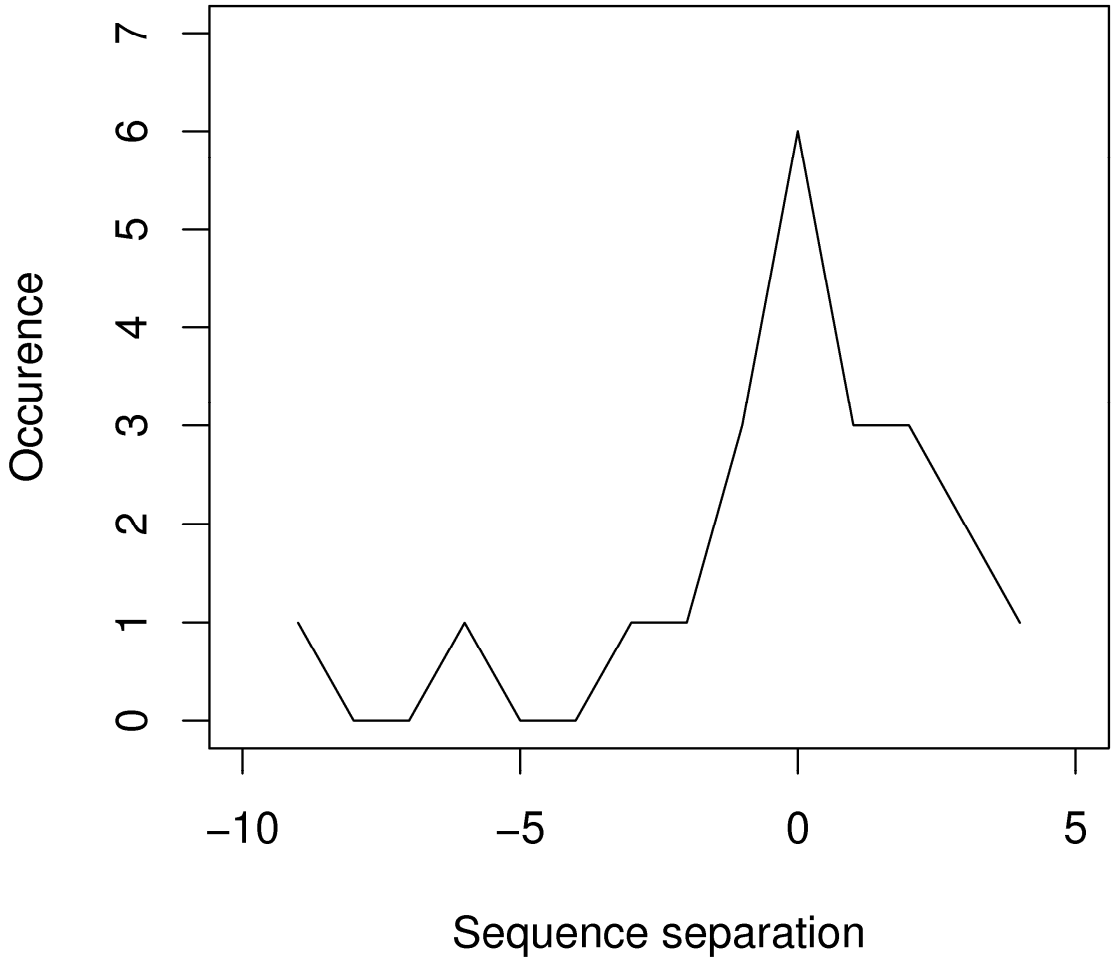


Figure 1.12.14. Séparation en séquence entre les positions connues pour faire partie du noyau par leur valeur de PHI et les MIR lissés, pour quatre domaines immunoglobulines.

1.12.10. Comparaison avec les CoC

Le concept de Conservatism of Conservatism est produit à partir d'un profil de conservation sur un alignement d'une famille protéique d'analogues partageant un repliement commun. Cela met en exergue les positions dont le conservatisme est relié à la stabilité structurale d'un point de vue évolutif. Cette technique renseigne sur la cinétique du repliement plus que sur la fonction. Aucune partie responsable de la fixation spécifique ou du placement du site actif n'a été utilisée dans cette étude. Ainsi les CoC de grande valeur ne peuvent pas être expliqués par la conservation de la fonction et reposent sur des caractéristiques structurales du repliement. Pour le domaine de code 1TEN (pris

comme représentant du repliement immunoglobuline) les six plus hautes valeurs de CoC sont aux positions : A17, I20, W22, L34, V70, et L72 (les résidus ont été renumérotés par rapport au papier original afin d'être cohérent avec les données présentées ici). Ces positions forment un regroupement dense et profondément enfoui dans le cœur protéique, témoin des contacts entre les brins canoniques B (positions 18, 20 et 22), C (position 34) et F (positions 70 et 72). Sur ces six positions, deux correspondent à des positions prédites par l'algorithme de SMIR (22 et 72) et une distante de 2 d'un SMIR (36). Il serait donc appréciable d'utiliser les données de CoC afin de valider les prédictions des SMIR ou les améliorer.

L'analyse de la corrélation entre les CoC et l'accessibilité au solvant donne des résultats en moyenne très probants, mais les valeurs les plus fortes en CoC ne peuvent pas être expliquées par ce seul critère. Selon Mirny et Shaknovich « La haute valeur de CoC ne peut pas être expliquée par l'accessibilité au solvant […] ; des facteurs autres que l'accessibilité au solvant ont contribué au conservatisme ». Ainsi nous avons appliqué la méthode à un autre repliement dont un des sujets 1CheY comporte le caractère particulier d'avoir des résidus du noyau du repliement accessibles au solvant. Ce repliement est celui de type flavodoxine.

1.13. Le repliement de type Flavodoxine

Nous avons testé la méthode développée sur le repliement immunoglobuline sur une autre famille structurale. Le repliement flavodoxine like présente l'avantage d'être de la famille mixte alpha bêta.

1.13.1. Banque

La banque de repliements de type flavodoxine a été constituée à partir des résultats donnés par la banque de données SCOP (Murzin et al., 1995).

1.13.2. Taux d'identité

Le taux d'identité a été calculé de la même façon que pour le repliement immunoglobuline. On retire une des deux entrées des couples montrant un taux d'identité supérieur à 70%, afin de limiter la redondance. Comme pour le repliement immunoglobuline, on retire du couple présentant une trop forte identité le partenaire ayant la moins bonne résolution.

La table ci-dessous indique les caractéristiques des protéines prises en compte dans cette banque

Protéine	Espèce	Code PDB	Résolution	Chaîne	Domaine
Glutamate mutase	Escherichia coli	1BE1	NMR	_	_
GTP-specific succynil-CoA synthetase	Escherichia coli	1EUC	2.1	A	01(9-131)
					02(132-301)
				B	01(1-20/112-246)
					02(21-111)
					03(247-392)
B-12-binding domains of methionine synthase		1BMT	3	A	01(651-737)
					02(738-895)
3-dehydroquinate dehydratase	Escherichia coli	1H05	1.5	_	_
S-adenosylhomocysteine hydrolase	Escherichia coli	1LI4		A	02(193-352)
N5-carboxyaminoimidazole Ribonucleotide Mutase	Escherichia coli	1U11	1.55	A	_
Acetylxylan Esterase	Escherichia coli	1ZMB	2.61	A	01(2-248)
Type II dehydroquinase	Escherichia coli	1GTZ	1.6	A	_
Nucleoside 2-deoxyribosyl transferase	Lactobacillus leichmannii	1F8Y	2.4	A	_
Methylmutase	Pseudomonas denitrificans	1I1H	2.6	_	_
CtBP dehydrogenase	Homo sapiens	1MX3	1.95	A	02(125-194)

Protéine	Espèce	Code PDB	Résolution	Chaîne	Domaine
Type II 3-dehydroquinate dehydratase	Actinobacillus pleuropneumoniae	1UQR	1.7	A	_
Product of gene AT4G34215 (function unknown)	Arabidopsis thaliana	2APJ	1.6	A	_
D-2-Hydroxyisocaproate dehydrogenase	Lactobacillus casei	1DXY	1.86	A	02(102-297)
Platelet activating factor acetylhydrolase IB	Bos taurus	1FXW	2.10	A	_
				F	_
D-lactase dehydrogenase	Lactobacillus delbrueckii subsp. bulgaricus	1J4A	1.9	A	02(104-299)
Phosphoribosylalino imidazole mutase	Thermotoga maritima	1O4V	1.77	_	_
S-Adenosyl-I-Homocysteine Htdrolase	Plasmodium falciparum	1V8B	2.4	A	02(237-397)
H.Pylori type II Dehydroquinase	Helictobacter pylori	2C4V	2.5	_	_
Succinyl-CoA synthetase	Escherichia coli	1JKJ	2.35	A	01(1-122)
					02(123-287)
				B	02(21-104)
					03(240-388)
Succinyl-CoA synthetase	Thermus thermophilus	1OI7	1.23	A	01(1-122)
					02(123-287)
PurE (Lyase)	Bacillus anthracis	1XMP	1.8	_	_

Protéine	**Espèce**	**Code PDB**	**Résolution**	**Chaîne**	**Domaine**
CheY (Signal Transfert Protein)	Escherichia coli	3CHY	1.66	_	_
Tir domain of TLR1	Homo sapiens	1FYV	2.9	_	_
D-3-Phosphoglycerate dehydrogenase	Mycobacterium tuberculosis	1YGY	2.3	A	02(99-182)
					03(321-453)
Methylmalonyl-CoA mutase	Propionibacterium freudenreichii suvsp. shermanii	7REQ	2.2	A	02(563-726)
Lysophospholiase L1/Acyl-CoA Thioesterase I/Protease I	Escherichia coli	1JRL	1.95	_	_
Nad-dependent D-Glycerate dehydrogenase	Hyphomicrobium methylovorum	1GDH	2.4	A	02(102-285)
PurE (Lyase)	Escherichia coli	1QCZ	1.5	_	_
Lipase	Escherichia coli	1Z8H	2.02	A	_
L-alanine dehydrogenase	Phormidium lapideum	1SAY	2.1	A	02(129-176)
Lipase/acylhydrolase	Enterococcus faecalis	1YZF	1.9	_	_

1.14. Discussion des résultats de corrélation

L'enfouissement suffit il à prédire le noyau ?

Le calcul des accessibilités au solvant a été calculé pour les 56 structures de la banque de domaines immunoglobulines. Les résidus pour lesquels la valeur d'accessibilité au solvant est nulle, ont été comparés aux SMIR. Le calcul d'accessibilité a été fait avec l'algorithme de Richmond (Richmond, 1984). Il est à noter que les résultats d'accessibilité diffèrent en général selon l'algorithme employé, ainsi la prédiction fondée sur l'accessibilité souffre-t-elle d'un manque de précision ? De plus un grand nombre de glycines montrent des valeurs nulles d'accessibilité, or les glycines ont un hydrogène en guise de chaîne latérale et il parait difficile qu'elles participent par leur interaction au noyau du repliement.

Un autre problème dû à l'utilisation de l'accessibilité pour la prédiction du noyau, est la condition indispensable de calculer l'accessibilité sur des chaînes peptidiques mono domaine. En effet sur une protéine multi domaines il peut arriver que des résidus aux surfaces d'interaction des domaines se retrouvent inaccessibles au solvant alors qu'ils font partie de la périphérie du domaine et ne participent donc pas au noyau du repliement.

Le nombre moyen de résidus ayant une accessibilité au solvant nulle est de 18 par domaine, alors que le nombre moyen de SMIR par domaine n'est que de 8. De plus en moyenne seuls deux SMIR par domaine ont des accessibilités nulles. Ainsi on peut dire que notre algorithme de prédiction du noyau est plus précis et en accord avec les études menées sur le noyau du repliement que le seul degré d'enfouissement.

Application de la méthode à l'aide à l'alignement à basse identité

A partir d'une séquence dont la structure n'est pas disponible, les méthodes classiques telles que BLAST ont du mal à proposer des apparentements à faible taux d'identité de séquence, et quand ils y parviennent l'utilisateur a du mal à mettre en évidence des résultats perdus au fin fond du listing de sortie. La prise en compte de résidus particuliers appartenant au noyau protéique pourrait aider à guider cet alignement. En effet, le noyau est partiellement conservé au sein d'un repliement donné, et si les MIR permettent de s'en approcher, on peut ainsi leur donner un poids plus fort puisqu'ils sont corrélés avec les positions de cœur et les positions de structures secondaires conservées. On peut envisager par ce moyen d'améliorer les algorithmes d'alignements mais au prix d'un prétraitement des banques de séquences car il faut pouvoir disposer du calcul des MIR sur un grand nombre de séquences.

1.15 Conclusion

Sur 56 structures de domaines immunoglobulines de longueur moyenne de 109 acides aminés, le cœur commun a été défini comme l'ensemble de résidus occupant une position qui montre un taux d'occupation au moins égal à 80% dans l'alignement multiple déduit de la superposition structurale de ces domaines. Ce cœur commun est composé de 74 résidus en moyenne par domaine, ce qui correspond à environ 75 % de la séquence. Ce résultat surestime apparemment le nombre de résidus strictement impliqués, de manière transitoire ou plus longue, dans le processus du repliement du domaine immunoglobuline. En effet, les résidus participant directement au repliement définissent le noyau du repliement. Cependant, on peut augmenter nos intuitions au sujet du noyau en ajoutant des informations supplémentaires. Ces informations peuvent être fondées sur l'analyse structurale ou sur la prédiction *ab initio*. L'analyse structurale des protéines par la méthode de TEF produit une bonne corrélation entre les extrémités de TEF et les cœurs conservés. On peut donc admettre que la limite des TEF est cohérente avec la physique du repliement protéique. L'inconvénient avec cette analyse structurale est qu'il faut disposer de la structure tridimensionnelle afin de délimiter le cœur. Cependant on peut penser que des protéines de haut taux d'identité partageront le même repliement. Ainsi, si l'on possède la structure d'une protéine homologue, on peut remonter au cœur de la protéine dont la structure est inconnue. Néanmoins, il faut rester prudent avec ce genre de méthodes, car il est avéré, dans de rares cas, que deux séquences très semblables peuvent adopter des repliements différents. Aussi il est avantageux de coupler les méthodes d'analyse structurale aux méthodes de prédiction *ab initio*, telles que celles utilisées au cours de notre étude : la prédiction des MIR et l'attribution des TEF. Cette approche permet une prédiction des positions des MIR qui se retrouvent à 73% dans le cœur conservé de notre banque de données. Notre outil de prédiction a été comparé à d'autres études utilisant différents algorithmes, comme par exemple l'étude de Gerstein et Altman (Gerstein et Altman, 1995), avec les HTOO résultant d'une analyse structurale des conservations. Le nombre moyen d'HTOO est d'environ 15 par séquence, la prédiction par les MIR ne donne pas de bonne corrélations, ce qui est partiellement dû au fait que les HTOO tendent à être agglutinés pour des raisons propres au fonctionnement de l'algorithme développé. Cependant si une procédure de lissage de la distribution des MIR est appliquée et que l'on compare les SMIR et les centres de gravités des regroupements d'HTOO, on s'aperçoit que l'on est capable de prédire la moitié des positions d'HTOO. La comparaison de notre schéma de prédiction du noyau s'avère être aussi compatible avec l'approche de CoC de Mirny et Shakhnovich (Mirny et Shakhnovich, 1999). Cette approche détecte, à partir d'un alignement multiple, l'ensemble des positions dont la conservation est reliée à la structure et non à la fonction. Il est encourageant de voir que nos résultats coïncident avec ceux de cette technique. L'analyse des valeurs de Phi est la seule technique expérimentale capable d'affecter à une

position son appartenance au noyau protéique. Bien que souffrant d'un manque de précision, cette technique a été choisie afin de tester nos prédictions. Sur les quatre domaines immunoglobulines possédant les valeurs de Phi, on peut conclure à une bonne correspondance entre résidus dont l'appartenance au noyau du repliement est hautement probable et les SMIR prédits.

Ainsi on peut dire que notre méthode permet de prédire les résidus formant le noyau du repliement avec une précision raisonnable, mais que l'on peut encore gagner en précision. Cet algorithme a aussi été testé sur un repliement de type flavodoxine. Les résultats sont sensiblement du même ordre que ceux obtenus sur l'analyse du repliement immunoglobuline, un peu moins bons dû au fait de l'originalité de ce repliement, de type mixte, alpha-beta. Le modèle que l'on propose est capable de gérer quelques aspects grossiers du processus de repliement conduisant à la formation du noyau.

2. Seconde partie : Classification structurale d'une banque de fragments de protéines

2.1. Introduction

Les protéines peuvent interagir entre elles et parfois la fonction ne prend naissance qu'à la condition de rassembler plusieurs chaînes, de séquences identiques ou non. La grande majorité des fonctions de la cellule ainsi que l'architecture de celle-ci nécessitent que les protéines se touchent, s'assemblent et se désassemblent. On peut voir l'interface protéine - protéine comme un cœur hydrophobe des résidus ayant perdu l'accessibilité au solvant. Cependant, il a été démontré que ce modèle n'est pas général (Cazals et al., 2007). La connectivité au sein des interfaces montre de grands changements (les résidus impliqués dans l'interaction varient selon la protéine et la taille de la surface d'interaction varie aussi) selon la famille d'appartenance du couple protéique. On peut observer les interfaces protéiques dans des cristaux de polymères dont les structures sont déposées dans la PDB. Les propriétés de connectivité sont dépendantes de l'eau cristallographique, celle dont les molécules sont liées aux acides aminés de la protéine. Les variations de connectivité observées en présence d'eau quantifient la propension de celle-ci à combler les interstices à l'interface. Quasiment toutes les interfaces possèdent des trous remplis par de l'eau. Il arrive que l'interface ne soit que l'addition de quelques composantes connexes d'interfaces. Ce modèle est nommé multi-patch et ces cas sont indépendants de l'eau cristallographique, pour 10% des cas environ. A l'échelle de l'interface, la courbure moyenne discrète corrèle avec l'aire de l'interface, alors qu'à une échelle locale, la même courbure suit une loi bimodale. La composition chimique des interfaces en termes de paires exhibe une proportion fixe d'interactions non définies, et l'on observe de subtiles variations entre les familles. Ces observations permettent d'apprécier les paramètres les plus à même de décrire précisément les interfaces de complexes protéine - protéine.

Les structures tridimensionnelles expérimentalement déterminées permettent de caractériser les interactions spécifiques protéine - protéine en terme de taille, de forme et de compacité. La comparaison avec les interfaces d'empilement cristallin représente des interactions non spécifiques protéine – protéine et montre à quel point les contacts spécifiques diffèrent des contacts non spécifiques de faible affinité que l'on peut rencontrer dans les cristaux.

La compréhension du mécanisme d'interfaçage protéine – protéine nécessite la connaissance des coordonnées atomiques du complexe protéique (Bahadur & Zacharias, 2008). Les techniques utilisées sont la cristallographie aux rayons X, la résonance magnétique nucléaire ou encore la cryo -

microscopie électronique. D'autres techniques peuvent venir confirmer ces résultats afin de valider la spécificité de l'interaction d'un complexe. Ces méthodes sont la chromatographie d'affinité ou encore les expériences de double-hybride. Des études récentes ont montré que la majorité des protéines de la cellule existent en tant que parties d'assemblages aux composantes multiples. L'analyse structurale de tels complexes est devenue une priorité pour bon nombre de structuralistes. L'analyse des interfaces des complexes connus est essentielle à la compréhension de la reconnaissance spécifique entre deux protéines et indispensable en vue de la prédiction.

La diversité structurale de la reconnaissance protéine–protéine est visible dans les complexes obtenus par cristallisation. Certains homo–dimères ne sont pas retrouvés en tant que complexe stable dans la cellule. La formation de cristaux implique des contacts non-spécifiques qui seront appelés interfaces non- spécifiques car elle n'interviennent pas dans le processus biologique et sont des artefacts de la méthode de cristallisation. Nous parlerons plus tard de la difficile tâche qui est de discriminer entre les multimères biologiques et cristallins.

2.1.1. Les propriétés structurales des interfaces protéine – protéine

La taille de l'interface peut être quantifiée par l'aire de la surface accessible au solvant (SASA). La surface d'interaction (B) pour un complexe constitué de deux chaînes peut ainsi être calculée par la différence entre l'accessibilité au solvant des deux sous – unités et celle du complexe.

$$B = SASA_{sous-unité1} + SASA_{sous-unité2} - SASA_{complexe}$$

La surface accessible au solvant (Richmond, 1984) est calculée à partir des coordonnées atomiques des sous unités isolées en faisant rouler une sonde au rayon de la molécule d'eau sur la surface de chaque protéine et du complexe. En moyenne les interfaces correspondant aux interactions spécifiques sont bien plus larges que les interfaces dues à l'entassement cristallin. Les homodimères présentent une moyenne de surface d'interaction deux fois supérieure aux hétérodimères et deux fois et demi plus grande que les multimères dus à l'empilement cristallin (Bahadur et Zaccharias, 2008). La taille moyenne des surfaces d'hétérodimères est de 1200-2000 Angstrom² alors que les interfaces dus à l'empilement cristallin présentent une moyenne de 570 Angstrom² pour un set de 1320 interfaces d'enpilements (Bahadur et Zaccharias, 2008). Il existe des cas particuliers de surfaces d'interfaces de dimères biologique mais en général on peut dire que la surface enfouie va de 800 a 10000 Angstrom². Cependant il arrive que les dimensions des surfaces d'interaction se chevauchent entre dimères biologiques ou d'empilement cristallin. Ces faits montrent que la seule information de surface enfouie n'est pas suffisante pour distinguer les contacts spécifiques des contacts non spécifiques provenant des contacts cristallins.

En plus de la taille de l'interface, la courbure de l'interface peut être un élément de reconnaissance des interactions spécifiques. Jones et Thornton (Jones & Thornton, 1996) ont développé un paramètre dénommé coefficient planaire. Il en résulte que les interfaces semblent être plutôt plates. A l'exception des inhibiteurs d'enzymes qui forment pour la plupart des surfaces convexes qui s'emboitent dans les surfaces concaves des sites actifs. Un autre paramètre caractérisant la forme des interfaces est la circularité. La circularité est le rapport entre les longueurs des axes principaux décrivant le plan des atomes en interaction. En moyenne les interfaces ne sont pas parfaitement circulaires, mais les valeurs ne diffèrent pas suffisamment pour permettre de distinguer les dimères biologiques des monomères entassés dans un cristal. On peut aussi caractériser les interfaces protéine – protéine par la complémentarité de forme. Laskowsky a développé un outil afin de quantifier ce paramètre (Laskowski, 1996). Il est appelé Gap Volume (GV) et est défini par le volume d'écart entre les deux molécules divisé par l'aire de la surface d'interaction. Les interfaces dues à l'empilement cristallin montrent un GV supérieur d'un facteur 2 aux multimères biologiques. Ceci suggère que les interfaces d'empilements contiennent un volume de cavité par unité d'aire bien plus grand que les dimères biologiques.

La composition chimique des interfaces protéine – protéine diffère entre les interactions spécifiques et non spécifiques. Proportionnellement à la surface d'accessibilité au solvant, les interfaces de multimères biologiques apparaissent enrichies en acides aminés aliphatiques (Leu, Val, Ile, Met) et aromatiques (His, Phe, Tyr, Trp) et dépourvues en acides aminés chargés (Asp, Glu, Lys) à l'exception de l'arginine, alors que les interfaces d'entassements présentent une abondance en acides aminés chargés et polaires, et une pauvreté en acides aminés hydrophobes. Plus précisément on peut diviser en deux groupes, polaires et non polaires, les résidus présents aux interfaces protéiques. Il en résulte que 58% des résidus sont apolaires dans les interfaces d'empilement alors que ce nombre augmente à un taux de 65% dans les dimères biologiques. Il en découle que les interfaces protéine – protéine sont majoritairement guidées par l'effet hydrophobe bien qu'il y ait une petite proportion d'acides aminés polaires pouvant former des liaisons hydrogènes.

2.1.2. Matrice de contacts et tesselation de Voronoï

La tesselation consiste à paver une surface autour d'un ensemble de points. Ici la tesselation est dans l'espace tridimensionnel, mais le concept reste le même. Soit une distribution de points dans un espace euclidien. A Chaque couple de points on définit un plan orthogonal au vecteur séparant les deux points. Pour chaque couple de points la même opération est effectuée. Ainsi les plans s'intersectent et forment des cellules. Les cellules sont donc composées par des faces et des arrêtes. L'application aux acides aminés d'une protéine est faite de la façon suivante : les points de la distribution sont les coordonnées tridimensionnelles des carbones alpha de la chaîne peptidique (on peut aussi bien utiliser le centroïde ou un autre atome). Ainsi les cellules partageant une face sont dites en contact. Une

matrice de contact peut alors être générée afin d'étudier la connectique de la chaîne peptidique dans l'espace. La figure 2.1.1. est une représentation en deux dimensions d'une tesselation de Voronoï.

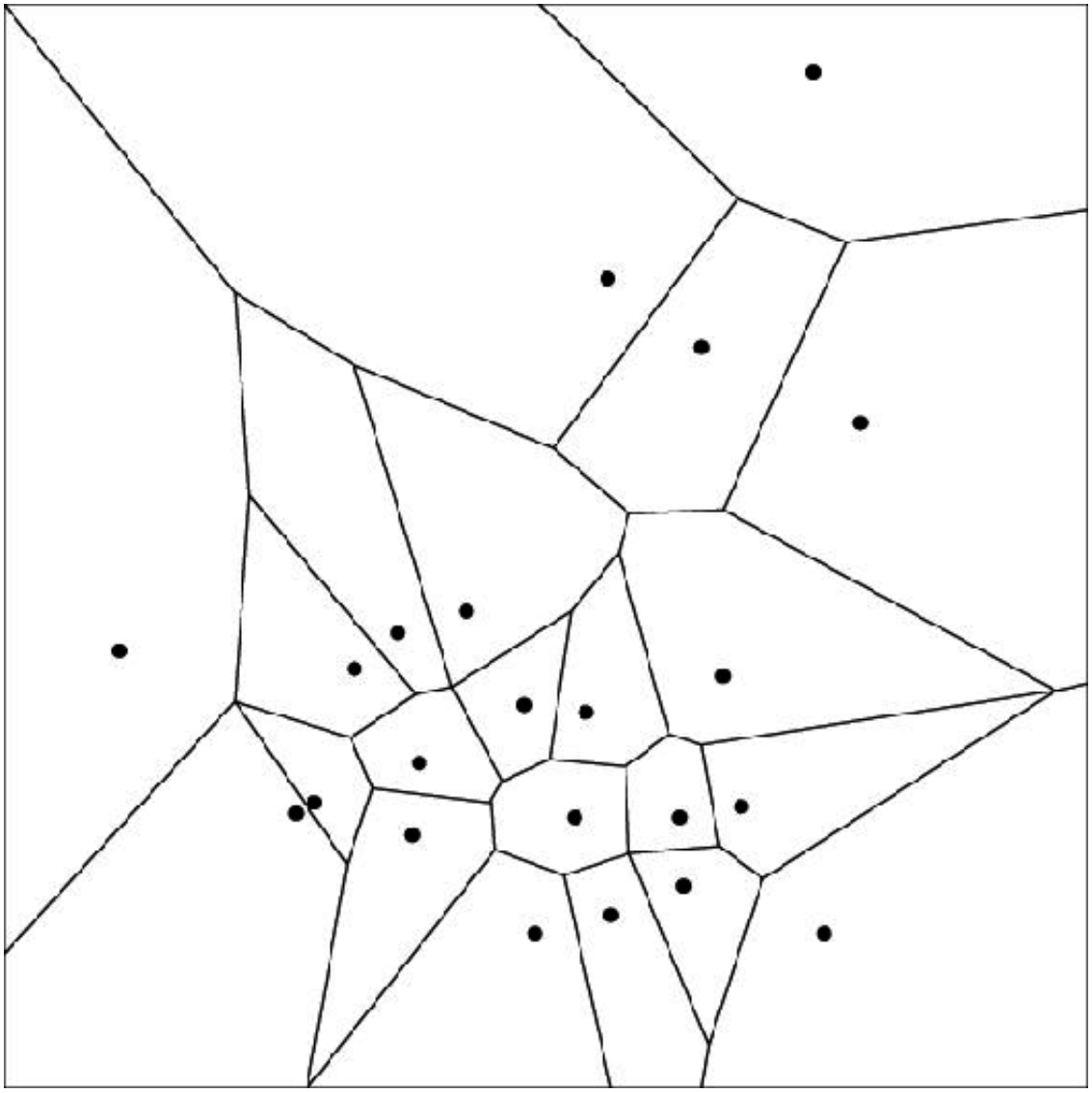


Figure 2.1.1. Cellules de Voronoï à deux dimensions d'un ensemble de points.

