UNIVERSITÉ DE REIMS CHAMPAGNE-ARDENNE

écoles .urca Doctorales

Outils d'aide à l'étude des protéines :

modélisation surfacique et visualisation sémantique des feuillets β

THÈSE

présentée et soutenue publiquement le 09 juillet 2010

pour l'obtention du

Doctorat de l'Université de Reims Champagne-Ardenne
(spécialité Informatique)

par

Loïc NOLIN

Composition du jury

Rapporteurs :	Mme. Isabelle CALLEBAUT	Directeur de recherche, CNRS Paris 6 et Paris 7
	M. Christophe RENAUD	Professeur, LISIC Calais
Examinateurs :	M. Jean-Paul MORNON	Directeur de recherche émérite, CNRS Paris 6 et Paris 7
	M. Andreas HILDEBRANDT	Assistant Professor, Universität des Saarlandes
	M. Aassif BENASSAROU	Maître de conférences, URCA (encadrant)
Directeurs :	M. Yannick RÉMION	Professeur, URCA
	M. Manuel DAUCHEZ	Professeur, URCA

Remerciements

Je souhaite remercier Isabelle Callebaut ainsi que Christophe Renaud, qui ont accepté d'être les rapporteurs de ma thèse. Je remercie également Jean-Paul Mornon et Andreas Hildebrandt d'avoir bien voulu être les examinateurs de mon jury.

Je remercie Manuel Dauchez de m'avoir accueilli au sein du laboratoire SiRMa en partageant son bureau avec moi. Merci Manu pour ton enthousiasme sans faille, ta bonhommie et ton entrain. Il est parfois difficile de t'attraper (tenter de te suivre est illusoire), mais tu réponds toujours présent...

Je remercie Yannick Rémion, qui m'a ouvert les portes du LERI (même si ce n'est plus sa dénomination officielle), dont les interventions sont toujours impressionnantes de perspicacité et d'efficacité. Même lorsqu'il ne s'agit pas de son domaine de prédilection.

Je remercie Aassif Benassarou qui a accepté la lourde tâche de me prendre en charge. Merci pour les claques amicales derrière la tête et pour ton exactitude.

Je remercie Antoine Jonquet qui n'hésite jamais à apporter son aide et son expérience, quitte parfois à y passer du temps.

Je remercie Jérôme Cutrona pour le temps consacré, et la patience dont il a fait preuve, à relire et corriger mon manuscrit.

Je remercie également l'ensemble du LERI et du SiRMa, pour les soirées (R.I.P. Stewball), les billards, les championnats du monde de Shifumi, les mastars, les p'tits jaunes et les grandes mousses. Les bons moments passés à la B.H., et le lot de découvertes qui s'y rattache... À Laurent, Manu, Yannick, Hervé, Aa!, Fred, Sylvia, Jérôme, Antoine, Jess, Jean-Mi, Olivier, Didier, Stéph, Barbara, Cédric, Philou, Cyril de Reims, Herman, Gauthier, David, Thomas, Manu (Pouet-Pouet), Dude, Callaghan, Nico, Nanaï, Puppet, la Coq', Stéphane, Laurent Martiny, les membres de la chaîne, et tout ceux que j'oublie de citer... Et bien sûr, le « Z », à qui je dois tout.

Je remercie Stéph qui a su m'accompagner durant ces années, et qui a supporté les aléas de la vie d'un thésard.

Enfin je souhaite remercier tout particulièrement Hervé Kaplan qui m'a mis le pied à l'étrier, et sans qui je n'aurais pas connu ce parcours, sans compter les personnes que j'y ai rencontrées dont certaines sont devenues des amis chers.

Table des matières

Index des illustrations

Index des tables

Chapitre 1
Introduction

« To start press any key... Where's the ANY key? »

Homer J. SIMPSON

Cette thèse présente les travaux réalisés entre 2006 et 2010. Ces travaux ont été financés par une allocation de la région Champagne-Ardenne sur un projet de recherche porté par le groupe « Signal, Image et Connaissance » du CReSTIC et le laboratoire de « Signalisation et Récepteurs Matriciels » de l'UMR CNRS 6237 MEDyC de l'Université de Reims.

Étant titulaire d'un DEA de biochimie et d'une licence professionnelle d'imagerie numérique, la nature atypique de mon cursus m'a permis d'accéder à ce sujet transversal en informatique graphique et biologie structurale. Tout au long de ce manuscrit, je m'adresserai aux deux communautés. C'est pourquoi certaines notions, parfois triviales, seront rappelées, voire détaillées, afin que chaque communauté puisse les appréhender.

Dans les activités associées à la biologie, l'arrivée relativement récente de la révolution génomique a obligé la communauté scientifique à élargir son champ d'investigation aux relations séquences – structures – fonctions tant pour des raisons cognitives qu'économiques. Dans ce triptyque, les approches actuelles de bio-informatique (archivage, « datamining », analyses de séquences, etc…) ne peuvent remonter directement à la fonction d'une protéine à partir de sa seule séquence d'acides aminés. La structure tridimensionnelle est donc encore l'intermédiaire obligé

pour accéder aux fonctions. Actuellement, les tentatives de résolutions expérimentales (cristallographie à rayons X, résonance magnétique nucléaire) quasi-automatiques et à haut débit (à l'image de la fourniture des séquences) n'ont pas encore montré leur complète efficacité. Il y a là un véritable verrou méthodologique contre lequel les techniques de calcul (modélisation moléculaire) sont sollicitées pour apporter des réponses. La modélisation moléculaire est donc confrontée à des exigences bien plus fortes que par le passé et elle doit aussi relever des défis technologiques (dans sa façon d'aborder les problèmes de structures) bien plus complexes. Ainsi, les projets transversaux et multidisciplinaires entre différentes communautés scientifiques, peuvent apporter un supplément d'informations très important et fondamental pour augmenter l'information existante mais non appréhendable ou visible. Cette thèse initiée entre informaticiens spécialisés dans l'imagerie et spécialistes de modélisation moléculaire s'inscrit dans cette démarche.

Mon travail a consisté en la création de nouveaux modes de représentation en modélisation moléculaire, afin de représenter des motifs structuraux réguliers des protéines que sont les feuillets β. Les représentations classiques de ces motifs ne sont pas satisfaisantes dans la mesure où seuls les éléments constitutifs sont représentés. Si nous prenons le parallèle avec un éventail, cela reviendrait à représenter celui-ci par sa seule armature, sans y ajouter sa feuille. C'est pourquoi nous nous sommes intéressés à représenter les feuillets β dans leur ensemble.

Le but d'une telle représentation est de pouvoir mieux appréhender les feuillets β afin de faciliter l'étude des protéines qui les contiennent. Ce mode de visualisation s'avère apporter de nombreuses informations structurales, topologiques et topographiques, sans pour autant complexifier la vision de la macromolécule étudiée. Ces informations sont obtenues lors de visualisations statiques, ou lors de visualisations de trajectoires issues de simulations de dynamique moléculaire. Pour cela il faut pouvoir :

- utiliser les données, concernant les feuillets β, disponibles dans les fichiers dédiés ;
- calculer les données manquantes, ou incomplètes ;
- visualiser les feuillets β à l'aide d'outils issus de l'informatique graphique ;
- interagir avec chaque feuillet de manière indépendante ;
- représenter des informations importantes (nature du feuillet, nature de ses composants, nature physico-chimique...) concernant le feuillet.

Une fois nos modèles de visualisation des feuillets β opérationnels, nous les utilisons afin d'en tirer des enseignements de différente nature :

- influence des feuillets β au sein d'une protéine,
- composition du feuillet : informations structurales, informations physico-chimiques au niveau atomique et résiduel,
- stabilité du feuillet au cours du temps lors de simulation de dynamique moléculaire,
- conséquences de la topologie et de la position du feuillet au sein de la protéine.

Dans un premier chapitre, nous présenterons les différents aspects de ce travail, avec une discussion historique des informations disponibles. Ainsi, une description des différents niveaux de complexité des protéines sera effectuée avec un accent particulier mis sur la nature des amino-acides qui composent ces macromolécules, et sur les éléments de structures secondaires, plus particulièrement les brins β et les feuillets β. Les méthodes d'obtention des structures seront présentées, ainsi que la base de données les contenant. A partir de ces informations, de nombreuses études ont pu être menées tant sur les aspects prédictifs que sur la définition de nouvelles bases de données issues des informations structurales. Ensuite, les différents éléments liés à la modélisation moléculaire et les méthodes de graphisme moléculaire associées seront présentés. Ce chapitre se terminera par les éléments de matériels et méthodes nécessaires à la réalisation du travail : en premier lieu, les langages informatiques et les choix méthodologiques effectués comme les courbes de Bézier et les splines de Catmull-Rom puis les informations structurales contenues dans les fichiers issus de la « *Protein Data Bank* » (PDB), et enfin la méthode « *Define Secondary Structure of Proteins* » (DSSP) d'attribution des structures secondaires.

Dans le chapitre suivant, nous décrirons les différents modèles de feuillets β que nous avons dû développer au cours de ce travail, ayant choisi l'environnement du logiciel BALLView. Nous avons entrepris d'utiliser les informations structurales définies dans les fichiers issus de la PDB par les auteurs qui les déposent. Cette première tentative s'étant avérée intéressante, nous avons continué en développant un modèle à base de splines de Catmull-Rom. Après avoir effectué une interpolation unidimensionnelle par brin puis bidimensionnelle, tenant compte des problématiques intrinsèques de l'organisation des feuillets β, nous avons obtenu un modèle tout à fait satisfaisant. Cependant, si ce modèle présente des vertus indéniables, il s'avère être dépendant de l'information contenue dans le fichier issu de la PDB et se révèle incapable de traduire correctement des

phénomènes d'apparition de trous au sein de grands feuillets, de déchirures, de réorientation de feuillets, voire d'invaginations. Nous avons alors choisi de développer un nouveau modèle en utilisant cette fois des carreaux de Bézier. Cette nouvelle représentation s'est avérée très efficace, que ce soit dans des conditions statiques de visualisation de structures tridimensionnelles ou lors de l'étude du comportement de structures durant des trajectoires de dynamique moléculaire. Dans ce dernier cas, notre modèle a pu s'apparenter à « un tapis volant ». Ces types de représentation étant définis, nous avons apporté un supplément d'information en plaquant, sur les surfaces nouvellement modélisées, des textures qu'il est possible de combiner avec de multiples modes de coloration prédéfinis dans le logiciel ou que nous avons développés pour nos besoins.