La figure 2.1.2. donne une représentation de la matrice de contact générée par le programme VORO3D (Dupuis et al., 2004) pour la protéine de code PDB 1em8.

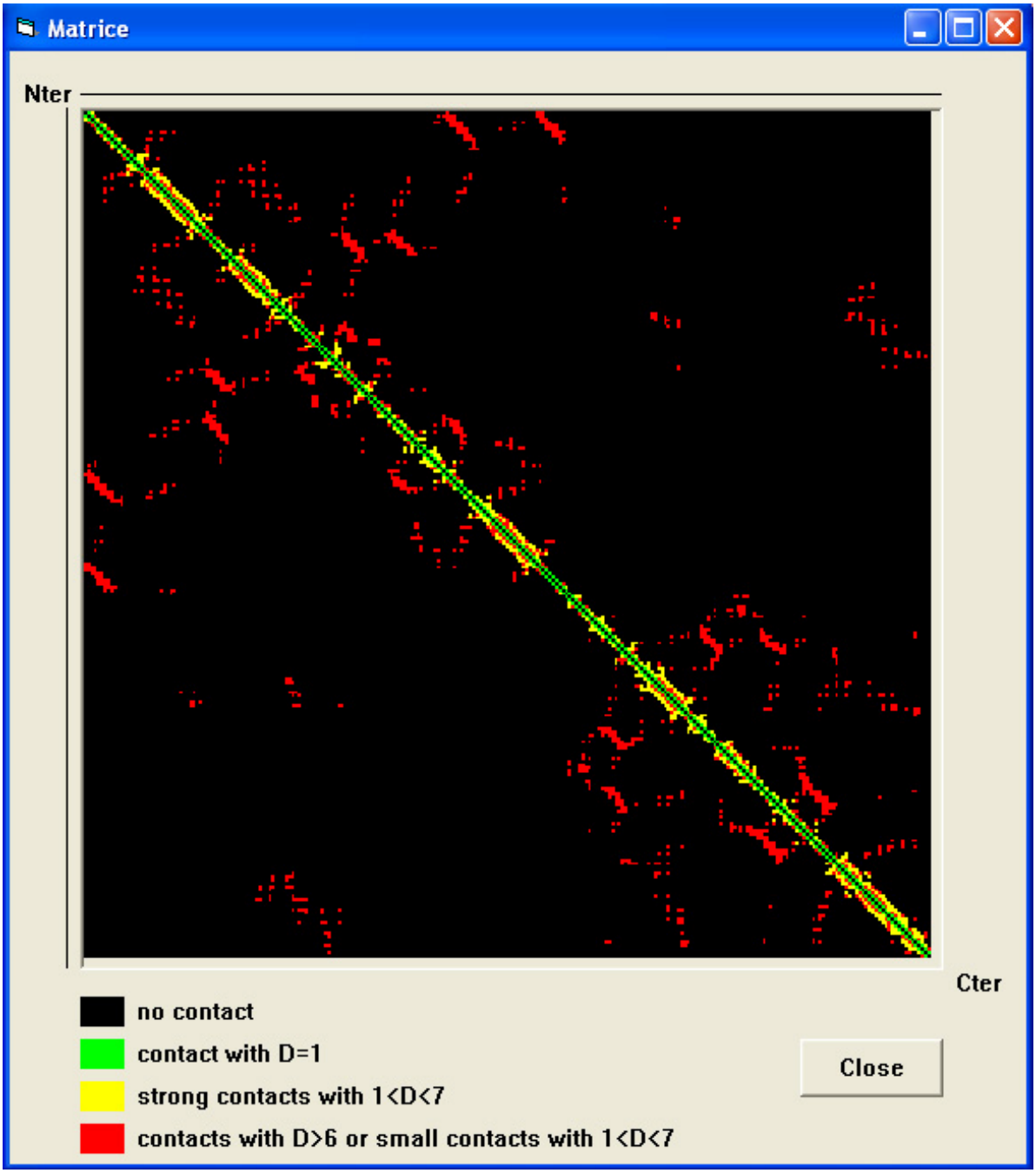


Figure 2.1.2. Matrice de contacts de la protéine de code PDB 1em8.

Pour les protéines, on utilise le programme VORO3D de Dupuis et Sadoc. La tessellation de Voronoï est définie par ces auteurs comme l'intersection de plans de contact à mi–distance entre les points, construisant ainsi les cellules de Voronoï, où dans notre cas les points représentant les acides aminés sont les centres géométriques des chaînes latérales des acides aminés. Pour un ensemble de points donnés, la tessellation de Voronoï est unique et absolue puisqu'il n'y a pas d'espace libre entre les cellules. Les surfaces des cellules de Voronoï peuvent définir de manière claire les contacts entre les points associés et leurs plus proches voisins. La principale difficulté à la construction de cellules de

Voronoï pour une chaîne peptidique est le problème des résidus qui se trouvent à la surface. En effet les cellules associées à ces résidus apparaîtraient très allongées voire ouvertes. Pour résoudre ce problème, VORO3D recouvre la protéine par un environnement modélisé. Le programme fonctionne en deux étapes : la première est de générer un environnement à la protéine afin que les résidus de surfaces n'aient pas une cellule de dimension infinie. Ainsi des « molécules » sont placées uniformément dans tout l'espace incluant la protéine. L'environnement est dispersé autour de la protéine comme une coquille avec une épaisseur constante (Angelov et al., 2002). Si une cavité existe au sein de la protéine, elle sera aussi comblée par l'environnement. Pour une protéine d'environ 200 résidus, le nombre de points représentant chaque acide aminé et les « molécules » d'environnement n'excède pas 2000. Ces points sont sélectionnés à partir de 8000 sphères entassées dans une boite cubique contenant la protéine (les points correspondant aux centres des sphères). Durant la relaxation où l'environnement est sélectionné, la protéine est considérée comme rigide. La méthode de relaxation consiste à retirer les sphères de l'environnement qui sont superposées aux acides aminés. Les molécules d'environnement qui ne pourraient pas cohabiter de façon stérique avec la protéine sont éliminées. La figure 2.1.3. représente une protéine (de code PDB : 1cus) avec son environnement relaxé. Les points en gras sont les centres géométriques des acides aminés de la protéine et les points plus minces sont les centres des sphères modélisant l'environnement. Les axes de la boîte sont quantifiés en Angstrom.

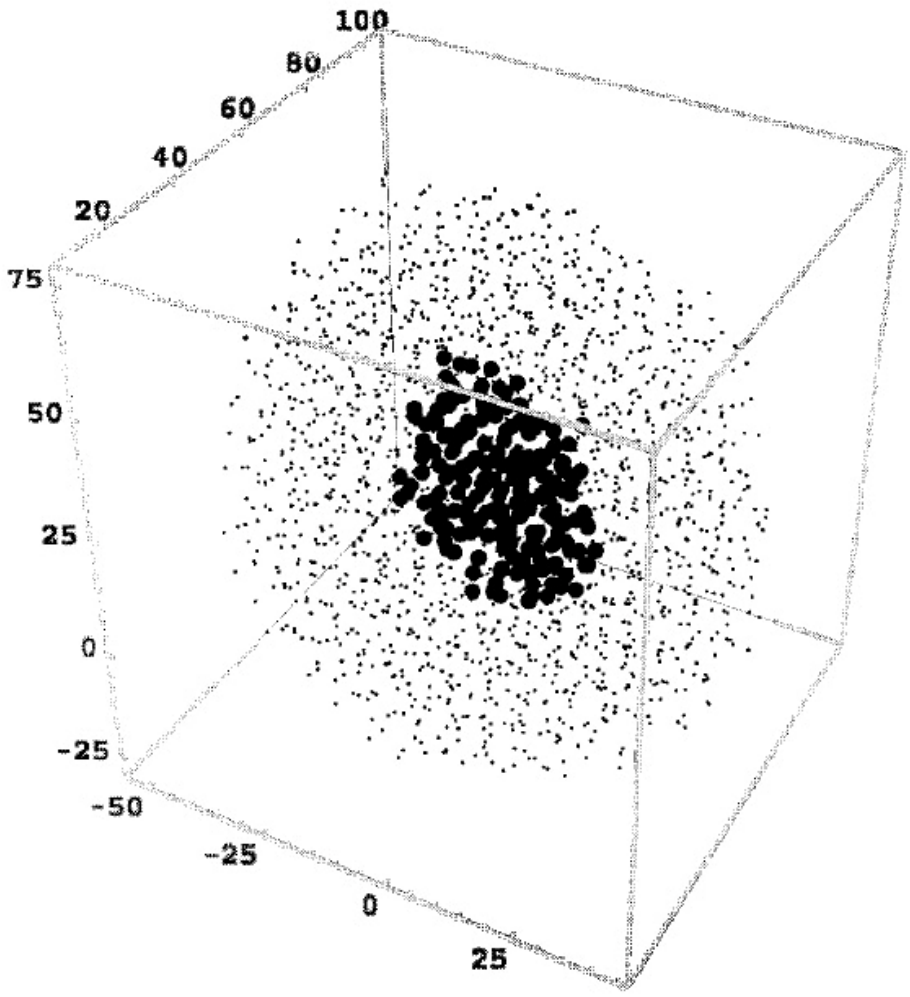


Figure 2.1.3. Schéma de la protéine 1cus (les points noirs) placée dans un cube avec son environnement représentatif (les points gris).

La qualité des cellules de Voronoï de l'environnement a été vérifiée (Angelov et al.,2002) et semble ne pas être influencée par la surface extérieure de l'agrégat formé par la protéine et son environnement dans lequel elle est plongée. Ces molécules d'environnement modélisées peuvent être paramétrées dans leur diamètre. Nous avons choisi les paramètres par défaut qui correspondent au diamètre moyen d'un résidu, c'est-à-dire 6,5 Angstrom. La deuxième phase consiste à créer les cellules de Voronoï décrivant l'ensemble des résidus.

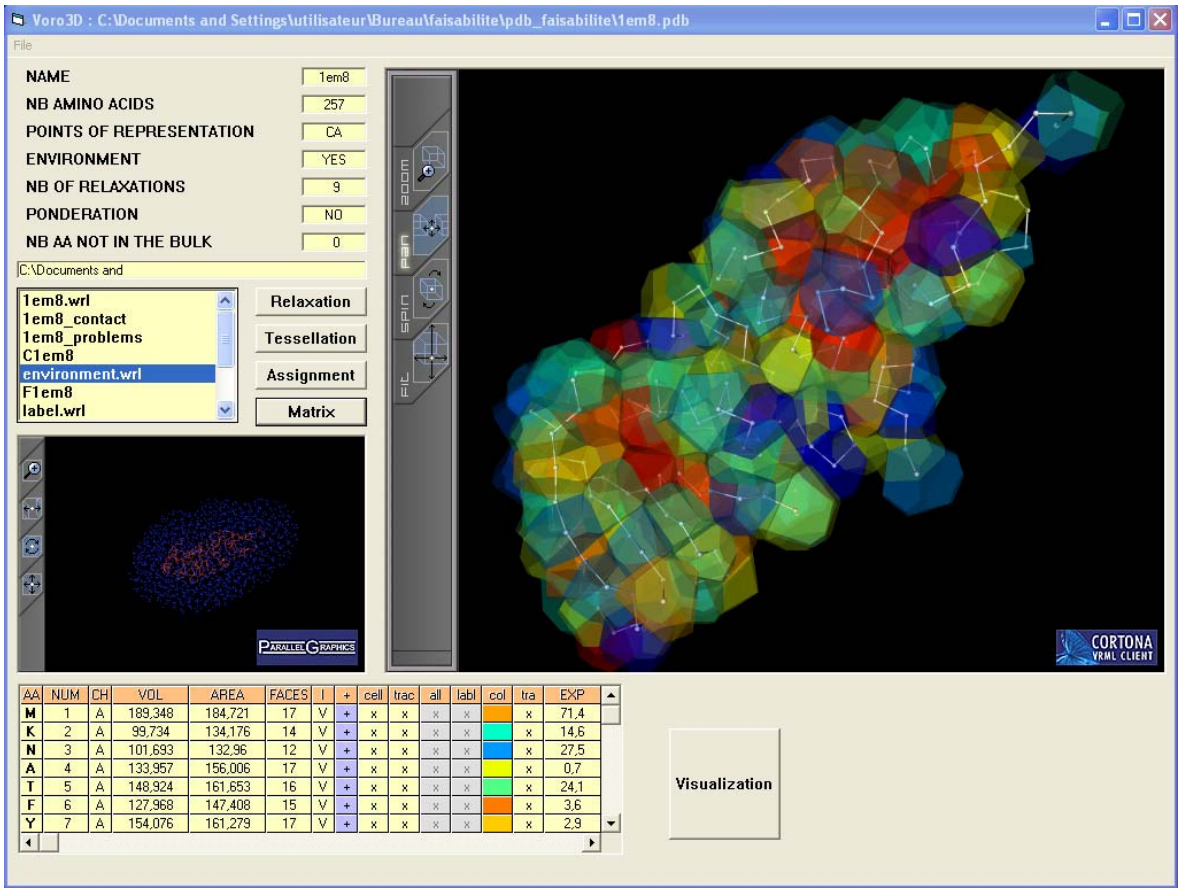


Figure 2.1.4. Exemple d'une fenêtre du logiciel Voro3D avec les différents paramètres accessibles pour calculer les cellules de Voronoï.

VORO3D est implémenté pour Windows et propose une interface graphique pour que l'utilisateur puisse paramétrer la tessellation. La figure 2.1.4. représente l'interface graphique de VORO3D avec comme exemple de protéine, celle correspondant au code PDB 1em8.

Pour notre part, le nombre d'entrées à traiter étant bien trop important pour les traiter une par une, nous avons pu avoir accès aux codes sources. Ainsi nous avons pu automatiser la technique et la faire tourner sous Linux afin de construire les tesssellations de notre banque de dimères.

2.1.3. Discriminer entre multimères biologiques et multimères d'entassement cristallin

2.1.3.1. Thornton

La cristallographie permet de profiter des coordonnées tridimensionnelles des atomes constituant les protéines cristallisées. Cependant lorsque l'on travaille sur des complexes, il n'est pas trivial qu'un oligomère se retrouve cristallisé alors qu'il n'existerait pas dans un milieu physiologique. Il est déjà difficile de déterminer le nombre exact de sous-unités pour des homopolymères, mais le manque

d'information disponible sur le comportement des protéines étudiées en solution augmente encore la difficulté quant à la résolution de ce problème. On pourrait croire que le problème n'existe pas pour les hétéro-polymères, or il existe des cas où les cristaux sont formés mais les sous-unités n'interagissent pas en milieu physiologique. De plus, la PDB ne donne pas accès aux opérations de symétrie qui génèrent la macromolécule à partir des coordonnées atomiques qui représentent l'unité asymétrique. La méthode habituelle pour évaluer la taille des contacts entre chaînes protéiques est de recourir à la mesure de la surface accessible au solvant. D'autres paramètres caractérisent les interactions quaternaires : L'hydrophobie, la complémentarité de forme à l'interface, la circularité et la planéité de la surface d'interaction. Ces paramètres doivent être examinés avec minutie afin de décrire l'interaction protéine-protéine, ils ne se prêtent donc pas à une automatisation en vue d'une prédiction quaternaire.

Henrick et Thornton (Henrick & Thornton, 1998) ont pensé à mieux estimer l'erreur associée à la prédiction quaternaire proposée par le serveur de l'EBI (http://pqs.ebi.ac.uk) par une méthode statistique plus rigoureuse fondée sur l'étude des surfaces de contact (Jones et al., 2000). 172 structures non-homologues dont l'état oligomérique est connu ont été étudiées et des dimères hypothétiques ont été générés. La surface de contact de ces dimères hypothétiques a été calculée par la manière courante de la différence d'accessibilité au solvant entre le complexe et les formes libres. La fonction de score fondée sur la fréquence en paires d'atomes est définie comme suit. Deux atomes appartenant au dimère hypothétique sont comptés dans une paire s'ils appartiennent à une chaîne polypeptidique différente et s'ils sont éloignés de plus de 8 Angstrom. Les atomes triés par leur connectivité forment 17 types pour les 20 acides aminés communs. Le calcul du score dépendant de la distance, à partir de la fréquence par paire, a été réalisé en accord aux études sur les potentiels statistiques ou « potentiels de paires ».

$$s_{ab}(r) = \ln(1 + \sigma N_{ab}) - \ln\left(1 + \sigma N_{ab}\frac{n_{ab}(r)}{n(r)}\right)$$

N_{ab} est le nombre de paires d'atomes de type *a* et *b*, $n_{ab}(r)$ représente la fraction de ces paires où les atomes de type *a* et *b* sont séparés d'une distance *r*. n(r) représente le nombre de paires séparées par la distance *r*. σ est un facteur pondérant proportionnel à l'effet de la suppression graduelle des petites fréquences par paires d'atomes. Ce facteur a été fixé à 0.02, les scores log-odds obtenus pour un monomère ou un homodimère sont la somme des termes $s_{ab}(r)$ pour tous les atomes appariés le long de la surface du dimère hypothétique.

L'approche habituelle pour estimer l'erreur de la classification généralisée est de réduire artificiellement la taille du groupe de données de manière aléatoire. Ceci afin de déduire un modèle et

une erreur généralisée à partir de l'erreur observée sur le reste du groupe de données qui devrait être plus cohérente avec l'erreur obtenue avec des données inconnues. L'équipe de Thornton a alors choisi une méthode dite « bootstrap » (Efron & Tibshirani, 1993) pour estimer le taux de divergence (Efron & Tibshirani, 1997). Ainsi un jeu de données dites d'entrainement a été réalisé à partir du jeu de données original (172 oligomères) et la meilleure valeur de seuil pour la discrimination a été déterminée à partir des scores calculés pour le jeu d'entrainement. Cette valeur seuil est ensuite appliquée à un jeu de données test afin d'obtenir le taux de divergence. Le taux de divergence est calculé pour les scores par paires et pour la différence de surface accessible au solvant. C'est grâce à ces valeurs seuil de divergences que sera attribuée la véracité d'un dimère ou non.

2.1.3.2. DiMoVo

La motivation de l'équipe de Poupon et Bernaueur (Bernauer et al., 2008), qui a développé le logiciel DiMoVo, était que la connaissance de l'état oligomérique d'une protéine est essentielle pour la compréhension du mécanisme fonctionnel de ladite protéine. Ainsi ils ont développé une méthode fondée sur la tesselation de Voronoï et une modélisation de type gros grain de la protéine permettant de discriminer entre les dimères biologiques et les dimères d'empilement cristallin.

L'algorithme

Bien qu'une structure protéique soit considérée comme unique, les atomes de surface sont faiblement contraints et on observe des mouvements lorsque deux protéines interagissent. Comme ces mouvements sont difficiles à prédire, un modèle simplifié à faible résolution est mieux approprié à la prédiction des interactions protéine–protéine. Les protéines sont donc représentées par des sphères correspondant à chaque résidu et une tesselation de Voronoï est construite à partir de ce modèle. La construction des tesselations de Voronoï a été faite en utilisant la librairie Computational Geometry Algorithms (CGAL) (http://www.cgal.org/) et a été optimisée pour ne prendre que quelques secondes par complexe.

Les paramètres utilisés sont :

- L'aire d'interface ;

- Le nombre de résidus au cœur de l'interface ;

- Leurs volumes de Voronoï ;

- La fréquence de chaque type de résidu à l'interface ;

- La fréquence de chaque paire de résidus en contact ;

- La distance entre les centres géométriques.

La méthode consiste ensuite à varier les différents paramètres en utilisant un algorithme d'apprentissage de type Support Vector Machine. L'apprentissage a été réalisé avec le package libSVM du programme de calcul R. Les paramètres sont ajustés sur différents sets de données tels que des multimères biologiques vérifiés dans la littérature et des multimères d'empilement cristallin ou encore un set d'apprentissage. La méthode obtient une précision de 0,97 pour un seuil de 0,5.

Aucun détail n'est donné quant au traitement des hétéromultimères. Il semble que DiMoVo puisse traiter plusieurs chaînes peptidiques différentes (hétérodimères ou multimères) aussi bien qu'une seule (homodimère).

Le programme DiMoVo a été créé dans l'optique d'un serveur, c'est-à-dire pour que les équipes valident le complexe étudié. L'équipe qui a développé DiMoVo n'a pas généré de banque de multimères.

2.1.4. Pistes de classification structurale qui n'ont pas été suivies.

2.1.4.1 Introduction au projet interdisciplinaire proposé

Cette analyse des interactions protéine – protéine a été proposée comme projet pour un programme interdisciplinaire Physique Chimie Biologie du CNRS, en collaboration avec J. Chomilier (Prédiction des Structures Protéiques IMPMC,UMR 7590), J.F. Sadoc (Equipe Organisation et dynamique de la matière condensée Laboratoire de Physique des Solides, UMR 8502) et P. Derreumaux (Laboratoire de Biochimie théorique IBPC, UMR 9080). Le titre du projet était : Prédiction de la topologie des structures de dimères protéiques. Il a été déposé le 12 novembre 2007. L'étude se fondait sur la connaissance de l'aspect modulaire des protéines que l'équipe a pu développer au cours des ans. Le souhait que portait le projet est que les éléments modulaires, c'est-à-dire les TEF, pourraient faciliter la prédiction des orientations relatives des domaines impliqués dans la formation d'un dimère. En effet, encore à ce jour, les algorithmes de prédiction supposent de tester de manière exhaustive toutes les orientations relatives des deux globules, ce qui conduit à une énumération très coûteuse en temps de calcul. Une indication sur les orientations relatives des domaines aurait permis de réduire drastiquement le nombre de conformations explorées dans la recherche du modèle du dimère. La figure 2.1.5. représente des TEF en interaction dans un dimère (code PDB 1em8). Les extrémités des TEF sont représentées par des sphères.

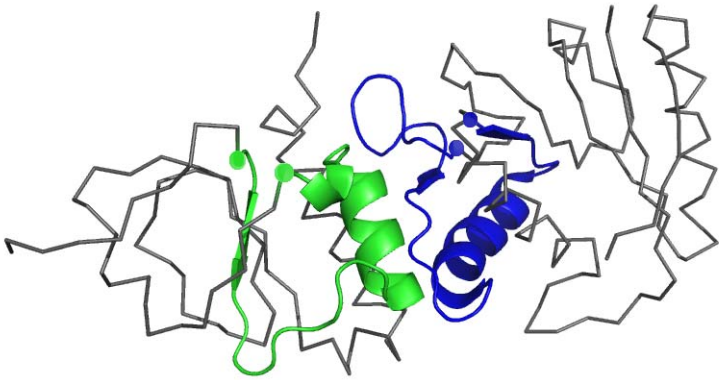


Figure 2.1.5. Exemple de TEF en interaction dans un dimère (1em8).

Le projet, qui n'a pas reçu de financement, va être décrit dans ses grandes lignes. Dans un premier temps une banque de TEF est constituée à partir des entrées de SCOP contenant plusieurs chaînes. L'extraction des TEF de cette banque doit produire de 5000 à 6000 TEF. Afin d'éviter les redondances, seules les séquences suffisamment divergentes sont traitées. La deuxième étape de la constitution de la banque propose de regrouper les TEF en familles sur des bases de ressemblances structurales. Des essais erreurs permettent de placer correctement le curseur entre un petit nombre de familles très divergentes ou un grand nombre aux membres plus proches. Il est à signaler que l'équipe porteuse de ce projet a une antériorité dans ce domaine, puisqu'elle a par le passé mis en œuvre une classification structurale des boucles protéiques, par un algorithme de tri hiérarchique qui s'est avéré pertinent (Kwasigroch et al., 1996; Wojcik et al., 1999). Une analyse topologique inscrit un TEF dans un cylindre qui contient tous ses atomes (méthodologie qui sera retenue et effectuée durant la thèse). Il a été convenu d'établir si l'on peut se contenter du squelette ou si on doit y ajouter les chaînes latérales. L'avantage d'utiliser un cylindre réside dans le fait que sa forme mathématiquement simple est compatible avec des calculs géométriques rapides. Le rayon ainsi que l'extension des TEF le long de l'axe de ces cylindres peuvent être utilisés pour caractériser chaque TEF de manière simple.

2.1.4.2. Descriptions des interactions

En parallèle, on collecte les structures des complexes disponibles pour constituer une base d'apprentissage. La banque de l'EBI, E-MSD (Boutselakis, 2003), est un bon outil pour sélectionner les dimères, qu'ils proviennent de la PDB ou de SCOP. A l'heure actuelle, on compte environ 400

complexes issus de SCOP, parmi lesquels il convient là encore de supprimer la redondance, et surtout de séparer les dimères biologiques des dimères formés par l'empilement cristallin.

Les sites d'interactions entre domaines peuvent être analysés soit par les méthodes classiques de distances inter-atomiques, soit par tessellation de Voronoï (Angelov et al., 2002) pour éviter les inconvénients des distances seuil. On se ramène dans l'interaction entre deux domaines aux TEF qui portent les résidus en interaction entre domaines. L'étape suivante consiste à modéliser leur orientation relative. Pour cela, on peut utiliser les outils classiques de la géométrie, mais aussi profiter des fortes contraintes imposées par la structure en chaîne de chaque TEF pour coder la géométrie par des matrices de type matrices de contact. La gestion des configurations étudiées est alors beaucoup plus légère. En effet une des pistes de classification structurale repose sur la description des TEF par tesselation de Voronoï et la production des matrices de contacts intra – TEF, la classification se faisant par analyse des valeurs propres des matrices.

L'hypothèse sous-jacente à cette entreprise est que les TEF appartenant à deux familles structurales données présentent les mêmes caractéristiques d'interaction. S'il n'y a pas de corrélation entre familles structurales de TEF et similitude de topologie des interactions, nous devrons nous passer de la première étape de ce projet, et mettre en place une classification des TEF spécifique aux complexes (vu que la première étape n'a pas pu être réalisée, c'est cette dernière méthodologie qui va être appliquée).

A l'issue de cette phase analytique, nous serons en possession d'une bibliothèque structurale de paires de TEF dont on pourra déduire les topologies de leurs interactions.

2.1.4.3. Prédiction topologique

Si l'on synthétise la problématique, le projet devait proposer à l'expérimentateur un outil qui lui permette d'aider à déterminer l'orientation relative des deux chaînes impliquées dans un dimère au niveau de la structuration quaternaire. Nous nous sommes placés volontairement dans l'hypothèse où l'état multimérique est connu expérimentalement, que ce soit par gel bidimensionnel d'électrophorèse, par spectrométrie de masse, ou par diffusion de lumière. L'amélioration récente de techniques de label de spin en RPE (Résonnance Paramagnétique Electronique) a permis d'obtenir des informations sur les résidus en interaction qui peuvent être introduites comme contraintes dans notre simulation. Pour notre prédiction, les TEF des deux domaines du complexe qui sont les principaux fragments impliqués dans l'interaction sont déterminés par similitude avec des chaînes connues. La famille à laquelle ils appartiennent est ensuite recherchée dans la banque de TEF sur des critères de ressemblance structurale. Dans un premier temps, nous avons envisagé un critère simple et rapide, comme le RMSD, mais une approche plus sophistiquée aurait été nécessaire dans un second temps. En particulier, on pourra envisager le score ProQres, mis au point par l'équipe d'Elofsson (Wallner et Elofsson, 2006).

La prédiction du positionnement relatif des deux domaines dans le complexe se ramènera à l'étude de celui des paires de TEF, un par domaine (dans une première approche en tout cas). Nous supposons que l'orientation relative des TEF est conservée au niveau des complexes. La complexité de la description de l'espace conformationnel est alors grandement réduite puisque l'orientation relative des TEF est supposée connue par référence aux TEF homologues disponibles dans la bibliothèque. On se ramènerait alors à la rotation de l'un des domaines autour de l'axe d'un des TEF. Si l'on admet un pas de 10°, cela fournira 36 solutions par paire de TEF. Sachant que pour un domaine moyen de 150 acides aminés de long, il y a environ 5 TEF, le nombre de paires de TEF sera de 5^5, soit de l'ordre de 3000 conformations, à comparer à 36^3 (plus de 46 000) conformations du calcul exhaustif avec le même pas angulaire. Ceci permet de gagner un ordre de grandeur sur le temps de calcul.

2.1.4.4. Sélection des modèles

L'étape suivante est la sélection de la meilleure solution parmi celles qui vont être proposées. Pour cela nous avons envisagé d'utiliser le serveur STREVAL (STRucture EVALuation serveur) récemment élaboré à l'IBPC. STREVAL utilise dans sa version actuelle plusieurs mesures fondées sur des représentations atomiques détaillées : énergie libre de solvatation par la méthode Sfe (Chiche et al., 1990), potentiels statistiques par les méthodes Anolea (Melo et al., 1997; Melo & Feytmans, 1998) et Dope, scores de compatibilité 1D/3D avec Verify3D (Luthy et al., 1992) et Eval23D (Gracy et al., 1993), types de contact (Errat). Le potentiel gros grain OPEP, un des plus efficaces pour discriminer les structures natives des états non natifs est en cours d'installation (Maupetit et al., 2007). Afin de faciliter la comparaison entre ces différentes méthodes, une procédure de standardisation a été proposée. Réalisée sur un ensemble d'apprentissage de 722 structures PDB, cette standardisation permet la comparaison des résultats des différentes méthodes selon une même métrique mais aussi de déterminer le degré d'éloignement d'un modèle par rapport à une structure idéale en fonction du nombre d'acides aminés. Un soin particulier a été apporté à l'interface pour faciliter l'interprétation d'une ou plusieurs centaines de modèles simultanément. Ainsi, les scores obtenus sont affichés selon un code couleur directement visualisable sur la structure 3D du modèle. Lorsque plusieurs modèles sont proposés, ils apparaissent classés selon leur qualité relative les uns par rapport aux autres. Un algorithme d'apprentissage automatique est nécessaire dans le cadre de ce projet pour combiner les scores des différentes méthodes et ainsi offrir un score consensus plus fiable. Un site WEB, fonctionnel pour les monomères, est disponible à l'adresse http://www.shaman.ibpc.fr/streval/.

Bien que très ambitieux, ce projet n'a pas été retenu, cependant une partie a été réalisée et elle sera expliquée plus en détail dans la suite.

2.2. Matériels et Méthodes

La classification structurale des TEF s'est faite sur la même banque que celles des multimères, contrairement aux premières intentions portées par le projet interdisciplinaire.

2.2.1. Nettoyage de la banque

Nous avons pris comme départ la liste entière de multimères étudiée par l'équipe de Thornton (Jones et al., 2000). Chaque entrée a été analysée par le programme DiMoVo. La validation du complexe comme multimère biologique est faite grâce à DiMoVo. Nous avons pu disposer des codes sources et ainsi grâce à un script que j'ai programmé, nous avons pu automatiser la procédure DiMoVo à toutes nos entrées PDB correspondant à la liste de base de Thornton. Nous avons utilisés les paramètres de filtre développé par l'équipe de Bernauer.

2.2.2. Attribution des TEF

L'attribution des TEF de la banque préliminaire des entrées extraites de SCOP et de la banque finale de dimères filtrée est faite par le même programme que celui qui a été utilisé dans la première partie, disponible sur le serveur RPBS (http://bioserv.rpbs.jussieu.fr/TEF/).

2.2.3. Découpage des fichiers PDB

Avant de découper les 9000 fichiers PDB pour en extraire les différents TEF, on répertorie le numéro des résidus commençant et finissant chaque TEF pour chaque entrée de la banque. Ceci est nécessaire car les TEF sont parfois chevauchants et sont présentés sur deux lignes en sortie du programme d'assignation. Traiter directement ce format aurait fait appel à une programmation lourde faisant de nombreux allers et retours sur des ouvertures et fermetures de fichiers et cela aurait demandé beaucoup de temps de calcul. La liste des débuts et fins de TEF est ensuite utilisée pour découper chaque structure en comparant les numéros des résidus du fichier PDB source aux limites des TEF listées. Les résidus compris entre ces limites sont copiés dans un nouveau fichier avec toutes les informations associées que contient le fichier PDB source. Les fichiers correspondant aux TEF comprennent dans leur dénomination plusieurs informations : le code PDB d'appartenance, le numéro du modèle de la structure si elle est issue de la RMN et X si elle issue de la cristallographie, la chaîne d'appartenance et le numéro du TEF dans la chaîne protéique. Un exemple : 1ACX_X_A_2

2.2.4 Détermination des cylindres enveloppant

Deux méthodes ont été testées pour décrire les cylindres englobant les TEF. Dans les deux méthodes, l'algorithme prend en entrée le fichier PDB contenant le TEF d'intérêt. Le fichier est parcouru est les coordonnées de chaque carbone alpha sont stockés.

Dans la première méthode on prend comme origine du repère le carbone alpha correspondant au milieu de la séquence qu'on nommera M. On parcourt ensuite les carbones alpha restants et le plus éloigné est dénommé F pour « further ». C'est par ces deux points que l'axe du cylindre passera et leur distance sera enregistrée comme étant la longueur du cylindre. Le rayon est déterminé en parcourant chaque carbone alpha et en calculant sa distance à l'axe du cylindre grâce au produit scalaire des vecteurs CaM et MF. Ce cas est représenté sur la figure 2.2.1.a.

Pour la seconde méthode le centre de gravité G des carbones alpha est calculé ainsi que le point M décrivant le milieu des deux carbones alpha situés aux extrémités du TEF. L'axe du cylindre englobant le TEF est la droite passant par le centre de gravité et le milieu des extrémités M. Le rayon du cylindre englobant est calculé en parcourant chaque carbone alpha. Grâce au produit scalaire des vecteurs CalphaG et GM on calcule la distance de chaque Carbone alpha à l'axe du cylindre. Ce cas est représenté sur la figure 2.2.1.b. Ces valeurs et l'équivalence en numéro de résidu dans le TEF sont stockées puis classées, la plus grande distance d'un carbone alpha à l'axe définit le rayon du cylindre englobant la totalité du squelette protéique du TEF. La longueur du cylindre est calculée par la distance entre les deux carbones alpha les plus proches de l'axe du cylindre. Certes, la distance ne correspond pas exactement à l'axe du cylindre, cependant vu la petitesse des angles en question, cette distance est là une bonne approximation de la hauteur et permet un gain en temps de calcul.

La détermination de la hauteur du cylindre est calculé par la distance des deux carbones alpha les plus proches de l'axe. En effet, les angles entre l'axe du cylindre et l'axe passant par chaque carbone alpha et le centre de gravité sont déjà stockés et classés à ce moment de l'itération. Il suffit de prendre les deux plus petites valeurs des angles, qui correspondent aux deux carbones alpha les plus proches de l'axe du cylindre et donc les plus à même d'estimer la hauteur du cylindre. Une condition a été introduite lorsque les deux plus petits angles sont égaux, ceci correspond aux extrémités du TEF lorsque l'axe du cylindre passe au milieu de deux carbones alpha opposés à ces extrémités. Dans ce cas on choisit le premier et le troisième plus petit angle (souvent le premier et le troisième, les deux premiers étant égaux puisqu'ils correspondent aux extrémités et que l'axe y passe en leur milieu).

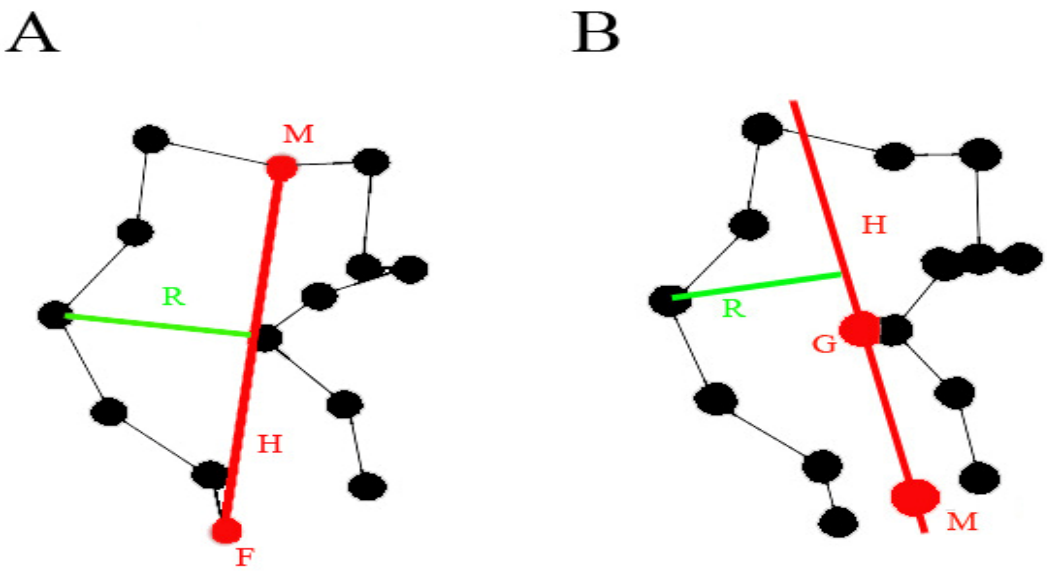


Figure 2.2.1. Deux méthodes pour déterminer le cylindre englobant un TEF. L'axe du cylindre est représenté en rouge. R est le rayon de ce cylindre.

L'élaboration de cette technique a été testée sur une banque de 626 structures extraites de SCOP et représentant des repliements variés. On obtient ainsi 7820 TEF sur lesquels nous avons fait nos tests préliminaires.

La distribution des TEF en fonction de la hauteur du cylindre englobant correspondant a été produite pour les deux méthodes. La figure 2.2.2.A a été faite à partir des données produites par la première méthode et la figure 2.2.2.B avec les données de la seconde méthode. Il y a une grande différence pour la hauteur entre les algorithmes ; dans la figure 2.2.2.A la distribution en hauteur apparait comme une courbe de type gaussien avec un maximum à 20 Angstrom alors que la figure 2.2.2.B montre un pic à 4 Angstrom et une décroissance régulière avec des pics plus petits lorsque la hauteur augmente. Les figures 2.2.2.C et 2.2.2.D décrivant les distributions respectives en fonction du rayon du cylindre montrent quant à elles à peu près la même forme. Les courbes sont toutes deux voisines de gaussiennes et présentent un maximum d'occurrence à 10 Angstrom et 8 Angstrom. La différence entre les deux méthodes décrivant le rayon concerne essentiellement le maximum d'occurrence qui est déplacé ainsi que le rayon maximal qui passe de 24 Angstom pour la première méthode à 29 Angstrom pour la seconde.

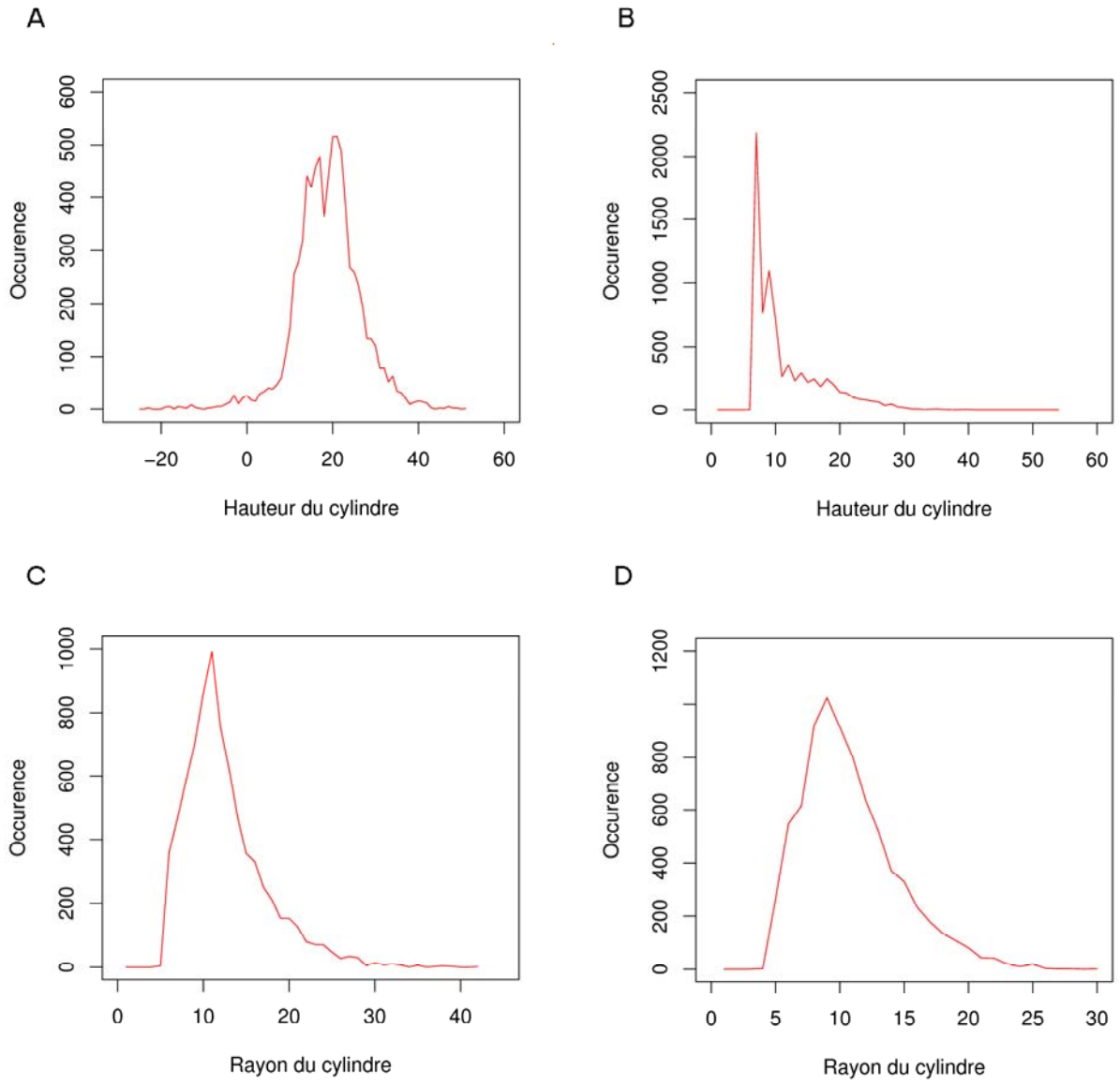


Figure 2.2.2. Distribution du nombre de cylindres englobant les TEF, en fonction de la hauteur (A et B) et en fonction du rayon (C et D) selon la méthode de détermination. A et C) méthode de la plus grande extension ; B et D méthode du centre de gravité.

Afin de mieux analyser la distribution des cylindres englobant les TEF, nous avons établi une carte 2D dans laquelle chaque cylindre est représenté par un point en fonction de son rayon en abscisse et de sa hauteur en ordonnée. La figure 2.2.3.A est la carte correspondant à la première méthode (plus grande extension) alors que la figure 2.2.3.B correspond aux résultats obtenus avec la seconde méthode (centre de gravité).

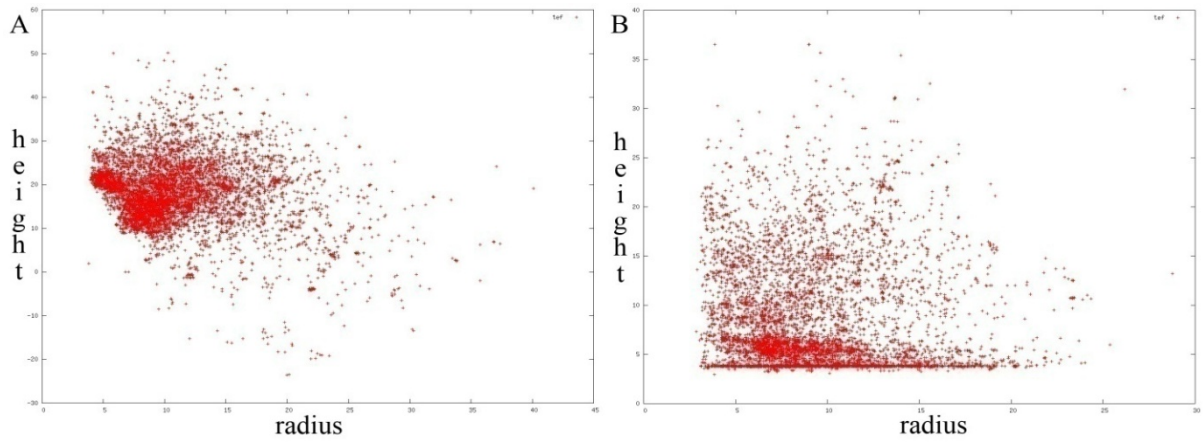


Figure 2.2.3. Cartes dans lesquelles sont représentés les cylindres englobant les TEF, selon la hauteur et le rayon, pour les deux méthodes de détermination des TEF. A) méthode de la plus grande extension ; B) méthode du centre de gravité.

Il apparaît que la distribution provenant de la première méthode est continue et peu avantageuse pour opérer une classification des fragments de protéines tels que les TEF sur des bases structurales. Alors que la distribution produite par la seconde méthode laisse apparaître des strates horizontales, correspondant à la hauteur, et sera donc retenue pour la description des cylindres englobants de TEF.

Les tests des cylindres englobant les TEF et les chaînes latérales n'a pas été étudiée par manque de temps.

2.2.5 Classification des TEF

Selon la répartition en hauteur des cylindres enveloppant

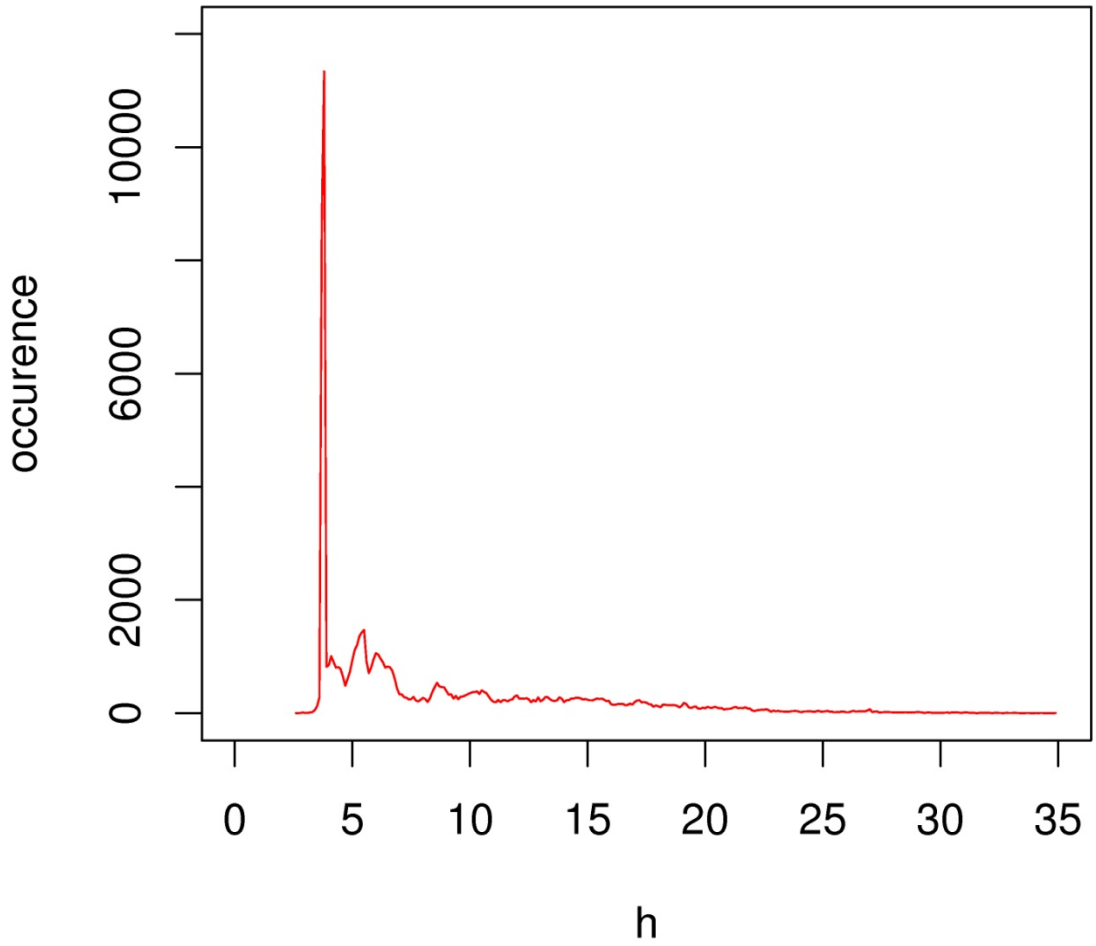


Figure 2.2.4. Répartition des cylindres enveloppant selon leur hauteur par la méthode des centres de gravités sur la banque entière des dimères filtrée par DiMoVo.

Ne sont indiquées sur la figure 2.2.4., qui représente les occurrences des cylindres englobant les TEF, que les hauteurs inférieures à 35 Angströms car les populations pour les tranches supérieures sont trop petites et très peu porteuses d'informations. De plus cela est cohérent avec le fait que les comparaisons

entre les résultats préliminaires sur les protéines de SCOP et les dimères filtrés par DiMoVo ont été faits sur cette même fenêtre de hauteur.

La distribution en hauteur des cylindres englobants a été analysée par pas de 0.1 Angstrom, les points de la courbe de la figure 2.2.4. représentent ainsi la population dans chaque tranche de 0,1 Angstrom.

On observe sur la figure 2.2.4. un maximum à 3,8 Angstrom d'une population de 11339 cylindres enveloppants (le nombre total de TEF 109297). On observe des pics à 5,5 Angstrom et 6 Angstrom mais beaucoup moins peuplés (environ un millier de TEF). Ainsi on s'aperçoit que le modèle utilisé décrit principalement les TEF par des cylindres de très petite hauteur et qui doivent avoir un rayon conséquent, les faisant ressembler à une boîte de cirage pour l'exemple. Les minima locaux ont été calculés. L'algorithme des minima locaux fonctionne de la manière suivante : la courbe est parcourue et dès qu'on observe un minimum local, c'est-à-dire que la pente de la courbe passe de négative à positive, la position est répertoriée. Ensuite on parcourt la liste des minima trouvés et les nouveaux minima de cette dernière liste sont calculés. Il faut ainsi quatre itérations pour définir 11 minima locaux de la distribution en hauteur par tranche de 0 ,1 Angstrom. Ce qui nous permet de produire 12 classes de cylindres qui seront notées de A à L.

2.2.6 Analyse des contacts par tesselation de Voronoï

Les contacts entre résidus d'une même chaîne et de plusieurs chaînes d'un complexe sont analysés grâce à la tesselation de Voronoï décrite plus haut. Ces contacts sont définis par toutes les cellules de Voronoï partageant une face. La tesselation de Voronoï a été réalisée par le programme VORO3D qui marche en deux grandes phases. La première correspond à l'installation d'un environnement pour délimiter les cellules de bords de la protéine. Elle produit des fichiers dénommés selon le modèle R (exemple R1enh_8 pour la huitième relaxation de 1enh) code pdb_numéro de la phase de relaxation. Seuls les fichiers correspondant à la neuvième relaxation sont pris en entrée pour la deuxième phase du programme. Cette dernière phase est la construction des cellules de Voronoï à proprement parler. Les fichiers résultats sont dénommés selon le modèle code pdb avec l'extension « .out ». La figure 2.2.5. donne un exemple de fichier de sortie fourni par VORO3D :

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
M	0	1		206.353	191.108	14	1	-> V	1	2		6	28.786	20.785	4.693
M	0	1		206.353	191.108	14	1	-> L	2	3		3	4.223	10.576	6.657
M	0	1		206.353	191.108	14	1	-> K	133	134		6	15.660	15.430	6.335
M	0	1		206.353	191.108	14	1	-> E	136	137		5	8.108	11.396	7.243
M	0	1		206.353	191.108	14	1	-> L	137	138		6	26.764	19.735	4.994
M	0	1		206.353	191.108	14	1	-> K	140	141		5	8.800	12.745	7.151
M	0	1		206.353	191.108	14	1	-> U	4	-1	&	5	16.966	16.908	6.301
M	0	1		206.353	191.108	14	1	-> U	242	-1	&	5	7.112	11.614	7.406
M	0	1		206.353	191.108	14	1	-> U	397	-1	&	6	19.221	17.243	7.189
M	0	1		206.353	191.108	14	1	-> U	477	-1	&	5	16.504	16.222	7.491
M	0	1		206.353	191.108	14	1	-> U	481	-1	&	4	2.786	6.831	9.619
M	0	1		206.353	191.108	14	1	-> U	550	-1	&	5	5.666	9.506	9.321
M	0	1		206.353	191.108	14	1	-> U	597	-1	&	5	18.519	16.649	7.003
M	0	1		206.353	191.108	14	1	-> U	661	-1	&	6	11.994	13.622	7.479
V	1	2		152.260	167.523	11	1	-> M	0	1		6	28.786	20.785	4.693
V	1	2		152.260	167.523	11	1	-> L	2	3		7	30.265	20.870	4.491
V	1	2		152.260	167.523	11	1	-> S	3	4		4	5.873	10.483	6.772
V	1	2		152.260	167.523	11	1	-> G	80	81		5	5.622	9.972	8.431
V	1	2		152.260	167.523	11	1	-> K	133	134		3	0.391	3.019	7.595
V	1	2		152.260	167.523	11	1	-> L	137	138		4	0.703	3.689	7.925
V	1	2		152.260	167.523	11	1	-> U	4	-1	&	5	15.043	15.800	6.737
V	1	2		152.260	167.523	11	1	-> U	213	-1	&	6	28.181	20.615	5.578
V	1	2		152.260	167.523	11	1	-> U	242	-1	&	6	28.905	21.009	4.865
V	1	2		152.260	167.523	11	1	-> U	621	-1	&	4	5.401	11.551	7.498
V	1	2		152.260	167.523	11	1	-> U	661	-1	&	4	18.353	17.640	5.838
L	2	3		139.858	154.308	14	1	-> M	0	1		3	4.223	10.576	6.657
L	2	3		139.858	154.308	14	1	-> V	1	2		7	30.265	20.870	4.491
L	2	3		139.858	154.308	14	1	-> S	3	4		6	20.290	17.574	4.362
L	2	3		139.858	154.308	14	1	-> E	6	7		5	14.563	15.803	4.724
L	2	3		139.858	154.308	14	1	-> W	7	8		7	24.947	18.951	4.869
L	2	3		139.858	154.308	14	1	-> K	79	80		3	0.161	1.944	7.712
L	2	3		139.858	154.308	14	1	-> G	80	81		6	8.158	11.618	8.093
L	2	3		139.858	154.308	14	1	-> A	130	131		4	0.916	4.088	7.209
L	2	3		139.858	154.308	14	1	-> K	133	134		8	21.569	17.889	5.279
L	2	3		139.858	154.308	14	1	-> A	134	135		5	8.868	12.864	6.765
L	2	3		139.858	154.308	14	1	-> L	137	138		5	6.366	10.440	7.584
L	2	3		139.858	154.308	14	1	-> U	226	-1	&	4	1.071	4.455	7.461
L	2	3		139.858	154.308	14	1	-> U	242	-1	&	4	5.461	10.052	6.861
L	2	3		139.858	154.308	14	1	-> U	621	-1	&	5	7.450	11.202	7.367
S	3	4		153.284	155.142	14	1	-> V	1	2		4	5.873	10.483	6.772

Figure 2.2.5. Exemple d'un fichier de sortie de VORO3D. La première ligne de numéro en gras correspond à la numérotation des colonnes.

Ainsi chaque résidu de la chaîne protéique est listé (colonne 1 à 3) et chaque voisin partageant une face de cellule est associé par une flèche selon le format VORO3D. On peut obtenir ainsi les informations suivantes : le nom du résidu (colonne 1), son numéro dans le fichier PDB (colonne 2), son numéro total dans la chaîne (colonne 3), la chaîne à laquelle il appartient (colonne 4), le volume de la cellule de Voronoï (colonne 5), la surface totale de la cellule (colonne 6), et le nombre de faces (colonne 7). Chaque résidu est répété autant de fois qu'il a de voisins différents, les mêmes informations sont disponibles pour le dit voisin. Y sont ajoutées des informations quant à la surface commune du couple de voisins, telles que : le nombre de côtés de la face (colonne 13), la surface (colonne 14), le périmètre de la face (colonne 15) et la distance entre les deux points représentant les résidus en contact (colonne 16). Ces fichiers sont utilisés par un script domestique afin de répertorier les interactions entre TEF au sein des complexes. Les concordances entre résidus listés dans les

fichiers de sortie de VORO3D et les résidus correspondant dans les fichiers de sortie d'attribution de TEF sont établies. On remonte ainsi à l'information sur les interactions entre TEF du complexe protéique. Les fichiers de sortie de VORO3D listent tous les résidus du complexe, de nombreuses redondances sont présentes.

2.2.7 Attribution des interactions correspondant aux différentes classes

J'ai programmé un module de nettoyage des redondances dans les interactions pour le script chargé de répertorier ces dernières, ce qui permet de produire des fichiers de sortie beaucoup moins lourds pour l'analyse ultérieure des interactions de TEF. De plus, les contacts de résidus adjacents sont éliminés. En effet, sur une banque de 9000 complexes, il est préférable que les fichiers de résultats d'interaction de TEF soient le plus concis possible. D'autre part, ils ne sont pas informatifs en ce qui concerne le repliement des protéines.

La phase suivante est de relier ces informations d'interactions entre TEF aux informations de classification des dits TEF. Pour ce faire, une liste des résultats des TEF enveloppés dans leur cylindre descriptif est produite avec l'information de leur appartenance à telle ou telle classe. Cette liste contient toutes les entrées de TEF qui ont été préalablement découpées et analysées, à chaque TEF correspond sa classe d'appartenance et selon le code de dénomination du TEF expliqué précédemment, on peut remonter au complexe, à la chaîne et au numéro du TEF. En croisant ces informations avec celles produites par l'analyse des interactions de TEF, on peut estimer quelle classe de TEF interagit avec telle autre classe et en quelle proportion.

2.2.8 Comptabilisation des interactions par couples de classes

Les fichiers de sortie du script interact_tef sont traités par un nouveau script (class_tef_interact) qui produit de nouveaux fichiers de sortie où sont listées les interactions mais seules les informations de classe y sont répertoriées. On remonte à ces informations de classe par la liste de TEF classifiée. Il suffit ensuite de parcourir tous ces fichiers et de stocker chaque type interaction, c'est à dire la paire de classes de TEF qui interagissent. Avec 12 classes de TEF, on peut avoir 78 types d'interactions dans une variable. Ensuite les propensions des interactions de chaque type de classe sont calculées et une matrice d'interaction de classe peut être produite.

2.3 Résultats

2.3.1 Résultats généraux

2.3.1.1. Numérotation des différents multimères

Sur les 18 626 multimères de la liste de Thornton, seulement 9424 ont été validés comme multimères biologiques par le programme DiMoVo.

Répartition multimérique :

	Homo multimères	Hétéro multimères
Dimères	4432	845
Trimères	508	562
Tetramères	1252	626
Pentamères	67	71
Héxamères	345	206
Octamères	161	77
Nonamères	1	17
Dodécamères	54	112

Tableau 2.3.1.1 Répartition multimérique.

Sur 9424 dimères biologiques, 9146 donnent des résultats de TEF dont 3249 ont dû être renumérotés. Ce qui donne un total de 109 297 TEF.

2.3.1.2 Énumération des TEF par classe

Classe	Nombre de TEF
A	35624
B	31793
C	6337
D	6628
E	5075
F	4309
G	6659
H	4024
I	2725
J	1353
K	2506
L	2264

Tableau 2.3.1.2 Population des différentes classes de TEF.

Ce tableau nous donne une idée des populations au sein de chaque classe de TEF et nous permettra de calculer les proportions dans les interactions de ces derniers.

2.3.2 Interactions entre TEF dans les complexes

Les interactions entre les différentes chaînes de protéines au sein des complexes biologiques montrent que l'on trouve souvent un TEF à l'interface protéine – protéine. Sur l'ensemble des TEF décrits, 16129 se retrouvent à des interfaces protéine - protéine et interagissent avec d'autres TEF à ces interfaces. Cependant il est rare que seuls deux TEF, un dans chaque chaîne en interaction, interagissent à l'interface. Ceci est un peu désappointant, car le modèle de prédiction est censé se fonder sur l'orientation relative de deux TEF en interaction pour la construction des modèles. Il est quand même envisageable de comptabiliser le nombre d'interactions pour chaque couple de TEF et de choisir le plus favorable pour base à la construction des modèles.

2.3.3 Interaction des différentes classes de TEF

Quantification

AA 29890	AL 4756	BL 4049	DD 1334	EG 2653	FK 426	HK 1351
AB 44401	BB 22817	CC 1200	DE 1498	EH 1243	FL 300	HL 274
AC 8612	BC 9620	CD 2441	DF 939	EI 220	GG 3681	II 150
AD 8931	BD 8347	CE 2228	DG 1549	EJ 619	GH 1428	IJ 207
AE 10242	BE 4771	CF 2207	DH 457	EK 417	GI 797	IK 1473
AF 5859	BF 6411	CG 2007	DI 467	EL 693	GJ 604	IL 198
AG 11639	BG 7252	CH 1817	DJ 322	FF 620	GK 1634	JJ 76
AH 4369	BH 4172	CI 890	DK 562	FG 1318	GL 649	JK 36
AI 4518	BI 3659	CJ 766	DL 756	FH 502	HH 651	JL 198
AJ 2239	BJ 2492	CK 882	EE 689	FI 345	HI 699	KK 563
AK 5582	BK 4206	CL 876	EF 996	FJ 389	HJ 240	KL 252
						LL 1032

Tableau 2.3.3.1 Population des différentes paires de classes en interaction.

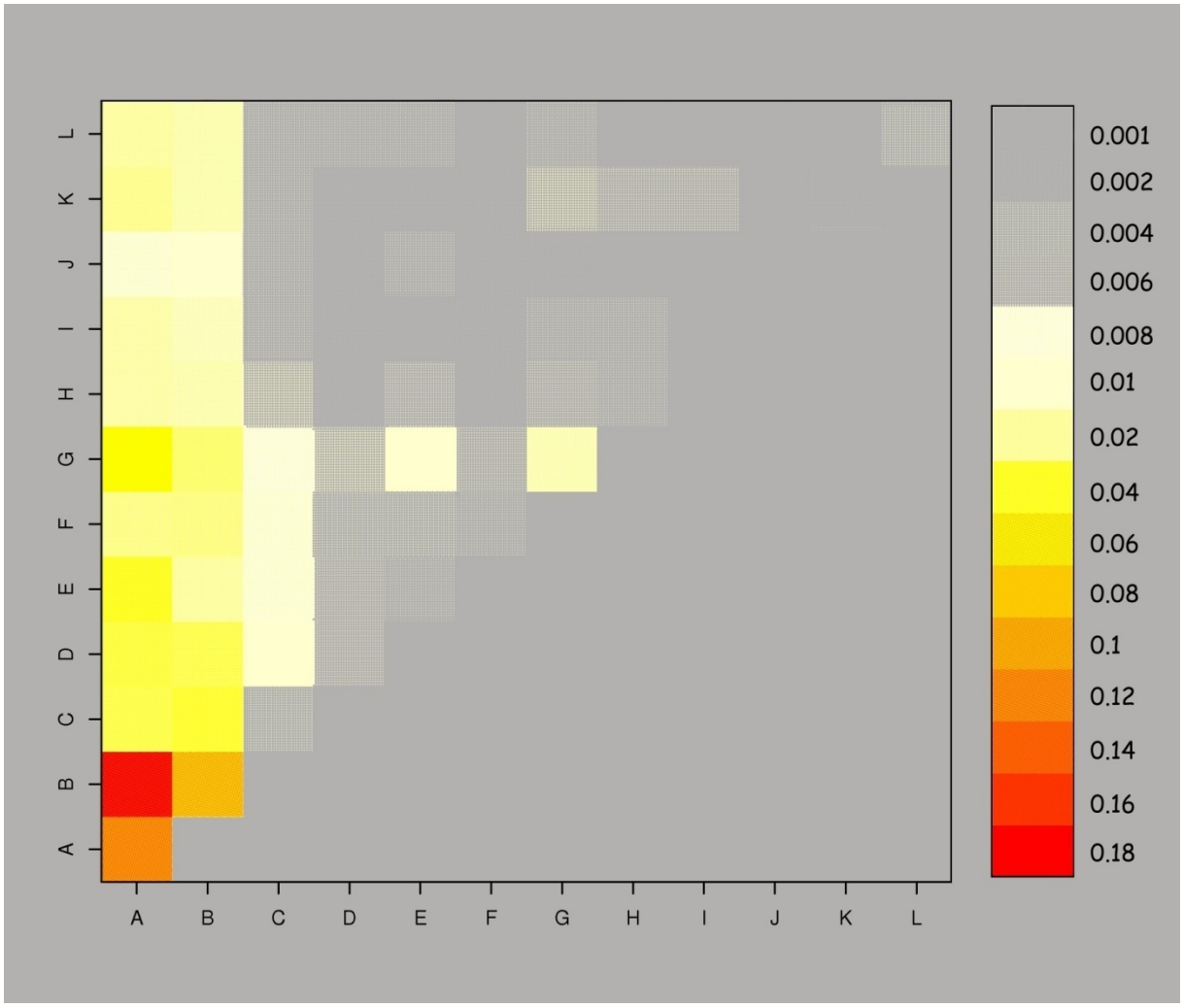


Figure 2.2.5. Carte d'interactions des classes de TEF. Le code couleur indiquant la propension des interactions est représenté sur la droite.

La carte d'interactions des différentes classes de TEF de dimères biologiques est représentée sur la figure 2.2.5. Comme on peut le voir, les classes interagissant le plus sont les classes A et B. Ceci est logique puisque conformément à la répartition en hauteur des cylindres enveloppants, ces deux classes correspondent aux deux premiers pics qui sont les plus peuplés. Chose un peu plus étonnante, c'est le comportement de la classe G qui interagit assez entre elle-même et avec les autres. Autre résultat étonnant c'est la zone de désert dans la carte correspondant aux interactions des classes allant de H à L avec les classes allant de C à F. Il en résulte que ces classes ne sont pas favorisées dans leurs interactions, probablement pour des motifs géométriques car ces classes ont des longueurs grandes dans la classification.

2.4 Discussion

Notre étude structurale des protéines est fondée sur l'analyse du squelette peptidique. C'est tout naturellement que nous n'avons considéré que le squelette pour la description structurale des TEF par la méthode des cylindres enveloppant. Il aurait été intéressant de prendre en compte les chaînes latérales et de comparer les résultats. Cependant, seul les résultats des rayons des cylindres auraient changé alors que la longueur serait restée sensiblement la même, or c'est sur le critère de la longueur qu'est fondé notre outil de classification. La prise en compte des chaînes latérales pour le calcul de la longueur aurait demandé une refonte complète de l'algorithme et nous aurions pu perdre notre critère de classification.

La matrice d'interaction de classes a été calculée en proportion des interactions par rapport à la population totale de TEF. Pour vraiment apprécier les interactions favorisées, il faudrait pondérer le nombre des interactions d'un couple de classes par les populations propres aux classes de ce couple.

Ce travail de classification structurale des TEF aux interfaces des protéines génère des travaux en perspective. L'analyse des orientations entre TEF en interaction sera grandement facilitée par l'algorithme des cylindre enveloppants. En effet l'orientation relative des axes des cylindres en interaction pourra servir de modèle de référence pour la prédiction de structures quaternaires de domaines homologues à ceux existant dans notre banque. Cependant il reste à tester la méthode de sélection du couple de TEF ayant le plus de poids dans l'interface (comme il est dit plus tôt, il arrive que plusieurs couples de TEF soient à l'interface). Un outil émanant de ces travaux de recherche pourrait être développé afin de construire les modèles et de sélectionner le meilleur en opérant des rotations autour des axes des cylindres enveloppant. La sélection se fera essentiellement par l'abolition des gènes stériques, mais d'autres paramètres comme la surface d'interface pourront être envisagés.

2.5. Conclusion

Au cours de ces travaux représentant la seconde partie de mon travail de thèse, on a pu constituer une banque de multimères biologiques. Cette banque devrait être disponible sur RPBS et ainsi être partagée et utilisée pour toute étude statistique sur les structures quaternaires. Une deuxième banque a été créée à partir de la première, c'est la banque de fragments protéiques correspondant aux TEF des différents multimères. De nombreuses pistes ont été étudiées afin de pouvoir classifier ces fragments

de manière structurale ou géométrique. Cependant, faute de budget pour une collaboration avec des équipes de physiciens et de chimistes, c'est celle qui restait dans notre domaine de compétence qui a été adoptée. Cette méthode est celle des cylindres enveloppants. Plusieurs algorithmes conduisant à l'enveloppage des fragments ont été testés et c'est celui donnant les résultats les plus enclins à amener une classification qui a été retenu. L'analyse de ces résultats a mené à une classification selon la hauteur des cylindres enveloppants. On s'est aperçu qu'une classe était bien plus peuplée que les autres et cela a un peu changé la vision que l'on avait des structures de TEF. Les interactions au niveau des surfaces protéines–protéines ont été étudiées relativement aux classes de fragments. Il est évident que la classe la plus peuplée interagit le plus, cependant les autres clases en interactions ne donnent pas des résultats directement proportionnels à la population des classes. En effet la diagonale de la matrice ne correspond pas à la population des classes. Il est ainsi permis de penser qu'il existe des géométries de TEF utilisées préférentiellement aux interfaces protéine – protéine.

Remerciements

Je remercie les membres du Jury de s'être déplacé et d'avoir eu un œil critique sur mon travail, le Pr. Jean-François Zagury, le Pr. Manuel Dauchez, le Dr. Nikolaos Papandreou, et le Dr. Joël Pothier. Je remercie tout particulièrement le Président du Jury, le Pr. Thierry Foulon, qui fut aussi mon responsable de Master Recherche et sans qui rien de tout ça n'aurait été possible. En effet c'est lui qui m'a défendu et soutenu lors de mon passage « chaotique » en Master Recherche.

Je tiens à remercier le Dr Jacques Chomilier pour m'avoir encadré pendant mon stage de Master, de m'avoir permis de présenter un projet de thèse sous sa direction et pour son encadrement tout au long de mon travail de recherche. Je remercie l'école doctorale B2M de m'avoir sélectionné pour la bourse ministérielle. Je remercie le Dr Bernard Capelle, directeur de l'IMPMC, pour son accueil au sein de l'institut depuis mon stage de Master.

Je tiens à remercier tout particulièrement le Dr. Mathieu Lonquety pour ses nombreux conseils et aides en programmation, mais aussi pour être devenu un véritable ami. Je remercie également les équipes avec qui j'ai pu collaboré, telle que celle du Dr. Anne Poupon et du Dr. Julie Bernauer, mais aussi celle du Dr. Jean-François Sadoc et du Dr. David Perahia. Ils ont mis a disposition des logiciels indispensables à mon travail et pour le dernier a été prêt à soumettre un projet commun.

Je remercie toute l'équipe de Prédiction de Structures Protéiques, le Dr. Isabelle Callebaut, le Dr. Jean-Paul Morno, le Dr. Françoise Schoentgen, le Dr. Elodie Duprat et le Dr. Stephanie Finet, ainsi que le Dr. Richard Eudes, le Dr. Denis Znamenskiy et le Dr. Guillaume Fourty qui on été en thèse dans l'équipe.

Je remercie le Dr. Eric Larquet pour ses conseils, mais aussi sa bonne humeur lors de discussions plus légères. Je tiens à rendre hommage au Dr. Nicolas Boisset qui instaurait une ambiance de travail la plus agréable que j'ai pu connaître. Je remercie le reste de l'équipe de Structure Assemblage des Macromoléculaires, le Dr. Magali Cottevielle, le Dr. Ricardo Aramayo, qui sont allé continer leurs recherches à l'étranger et avec qui j'ai eu beaucoup de plaisir lors de nos fréquentes joutes verbales. Ainsi que le Dr. Slavica Jonić, qui malgré la prise de responsabilité toujours croissante a su rester tout aussi accessible. Je remercie Zaelle Devaux pour son soutient rythmé par nos pauses cigarettes et les très bons moments passés lors de nos pots du vendredi.

Je remercie le Dr. Caroline Magnain pour m'avoir soutenu pendant la rédaction du manuscrit alors qu'elle était elle aussi en train de rédiger, pour sa bonne humeur même pas altérée par une épaule cassée.

Je remercie aussi les stagiaires qui sont passés par le laboratoire, Vincent Leduc, Anne Campagna, Jemila Houacine, Stephanie Pérot, Anthony Cathala et Florence Dol. Ainsi que l'ensemble du laboratoire pour son accueil et la bonne ambiance qui règne (surtout à la cafétéria).

Je tiens à remercier ma famille pour m'avoir soutenu et encouragé dans mes démarches, tout particulièrement mes parents qui se sont toujours occupé de moi au moment où j'en avais le plus besoin.

Je remercie Coraline pour les câlins et les jeux pendant les pauses rédaction à la maison.

Je remercie enfin Tiana Popović pour m'avoir supporté et soutenu depuis mon Master et avec j'espère faire ma vie de chercheur.

BIBLIOGRAPHIE

Abkevich, V., Gutin, A. & Shakhnovich, E. (1994). Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry* **33**, 10026-10036.

Alland, C., Moreews, F., Boens, D., Carpentier, M., Chiusa, S., Lonquety, M., Renault, N., Wong, Y., Cantalloube, H., Chomilier, J., Hochez, J., Pothier, J., Villoutreix, B., Zagury, J.-F. & Tufféry., P. (2005). RPBS: a web resource for structural bioinformatics. *Nucleic Acids Res* **33**, W44-W49.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science* **181**, 223-30.

Anfinsen, C. B., Haber, E., Sela, M. & White, F. H. (1961). The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci. USA* **47**, 1309-1314.

Angelov, B., Sadoc, J.-F., Jullien, R., Soyer, A., Mornon, J.-P. & Chomilier, J. (2002). Voronoï tessellation of proteins: a novel concept for analysis of protein folding. *Proteins* **49**, 446-452.

Bahadur, R. & Zacharias, M. (2008). The interface of protein-protein complexes: analysis of contacts and prediction of interactions. *Cell Mol Life Sci* **65**, 1059-1072.

Berezovsky, I. & Trifonov, E. (2002). Loop fold structure of proteins: resolution of Levinthal's paradox. *J Biomolec Struct Dynamics* **20**, 5-6.

Berezovsky, I. N., Grosberg, A. Y. & Trifonov, E. N. (2000). Closed loops of nearly standard size : common basic element of protein structure. *FEBS Letters* **466**, 283-286.

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

Bernauer, J., Bahadur, R. P., Rodier, F., Janin, J. & Poupon, A. (2008). DiMoVo: a Voronoi tessellation based method for discriminating crystallographic and biological protein protein interactions. *Bioinfomatics* **24**, 652-658.

Boutselakis, H. (2003). E-MSD: the European Bioinformatics Institute Macromolecular Structure Database. *NAR* **31**, 458-462.

Bresler, S. & Talmud, D. (1944a). On the nature of globular proteins. I. *comptes rendus (Dokady) de l'Acdémie des sciences de l'URSS* **43**, 310, 349.

Bresler, S. & Talmud, D. (1944b). On the nature of globular proteins. II A few consequences of the new hypothesis. *comptes rendus (Dokady) de l'Acdémie des sciences de l'URSS* **43**, 349-350.

Callebaut, I., Courvalin, J., Worman, H. & Mornon, J. (1997). Hydrophobic cluster analysis reveals a third chromodomain in the Tetrahymena Pdd1p protein of the chromo superfamily. *Biochem Biophys Res Commun.* **235**, 103-107.

Carpentier, M., Brouillet, S., & Pothier, J., Alignement multiple de familles protéiques, JOBIM 2005.

Cazals, F., Proust, F., bahadur, R. & Janin, J. (2007). Revisiting the Voronoi description of protein protein interfaces. *Prot. Sci.* **15**, 2082-2092.

Chen, J., Bryngelson, J. D. & Thirumalai, D. (2008). Estimations of the Size of Nucleation Regions in Globular Proteins. *J. Phys. Chem. B* **112**, 16115-16120.

Chiche, L., Gregoret, L., Cohen, F. & Kollman, P. (1990). Protein model structure evaluation using the solvation free energy of folding. *Proc Natl Acad Sci U S A* **87**, 3240-3243.

Chomilier, J., Lamarine, M., Mornon, J.-P., Torres, J. H., Eliopoulos, E. & Papandreou, N. (2004). Analaysis of fragments induced by simulated lattice protein folding. *Comptes Rendus Acad Sci* **327**, 431-443.

Chothia, C., Gelfand, I. & Kister, A. (1998). Structural deteminants in the sequences of immunoglobulin variable domains. *J. Mol. Biol.* **278**, 457-479.

Chothia, C., Lesk, A., Tramontano, A., Levitt, M., Smith-Gill, S., Air, G., Padlan, E., Davies, D., Tulip, W., Colman, P., Spinelli, S., Alzari, P. & Poljak, R. (1989). Conformations of immunoglobulin hypervariable regions. *Nature* **342**, 877-883.

Cota, E., Steward, A., Fowler, S. & Clarke, J. (2001). The folding nucleus of a fibronectin type III domain is composed of core residues of the immunoblobulin-like fold. *J. Mol. Biol.* **305**, 1185-1194.

Deret, S., Denoroy, L., Lamarine, M., Vidal, R., Mougenot, B., Frangione, B., Stevens, F. J., Ronco, P. M. & Aucouturier, P. (1999). Kappa light chain-associated Fanconi's syndrome: molecular analysis of monoclonal immunoglobulin light chains from patients with and without intracellular crystals. *Protein Eng* **12**, 363-369.

Dill, K. & Chan, H. (1997). Perspective: From Levinthal to pathways to funnel. *Nat. Strcut. Bio.* **4**, 4.

Ding, F., Dokholyan, N., Buldyrev, S., Stanley, E. & Shakhnovich, E. (2002). Molecular dynamics simulation of the SH3 domain aggregation suggests a generic amyloidogenesis mechanism. *J . Mol. Biol.* **324**, 851-857.

Dupuis, F., Sadoc, J. F. & Mornon, J. P. (2004). Protein secondary structure assignment through Voronoï tessellation. *Proteins* **55**, 519-528.

Efron, B. & Tibshirani, R. (1993). *An introduction to the bootsrap*. Monographs on statistics and applied probability (Hall, C., Ed.).

Efron, N. & Tibshirani, R. (1997). Improvements on cross-validation. *J Am Stat Assoc* **92**, 548-550.

Fersht, A. & Sato, S. (2004). Φ-value analysis and the nature of protein folding transition states. *Proceedings Natl. Acad. Sci. USA* **101**, 7976-7981.

Fersht, A. R. (1997). Nucleation mechanisms in protein folding. *Curr Opinion Struct Biol* **7**, 3-9.

Fersht, A. R. (2000). Transition state structure as a unifying basis in protein folding mechanisms: contact order, chain topology, stability and the extended nucleus mechanism. *Proceedings Natl. Acad. Sci. USA* **97**, 1525-1529.

Fourty, G. (2006). Recherche de contraintes structurales pour la modélisation "ab initio" du repliement protéique, Paris 7.

Fowler, S. & Clarke, J. (2001). Mapping the folding pathway of an immunoglobulin domain: structural detail from phi value analysis and movement of the transition state. *Structure* **9**, 355-366.

Geierhaas, C., Paci, E., Vendruscolo, M. & Clarke, J. (2004). Comparison of the transition state for folding ot two Ig like proteins from different superfamilies. *J. Mol. Biol.* **343**, 1111-1123.

Gerstein, M. & Altman, R. (1995). Average core structures and variability measures for protein families: application to the immunoglobulins. *J. Mol. Biol* **251**, 161-175.

Gracy, J., Chiche, L. & Sallantin, J. (1993). Improved alignment of weakly homologous protein sequences using structural information. *Prot Engng* **6**, 821-829.

Halaby, D. M., Poupon, A. & Mornon, J. P. (1999). The immunoglobulin superfamily: sequence analysis and 3D structure comparisons. *Protein Eng* **12**, 563-571.

Hamill, S., Steward, A. & Clarke, J. (2000). The folding of an immunoglobulin like Greek key protein is defined by a common core nucleus and regions constrained by topology. *J. Mol. Biol.* **297**, 165-178.

Haspel, N., Tsai, C., Wolfson, H. & Nussinov, R. (2003a). Hierarchical protein folding pathways: a computational study of protein fragments. *Proteins* **51**, 203-215.

Haspel, N., Tsai, C.-J., Wolfson, H. & Nussinov, R. (2003b). Reducing the computational complexity of protein folding via fragment folding and assembly. *Prot Sci* **12**, 1177-1187.

Henrick, K. & Thornton, J. (1998). PQS: a protein quaternary structure file server. *Trends Biochem Sci* **23**, 358-361.

Ikura, T., Go, N., Kohda, D., Inagaki, F., Yanagawa, H., Kawabata, M., Kawabata, S., Iwanaga, S., Noguti, T. & Go, M. (1993). Secondary structural features of modules M2 and M3 of barnase in solution by NMR experiment and distance geometry calculation. *Proteins* **16**, 341-356.

Ittah, V. & Haas, E. (1995). Nonlocal interactions stabilize long range loops in the initial folding intermediates of reduced bovine pancreatic trypsin inhibitor. *Biochemistry* **34**, 4493-4506.

Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. (1995). The sutructure of the transition state for folding of chmotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation condensation mechanism for protein folding. *J. Mol. Biol.* **25**, 260-288.

Jennings, P. & Wright, P. (1993). Formation of a molten globule intermediate early in the kinetic folding pathway of apomyoglobin. *Scinece* **262**, 848-849.

Jones, S., Marin, A. & Thornton, J. (2000). Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Eng* **13**, 77-82.

Jones, S. & Thornton, J. (1996). Principles of protein-protein interactions. *Proc Natl Acad Sci U S A* **93**, 13-20.

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

Karplus, M. & Weaver, D. L. (1996). Protein folding dynamics : the diffusion collision model and experimental data. *Prot. Sci.* **3**, 650-668.

Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. *Adv Protein Chem* **14**, 1-63.

Kawabata, T. (2003). MATRAS: a program for protein 3D structure comparison. *NAR* **31**, 3367-3369.

Kifer, I., Nussinov, R. & Wolfson, H. (2008). Constructing templates for protein structure prediction by simulation of protein folding pathways. *Proteins* **73**, 380-394.

Kim, P. & Baldwin, R. (1982). Specific intermediates in the folding ractions fo small proteins and the mechanism of protein folding. *Annu Rev Biochem* **51**, 459-489.

Kippen, A., Sancho, J. & Fersht, A. (1994). Folding of barnase in parts. *Biochemistry* **33**, 3778-3786.

Kolinski, A. & Skolnick, J. (1994). Monte Carlo simulation of protein folding. I Lattice model and interaction scheme. *Proteins* **18**, 338-352.

Konagurthu, A. S., Whisstock, J. C., Stuckey, P. J. & Lesk, A. M. (2006). MUSTANG: A multiple structural alignment algorithm. *Proteins* **64**, 559-574.

Koshi, J. M. & Goldstein, R. A. (1997). Mutation matrices and physical-chemical properties: correlations and implications. *Proteins* **27**, 336-44.

Kwasigroch, J.-M., Chomilier, J. & Mornon, J.-P. (1996). A global taxonomy of loops in globular proteins. *J. Mol. Biol.* **259**, 855-872.

Ladunga, I. & Smith, R. F. (1997). Amino acid substitutions preserve protein folding by conserving steric and hydrophobicity properties. *Protein Eng* **10**, 187-96.

Lamarine, M., Mornon, J.-P., Berezovsky, I. N. & Chomilier, J. (2001). Distribution of tightened end fragments of globular proteins statistically match that of topohydrophobic positions: towards an efficient punctuation of protein folding? *Cell. Mol. Life sci.* **58**, 492-498.

Lappalainen, I., Hurley, M. & Clarke, J. (2008). Plasticity within the obligatory folding nucleus of an immunoglobulin like domain. *J. Mol. Biol.* **375**, 547-559.

Laskowski, R. (1996). SURFNET. A program for visualizing moleuclar surfaces, cavities and intermolecular interactions. *J Mol Graph* **13**, 323-330.

Lesk, A. & Rose, G. (1981). Folding unit in globular proteins. *Proceedings Natl. Acad. Sci. USA* **78**, 4304-4308.

Levinthal, C. (1968). Are there pathways for protein folding? *J. Chim. Phys.* **65**, 44-45.

Li, L. & Shakhnovich, E. I. (2001). Different circular permutations produced different folding nuclei in proteins : a computational study. *J. Mol. Biol.* **306**, 121-132.

Liu, Y. & Eisenberg, D. (2002). 3D domain swapping: as domains continue to swap. *Prot Sci* **11**, 1285-1299.

Luthy, R., Bowie, J. U. & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature* **356**, 83-85.

Maupetit, J., Tufféry, P. & Derreumaux, P. (2007). A corase grained protein force field for folding and structure prediction. *Proteins* **69**, 394-408.

McLachlan, A. (1972). Mathematical procedure for superimposing atomic coordinates of proteins. *Acta Cryst A* **28**, 656-657.

Melo, F., Devos, D., Depiereux, E. & Feytmans, E. (1997). ANOLEA: a www server to assess protein structures. *Proc Int Conf Intell Syst Mol Biol* **5**, 187-190.

Melo, F. & Feytmans, E. (1998). Assessing protein structures with a non local atomic interaction energy. *J Mol Biol* **277**, 1141-1152.

Mirny, L. & Shakhnovich, E. (1999). Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol.* **291**, 177-191.

Miyazawa, S. & Jernigan, R. L. (1996). Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term for simulation and threading. *J. Mol. Biol* **256**, 623-644.

Moult, J. (2005). A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *COSB* **15**, 285-289.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.

Notredame, C., Higgins, D. & Heringa, J. (2000). T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205-217.

Orengo, C., Michie, A., Jones, S., Jones, D., Swindells, M. & Thornton, J. (1997). CATH: A hierarchical classification of protein domain structures. . *Structure* **5**, 1093-1108.

Orengo, C. A., Martin, A. M., Hutchinson, G., Jones, S., Jones, D. T., Michie, A. D., Swindells, M. B. & Thornton, J. M. (1998). Classifying a protein in the CATH database of domain structures. *Acta Crystallogr D Biol Crystallogr* **54**, 1155-1167.

Ortiz, A., Strauss, C. & Olmea, O. (2002). MAMMOTH (Matching molecular models obtained from theory): an automated method for model comparison. *Protein Science* **11**, 2606-2621.

Panchenko, A., Luthey-Shulten, Z. & Wolynes, P. (1996). Foldons, protein structural modules, and exons. *Proceedings Natl. Acad. Sci. USA* **93**, 2008-2013.

Papandreou, N., Eliopoulos, E., Berezovsky, I., Lopes, A. & Chomilier, J. (2004). Universal positions in globular proteins : observation to simulation. *Eur. J. Biochem.* **271**, 4762-4768.

Poupon, A. & Mornon, J. P. (1998). Populations of hydrophobic amino acids within protein globular domains; identification of conserved "topohydrophobic" positions. *Proteins* **33**, 329-342.

Poupon, A. & Mornon, J. P. (1999a). Predicting the protein folding nucleus from sequences. *FEBS Lett.* **452**, 283-289.

Poupon, A. & Mornon, J. P. (1999b). "Topohydrophobic positions" as key markers of globular protein folds. *Theoretical Chemistry Accounts* **101**, 2-8.

Qiu, J., Hue, M., Vert, J. & Noble, W. (2007). A structural alignment kernel for protein structures. *Bioinformatics* **23**, 1090-1098.

Ramachandran, G. N. & Sasisekharan, V. (1968). Conformation of polypeptides and proteins. *Adv. Prot. Chem.* **23**, 283-437.

Ramakrishnan, C. & Ramachandran, G. N. (1965). Stereochemical criteria for polypeptide and protein chain conformations. II. Allowed conformations for a pair of peptide units. *Biophys. J.* **5**, 909-932.

Richmond, T. (1984). Solvent accessible surface area and excluded volume in proteins. *J. Mol. Biol* **178**, 63-89.

Rooman, M., Dehouck, Y., Kwasigroch, J. M., Biot, C. & Gilis, D. (2002). What is paradoxical about the Levinthal paradox. *J Biomolec Struct Dynamics* **20**, 327-329.

Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H. & Zehfus, M. H. (1985). Hydrophobicity of amino acid residues in globular proteins. *Science* **229**, 834-8.

Schneider. (2002). A genetic algorithm for the identification of conformationally invariant regions in protein molecules. *Acta Crystallogr D Biol Crystallogr* **58**, 195-208.

Schuler G. D., Altschul S. F., Lipman D. J. (1991). A workbench for multiple alignment construction and analysis. Proteins **9**, 180-190

Shakhnovich, E., Abkevich, V. & Ptitsyn, O. (1996). Conserved residues and the mechanism of protein folding. Nature **379**, 96-98

Shi, J., Blundell, T. & Mizuguchi, K. (2001). FUGUE: sequence structure homology recognition using environment specific substitution tables and strcuture dependent gap penalties. *J . Mol. Biol.* **310**, 243-257.

Skolnick, J. & Kolinski, A. (1991). Dynamic Monte Carlo simulations of a new lattice model of globular protein folding, structure and dynamics. *J. Mol. Biol.* **221**, 449-531.

Taylor, T. & Vaisman, I. (2006). Graph theoretic properties of networks formed by the Delaunay tessellation of protein structures. *PR* **E 73**, 041925.

Taylor, W. (1986). The classification of amino acid concervation. *J Theor Biol* **119**.

Taylor, W., Flores, T. & Orengo, C. (1994). Mutliple protein structure alignment. *Prot Sci* **3**, 1858-1870.

Thomsen, J., Kragelund, B., Teilum, K., Knudsen, J. & Poulsen, F. (2002). Transient intermediary states with high and low folding probabilities in the apparent two-state folding equilibrium of ACBP at low pH. *J Mol Biol* **318**, 805-814.

Trifonov, E. N., Kirshner, A., Kirzhner, V. M. & Berezovsky, I. N. (2001). Distinct stages of protein evolution as suggested by protein sequence analysis. *J. Mol. Evol.* **53**, 394-401.

Tsai, C. J., Maizel, J. V. & Nussinov, R. (2000). Anatomy of protein structures: visualizing how a one dimensional protein chain folds into a three dimensional shape. *Proceedings Natl. Acad. Sci. USA* **97**, 12038-12043.

Tsai, C. J. & Nussinov, R. (2001). The building block folding model and the kinetics of protein folding. *Prot. Eng.* **14**, 723-733.

Wallner, B. & Elofsson, A. (2006). Identification of correct regions in protein models using structural, alignment, and consensus information. *Prot Sci* **15**, 900-913.

Wetlaufer, D. (1981). Folding of protein fragments. *Adv Protein Chem* **34**, 61-92.

Wojcik, J., Mornon, J.-P. & Chomilier, J. (1999). New efficient statistical sequence dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. *J. Mol. Biol.* **289**, 1469-1490.

Yew, B., Chintapalli, S., Upton, G. & Reynolds, C. (2007). Conservation of closed loops. *J MOL Graph Model* **26,** 652-655.

Yue, K. & Dill, K. A. (1992). Inverse protein folding problem: designing polymer sequences. *Proc Natl Acad Sci U S A* **89**, 4163-4167.

Zwanzig, R., Szabo, A. & Bagghi, B. (1992). Levinthal'paradox. *PNAS* **89**, 20-22.

Abréviations

AA : acide amine

CA : Carbone Alpha

RMN : Résonnance Magnétique Nucléaire

RMSD : Root Mean Square Deviation

MIR : Most Interacting Residue

TEF : Tightened End Fragments

PDB : Protein Data Bank

HTOO : High Throw Out Order

CoC : Conservatism of Conservatism

SSR : Structure Secondaire Régulière

CATH : Class Architecture Topology and Homologous superfamily

SCOP : Structural Classification Of Proteins

ANNEXES

A : Alignement des 56 domaines immunoglobuline :

```
1ACX.pdb       1  ------------------------------------------------A--PAF-----    4
1CTM01.pdb     1  ----YP--IFAQQNYENPREA-TGRIV-CANCHL-A------------S--KPV-----   31
1HLAA02.pdb    1  ---------------------------------------------------DAP-----    3
1NCIA.pdb      1  -------------------------------------------------GSDW-----    4
2FB4H02.pdb    1  ---------------------------------A-S---T--------KG--PSV-----    8
3hfmL01.pdb    1  ------------------------------------------------D--IVL-----    4
1BEC01.pdb     1  ----------------------------------------------------AV-----    2
1CTN01.pdb     1  ---------------------------------A-A------------P--GKP-----    6
1HLAM.pdb      1  ----------------------------------I---Q--------RT--PKI-----    7
1NCOA.pdb      1  ------------------------------------------------AA--PTA-----    5
3HHRB02.pdb    1  --------------------------------D-P------------P--IAL-----    6
1BEC02.pdb     1  ------D--------------L-----------R-Q---V--------TP--PKV-----   10
1CYG02.pdb     1  --------N-PAL-----------------------------------AYG-----    7
1HNF01.pdb     1  ----------------------------------------------------T-----    1
1RSY.pdb       1  --------G-GGI--------LDSM--------V-E------------KL--GKL-----   15
2FB4L02.pdb    1  ---------------------------------Q-P---K--------AN--PTV-----    8
4fabH01.pdb    1  ------------------------------------------------E--VKL-----    4
1BGLA02.pdb    1  --------------------------T-----T-Q------------I--SDF-----    7
1FC1A01.pdb    1  ----------------------------------------------------PSV-----    3
1HNF02.pdb     1  ---------------------------------E-R------------V--SKP-----    6
1SVC01.pdb     1  ----------------------------------------------------PYL-----    3
2fbjH01.pdb    1  ------------------------------------------------E--VKL-----    4
4fabL01.pdb    1  ------------------------------------------------D--VVM-----    4
1CD8.pdb       1  ----------------------------------------------------SQF-----    3
1FC1A02.pdb    1  -------K--------------AK---------G-Q---P--------RE--PQV-----   11
1HNGA01.pdb       ------------------------------------------------------------
1SVC02.pdb     1  ----------------------------------------------------SNL-----    3
2fbjL01.pdb    1  ------------------------------------------------E--IVL-----    4
6fabH01.pdb    1  ------------------------------------------------E--VQL-----    4
1CFB01.pdb     1  -IV-----------------------Q-----D-V------------P--NAP-----    9
1fdlH01.pdb    1  ------------------------------------------------Q--VQL-----    4
1HNGA02.pdb    1  ---------------------------------E-M------------V--SKP-----    6
1TEN.pdb       1  R-------------------------L-----D-A------------P--SQI-----    8
2MCM.pdb       1  ------------------------------------------------A--PGV-----    4
6fabL01.pdb    1  ------------------------------------------------D--IQM-----    4
1CFB02.pdb     1  -QP--------------------DVPF-----K-N------------P--DNV-----   12
1fdlL01.pdb    1  ------------------------------------------------D--IQM-----    4
1igfH01.pdb    1  ------------------------------------------------E--VQL-----    4
1TLK.pdb       1  -VAE------------------E---------K-P------------H--VKP-----   10
2SODO.pdb      1  ----------------------------------ZATKAVCVLKGDGP--VQG-----   17
7FABH01.pdb    1  ------------------------------------------------A--VQL-----    4
1CID01.pdb        ------------------------------------------------------------
1GGTA03.pdb    1  ---------------------------------S------------N--VDM-----    5
1mcpH01.pdb    1  ------------------------------------------------E--VKL-----    4
1VCAA01.pdb    1  ----------------------------------------------------FKI-----    3
3CD401.pdb     1  ----------------------------------------------------K-----    1
7FABL01.pdb    1  ------------------------------------------------AS-VL-----    4
1CID02.pdb     1  ----------------------------------------------------M-----    1
1GGTA04.pdb    1  ----------------------------------------------------EI-----    2
1ncaH01.pdb    1  ------------------------------------------------Q--IQL-----    4
1VCAA02.pdb    1  ---------------------------------SF------------P--KDP-----    6
3CD402.pdb     1  ----------------------------------------------------GL-----    2
1CLC01.pdb     1  ------------------------------------------------------IETKV    5
1ncaL01.pdb    1  ------------------------------------------------D--IVM-----    4
2FB4H01.pdb    1  ------------------------------------------------E--VQL-----    4
3hfmH01.pdb    1  ------------------------------------------------D--VQL-----    4
```

```
1ACX.pdb        5  --------------SV--SP-A---S-----G---AS----------------------   13
1CTM01.pdb     32  --------------DI--EV-P---Q-----A---VL----------------------   40
1HLAA02.pdb     4  --------------KT--HM-T-H-H-----A---VS----------------------   13
1NCIA.pdb       5  --------------V----I-P-P-I-----N---LP----------------------   12
2FB4H02.pdb     9  --------------FP--LA-P---S-----SK--ST--S--------------------   19
3hfmL01.pdb     5  --------------TQ--SP-A-T-L-----S---VT----------------------   14
1BEC01.pdb      3  --------------TQ--SP-R-N-K-----V---AV----------------------   12
1CTN01.pdb      7  --------------TI--AW-G---N-T---K---FA--I-VEVDQAA-----------T   25
1HLAM.pdb       8  --------------QV--YS-R---H-----P---AE----------------------   16
1NCOA.pdb       6  --------------TV--TP-S---S-----G---LS----------------------   14
3HHRB02.pdb     7  --------------NW--TL-L---N-----V----S--L-TG----------------   17
1BEC02.pdb     11  --------------SL--FE-P---S-----KA--EI--A--------------------   21
1CYG02.pdb      8  --------------DTE-QR-W---------------IN----------------------   15
1HNF01.pdb      2  --------------N----A-L-E-T-----W---GA----------------------    9
1RSY.pdb       16  --------------QY--SLDY---D-----F------------------------    23
2FB4L02.pdb     9  --------------TL--FP-P---S-----SE--EL----------------------   18
4fabH01.pdb     5  --------------DE--TG-G---G-----L---VQ----------------------   13
1BGLA02.pdb     8  --------------HV--AT-R---F-----N---D-----------------------   15
1FC1A01.pdb     4  --------------FL--FP-P---K-----PKDTLM----------------------   15
1HNF02.pdb      7  --------------KI--SW-T---C-------------------------------   12
1SVC01.pdb      4  --------------QI--LE-Q---PKQRGFR---FR--YVA-----EGPSHGGLPGASS   33
2fbjH01.pdb     5  --------------LE--SG-G---G-----L---VQ----------------------   13
4fabL01.pdb     5  --------------TQ--TP-L-S-L-----P---VS----------------------   14
1CD8.pdb        4  --------------RV--SP-L-D-R-----T---WN----------------------   13
1FC1A02.pdb    12  --------------YT--LP-P---S-----RE--EM----------------------   21
1HNGA01.pdb     1  --------------D----S-G-T-V-----W---GA----------------------    8
1SVC02.pdb      4  --------------KI--VR-M-DRT-----A---GC----------------------   14
2fbjL01.pdb     5  --------------TQ--SP-A-I-T-----A---AS----------------------   14
6fabH01.pdb     5  --------------QQ--SG-V---E-----L---VR----------------------   13
1CFB01.pdb     10  --------------KL--TG-I---T-----C----Q----------------------   17
1fdlH01.pdb     5  --------------KE--SG-P---G-----L---VA----------------------   13
1HNGA02.pdb     7  --------------MI--YW-E---C-------------------------------   12
1TEN.pdb        9  --------------EV--KD-----V----------T----------------------   14
2MCM.pdb        5  --------------TV--TP-A---T-----G---LS----------------------   13
6fabL01.pdb     5  --------------TQ--IP-S-S-L-----S---AS----------------------   14
1CFB02.pdb     13  --------------VG--QG-----T----------E----------------------   18
1fdlL01.pdb     5  --------------TQ--SP-A-S-L-----S---AS----------------------   14
1igfH01.pdb     5  --------------VE--SG-G---D-----L---VK----------------------   13
1TLK.pdb       11  --------------YFTKTI-L-D-M-----D---VV----------------------   22
2SODO.pdb      18  --------------TI--HFEA---K-----G-----------------------    25
7FABH01.pdb     5  --------------EQ--SG-P---G-----L---VR----------------------   13
1CID01.pdb      1  --------------T----S-I-T-A-----Y---KS----------------------    8
1GGTA03.pdb     6  --------------DF--EV-E---N---------AV----------------------   13
1mcpH01.pdb     5  --------------VE--SG-G---G-----L---VQ----------------------   13
1VCAA01.pdb     4  --------------ET--TP-ESR-Y-----L---AQ----------------------   14
3CD401.pdb      2  --------------KV---------V-----L---GK----------------------    7
7FABL01.pdb     5  --------------TQ---P-P-S-V-----S---GA----------------------   13
1CID02.pdb      2  --------------KV--TQ-P---D-----S-----------------------     8
1GGTA04.pdb     3  --------------II--KV-R---G-----T---QV----------------------   11
1ncaH01.pdb     5  --------------VQ--SG-P---E-----L---KK----------------------   13
1VCAA02.pdb     7  --------------EI--HL-S---G-----P---LE----------------------   15
3CD402.pdb      3  --------------TA--NS-D---T-----H---LL----------------------   11
1CLC01.pdb      6  SAAKITENYQFDSRIR--LN-S---I-----G---FI----------------------   28
1ncaL01.pdb     5  --------------TQ--SP-K-F-M-----S---TS----------------------   14
2FB4H01.pdb     5  --------------VQ--SG-G---G-----V---VQ----------------------   13
3hfmH01.pdb     5  --------------QE--SG-P---S-----L---VK----------------------   13
```

```
1ACX.pdb       14  --------D--GQ-----SV---SVSVAA-------A-----G----------------   26
1CTM01.pdb     41  --------P--DT-----VF---EAVVKI-P-YDMQL-----K---QVLANGKKGA--LN   70
1HLAA02.pdb    14  --------D---H-----EA---TLRCWA-L-S---F-----Y------------P--AE   30
1NCIA.pdb      13  --------E--NSRGPFPQE---LVRIRS-G-R---D-----K----------------N   33
2FB4H02.pdb    20  --------G---G-----TA---ALGCLV-K-D---Y-----F------------P----   34
3hfmL01.pdb    15  --------P--GN-----SV---SLSCRA-S-Q---S----------------------IG   30
1BEC01.pdb     13  --------T--GG-----KV---TLSCQQ-T-N---N----------------------H   27
1CTN01.pdb     26  AYNNLVKVKN-AA-----DV---SVSWNL-W-N---G-----D------------T---G   51
1HLAM.pdb      17  --------N--GK-----SN---FLNCYV-S-G---F-----H------------P--SD   34
1NCOA.pdb      15  --------D--GT-----VV---KVAGAG-------L-----Q----------------AG   29
3HHRB02.pdb    18  --------I--HA-----DI---QVRWEA-PRN---A-----DI-Q---------KGWMV   40
1BEC02.pdb     22  ------------N------KQKATLVCLA-R-G---F-----F------------P--DH   39
1CYG02.pdb     16  --------G---------DV---YVYERQ-F-G-----------------------KD   28
1HNF01.pdb     10  --------L--GQ-----DI---NLDIPSFQ-M---S-----D----------------DI   27
1RSY.pdb       24  ------------Q-----NN---QLLVGI-I-Q---AAE-L-PA-LD---MG--GT--SD   48
2FB4L02.pdb    19  -------QA---N-----KA---TLVCLI-S-D---F-----Y------------P--GA   36
4fabH01.pdb    14  --------P--GR-----PM---KLSCVA-S-G---F-----T------------F--SD   31
1BGLA02.pdb    16  --------D--FS-----RA---VLEAEV-Q-MCG-E-----L------------R--DY   35
1FC1A01.pdb    16  --------I--SR-----TP---EVTCVV-V-D---V-----SH----------ED--PQ   35
1HNF02.pdb     13  ------------IN------T---TLTCEV-M-NG-----------------------TD   26
1SVC01.pdb     34  -------EKNKKS-----YP---QVKICN-------Y-----V----------------GP   51
2fbjH01.pdb    14  --------P--GG-----SL---KLSCAA-S-G---F-----D------------F--SK   31
4fabL01.pdb    15  --------L--GD-----QA---SISCRS-S-Q---S----------------------LV   30
1CD8.pdb       14  --------L--GE-----TV---ELKCQV-L-L---S-----N----------------PT   30
1FC1A02.pdb    22  -------TK---N-----QV---SLTCLV-K-G---F-----Y------------P--SD   39
1HNGA01.pdb     9  --------L--GH-----GI---NLNIPNFQ-M---T-----D----------------DI   26
1SVC02.pdb     15  --------V--TG-----GE---EIYLLC-D-K---V-----Q----------------KD   31
2fbjL01.pdb    15  --------L--GQ-----KV---TITCSA-S-S---S----------------------V   29
6fabH01.pdb    14  --------A--GS-----SV---KMSCKA-S-G---Y-----T------------F--TS   31
1CFB01.pdb     18  --------A---D-----KA---EIHWEQ-Q-G---DNRS--P----------------I   35
1fdlH01.pdb    14  --------P--SQ-----SL---SITCTV-S-G---F-----S------------L--TG   31
1HNGA02.pdb    13  ------------SN------A---TLTCEV-L-EG-----------------------TD   26
1TEN.pdb       15  --------D---T-----TA---LITWFK-P-L---A-----E----------------I   29
2MCM.pdb       14  --------N--GQ-----TV---TVSATG-------L-----T----------------PG   28
6fabL01.pdb    15  --------L--GD-----RV---SISCRA-S-Q---D----------------------IN   30
1CFB02.pdb     19  --------P---N-----NL---VISWTP-M-PE--I-----E-HN---------A-P-N   38
1fdlL01.pdb    15  --------V--GE-----TV---TITCRA-S-G---N----------------------IH   30
1igfH01.pdb    14  --------P--GG-----SL---KLSCAA-S-G---F-----T------------F--SR   31
1TLK.pdb       23  --------E--GS-----AA---RFDCKV-E-G---Y----------------------PD   38
2SODO.pdb      26  ------------D-----TV---VVTGSI-T-G---L-----T----------------E   39
7FABH01.pdb    14  --------P--SQ-----TL---SLTCTV-S-G---T-----S------------F--DD   31
1CID01.pdb      9  --------E--GE-----SA---EFSFPL-N-L---G-----E----------------ES   25
1GGTA03.pdb    14  --------L--GK-----DF---KLSITF-R-NNS-H-----N------------R--YT   33
1mcpH01.pdb    14  --------P--GG-----SL---RLSCAT-S-G---F-----T------------F--SD   31
1VCAA01.pdb    15  --------I--GD-----SV---SLTCST-T-G---C----------------------ES   30
3CD401.pdb      8  --------K--GD-----TV---ELTCTA-S-Q---K----------------------KS   23
7FABL01.pdb    14  --------P--GQ-----RV---TISCTG-S-SS--N----IG---------------AG   32
1CID02.pdb      9  ------------------N---TLTCEV-M-GP--------T------------S--PK   22
1GGTA04.pdb    12  --------V--GS-----DM---TVTVEF-T-NPL-K--E--T------------L--RN   32
1ncaH01.pdb    14  --------P--GE-----TV---KISCKA-S-G---Y-----T------------F--TN   31
1VCAA02.pdb    16  --------A--GK-----PI---TVKCSV-A-D---VYPF--D----------------R   34
3CD402.pdb     12  --------Q--GQ-----SL---TLTLES-P-P---G----------------------SS   27
1CLC01.pdb     29  --------P--NH-----SK---KATIAA---N------------------------C   41
1ncaL01.pdb    15  --------V--GD-----RV---TITCKA-S-Q---D----------------------VS   30
2FB4H01.pdb    14  --------P--GR-----SL---RLSCSS-S-G---F-----I------------F--SS   31
3hfmH01.pdb    14  --------P--SQ-----TL---SLTCSV-T-G---D-----S------------I--TS   31
```

```
1ACX.pdb      27  -----ETYYIAQCAP----VG---G--Q-D---AC-------N------------P---   45
1CTM01.pdb    71  -----VGAVLILPE-----G----F---------ELAPPDRIS----------------   90
1HLAA02.pdb   31  -----ITLTWQRD------G----E---------D-------Q----------------   42
1NCIA.pdb     34  -----LSLRYSVTGPGADQ--------P-P-----------------------------   49
2FB4H02.pdb   35  -----QPVTVSWN------S----G---------A-------L----------------   46
3hfmL01.pdb   31  -----NNLHWYQQKS----H----E--S-P---RL-------L-------------IKY-   50
1BEC01.pdb    28  -----NNMYWYRQDT----G----H--G-L---RL-------I-------------HYS-   47
1CTN01.pdb    52  -----TTAKILLN----------------G---KE-------A-------------W---   64
1HLAM.pdb     35  -----IEVDLLKN------G----E---------R-------I----------------   46
1NCOA.pdb     30  -----TAYDVGQCAW----VDT--G--V-L---AC-------N------------P---   49
3HHRB02.pdb   41  -----LEYELQYKEV----N----E--T-----KW-------K------------M---   57
1BEC02.pdb    40  -----VELSWWVN------G----K---------E-------V----------------   51
1CYG02.pdb    29  -----VVLVAVNRSS----S----S---------NYSIT-GLF----------------   48
1HNF01.pdb    28  -----DDIKWEKTS------------D-K---KK-------I-------------AQF-   44
1RSY.pdb      49  -----PYVKVFLL------P----D---------K-------K----------------   60
2FB4L02.pdb   37  -----VTVAWKAD------G----S---------P-------V----------------   48
4fabH01.pdb   32  -----YWMNWVRQSP----E----K--G-L---EW-------V-------------AQI-   51
1BGLA02.pdb   36  -----LRVTVSLWQG--------------E---TQ-------V-------------A---   50
1FC1A01.pdb   36  -----VKFNWYVD------G----V---------Q-------V----------------   47
1HNF02.pdb    27  -----PELNLYQD----------------G---KH-------L-------------K---   39
1SVC01.pdb    52  -----AKVIVQLVTN----GKN--I--H-LHAHSL-------VGK-HC--------E---   78
2fbjH01.pdb   32  -----YWMSWVRQAP----G----K--G-L---EW-------I-------------GEI-   51
4fabL01.pdb   31  HSQGNTYLRWYLQKP----G----Q--S-P---KV-------L-------------IYK-   55
1CD8.pdb      31  -----SGCSWLFQPR--GAA----A--S-P---TF-------L-------------LYL-   52
1FC1A02.pdb   40  -----IAVEWESN------G----Q---------P-------E----------------   51
1HNGA01.pdb   27  -----DEVRWERG----------------S---TL-------V-------------AEF-   41
1SVC02.pdb    32  ------DIQIRFYEE----EEN--GGVW-E---GF-------G------------D---   52
2fbjL01.pdb   30  -----SSLHWYQQKS----G----T--S-P---KP-------W-------------IYE-   49
6fabH01.pdb   32  -----NGINWVKQRP----G----Q--G-L---EW-------I-------------GYN-   51
1CFB01.pdb    36  -----LHYTIQFNTS----F--TPA--S-W---DA-------A-------------Y---   55
1fdlH01.pdb   32  -----YGVNWVRQPP----G----K--G-L---EW-------L-------------GMI-   51
1HNGA02.pdb   27  -----VELKLYQG----------------K---EH-------L-------------R---   39
1TEN.pdb      30  -----DGIELTYGIK----D----V--P-G---DR-------T-------------T---   47
2MCM.pdb      29  -----TVYHVGQCAV----VEP--G--V-I---GC-------D------------A---   48
6fabL01.pdb   31  -----NFLNWYQQKP----D----G--T-I---KL-------L-------------IYF-   50
1CFB02.pdb    39  -----FHYYVSWKRD----I----P--A-A---AW-------E-------------N---   56
1fdlL01.pdb   31  -----NYLAWYQQKQ----G----K--S-P---QL-------L-------------VYY-   50
1igfH01.pdb   32  -----CAMSWVRQTP----E----K--R-L---EW-------V-------------AGI-   51
1TLK.pdb      39  -----PEVMWFKD------D----N---------P-------V----------------   50
2SODO.pdb     40  -----GDHGFHVH------Q-----------------------FGDNTQGCTSAGPHFN   64
7FABH01.pdb   32  -----YYWTWVRQPP----G----R--G-L---EW-------I-------------GYV-   51
1CID01.pdb    26  -----LQGELRWKAE----K-A--P--S-S---QS-------W-------------ITF-   46
1GGTA03.pdb   34  -----ITAYLSANIT----FYT--G--VPK---AE-------F-------------K---   54
1mcpH01.pdb   32  -----FYMEWVRQPP----G----K--R-L---EW-------I-------------AAS-   51
1VCAA01.pdb   31  -----PFFSWRTQ------I----D---------SP-------L----------------   43
3CD401.pdb    24  -----IQFHWKNS------N----Q---------IK-------I-------------LGN-   39
7FABL01.pdb   33  -----HNVKWYQQLP----G----T--A-P---KL-------L-------------I-F-   51
1CID02.pdb    23  -----MRLILKQE-N--------------Q---EA-------R-------------V---   36
1GGTA04.pdb   33  -----VWVHLDGPGV-----------------TR-------P-------------M---   46
1ncaH01.pdb   32  -----YGMNWVKQAP----G----K--G-L---KW-------M-------------GWI-   51
1VCAA02.pdb   35  -----LEIDLLKG----------------D---HL-------M-------------K---   47
3CD402.pdb    28  -----PSVQCRSP------R----G---------K-------N----------------   39
1CLC01.pdb    42  -----STFYVVKE------D----G---------T-------I----------------   53
1ncaL01.pdb   31  -----TAVVWYQQKP----G----Q--S-P---KL-------L-------------IYW-   50
2FB4H01.pdb   32  -----YAMYWVRQAP----G----K--G-L---EW-------V-------------AII-   51
3hfmH01.pdb   32  -----DYWSWIRKFP----G----N--R-L---EY-------M-------------GYV-   51
```

```
1ACX.pdb       46  --------------------------A-----T-ATSF------------T--TD--A-   55
1CTM01.pdb     91  ----------------------PEMK-EK-IG-NLSFQNYRPNKKN--ILVI--GP--V-  118
1HLAA02.pdb    43  ----------------------T----Q-----D-TELV------------E--TR--P-   53
1NCIA.pdb      50  ----------------------T----G-----I-FIIN------------P--I-----   58
2FB4H02.pdb    47  ----------------------T----S-----G-VHTF------------P--AV--L-   57
3hfmL01.pdb    51  ---------A-SQS--ISG-I-P----S-----R-FSGS------------G--S-----   67
1BEC01.pdb     48  -Y-GA---GS-TEK--GD--I-P----D-----G-YKAS------------R--P---S-   68
1CTN01.pdb     65  -------------------------------------SG------------P--ST----   69
1HLAM.pdb      47  ----------------------E----------K-VEHS------------D--LS--F-   56
1NCOA.pdb      50  --------------------------A-----D-FSSV------------T--AD--A-   59
3HHRB02.pdb    58  -------------------------------------MD--------------P---I-   61
1BEC02.pdb     52  ---------------------H----S-----G-VSTD------------PQ-AY--K-   63
1CYG02.pdb     49  ----------------------T-----ALPA-G-TYTDQLGG---LLDGNT--IQV-G-   72
1HNF01.pdb     45  -R-KE---KE-TFK------E-K----D-----T-YKLF------------K--------   61
1RSY.pdb       61  --------------------------------K-KFET------------K--VH--R-   69
2FB4L02.pdb    49  ----------------------K----A-----G-VETT------------K--PS--K-   59
4fabH01.pdb    52  -R-NKPYNYE-TYY--SDS-V-K----G-----R-FTIS------------R--DD----   76
1BGLA02.pdb    51  -------------------------------------SG------------T--AP--F-   56
1FC1A01.pdb    48  ----------------------H----N----AK-TKPR--------------EQ--Q-   58
1HNF02.pdb     40  -----------------------------------LS------------Q--------   42
1SVC01.pdb     79  --------------------------D-----G-ICTV------------T--AG--P-   88
2fbjH01.pdb    52  -H-PD-S-GT-INY--TPS-L-K----D-----K-FIIS------------R--DN----   74
4fabL01.pdb    56  ---------V-SNR--FSG-V-P----D-----R-FSGS------------G--S-----   72
1CD8.pdb       53  -S-Q----NK-PKA--AEG-LDT----Q-----R-FSGK------------R--L-----   73
1FC1A02.pdb    52  ----------------------N----------N-YKTT------------P--PV--L-   61
1HNGA01.pdb    42  -K-RK---MK-PFL------K-S----G-----A-FEIL------------A--------   58
1SVC02.pdb     53  -----------------------------------FS------------PTDVH--R-   60
2fbjL01.pdb    50  ---------I-SKL--ASG-V-P----A-----R-FSGS------------G--S-----   66
6fabH01.pdb    52  -N-PG-N-GY-IAY--NEK-F-K----G-----K-TTLT------------V--DK----   74
1CFB01.pdb     56  -------------------------------------EK-------------------VP   59
1fdlH01.pdb    52  -W-GD-G--N-TDY--NSA-L-K----S-----R-LSIS------------K--DN----   73
1HNGA02.pdb    40  -------------------------------------SL------------R--Q-----   43
1TEN.pdb       48  -------------------------------------ID--------------LT--E-   52
2MCM.pdb       49  --------------------------T-----T-STDV------------T--AD--A-   58
6fabL01.pdb    51  ---------TSRSQ--SG--V-P----S-----R-FSGS------------G--S-----   67
1CFB02.pdb     57  -------------------------------------NN--------------IF--D-   61
1fdlL01.pdb    51  ---------T-TTL--ADG-V-P----S-----R-FSGS------------G--S-----   67
1igfH01.pdb    52  -S-SG-G-SY-TFY--PDT-V-K----G-----R-FIIS------------R--NN----   74
1TLK.pdb       51  ----------------------K--ESR-----H-FQID------------Y--DE----   62
2SODO.pdb      65  PL-S----KK-HGGPKDEE-R-H----------V-GDLG------------N--VT--A-   88
7FABH01.pdb    52  -F-YT---GT-TLL--DPS-L-R----G-----R-VTML------------V--NT----   73
1CID01.pdb     47  -SLKN---QK-VSV--QK--S-T----S---NPK-FQLS------------E--T-----   69
1GGTA03.pdb    55  -------------------------------------KE------------T--FDV-T-   61
1mcpH01.pdb    52  -R-NKGNKYT-TEY--SAS-V-K----G-----R-FIVS------------R--DT----   76
1VCAA01.pdb    44  --------------------------------N-GKVT------------N--EG----   51
3CD401.pdb     40  -Q-G----SF-LTK--GPSKL-N----D-----R-ADSR------------R------S-   60
7FABL01.pdb    52  ---H----------------N-N----A-----R-FSVS------------K--S-----   62
1CID02.pdb     37  -------------------------------------SR------------Q--E-----   40
1GGTA04.pdb    47  -----------------------------------KK------------M--FR--E-   52
1ncaH01.pdb    52  -N-TN-T-GE-PTY--GEE-F-K----G-----R-FAFS------------L--ET----   74
1VCAA02.pdb    48  -------------------------------------SQ------------E--FLEDAD   56
3CD402.pdb     40  ----------------------------------IQGG--------------------   43
1CLC01.pdb     54  --------------------------------VYTG------------T--AT--S-   61
1ncaL01.pdb    51  ---------A-STR--HIG-V-P----D-----R-FAGS------------G--S-----   67
2FB4H01.pdb    52  -W-DD-G-SD-QHY--ADS-V-K----G-----R-FTIS------------R--ND----   74
3hfmH01.pdb    52  -S-YS---GS-TYY--NPS-L-K----S-----R-ISIT------------R--DT----   73
```

```
1ACX.pdb       56  ---------------S------GAASFSF-T-------VR---K--------S---Y---   69
1CTM01.pdb    119  -P---------G-Q-K-----YSEITFPI-LA-------P-D-P----ATNKD-------  140
1HLAA02.pdb    54  -A---------G---D-G--T-FQKWAAV-VV-------P---S----G---Q-------   71
1NCIA.pdb      59  ----------------------SGQLSV-TK------PL---D--------R-------   70
2FB4H02.pdb    58  -Q---------S---S-G--L-YSLSSVV-TV-------P---S----S---S-------   75
3hfmL01.pdb    68  ---------------G------TDFTLSI-NS------VE---T--------E-------   81
1BEC01.pdb     69  ---------------Q------EQFSLIL-EL------AT---P--------S-------   82
1CTN01.pdb     70  ---------------G------SSGTANF-KV---------NK---------G-------   82
1HLAM.pdb      57  -S---------K---D-W--S-FYLLYYT-EF-------T---P----T-----------   73
1NCOA.pdb      60  ---------------N------GSASTSL-T-------VR---R--------S---F---   73
3HHRB02.pdb    62  -----------L---T------TSVPVYS---------LK---V--------D-------   74
1BEC02.pdb     64  -E---------S---N-Y--S-YCLSSRL-RV-------S---A----T---F-------   81
1CYG02.pdb     73  ----------S----------NGSVNAF-DL-------------------------------   82
1HNF01.pdb     62  ----------------------NGTLKI-KH------LK---T--------D-------   73
1RSY.pdb       70  -K--------T------LNPV-FNEQFTF-KV-------P---Y----S---E-------   88
2FB4L02.pdb    60  -Q---------S---N-N--K-YAASSYL-SL-------T---P----E---Q-------   77
4fabH01.pdb    77  -----------S---K------SSVYLQM-NN------LR---V--------E-------   91
1BGLA02.pdb    57  -GGEIIDERG-GY--A-D--R-VTLRLNV-EN------PKLWSA--------E-------   86
1FC1A01.pdb    59  -Y-------------N-S--T-YRVVSVL-TV-------L---H----Q---N-------   75
1HNF02.pdb     43  ----------------------RVITHKW-TT-------------------S-------   52
1SVC01.pdb     89  -----------K---D------MVVGFANLG-------IL---H--------VTKKKVFE  110
2fbjH01.pdb    75  -----------A---K------NSLYLQM-SK------VR---S--------E-------   89
4fabL01.pdb    73  ---------------G------TDFTLKI-SR------VE---A--------E-------   86
1CD8.pdb       74  ---------------G------DTFVLTL-SD------FR---R--------E-------   87
1FC1A02.pdb    62  -D---------S---D-G--S-FFLYSKL-TV-------D---K----S---R-------   79
1HNGA01.pdb    59  ----------------------NGDLKI-KN------LT---R--------D-------   70
1SVC02.pdb     61  ---------------Q------FAIVFKT-P-------KY---K--------D---I---   74
2fbjL01.pdb    67  ---------------G------TSYSLTI-NT------ME---A--------E-------   80
6fabH01.pdb    75  -----------S---S------STAYMQL-RS------LT---S--------E-------   89
1CFB01.pdb     60  -N---------T---D------SSFVVQM-S------------P--------W-------   72
1fdlH01.pdb    74  -----------S---K------SQVFLKM-NS------LH---T--------D-------   88
1HNGA02.pdb    44  ----------------------KTMSYQW-T---------------------N-------   52
1TEN.pdb       53  -----------D---E------NQYSIGN---------LK---P--------D-------   65
2MCM.pdb       59  ---------------A------GKITAQL-K-------VH---S--------S---F---   72
6fabL01.pdb    68  ---------------G------TDYSLTI-SN------LE---Q--------E-------   81
1CFB02.pdb     62  -W---------R---Q------NNIVIAD---------QP---T--------F-------   75
1fdlL01.pdb    68  ---------------G------TQYSLKI-NS------LQ---P--------E-------   81
1igfH01.pdb    75  -----------A---R------NTLSLQM-SS------LR---S--------E-------   89
1TLK.pdb       63  ---------------E------GNCSLTI-SE------VC---G--------D-------   76
2SODO.pdb      89  -D---------K---N-G--V-AIVDIVD-PL-------I---SLSGEY---S-------  110
7FABH01.pdb    74  -----------S---K------NQFSLRL-SS------VT---A--------A-------   88
1CID01.pdb     70  ----------------------LPLTLQI-PQ------VS---L--------Q-------   82
1GGTA03.pdb    62  -L---------EP--L-S--F-KKEAVLI-QAGEYMGQLL---E--------Q-------   86
1mcpH01.pdb    77  -----------S---Q------SILYLQM-NA------LR---A--------E-------   91
1VCAA01.pdb    52  ----------------------TTSTLTM-NP------VS---F--------G-------   64
3CD401.pdb     61  LW---------D---Q------GNFPLII-KN------LK---I--------E-------   77
7FABL01.pdb    63  ---------------G------TSATLAI-TG------LQ---A--------E-------   76
1CID02.pdb     41  ----------------------KVIQVQA---------P--------------------   48
1GGTA04.pdb    53  -I---------RP--N-S--T-VQWEEVC-RP---------WV-----------------   69
1ncaH01.pdb    75  -----------S---A------STANLQI-NN------LK---N--------E-------   89
1VCAA02.pdb    57  -R---------K---SLE--T-KSLEVTF-T-------PV---I--------E-------   74
3CD402.pdb     44  ----------------------KTLSV-SQ------LE---L--------Q-------   54
1CLC01.pdb     62  -M---------FDNDT-K--E-TVYIADF-SS------VN---------------------   80
1ncaL01.pdb    68  ---------------G------TDYTLTI-SS------VQ---A--------E-------   81
2FB4H01.pdb    75  -----------S---K------NTLFLQM-DS------LR---P--------E-------   89
3hfmH01.pdb    74  -----------S---K------NQYYLDL-NS------VT---T--------E-------   88
```

```
1ACX.pdb       70  ----------------A------------GQTP-SG-----------------TPVGSVD   83
1CTM01.pdb         ------------------------------------------------------------
1HLAA02.pdb        ------------------------------------------------------------
1NCIA.pdb          ------------------------------------------------------------
2FB4H02.pdb        ------------------------------------------------------------
3hfmL01.pdb        ------------------------------------------------------------
1BEC01.pdb         ------------------------------------------------------------
1CTN01.pdb         ------------------------------------------------------------
1HLAM.pdb          ------------------------------------------------------------
1NCOA.pdb      74  ----------------E------------GFLF-DG-----------------TRWGTVD   87
3HHRB02.pdb        ------------------------------------------------------------
1BEC02.pdb         ------------------------------------------------------------
1CYG02.pdb         ------------------------------------------------------------
1HNF01.pdb         ------------------------------------------------------------
1RSY.pdb           ------------------------------------------------------------
2FB4L02.pdb        ------------------------------------------------------------
4fabH01.pdb        ------------------------------------------------------------
1BGLA02.pdb    87  ------------------------------------------------------------I   87
1FC1A01.pdb        ------------------------------------------------------------
1HNF02.pdb         ------------------------------------------------------------
1SVC01.pdb    111  TLEARMTEACIRGYNPGLLVHPDLAYLQAEGGGDRQLGDREKELIRQAALQQT---KEMD  167
2fbjH01.pdb        ------------------------------------------------------------
4fabL01.pdb        ------------------------------------------------------------
1CD8.pdb           ------------------------------------------------------------
1FC1A02.pdb        ------------------------------------------------------------
1HNGA01.pdb        ------------------------------------------------------------
1SVC02.pdb     75  ----------------------------------------------------N------   75
2fbjL01.pdb        ------------------------------------------------------------
6fabH01.pdb        ------------------------------------------------------------
1CFB01.pdb         ------------------------------------------------------------
1fdlH01.pdb        ------------------------------------------------------------
1HNGA02.pdb        ------------------------------------------------------------
1TEN.pdb           ------------------------------------------------------------
2MCM.pdb       73  ----------------Q------------AVVGADG----------------TPWGTVN   87
6fabL01.pdb        ------------------------------------------------------------
1CFB02.pdb         ------------------------------------------------------------
1fdlL01.pdb        ------------------------------------------------------------
1igfH01.pdb        ------------------------------------------------------------
1TLK.pdb           ------------------------------------------------------------
2SODO.pdb          ------------------------------------------------------------
7FABH01.pdb        ------------------------------------------------------------
1CID01.pdb         ------------------------------------------------------------
1GGTA03.pdb        ------------------------------------------------------------
1mcpH01.pdb        ------------------------------------------------------------
1VCAA01.pdb        ------------------------------------------------------------
3CD401.pdb         ------------------------------------------------------------
7FABL01.pdb        ------------------------------------------------------------
1CID02.pdb         ------------------------------------------------------------
1GGTA04.pdb        ------------------------------------------------------------
1ncaH01.pdb        ------------------------------------------------------------
1VCAA02.pdb        ------------------------------------------------------------
3CD402.pdb         ------------------------------------------------------------
1CLC01.pdb         ------------------------------------------------------------
1ncaL01.pdb        ------------------------------------------------------------
2FB4H01.pdb        ------------------------------------------------------------
3hfmH01.pdb        ------------------------------------------------------------
```

```
1ACX.pdb       84  C--A--TDA--CN--LGAGN-S-------------------G----------------   97
1CTM01.pdb    141  V--H--FLK--YPIYVGGNR-GRG-QIYP------------D-----------G-----  163
1HLAA02.pdb    72  E------QR--YT--CHVQH-E-------------------G-----------L-----   84
1NCIA.pdb      71  -E-LIARFH--LR--AHAVD-I-------------------N------G----N--Q--   89
2FB4H02.pdb    76  L--G--TQT--YI--CNVNH-K------P------------S-----------N-----   91
3hfmL01.pdb    82  ---D--FGM--YF--CQQSN-S-------------------W-----------P--Y--   96
1BEC01.pdb     83  ---Q--TSV--YF--CASGGGR-------------------G--SY--A----E--Q--  101
1CTN01.pdb     83  ---G--RYQ--MQ--VALCN-A-------------------D-----------G--C--   97
1HLAM.pdb      74  E-----KDE--YA--CRVNH-V-------------------T-----------L--S--   88
1NCOA.pdb      88  C--T--TAA--CQ--VGLSD-A-------------------A-----------G-----  102
3HHRB02.pdb    75  ---K--EYE--VR--VRSKQ-R-------------------N------S--G-NY-G--   92
1BEC02.pdb     82  WHNP--RNH--FR--CQVQF-H------GLSEEDKW--PE--G---SPKP----V--T--  113
1CYG02.pdb     83  ---G-----PGEVG-VWAYS-AT-------------------------------------   95
1HNF01.pdb     74  ---D--QDI--YK--VSIYD-T-------------------K------G----K--N--   89
1RSY.pdb       89  L--G--GKT--LV--MAVYD-F---------------D-R-F-----------S--K--  106
2FB4L02.pdb    78  WK-S--HRS--YS--CQVTH-E---------------------------------G-----   92
4fabH01.pdb    92  ---D--MGI--YY--CTGSY-Y-------------------G-----------M--D--  106
1BGLA02.pdb    88  ---P-NLYR--AV--VELHT-A-------------------D------G--T-L--I--  105
1FC1A01.pdb    76  WL-D--GKE--YK--CKVSN-K-------------------A-----------L--P--   92
1HNF02.pdb     53  ---L--SAK--FK--CTAGN-K-------------------V-----------S--K--   67
1SVC01.pdb    168  ---L-SVVR--LM--FTAFL-P------DS----------TGS-----------F--T--  187
2fbjH01.pdb    90  ---D--TAL--YY--CARLH-Y-------------------Y------G----YN-A--  106
4fabL01.pdb    87  ---D--LGV--YF--CSQST-H-------------------V-----------P--W--  101
1CD8.pdb       88  ---N--EGY--YF--CSALS-N-------------------S-----------I--M--  102
1FC1A02.pdb    80  WQ-Q--GNV--FS--CSVMH-E-------------------A-----------L-HN--   97
1HNGA01.pdb    71  ---D--SGT--YN--VTVYS-T-------------------N------G----T--R--   86
1SVC02.pdb     76  ---ITKPAS--VF--VQLRR-K-------------------S-----------D--L--   92
2fbjL01.pdb    81  ---D--AAI--YY--CQQWT-Y-------------------P-----------L--I--   95
6fabH01.pdb    90  ---D--SAV--YF--CARSE-Y-------------------YGGS---Y----KF-D--  109
1CFB01.pdb     73  ---A--NYT--FR--VIAFN-K-------------------I-----------GA-SP-   89
1fdlH01.pdb    89  ---D--TAR--YY--CARER-D-------------------Y-----------RL-D--  104
1HNGA02.pdb    53  ---L--RAP--FK--CKAVN-R-------------------V-----------S--Q--   67
1TEN.pdb       66  ---T--EYE--VS--LISRR-G-------------------D-----------MS-S--   81
2MCM.pdb       88  C--K--VVS--CS--AGLGS-D-------------------S-----------G-----  102
6fabL01.pdb    82  ---D--IAT--YF--CQQGN-A-------------------L-----------P--R--   96
1CFB02.pdb     76  ---V--KYL--IK--VVAIN-D-------------------R-----------GE-S--   91
1fdlL01.pdb    82  ---D--FGS--YY--CQHFW-S-------------------T-----------P--R--   96
1igfH01.pdb    90  ---D--TAI--YY--CTRYS-S-------------------D--P---F----YF-D--  107
1TLK.pdb       77  ---D--DAK--YT--CKAVN-S-------------------L-----------G--E--   91
2SODO.pdb     111  I--I----G--RT--MVVHE-K-P-DDLGR---GGNEES-T-K---T--G----N--A--  139
7FABH01.pdb    89  ---D--TAV--YY--CARNL-I-------------------A------G----GI-D--  105
1CID01.pdb     83  ---F--AGS--GN--LTLTL-D-------------------R-----------G-----   96
1GGTA03.pdb    87  ---A--SLH--FF--VTARI-N-------------------E------T--R-D--V--  103
1mcpH01.pdb    92  ---D--TAI--YY--CARNY-Y------G------------S--T---W----YF-D--  110
1VCAA01.pdb    65  ---N--EHS--YL--CTATC-E-------------------S-----------R--K--   79
3CD401.pdb     78  ---D--SDT--YI--CEVE-------------------------------------D--   88
7FABL01.pdb    77  ---D--EAD--YY--CQSYD-R-------------------S-----------L--R--   91
1CID02.pdb     49  ---E--AGV--WQ--CLLSE-G-------------------E-----------E--V--   63
1GGTA04.pdb    70  ---S-GHRK--LI--ASMSS-D-------------------S------L----R--H--   86
1ncaH01.pdb    90  ---D--TAT--FF--CARGE-D-------------------N-FG---S--L-S--D--  108
1VCAA02.pdb    75  ---D-IGKV--LV--CRAKL-H-------------------I--D---EMDS-V----P   95
3CD402.pdb     55  ---D--SGT--WT--CTVLQ-N-------------------Q-----------K--K--   69
1CLC01.pdb     81  ---E--EGT--YY--LAVPG------------------------------------V--   92
1ncaL01.pdb    82  ---D--LAL--YY--CQQHY-S-------------------P-----------P--W--   96
2FB4H01.pdb    90  ---D--TGV--YF--CARDG-G-HGFCSS------------A------S--CFGP-D--  114
3hfmH01.pdb    89  ---D--TAT--YY--CANW-------------------------------DG-D--  101
```

```
1ACX.pdb       98  -LNLGHVALTF-----G--------------------------  108
1CTM01.pdb    164  ---SKSN---N---TV--------------------------  170
1HLAA02.pdb    85  -P-KP-L-------------------------------------   88
1NCIA.pdb      90  VENPIDIVINV-----I---D----------------------  102
2FB4H02.pdb    92  -T-KVDKRVEP--------------------------------  100
3hfmL01.pdb    97  -TFGGGTKLEI-----K---R----------------------  108
1BEC01.pdb    102  -FFGPGTRLTV-----L---E----------------------  113
1CTN01.pdb     98  -TASDATEIVV--------------------------------  107
1HLAM.pdb      89  -Q-PKIVKWDR--------------------------------   97
1NCOA.pdb     103  -NGPEGVAISF-----N--------------------------  113
3HHRB02.pdb    93  -EFSEVLYVTL-----P--------------------------  103
1BEC02.pdb    114  -Q-NISAEAWG-----R---A--------------------D-  125
1CYG02.pdb         ---------------------------------------------
1HNF01.pdb     90  -VLEKIFDLKI-----Q--------------------------  100
1RSY.pdb      107  HDIIGEFKVPM-----N------TVDFG--HVTEEWRDLQSA--  135
2FB4L02.pdb    93  -S-TVEKTVAP--------------------------------  101
4fabH01.pdb   107  -YWGQGTSVTV-----S---S----------------------  118
1BGLA02.pdb   106  -E-AEACDVGF-----R--------------------------  115
1FC1A01.pdb    93  -A-PIEKTIS---------------------------------  100
1HNF02.pdb     68  -E-SSVEPVSC-----P---E--------------------K-   79
1SVC01.pdb    188  -RRLEPVVSDAIYD--S---K----------------------  202
2fbjH01.pdb   107  -YWGQGTLVTV-----S---A----------------------  118
4fabL01.pdb   102  -TFGGGTKLEI-----K---R----------------------  113
1CD8.pdb      103  -YFSHFVPVFL-----P---A----------------------  114
1FC1A02.pdb    98  -H-YTQKSLSL--------------------------------  106
1HNGA01.pdb    87  -ILNKALDLRI-----L--------------------------   97
1SVC02.pdb     93  -ETSEPKPFLY-----Y---P--------------------E-  105
2fbjL01.pdb    96  -TFGAGTKLEL-----K---R----------------------  107
6fabH01.pdb   110  -YWGQGTTLTV-----S---S----------------------  121
1CFB01.pdb     90  -P-SAHSDSCT-----T--------------------------   99
1fdlH01.pdb   105  -YWGQGTTLTV-----S---S----------------------  116
1HNGA02.pdb    68  -E-SEMEVVNC-----P---E----------------------   78
1TEN.pdb       82  --NPAKETFTT--------------------------------   90
2MCM.pdb      103  --EGAAQAITF-----A--------------------------  112
6fabL01.pdb    97  -TFGGGTKLEI-----K--------------------------  107
1CFB02.pdb     92  NVAAEEVVGYS-----GEDR-----------------------  106
1fdlL01.pdb    97  -TFGGGTKLEI-----K---R----------------------  108
1igfH01.pdb   108  -YWGQGTTLTV-----S---S----------------------  119
1TLK.pdb       92  -A-TCTAELLV-----E---T--------------------M-  103
2SODO.pdb     140  GSRLACGVIGI-----A------------------------K-  152
7FABH01.pdb   106  -VWGQGSLVTV-----S---S----------------------  117
1CID01.pdb     97  -ILYQEVNLVV--------------------------------  106
1GGTA03.pdb   104  -L-AKQKSTVL-----T---I--------------------P-  115
1mcpH01.pdb   111  -VWGAGTTVTV-----S---SES---ARNP-------------  128
1VCAA01.pdb    80  -L-EKGIQVEI-----Y--------------------------   89
3CD401.pdb     89  -Q-KEEVQLLV-----F--------------------------   98
7FABL01.pdb    92  -VFGGGTKLTV-----L---R----------------------  103
1CID02.pdb     64  ---KMDSKIQV--------------------------------   71
1GGTA04.pdb    87  -V-YGELDVQI-----Q---R--------------------RP   99
1ncaH01.pdb   109  -YWGQGTTVTV-----S---SA---------------------  121
1VCAA02.pdb    96  TVRQAVKELQV-----Y---I--------------------SP  110
3CD402.pdb     70  -VEF-KID-----------------------------------   75
1CLC01.pdb     93  -GKSVNFKIAM--------------------------------  102
1ncaL01.pdb    97  -TFGGGTKLEI-----K---R----------------------  108
2FB4H01.pdb   115  -YWGQGTPVTV-----S---S----------------------  126
3hfmH01.pdb   102  -YWGQGTLVTV-----S---A----------------------  113
```

Alignement des 43 domaines de types flavodoxine

```
1BE1.pdb      1  ---M-----------------------------------------------------    1
1EUCB02.pdb      -----------------------------------------------------
1GTZA.pdb        -----------------------------------------------------
1JRL.pdb      1  ---------------AD-------------------------------------    2
1Z8HA.pdb     1  ------------SKTQI-------------------------------------    5
1BMTA01.pdb      -----------------------------------------------------
1EUCB03.pdb   1  ----------------------------PIE-----------------NEAAKYDLK--Y   13
1H05.pdb         -----------------------------------------------------
1LI4A02.pdb   1  --------------------------YGCR----------------ESLIDGI--KRA   14
1U11.pdb         -----------------------------------------------------
1ZMBA01.pdb      -----------------------------------------------------
1BMTA02.pdb   1  --A--------------------------SKEQG-----------------------    6
1F8YA.pdb        -----------------------------------------------------
1I1H.pdb         -----------------------------------------------------
1MX3A02.pdb   1  -------------------------VEETA----------------DSTLCHI--LNL   15
1UQRA.pdb        -----------------------------------------------------
2APJA.pdb     1  ------------SPIPP-------------------------------------    5
1DXYA02.pdb   1  --------------------------PAAIA----------------EFALTDT--LYL   15
1FXWA.pdb     1  ----ENPA-SKPTPVQDVQGDGR---WMSLH----------------HRFVADS--KDK   33
1J4AA02.pdb   1  -------------------------PNAIA----------------EHAAIQA--ARI   15
1O4V.pdb         -----------------------------------------------------
1V8BA02.pdb   1  --------------------------YGCR----------------HSLPDGL--MRA   14
2C4V.pdb         -----------------------------------------------------
1FXWF.pdb     1  ----SNPA-AIPHAAEDIQGDDR---WMSQH----------------NRFVLDC--KDK   33
1JKJA01.pdb      -----------------------------------------------------
1OI7A01.pdb      -----------------------------------------------------
1XMP.pdb         -----------------------------------------------------
3CHY.pdb      1  AD----------------------------------------------------    2
1EUCA01.pdb      -----------------------------------------------------
1FYV.pdb         -----------------------------------------------------
1JKJA02.pdb   1  --------CPGVI-------------------------------------------TP    7
1OI7A02.pdb   1  --------CPGII-------------------------------------------SA    7
1YGYA02.pdb   1  ----------------------N---IHSAA----------------EHALALL--LAA   16
7REQA02.pdb   1  ---------IRT---------------ISGVYS-KEVKNTPEVEEARELVEEFE--QAE   32
1EUCA02.pdb   1  --------CPGVI-------------------------------------------NP    7
1GDHA02.pdb   1  -------------------------TVATA----------------EIAMLLL--LGS   15
1JKJB02.pdb      -----------------------------------------------------
1QCZ.pdb         -----------------------------------------------------
1YGYA03.pdb   1  -----N--------------E---EVAPWL----------------DLVRKLG--VLA   18
1EUCB01.pdb      -----------------------------------------------------
1JKJB03.pdb   1  ----------------------------PRE----------------AQAAQWELN--Y   13
1SAYA02.pdb   1  ---------------------LTPMSIIAG----------------RLSVQFG--ARF   19
1YZF.pdb      1  ---------------MR-------------------------------------    2
```

```
1BE1.pdb      2  ---E-------------------------------------------------    2
1EUCB02.pdb      ---------------------------------------------------------
1GTZA.pdb     1  ---------------------------------------------RSLA----------    4
1JRL.pdb         ---------------------------------------------------------
1Z8HA.pdb        ---------------------------------------------------------
1BMTA01.pdb   1  ---------------------------------------------QAEWR----------    5
1EUCB03.pdb  14  IGLD-----------------------------------------G----------   18
1H05.pdb         ---------------------------------------------------------
1LI4A02.pdb  15  ---T-------D-------------------------------V--MIA----------   20
1U11.pdb         ---------------------------------------------------------
1ZMBA01.pdb      ---------------------------------------------------------
1BMTA02.pdb   7  ---K-----------------------------------------T----------    8
1F8YA.pdb     1  ------P-------------------------------------------------    1
1I1H.pdb      1  ----PE------------------------------------------------    2
1MX3A02.pdb  16  ---Y-------RRATWLHQALREGTRVQ-SVEQIREVASGA---A--RIR----------   49
1UQRA.pdb        ---------------------------------------------------------
2APJA.pdb        ---------------------------------------------------------
1DXYA02.pdb  16  ---L-------RNMGKVQAQLQAGDYE--K----AGTF-IG---K--ELG----------   43
1FXWA.pdb    34  ---E-----------------------------------------P----------   35
1J4AA02.pdb  16  ---L-------RQDKAMDEKVARHDL---R----WAPT-IG---R--EVR----------   42
1O4V.pdb         ---------------------------------------------------------
1V8BA02.pdb  15  ---T-------D-------------------------------F--LIS----------   20
2C4V.pdb         ---------------------------------------------------------
1FXWF.pdb    34  ---E-----------------------------------------P----------   35
1JKJA01.pdb   1  ----------SI-------------------------------L--IDK----------    6
1OI7A01.pdb   1  ----------MI-------------------------------L--VNR----------    6
1XMP.pdb         ---------------------------------------------------------
3CHY.pdb      3  ---K-------------------------------------------------    3
1EUCA01.pdb   1  --------HL---------------------------------Y--VDK----------    6
1FYV.pdb      1  -------N-----------------------------------I--PLE-------ELQ    8
1JKJA02.pdb   8  ---G-------ECKI-------------------------GIQ--PGH----------   18
1OI7A02.pdb   8  ---E-------ETKI-------------------------GIM--PGH----------   18
1YGYA02.pdb  17  ---S-------RQIPAADASLREHTW---K----RSSF-SG---T--EIF----------   43
7REQA02.pdb  33  ---G-----------------------------------------R----------   34
1EUCA02.pdb   8  ---G-------ECKI-------------------------GIM--PGH----------   18
1GDHA02.pdb  16  ---A-------RRAGEGEKMIRTRSWPGWE----PLEL-VG---E--KLD----------   45
1JKJB02.pdb      ---------------------------------------------------------
1QCZ.pdb         ---------------------------------------------------------
1YGYA03.pdb  19  ---G-------VL-------------------------SDE------------LPV---   27
1EUCB01.pdb      ---------------------------------------------------------
1JKJB03.pdb  14  VALD---------------------------------------G----------   18
1SAYA02.pdb  20  ---L-------ER--------QQGG-------------RG---V--LLGGVPGV-K---   38
1YZF.pdb         ---------------------------------------------------------
```

```
1BE1.pdb      3  ---K-K---TIV--L-GV----I-----------------------GS----------   13
1EUCB02.pdb   1  ---------RFF--V-AD----T-------------------------------------    7
1GTZA.pdb     5  ---N-A---PIM--I-LN----G----------------PN-------LNLLGQAQPEIY   27
1JRL.pdb      3  ---------TLL--I-LG----D----------------SL-------S-----------   12
1Z8HA.pdb     6  ---------RIC--F-VG----D----------------SF-------V-----------   15
1BMTA01.pdb   6  ---SW----------------------------------------------------     7
1EUCB03.pdb  19  ---------NIA--C-FV----N-------------------------------------   25
1H05.pdb      1  -----L---IVN--V-IN----G----------------PN-------LGRLGRR-----   17
1LI4A02.pdb  21  ---G-K---VAV--V-AG----------------------Y-------------------   29
1U11.pdb      1  ------SAPVVG--I-IM----G----------------SQ-------------------   12
1ZMBA01.pdb   1  ---------VKSF-L-LG----Q----------------S-------N-----------   10
1BMTA02.pdb   9  ---N-G---KMV--I-AT----V----------------KG-------DV----------   21
1F8YA.pdb     2  ---K-K---TIY--F-GA----G----------------WF----TDR-----------   15
1I1H.pdb      3  -----Y---DYI--R-DGNAIYERSFAIIRA-EADLSR--F-------------------   29
1MX3A02.pdb  50  ---G-E---TLG--I-IG----------------------L-------------------   58
1UQRA.pdb     1  ---M-K---KIL--L-LN----G----------------PN-------LNMLGKREPHIY   23
2APJA.pdb     6  ---------NQIFIL-SG----Q----------------N-------M-----------   16
1DXYA02.pdb  44  ---Q-Q---TVG--V-MG----------------------T-------------------   52
1FXWA.pdb    36  ---------EVV--F-IG----D----------------SL-------V-----------   45
1J4AA02.pdb  43  ---D-Q---VVG--V-VG----------------------T-------------------   51
1O4V.pdb      1  -----P---RVG--I-IM----G----------------SD-------------------   10
1V8BA02.pdb  21  ---G-K---IVV--I-CG----------------------Y-------------------   29
2C4V.pdb      1  -----M---KIL--V-IQ----G----------------PN-------LNMLGHRDPRLY   22
1FXWF.pdb    36  ---------DVL--F-VG----D----------------SM-------V-----------   45
1JKJA01.pdb   7  ---N-T---KVI--C-QG----------------------FT-------------------   16
1OI7A01.pdb   7  ---E-T---RVL--V-QG----------------------IT-------------------   16
1XMP.pdb      1  -------KSLVG--V-IM----G----------------ST-------------------   11
3CHY.pdb      4  ---E-L---KFL--V-VD----D-------------------------------------   12
1EUCA01.pdb   7  ---N-T---KVI--C-QG----------------------FT-------------------   16
1FYV.pdb      9  RNLQ-F---HAF--I-SY----S----------------GH-------D-----------   23
1JKJA02.pdb  19  ---IH-KPGKVG--I-VS----R-------------------------------------   30
1OI7A02.pdb  19  ---VF-KRGRVG--I-IS----R-------------------------------------   30
1YGYA02.pdb  44  ---G-K---TVG--V-VG----------------------L-------------------   52
7REQA02.pdb  35  ---R-P---RIL--L-AK----M----------------GQ-------DG----------   47
1EUCA02.pdb  19  ---IH-KKGRIG--I-VS----R-------------------------------------   30
1GDHA02.pdb  46  ---N-K---TLG--I-YG----------------------F-------------------   54
1JKJB02.pdb   1  ---------VGY--A-CT----T-------------------------------------    7
1QCZ.pdb      1  -------PARVA--I-VG----S----------------K-------------------   10
1YGYA03.pdb  28  ---S-L---SVQ--V-RG----------------------ELAAE--E------------   41
1EUCB01.pdb   1  ---S-R---ETY--LAILM---DRSCN--G-PVLVGSPQGG-------------------   26
1JKJB03.pdb  19  ---------NIG--C-MV----N-------------------------------------   25
1SAYA02.pdb  39  ---P-G---KVV--I-LG----------------------G-------------------   47
1YZF.pdb      3  ---------KIV--L-FG----D----------------SI-------T-----------   12
```

```
1BE1.pdb     14  ----D-------------------------------------------------   14
1EUCB02.pdb   8  -----A-------------------------------------------------    8
1GTZA.pdb    28  GSDTLA------------------------------------------------   33
1JRL.pdb     13  -----AGY---------------------------------------------------  15
1Z8HA.pdb    16  -----NGT---------------------------------------------------  18
1BMTA01.pdb   8  -----E-------------------------------------------------    8
1EUCB03.pdb  26  ----GA------------------------------------------------   27
1H05.pdb     18  -GTTHD-----------------------------------------------------  22
1LI4A02.pdb  30  -----G-----------------------------------------------------  30
1U11.pdb     13  ---SDW-----------------------------------------------------  15
1ZMBA01.pdb  11  -----AGRGFINEVP-------------------IYNERIQ-LR-NGRWQ--TE----P   37
1BMTA02.pdb  22  ----HD------------------------------------------------   23
1F8YA.pdb    16  ----QN------------------------------------------------   17
1I1H.pdb     30  -----S-----------------------------------------------------  30
1MX3A02.pdb  59  -----G-----------------------------------------------------  59
1UQRA.pdb    24  GSQTLS------------------------------------------------   29
2APJA.pdb    17  -----AGRGG-----VFKDHHNNRWVWDKILPPECAPNSSILRLSADLRWEEAHEPLHVD   66
1DXYA02.pdb  53  -----G-----------------------------------------------------  53
1FXWA.pdb    46  -----QL----------------------------------------------------  47
1J4AA02.pdb  52  -----G-----------------------------------------------------  52
1O4V.pdb     11  ---SDL-----------------------------------------------------  13
1V8BA02.pdb  30  -----G-----------------------------------------------------  30
2C4V.pdb     23  GMVTLD------------------------------------------------   28
1FXWF.pdb    46  -----QL----------------------------------------------------  47
1JKJA01.pdb  17  -----G-----------------------------------------------------  17
1OI7A01.pdb  17  -----G-----------------------------------------------------  17
1XMP.pdb     12  ---SDW-----------------------------------------------------  14
3CHY.pdb     13  ----FS------------------------------------------------   14
1EUCA01.pdb  17  -----G-----------------------------------------------------  17
1FYV.pdb     24  ----SF------------------------------------------------   25
1JKJA02.pdb  31  ----SG------------------------------------------------   32
1OI7A02.pdb  31  ----SG------------------------------------------------   32
1YGYA02.pdb  53  -----G-----------------------------------------------------  53
7REQA02.pdb  48  ----HD------------------------------------------------   49
1EUCA02.pdb  31  ----SG------------------------------------------------   32
1GDHA02.pdb  55  -----G-----------------------------------------------------  55
1JKJB02.pdb   8  -----P-------------------------------------------------    8
1QCZ.pdb     11  ---SDW-----------------------------------------------------  13
1YGYA03.pdb  42  ----VE------------------------------------------------   43
1EUCB01.pdb  27  -----V-----------------------------------------------------  27
1JKJB03.pdb  26  ----GA------------------------------------------------   27
1SAYA02.pdb  48  -----G-----------------------------------------------------  48
1YZF.pdb     13  -----AGY---------------------------------------------------  15
```

```
1BE1.pdb     15  -------------CHAVGNKI----LD--HSFTN-AG-------F------NVV--N-I-   37
1EUCB02.pdb   9  -------------NE------ALEAAK--RL-N-----------A--K---EIV--L-KA   27
1GTZA.pdb    34  -------------DV------EALCVK--AA-AA-HG-------G------TVD--F-R-   53
1JRL.pdb     16  ---R--MSASA--AW------PALLND--KW-S------------KT---SVV--N-A-   39
1Z8HA.pdb    19  ---GDPE-C-L--GW------TGRVCV--NA-NK-KG-------Y-DV---TYY--N-L-   46
1BMTA01.pdb   9  -------------VN------KRLEYS--LV-KGI------------------------   21
1EUCB03.pdb  28  -------------GL------AMATCD--II-FL-NG-------G------KPA--NFL-   48
1H05.pdb     23  -------------EL------VALIER--EA-AE-LG-------L------KAV--V-R-   42
1LI4A02.pdb  31  -------------DV------GKGCAQ--AL-RG-FG-------A------RVI--I-T-   50
1U11.pdb     16  -------------ET------MRHADA--LL-TE-LE-------I------PHE--T-L-   35
1ZMBA01.pdb  38  INYDRPV-SGI--SL------AGSFAD--AW-SQ-KN---QE--D-II---GLI--P-C-   71
1BMTA02.pdb  24  -------------IG------KNIVGV--VL-QC-NN-------Y------EIV--D-L-   43
1F8YA.pdb    18  -------------KA------YKEAME--AL-KE-NP-------T---IDL--E--N-SY   39
1I1H.pdb     31  -------------EE------EADLAV--RM-VH-AC-GSVEAT---R---QFV--F-SP   57
1MX3A02.pdb  60  -------------RV------GQAVAL--RA-KA-FG-------F------NVL--F-Y-   79
1UQRA.pdb    30  -------------DI------EQHLQQ--SA-QA-QG-------Y------ELD--Y-F-   49
2APJA.pdb    67  IDTGK-V-CGV--GP------GMAFAN--AV-KN-RL-ETDS--A-VI---GLV--P-C-  101
1DXYA02.pdb  54  -------------HI------GQVAIK--LF-KG-FG-------A------KVI--A-Y-   73
1FXWA.pdb    48  ------------------------MHQCEIW-RE-LF-------S-PL---HAL--N-F-   66
1J4AA02.pdb  53  -------------HI------GQVFMQ--IM-EG-FG-------A------KVI--T-Y-   72
1O4V.pdb     14  -------------PV------MKQAAE--IL-EE-FG-------I------DYE--I-T-   33
1V8BA02.pdb  31  -------------DV------GKGCAS--SM-KG-LG-------A------RVY--I-T-   50
2C4V.pdb     29  -------------QI------HEIMQT--FV-KQ-GN-------L------DVELEF-F-   50
1FXWF.pdb    48  ------------------------MQQYEIW-RE-LF-------S-PL---HAL--N-F-   66
1JKJA01.pdb  18  -------------SQ------GTFHSE--QA-IA-YG-------T------KMV--G-GV   38
1OI7A01.pdb  18  -------------RE------GQFHTK--QM-LT-YG-------T------KIV--A-GV   38
1XMP.pdb     15  -------------ET------MKYACD--IL-DE-LN-------I------PYE--K-K-   34
3CHY.pdb     15  -------------TM------RRIVRN--LL-KE-LG-------F------NNV--E-E-   34
1EUCA01.pdb  18  -------------KQ------GTFHSQ--QA-LE-YG-------T------NLV--G-GT   38
1FYV.pdb     26  -------------WV------KNELLP--NL-EK-EG------------------Q-IC   42
1JKJA02.pdb  33  -------------TL------TYEAVK--QT-TD-YG-------F------GQS--TCV-   53
1OI7A02.pdb  33  -------------TL------TYEAAA--AL-SQ-AG-------L------GTT--TTV-   53
1YGYA02.pdb  54  -------------RI------GQLVAQ--RI-AA-FG-------A------YVV--A-Y-   73
7REQA02.pdb  50  -------------RG------QKVIAT--AY-AD-LG-------F------DVD--V-G-   69
1EUCA02.pdb  33  -------------TL------TYEAVH--QT-TQ-VG-------L------GQS--LCV-   53
1GDHA02.pdb  56  -------------SI------GQALAK--RA-QG-FD-------M------DID--Y-F-   75
1JKJB02.pdb   9  -------------RE------AEEAAS--KI-G-----------A--G---PWV--V-KC   27
1QCZ.pdb     14  --------------A------TQFAAE--IF-EI-LN-------V------PHH--V-E-   32
1YGYA03.pdb  44  -------------VL------RLSALR--GL-FS-AVI-EDAV----------T--F-V-   65
1EUCB01.pdb  28  --------------D------IEEVAA--SN-P-----------E--L---IFK--E-QI   45
1JKJB03.pdb  28  -------------GL------AMGTMD--IV-KL-HG-------G------EPA--NFL-   48
1SAYA02.pdb  49  -------------VV------GTEAAK--MA-VG-LG-------A------QVQ--I-F-   68
1YZF.pdb     16  ---L-DE-A-VSPVL------VDLVKR--DI-AA-MG-------LEEV---AVI--N-A-   45
```

```
1BE1.pdb     38  G---------------V--LS-S-Q---------------------------------   43
1EUCB02.pdb  28  Q---------------ILA-G-------------------------------------   32
1GTZA.pdb    54  Q---------------S---N-H-E-----------------------------------   58
1JRL.pdb     40  S---------------I-SGD-T-SQQ---G----L----------------------   50
1Z8HA.pdb    47  G---------------I-RRD-T-SSD---I----A----------------------   57
1BMTA01.pdb  22  ------------------------T---------------------------------   22
1EUCB03.pdb  49  D--------------LGG--GV-K-E---------------------------------   56
1H05.pdb     43  Q---------------S---D-S-E---------------------------------   47
1LI4A02.pdb  51  E---------------I-D-PIN-ALQ---A----A--ME--------G---Y---EVT   68
1U11.pdb     36  I---------------VSAHR-T-P---------------------------------   43
1ZMBA01.pdb  72  A---------------E-GGS-S-IDE---W----ALD----------------GVLF   88
1BMTA02.pdb  44  G---------------V--MV-P-A-------------------------------   49
1F8YA.pdb    40  V---------------P-L-D-NQY-------KGIR--VDEHPEYLHDK---VWATATY   68
1I1H.pdb     58  DFVSSARAALKAGAPILCD--A-E---------------------------------   78
1MX3A02.pdb  80  D---------------P-Y-L-S-DGV---E----R--AL--------G---L---QRV   96
1UQRA.pdb    50  Q---------------A---N-G-E---------------------------------   54
2APJA.pdb   102  A---------------S-GGT-A-IKE---W----ERG----------------SHLY  118
1DXYA02.pdb  74  D---------------P-Y-P---MKGDHPD--------------------F---D-Y   87
1FXWA.pdb    67  G---------------I-GGD-S-TQH---V----L----------------------   77
1J4AA02.pdb  73  D---------------I-F-R---NPE----L---E--KK--------G---Y---Y-V   87
1O4V.pdb     34  I---------------VSAHR-T-P---------------------------------   41
1V8BA02.pdb  51  E---------------I-D-PIC-AIQ---A----V--ME--------G---F---NVV   68
2C4V.pdb     51  Q---------------T---N-F-E---------------------------------   55
1FXWF.pdb    67  G---------------I-GGD-T-TRH---V----L----------------------   77
1JKJA01.pdb  39  T---------------P-G-K-G-G---------TT--HL--------G---L---PVF   53
1OI7A01.pdb  39  T---------------P-G-K-G-G---------ME--VL--------G---V---PVY   53
1XMP.pdb     35  V---------------VSAHR-T-P---------------------------------   42
3CHY.pdb     35  A-------------------E-D-G---------------------------------   38
1EUCA01.pdb  39  T---------------P-G-K-G-G---------KT--HL--------G---L---PVF   53
1FYV.pdb     43  L---------------H-E-R-NFV-------P--------------G---K---SIV   55
1JKJA02.pdb  54  G---------------I---G-GDP---------------------------------   59
1OI7A02.pdb  54  G---------------I---G-GDP---------------------------------   59
1YGYA02.pdb  74  D---------------P-Y-V-S-PAR---A----A--QL--------G---I---E-L   89
7REQA02.pdb  70  P---------------L--FQ-T-P---------------------------------   75
1EUCA02.pdb  54  G---------------I---G-GDP---------------------------------   59
1GDHA02.pdb  76  D---------------T-H-R-A-SSS---DE---A--SY--------Q---A---TFH   93
1JKJB02.pdb  28  Q---------------VHA-G-------------------------------------   32
1QCZ.pdb     33  V---------------VSAHR-T-P---------------------------------   40
1YGYA03.pdb  66  --------------------------N---A----P--AL--------A----------   71
1EUCB01.pdb  46  D---------------I-I-E-G-----------I----------------------   51
1JKJB03.pdb  49  D--------------VGG--GA-T-K---------------------------------   56
1SAYA02.pdb  69  D---------------I-N-VER-LSY---L----E--TL--------FGSRV---ELL   89
1YZF.pdb     46  G---------------M-PGD-T-TED---G----L----------------------   56
```

```
1BE1.pdb     44  --E------D-FINA-AI-------E-------------------------------TK   54
1EUCB02.pdb  33  -------------GR-GKGVF--SSGLK-----------------------------G-   45
1GTZA.pdb    59  --G------E-LVDW-IH-------E-------------------------------AR   69
1JRL.pdb     51  --A------R-LPAL-LK-------Q-------------------------------HQ   61
1Z8HA.pdb    58  --K------R-WLQE-VS-------L-------------------------------RL   68
1BMTA01.pdb  23  -E-------FI-EQD-TE-------E-------------------------------AR   33
1EUCB03.pdb  57  --S------Q-VYQA-FK-------L-------------------------------LT   67
1H05.pdb     48  --A------Q-LLDW-IH-------Q-------------------------------AA   58
1LI4A02.pdb  69  ---------T-MDEA-CQ---------------------------------------E   76
1U11.pdb     44  --D------R-LADY-AR-------T-------------------------------AA   54
1ZMBA01.pdb  89  --R------H-ALTE-AK-------F-------------------------------AE   99
1BMTA02.pdb  50  --E------K-ILRT-AK-------E-------------------------------VN   60
1F8YA.pdb    69  --N------N-DLNG-IK-------T-------------------------------N-   78
1I1H.pdb     79  -------------MV-AHGVTRARLPAGNEVICTLRDPRTPALAAEIGNTRSAAALKLWS  124
1MX3A02.pdb  97  -----S---T-LQDL-LF---------------------------------------H  105
1UQRA.pdb    55  --E------S-LINR-IH-------Q-------------------------------AF   65
2APJA.pdb   119  --E------R-MVKR-TE-------E-------------------------------SR  129
1DXYA02.pdb  88  -----V---S-LEDL-FK---------------------------------------Q   96
1FXWA.pdb    78  --W------R-LENGELE-------H-------------------------------IR   89
1J4AA02.pdb  88  -----D---S-LDDL-YK---------------------------------------Q   96
1O4V.pdb     42  --D------R-MFEY-AK-------N-------------------------------AE   52
1V8BA02.pdb  69  ---------T-LDEI-VD---------------------------------------K   76
2C4V.pdb     56  --G------E-IIDK-IQ-------E-------------------------------SV   66
1FXWF.pdb    78  --W------R-LKNGELE-------N-------------------------------IK   89
1JKJA01.pdb  54  -----N---T-VREA-VA-------A-------------------------------TG   64
1OI7A01.pdb  54  -----D---T-VKEA-VA-------H-------------------------------HE   64
1XMP.pdb     43  --D------Y-MFEY-AE-------T-------------------------------AR   53
3CHY.pdb     39  --V------D-ALNK-LQ-------A-------------------------------GG   49
1EUCA01.pdb  54  -----N---T-VKEA-KE-------Q-------------------------------TG   64
1FYV.pdb     56  --E------N-IITC-IE-------K-------------------------------S-   65
1JKJA02.pdb  60  --IPGS---N-FIDI-LE-------M-------------------------------FE   73
1OI7A02.pdb  60  --VIGT---T-FKDL-LP-------L-------------------------------FN   73
1YGYA02.pdb  90  -----L---S-LDDL-LA---------------------------------------R   98
7REQA02.pdb  76  --E------E-TARQ-AV-------E-------------------------------AD   86
1EUCA02.pdb  60  --FNGT---D-FTDC-LE-------I-------------------------------FL   73
1GDHA02.pdb  94  -----D---S-LDSL-LS---------------------------------------V  102
1JKJB02.pdb  33  -------------GR-GK-------A-------------------------------G-   38
1QCZ.pdb     41  --D------K-LFSF-AE-------S-------------------------------AE   51
1YGYA03.pdb  72  -----------------A---------------------------------------E   73
1EUCB01.pdb  52  -----------KDSQ-AQ-------R-------------------------------MA   60
1JKJB03.pdb  57  --E------R-VTEA-FK-------I-------------------------------IL   67
1SAYA02.pdb  90  Y----SNSAE-IETA-VA---------------------------------------E  102
1YZF.pdb     57  --K------R-LNKE-VL-------I-------------------------------EK   67
```

```
1BE1.pdb     55  --------A-DLICVS-S--LYG---------------QG------E-IDCKG-------   73
1EUCB02.pdb  46  ------------GVHL-T--K-----------------D------P-EVVGQ-------   58
1GTZA.pdb    70  L-----N-H-CGIVIN-P--AAY-SHT-------------------S-VAILD-------   91
1JRL.pdb     62  --------P-RWVLVE-L--GGNDGL------RGFQPQQT------E-QTLRQ-------   89
1Z8HA.pdb    69  HKE-Y---N-SLVVFS-F--GLNDTTLENGKPR-VSIAET------I-KNTRE-------  105
1BMTA01.pdb  34  Q-Q---A-------------------------------------------------------   36
1EUCB03.pdb  68  A-D---PKV-EAILVN-I--FGG----------IV---ZN------A-IIANG-------   92
1H05.pdb     59  D-----A-A-EPVILN-A--GGL-THT-----------------S-VALRD-------   80
1LI4A02.pdb  77  --------G-NIFVTT-T--G------------------------C----ID-------   88
1U11.pdb     55  E-R---G-L-NVIIAG-A--GGA----------------------A-HLPGM-------   74
1ZMBA01.pdb 100  S----SE-L-TGILWH-Q--GESDSL-------NGNYKVY------Y-KKLLL-------  129
1BMTA02.pdb  61  --------A-DLIGLS-G--LIT---------------PS------L-DEMVN-------   79
1F8YA.pdb    79  ----------DIMLGV-Y--I-----------------P--D-EED-VGLGM-------   96
1I1H.pdb    125  E-RLA---G---SVVA-I--GN-----------------A------P-TALFF-------  143
1MX3A02.pdb 106  --------S-DCVTLH-C--G------------------------LNEHNHH-------  121
1UQRA.pdb    66  Q-----N-T-DFIIIN-P--GAF-THT-----------------S-VAIRD-------   87
2APJA.pdb   130  K-C-GGE-I-KAVLWY-Q--GESDVLD------IHDAESY------G-NNMDR-------  162
1DXYA02.pdb  97  --------S-DVIDLH-V--P------------------------GIEQNTH-------  112
1FXWA.pdb    90  --------P-KIVVVW-V--GTNN-H-------GHTAEQV------T-GGIKA-------  115
1J4AA02.pdb  97  --------A-DVISLH-V--P------------------------DVPANVH-------  112
1O4V.pdb     53  E-R---G-I-EVIIAG-A--GGA----------------------A-HLPGM-------   72
1V8BA02.pdb  77  --------G-DFFITC-T--G------------------------N----VD-------   88
2C4V.pdb     67  G-S---D-Y-EGIIIN-P--GAF-SHT-----------------S-IAIAD-------   89
1FXWF.pdb    90  --------P-KVIVVW-V--GTNN-H-------ENTAEEV------A-GGIEA-------  115
1JKJA01.pdb  65  --------A-TASVIY-V--P------------------------A-PFCKD-------   79
1OI7A01.pdb  65  --------V-DASIIF-V--P------------------------A-PAAAD-------   79
1XMP.pdb     54  E-R---G-L-KVIIAG-A--GGA----------------------A-HLPGM-------   73
3CHY.pdb     50  --------Y-GFVISD-W--N-M---------PNM--------------DGLE-------   66
1EUCA01.pdb  65  --------A-TASVIY-V--P------------------------P-PFAAA-------   79
1FYV.pdb     66  ----------YKSIFV-L--S-----------------P--------NFVQS-------   79
1JKJA02.pdb  74  K-D---PQT-EAIVMI-G--EIG----------------------G-SAEEE-------   94
1OI7A02.pdb  74  E-D---PET-EAVVLI-G--EIG----------------------G-SDEEE-------   94
1YGYA02.pdb  99  --------A-DFISVH-L--P------------------------KTPETAG-------  114
7REQA02.pdb  87  --------V-HVVGVS-S--LAG---------------GH------L-TLVPA-------  105
1EUCA02.pdb  74  N-D---PAT-EGIILI-G--EIG----------------------G-NAEEN-------   94
1GDHA02.pdb 103  --------S-QFFSLN-A--P------------------------STPETRY-------  118
1JKJB02.pdb  39  ------------GVKV-V--N-----------------S------K-EDIRA-------   51
1QCZ.pdb     52  E-N---G-Y-QVIIAG-A--GGA----------------------A-HLPG--------   70
1YGYA03.pdb  74  --------RGVTAEIC-K--A-----------------S--E---S----PNHRSVVDV   95
1EUCB01.pdb  61  E-N-----L------GFLGP------------------L------Q-NQAAD-------   75
1JKJB03.pdb  68  S-D---DKV-KAVLVN-I--FGG----------IV---RC------D-LIADG-------   92
1SAYA02.pdb 103  --------A-DLLIGA-V--L-----------------VPGRR--A----PI-------  119
1YZF.pdb     68  --------P-DEVVIF-F--GANDASLD----RNITVATF------R-ENLET-------   97
```

```
1BE1.pdb     74  -L-R--EKCDEA---G------------LK--G-I---K--LFVGGNI--V---------   95
1EUCB02.pdb  59  -L-A--KQMIGYNLATKQT----------PK-E-G---V-KV-----------------   80
1GTZA.pdb    92  -A-L--NTC--------------------D--G-L---P--VVEVH-I--SN---IH-Q-  111
1JRL.pdb     90  -I-L--QDVKAA---N-----------------A---E--PLLMQ-IR-PP---AN---  111
1Z8HA.pdb   106  -I-L--TQAKKL---Y-------------P---V----L--ISPAP-----Y---IE---  125
1BMTA01.pdb  37  ----------------------------T-R------P--------------------   39
1EUCB03.pdb  93  -I-T--KACREL---E------------L--K-V---P--LVVRL-E--GT--------  113
1H05.pdb     81  -A-C--AEL--------------------S----A---P--LIEVH-I--SN---VH-A-   99
1LI4A02.pdb  89  -IIL--GRHFEQ---M------------KD-D-A---I--VCNIG-H------------  109
1U11.pdb     75  -C-A--AWT--------------------R----L---P--VLGVP-V--ES---RA-L-   93
1ZMBA01.pdb 130  -I-I--EALRKE---L---------N---VP-D-I---P--IIIGG-LG-DF-L--G-K-  156
1BMTA02.pdb  80  -V-A--KEMERQ---G------------F--T-I---P--LLIGG-A--T---------   99
1F8YA.pdb    97  -E-L--GYALSQ---G------------K--------Y--VLLVI-P--D---------  114
1I1H.pdb    144  -L-L--EMLR-------DG-------------A-P-KPA-AILGMP-V--G---------  164
1MX3A02.pdb 122  -LIN--DFTVKQ---M------------RQ-G-A---F--LVNTA-R------------  142
1UQRA.pdb    88  -A-L--LAV--------------------S----I---P--FIEVH-L--SN---VH-A-  106
2APJA.pdb   163  -L-I--KNLRHD---L---------N---LP-S-L---P--IIQVA-IA--S--------  185
1DXYA02.pdb 113  -IIN--EAAFNL---M------------KP-G-A---I--VINTA-R------------  133
1FXWA.pdb   116  -I-V--QLVNER---Q------------P--Q-A---R--VVVLG-LL-PR---G----  138
1J4AA02.pdb 113  -MIN--DESIAK---M------------KQ-D-V---V--IVNVS-R------------  133
1O4V.pdb     73  -V-A--SIT--------------------H----L---P--VIGVP-V--KT---ST-L-   91
1V8BA02.pdb  89  -VIK--LEHLLK---M------------KN-N-A---V--VGNIG-H------------  109
2C4V.pdb     90  -A-I--MLA--------------------G----K---P--VIEVH-L--TN---IQ-A-  108
1FXWF.pdb   116  -I-V--QLINTR---Q------------P--Q-A---K--IIVLG-LL-PR---G----  138
1JKJA01.pdb  80  -S-I--LEAIDA---G------------I----K---L--IITIT-E--G---------   98
1OI7A01.pdb  80  -A-A--LEAAHA---G------------I----P---L--IVLIT-E--G---------   98
1XMP.pdb     74  -V-A--AKT--------------------N----L---P--VIGVP-V--QS---KA-L-   92
3CHY.pdb     67  -L-L--KTIRAD---G---------A--MS--A-L---P--VLMVT-A--E---------   88
1EUCA01.pdb  80  -A-I--NEAIDA---E------------V----P---L--VVCIT-E--G---------   98
1FYV.pdb     80  -E-W--CHYELY---FAHH-NLFHEG---SN-------S--LILIL-L--E---------  107
1JKJA02.pdb  95  -A-A--AYIKEH---V------------T----K---P--VVGYI-A--GVTAPKG-KR  121
1OI7A02.pdb  95  -A-A--AWVKDH---M------------K----K---P--VVGFI-G--GR--------  114
1YGYA02.pdb 115  -LID--KEALAK---T------------KP-G-V---I--IVNAA-R------------  135
7REQA02.pdb 106  -L-R--KELDKL---G------------RP--D-I---L--ITVGG-V--I---------  126
1EUCA02.pdb  95  -A-A--EFLKQH---N---------SGPKS----K---P--VVSFI-A--GLTAPPG-RR  125
1GDHA02.pdb 119  -FFN--KATIKS---L------------PQ-G-A---I--VVNTA-R------------  139
1JKJB02.pdb  52  -F-A--ENWLGKRLVTYQT----------DA-N-G---Q-PV-----------------   73
1QCZ.pdb     71  -I-A--AKT--------------------L----V---P--VLGVP-V--QS---AA-L-   89
1YGYA03.pdb  96  RAVGAD-----------------------G-S-V---V--TVSGT--------------  110
1EUCB01.pdb  76  -Q-I--KKLYNL---FLKID---------AT-Q-V---EVNP--FG-E--T---------  100
1JKJB03.pdb  93  -I-I--GAVAEV---G------------V--N-V---P--VVVRL-E--GN--------  113
1SAYA02.pdb 120  -LVP--ASLVEQ---M------------RT-G-S---V--IVDVA-V-D----------  141
1YZF.pdb     98  -M-I--HEIG-----------------------SE--K--VILIT-PP-YA---DSGR-  119
```

```
1BE1.pdb     96  ----------------VGKQNWP-D--------V----E---QRF--KAM-----G---  112
1EUCB02.pdb  81  ------------------------------------------------------N----   81
1GTZA.pdb   112  ---------------R------E-P--------F----R---HHSYVSQR---------  124
1JRL.pdb    112  -----Y-------GRRYN-----E-A--------F----S---AIY--PKL-----A---  128
1Z8HA.pdb   126  ---QQDPG-----RRRRT-----I-D--------L----S---QQL--ALV-----C---  146
1BMTA01.pdb  40  -----------------I-----E-V--------I----E--------------------   44
1EUCB03.pdb 114  ----------------NV-----H-E--------A----Q---NIL--TNS-----G---  126
1H05.pdb    100  ---------------R------E-E--------F----R---RHSYLSPI---------  112
1LI4A02.pdb 110  ---------------F------D-V------EID----V---KWL--NEN-----A-VE  125
1U11.pdb     94  ----------------KG-----M-D--------S----L---LSI--VQM-----P---  106
1ZMBA01.pdb 157  --ERFGKG------CTEY-----N-F--------I----N---KEL--QKF-----A---  177
1BMTA02.pdb 100  ----------------TS----KA-H--------T----A---VKI--EQN-----Y---  113
1F8YA.pdb   115  ---------------------ED---------YGKPIN---LMS--WGV-----S---  129
1I1H.pdb    165  -------------FVGA-----A-E--------S----K---DAL--AEN-----S---  179
1MX3A02.pdb 143  ----------------G--------G------LVD----E---KAL--AQA-----L---  155
1UQRA.pdb   107  ---------------R------E-P--------F----R---HHSYLSDV---------  119
2APJA.pdb   186  ------------G-GGYI-----D-K--------V----R---EAQ--LGL-----K---  201
1DXYA02.pdb 134  ----------------P--------N------LID----T---QAM--LSN-----L---  146
1FXWA.pdb   139  ---QHPNP-----LREKN-----R-R--------V----N---ELV--RAA-----L---  159
1J4AA02.pdb 134  ----------------G--------P------LVD----T---DAV--IRG-----L---  146
1O4V.pdb     92  ----------------NG-----L-D--------S----L---FSI--VQM-----P---  104
1V8BA02.pdb 110  ----------------F------D-D------EIQ----V---NEL--FNY-----KGIH  126
2C4V.pdb    109  ----------------R------E-E--------F----R---KNSYTGAA---------  121
1FXWF.pdb   139  ---EKPNP-----LRQKN-----A-K--------V----N---QLL--KVS-----L---  159
1JKJA01.pdb  99  ----------------IP-----T-L--------D----M---LTV--KVK-----L---  111
1OI7A01.pdb  99  ----------------IP-----T-L--------D----M---VRA--VEE-----I---  111
1XMP.pdb     93  ----------------NG-----L-D--------S----L---LSI--VQM-----P---  105
3CHY.pdb     89  ----------------AK-----K-E--------N----I---IAA--AQA-----G---  101
1EUCA01.pdb  99  ----------------IP-----Q-Q--------D----M---VRV--KHR-----L---  111
1FYV.pdb    108  ---------------PIP-----Q-Y--------S----IPSSYHK--LKS-----L---  124
1JKJA02.pdb 122  MGHAGAIIAGGKG---TA-----D-E--------K----F---AAL--EAA---------  146
1OI7A02.pdb 115  -----------VG---TP-----E-S--------K----L---RAF--AEA---------  128
1YGYA02.pdb 136  ----------------G--------G------LVD----E---AAL--ADA-----I---  148
7REQA02.pdb 127  -----------------P-----E-Q--------D----F---DEL--RKD-----G---  138
1EUCA02.pdb 126  MGHAGAIIAGGKG---GA-----K-E--------K----I---TAL--QSA---------  150
1GDHA02.pdb 140  ----------------G--------D------LVD----N---ELV--VAA-----L---  152
1JKJB02.pdb  74  ------------------------------------------------------N----   74
1QCZ.pdb     90  ----------------SG-----V-D--------S----L---YSI--VQ------P---  101
1YGYA03.pdb 111  -------------------------------L--Y----G---PQ-------------L-  116
1EUCB01.pdb 101  ----------------------P--------------------------EGQVVC---  107
1JKJB03.pdb 114  ----------------NA-----E-L--------G----A---KKL--ADS-----G---  126
1SAYA02.pdb 142  ----------------Q------G-GCVETLHPTS----H---TQP-------------  157
1YZF.pdb    120  --RPER-------PQTRI-----K-E--------L----V---KVA--QEV-----G---  139
```

```
1BE1.pdb    113  ------------------FDRVYPP-----------------------G---------  120
1EUCB02.pdb  82  ------------------KVMVAEA---------------------------------   88
1GTZA.pdb   125  -----------------ADGVVAG-------------------------------CGV  134
1JRL.pdb    129  K-EF-------------D-VPLLPFFME-EVYL----KPQW-MQDDGIHPNRDAQPFIA  166
1Z8HA.pdb   147  Q-DL-------------D-VPYLDV-FP-LLEKPSV-WLHEAKANDGVHPQA----GGY  183
1BMTA01.pdb      ------------------------------------------------------------
1EUCB03.pdb 127  -----------------LPITSAV---------------------------------  133
1H05.pdb    113  -----------------ATGVIVG-------------------------------LGI  122
1LI4A02.pdb 126  K-VNIKPQVDRYR-LKNGRRIILLA-----------------------------------  148
1U11.pdb    107  ---G-------------GVPVGTLA-------------------------------IGA  118
1ZMBA01.pdb 178  F-EQ-------------DNCYFVTA-SG-L----------T-CNPDGIHIDA----ISQ  205
1BMTA02.pdb 114  -----------------SGPTVYVQ---------------------------------  121
1F8YA.pdb   130  ------------------DNV--I---------------------------------  133
1I1H.pdb    180  ---Y-------------GVPFAIV-RG---------------------RLG------  192
1MX3A02.pdb 156  K-EG-------------RIRGAAL---------------------------------  165
1UQRA.pdb   120  -----------------AKGVICG-------------------------------LGA  129
2APJA.pdb   202  ---L-------------SNVVCVDA-KG-L----------P-LKSDNLHLTT----EAQ  227
1DXYA02.pdb 147  K-SG-------------KLAGVGI---------------------------------  156
1FXWA.pdb   160  A-GH-------------PRAHFLDA-DPGFVHSDGTISHH--DMYDYLHLSR----LGY  197
1J4AA02.pdb 147  D-SG-------------KIFGYAM---------------------------------  156
1O4V.pdb    105  ---G-------------GVPVATVA-------------------------------I-N  115
1V8BA02.pdb 127  I-ENVKPQVDRIT-LPNGNKIIVLA---------------------------------  149
2C4V.pdb    122  -----------------CGGVIMG-------------------------------FGP  131
1FXWF.pdb   160  P-KL-------------ANVQLLDT-DGGFVHSDGAISCH--DMFDFLHLTG----GGY  197
1JKJA01.pdb 112  D-EA-------------GVRMIG---------------------------------  120
1OI7A01.pdb 112  K-AL-------------GSRLIG---------------------------------  120
1XMP.pdb    106  ---G------------GVPVATVA-------------------------------IGK  117
3CHY.pdb    102  -----------------ASGYVVK----------------------PF-------TA  112
1EUCA01.pdb 112  L-RQ-------------GKTRLIG---------------------------------  121
1FYV.pdb    125  AR-R-------------TYLEWPK-------------------------EK----SKR  139
1JKJA02.pdb 147  -----------------GVKTVR---------------------------------  152
1OI7A02.pdb 129  -----------------GIPVAD---------------------------------  134
1YGYA02.pdb 149  T-GG-------------HVRAAGL---------------------------------  158
7REQA02.pdb 139  -----------------AVEIYTP----------------------GT---------  147
1EUCA02.pdb 151  -----------------GVVVSM---------------------------------  156
1GDHA02.pdb 153  E-AG-------------RLAYAGF---------------------------------  162
1JKJB02.pdb  75  -----------------QILVEAA---------------------------------   81
1QCZ.pdb    102  ---R------------GIPVGTLA-------------------------------IGK  113
1YGYA03.pdb 117  -------SQKIVQI----NGRHFDL---------------------------------  130
1EUCB01.pdb 108  ---F-------------DAKINFD-DN---------------------A--------  118
1JKJB03.pdb 127  -----------------LNIIAAK---------------------------------  133
1SAYA02.pdb 158  --------TYEV------FGVVHYG---------------------------------  168
1YZF.pdb    140  A-AH-------------N-LPVIDL-YK-AMTVYPG-TDEF-LQADGLHFSQ----VGY  175
```

```
1BE1.pdb    121  --TS--------------------PE---TT-IA--DM-KE------------------  132
1EUCB02.pdb      ------------------------------------------------------------
1GTZA.pdb   135  Q-GY--------------------VF---GV-ER--IA-AL------------------  147
1JRL.pdb    167  DWMA--------------------KQ---LQ-PL--VN--------------------  178
1Z8HA.pdb   184  TEFA--------------------RI---VE-NWDAWL-NW------------------  199
1BMTA01.pdb  45  --GP--------------------LM---DG-MN--VV-GD------------------   56
1EUCB03.pdb 134  --DL--------------------ED---AA-KK--AV-AS------------------  145
1H05.pdb    123  Q-GY--------------------LL---AL-RY--LA-EH------------------  135
1LI4A02.pdb 149  --EGR-----------------------------LV--NL-GC------------------  157
1U11.pdb    119  S-GA--------------------KN---AA-LL--AA-SI------------------  131
1ZMBA01.pdb 206  RKFG--------------------LR---YF-EA--FF-NRKHVLEPLINENELLNLNY  237
1BMTA02.pdb 122  --NA--------------------SR---TV-GV--VA-AL------------------  133
1F8YA.pdb   134  --KM--------------------SQ---LK-DF--N-F-N------------------  144
1I1H.pdb    193  --GS--------------------AM---TA-AA--L--NS------------------  203
1MX3A02.pdb 166  ---DVHESEP----F-S--------FS----Q-GP--LK-DA------------------  183
1UQRA.pdb   130  K-GY--------------------DY---AL-DF--AI-SE------------------  142
2APJA.pdb   228  VQLG--------------------LS---LA-QA--YL-SN------------------  241
1DXYA02.pdb 157  ---DTYEYETEDLLNL-AKHGSF-KDP---LW-DE--LL-GM------------------  186
1FXWA.pdb   198  TPVC--------------------RA---LH-SL--LL-RL------------------  211
1J4AA02.pdb 157  ---DVYEGEVGIFNEDWEGKE--FPDA---RL-AD--LI-AR------------------  186
1O4V.pdb    116  --NA--------------------KN---AG-IL--AA-SI------------------  127
1V8BA02.pdb 150  --RGR-----------------------------LL--NL-GC------------------  158
2C4V.pdb    132  L-GY--------------------NM---AL-MA--MV-NI------------------  144
1FXWF.pdb   198  AKIC--------------------KP---LH-EL--IM-QL------------------  211
1JKJA01.pdb 121  ---P------------------------------------------------------  121
1OI7A01.pdb 121  ---G------------------------------------------------------  121
1XMP.pdb    118  A-GS--------------------TN---AG-LL--AA-QI------------------  130
3CHY.pdb    113  A-TL--------------------EE---KL-NK--IF-EK------------------  125
1EUCA01.pdb 122  ---P------------------------------------------------------  122
1FYV.pdb    140  GLFW--------------------AN---LR-AA--I--NI------------------  152
1JKJA02.pdb 153  --SL--------------------AD---IG-EA--LK-TV------------------  164
1OI7A02.pdb 135  --TI--------------------DE---IV-EL--VK-KA------------------  146
1YGYA02.pdb 159  ---DVFATEP----C-T-------------D-SP--LF-EL------------------  174
7REQA02.pdb 148  --VI--------------------PE---SA-IS--LV-KK------------------  159
1EUCA02.pdb 157  --SP--------------------AQ---LG-TT--IY-KE------------------  168
1GDHA02.pdb 163  ---DVFAGEP----N-I-------------N-EG--YY-DL------------------  178
1JKJB02.pdb      ------------------------------------------------------------
1QCZ.pdb    114  A-GA--------------------AN---AA-LL--AA-QI------------------  126
1YGYA03.pdb 131  ----R-------------------A------Q---------------------------  133
1EUCB01.pdb 119  -------------------------EFRQK-DI--F--AM------------------  128
1JKJB03.pdb 134  --GL--------------------TD---AA-QQ--VV-AA------------------  145
1SAYA02.pdb 169  ---VP--------------------------NM--PG-AV------------------  176
1YZF.pdb    176  ELLG--------------------AL---IV-RE--IK-GR------------------  189
```

```
1BE1.pdb    133  ---V----L-G----V-----------------E-------------------------  137
1EUCB02.pdb  89  ------------------------L----DI----------------------------   91
1GTZA.pdb   148  ---A----G-------------------------------------------------  149
1JRL.pdb         ------------------------------------------------------------
1Z8HA.pdb   200  ---FR----------------------------------------------------  201
1BMTA01.pdb  57  ---L----F-G----EGKMFL----PQV-----VKSARVMKQAVAYLEPFIE--------   87
1EUCB03.pdb 146  ---V-------------------------------------------------------  146
1H05.pdb    136  ---V----G-T-----------------------------------------------  138
1LI4A02.pdb 158  ---A----------------------------MG-------------------------  160
1U11.pdb    132  ---L----A-L----Y-----------N-----PALAARLETWRALQTASVPN-------  156
1ZMBA01.pdb 238  ART-------------------------------------------------------  240
1BMTA02.pdb 134  ---L----S-DT---------------------QRDDFVARTRKEYETVRIQHG-----  158
1F8YA.pdb   145  ------KP-------R----FDFYE----G-----------------------A-----  154
1I1H.pdb    204  ------LAR-P-GL--------------------------------------------  209
1MX3A02.pdb 184  ---P-NLI--------------CTP----HAAW--------------------------  194
1UQRA.pdb   143  ---L----Q-K----I-------------------------------------------  146
2APJA.pdb   242  ---F----C-------------------------------------------------  243
1DXYA02.pdb 187  ---P-NVV--------------LSP----HIA---------------------------  196
1FXWA.pdb        ------------------------------------------------------------
1J4AA02.pdb 187  ---P-NVL--------------VTP----KTA---------------------------  196
1O4V.pdb    128  ---L----G-I----K-----------Y-----PEIARKVKEYKERMKREVLEKA-QRLE  158
1V8BA02.pdb 159  ---A----------------------------TG-------------------------  161
2C4V.pdb    145  ---L----A-E----M-----------K-------------AFQEAQKNN----------  158
1FXWF.pdb   212  ---L-------------------------------------------------------  212
1JKJA01.pdb 122  ----------------------------N-------------------------------  122
1OI7A01.pdb 122  ----------------------------N-------------------------------  122
1XMP.pdb    131  ---L----G-S----F-----------H-----DDIHDALELRREAIEKDVRE-------  155
3CHY.pdb    126  ---L----GM------------------------------------------------  128
1EUCA01.pdb 123  ----------------------------N-------------------------------  123
1FYV.pdb    153  ------KL-------------TEQA----K-------------------------------  159
1JKJA02.pdb 165  ---L-------------------------------------------------------  165
1OI7A02.pdb 147  ---L-------------------------------------------------------  147
1YGYA02.pdb 175  ---A-QVV--------------VTP----H-------------------------------  182
7REQA02.pdb 160  ---L----R-A----S-----------------------------------L----  164
1EUCA02.pdb 169  ---F----E-------------------------------------------------  170
1GDHA02.pdb 179  ---P-NTF--------------LF-----------------------------------  184
1JKJB02.pdb  82  ------------------------T----DI----------------------------   84
1QCZ.pdb    127  ---L----A-T----H-----------D-----KELHQRLNDWRKAQTDEVLENP----D  154
1YGYA03.pdb      ------------------------------------------------------------
1EUCB01.pdb 129  ------DD------K--------SE--NE-------------------------------  135
1JKJB03.pdb 146  ---V----E-G----K-----------------------------------------  149
1SAYA02.pdb      ------------------------------------------------------------
1YZF.pdb    190  ---L----K-------------PKQ----A-------------------------------  195
```

```
1BE1.pdb          ---------------
1EUCB02.pdb       ---------------
1GTZA.pdb         ---------------
1JRL.pdb          ---------------
1Z8HA.pdb         ---------------
1BMTA01.pdb       ---------------
1EUCB03.pdb       ---------------
1H05.pdb          ---------------
1LI4A02.pdb       ---------------
1U11.pdb    157   SP----------I--   159
1ZMBA01.pdb       ---------------
1BMTA02.pdb       ---------------
1F8YA.pdb   155   ------------VY-   156
1I1H.pdb          ---------------
1MX3A02.pdb       ---------------
1UQRA.pdb         ---------------
2APJA.pdb         ---------------
1DXYA02.pdb       ---------------
1FXWA.pdb         ---------------
1J4AA02.pdb       ---------------
1O4V.pdb    159   QIGYKEYLNQK----   169
1V8BA02.pdb       ---------------
2C4V.pdb          ---------------
1FXWF.pdb         ---------------
1JKJA01.pdb       ---------------
1OI7A01.pdb       ---------------
1XMP.pdb          ---------------
3CHY.pdb          ---------------
1EUCA01.pdb       ---------------
1FYV.pdb          ---------------
1JKJA02.pdb       ---------------
1OI7A02.pdb       ---------------
1YGYA02.pdb       ---------------
7REQA02.pdb       ---------------
1EUCA02.pdb       ---------------
1GDHA02.pdb       ---------------
1JKJB02.pdb       ---------------
1QCZ.pdb    155   PR---------GA-A   159
1YGYA03.pdb       ---------------
1EUCB01.pdb       ---------------
1JKJB03.pdb       ---------------
1SAYA02.pdb       ---------------
1YZF.pdb          ---------------
```