Le dernier chapitre présentera les possibilités de ces modes de représentation graphique. Dans un premier temps, les possibilités créées seront étudiées sur les différents niveaux de complexité des protéines. Une attention particulière sera prêtée aux éléments considérés seuls mais aussi au sein de structures tertiaires et quaternaires. Nous avons ainsi testé plus de 1500 fichiers pour, dans une première étape, vérifier la robustesse de nos algorithmes, mais aussi pour appréhender la plus-value apportée par nos modèles. Ce chapitre comprendra une utilisation sur le niveau le plus simple de CATH, base de données de classification structurale, ainsi qu'une application aux amyloïdes. S'ensuit une étude des structures amyloïdes, qui se prêtent particulièrement bien aux représentations simplifiées de feuillets β. Nous avons donc dans une première étape évalué l'apport de cette visualisation sur un feuillet β type issu de l'amyloïde β 1-42, puis nous avons envisagé les différents types de β-solénoïdes qui existent, pour enfin tester les différentes topologies de structures amyloïdes.

Enfin, ce travail se terminera par une conclusion dans laquelle nous rappellerons les intérêts et les apports des modèles que nous avons développés. Pour finir nous discuterons des perspectives envisagées.

Chapitre 2

Contexte scientifique

« How is education supposed to make me feel smarter? Besides, every time I learn something new, it pushes some old stuff out of my brain. Remember when I took that home winemaking course, and I forgot how to drive? »

Homer J. SIMPSON

2.1 De l'atome à la protéine macromoléculaire

En 1805, John Dalton [Dalton1808] établit la théorie atomique, selon laquelle la matière est composée d'atomes de masses différentes qui se combinent. En 1808, il publie les bases de la table périodique des éléments qui forme la base de la chimie moderne. Le terme atome vient du grec ancien ατομος (atomos) « que l'on ne peut diviser ».

Le concept de molécule est introduit en 1811 par Amedeo Avogadro [Avogadro1811] qui se base sur les travaux de Dalton pour énoncer la loi d'Avogadro, spécifiant que deux volumes égaux de gaz parfaits différents, dans les mêmes conditions de température et de pression, contiennent un nombre identique de molécules. Il définit un nombre correspondant au nombre d'atomes dans 12 grammes de carbone 12, le nombre d'Avogadro. Le nom molécule provient du latin scientifique *molecula* signifiant « masse ». Une molécule est un ensemble, électriquement neutre, d'atomes liés par des liaisons chimiques très fortes : les liaisons covalentes.

Il faudra attendre 1911 et la publication des travaux de Jean Perrin [Perrin1911] pour que l'existence des molécules ne fasse plus débat au sein de la communauté scientifique. Perrin démontre la théorie et les expériences publiées en 1905 par Albert Einstein [Einstein1905] sur le mouvement aléatoire des molécules, il détermine également une valeur précise du nombre d'Avogadro.

Les molécules possédant au moins plusieurs dizaines d'atomes sont appelées macromolécules. Les macromolécules sont impliquées dans tous les processus de la vie, ce sont les macromolécules biologiques. Leur taille varie de quelques dizaines à plusieurs centaines de milliers d'atomes. Les lipides, les sucres, les vitamines, les acides nucléiques tels ADN ou ARN et les protéines sont des macromolécules biologiques.

C'est Gerardus J. Mulder en 1838 [Mulder1838] qui prouva que les protéines sont des entités chimiques finies. Le terme protéine, qui lui a été proposé par Jöns J. Berzelius, provient du grec πρωτος (prôtos) qui signifie « premier, essentiel ». L'étude des protéines a montré que ces macromolécules biologiques sont composées d'entités, les acides aminés, liées entre elles par des liaisons covalentes. Une protéine est donc une séquence d'acides aminés. Ils furent progressivement identifiés entre 1820 et 1918.

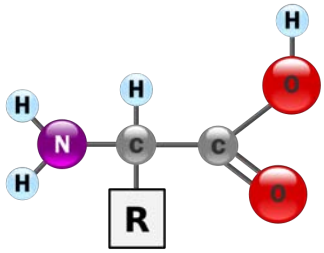


Figure 2.1.1 – Structure commune à tous les acides aminés

Les acides aminés sont des molécules caractérisées par un carbone α (Cα) qui porte quatre groupements : un hydrogène, une fonction amine (-NH_2), un fonction acide carboxylique (-COOH) et une chaîne latérale R (Fig. 2.1.1). La molécule sans R est appelée chaîne principale, elle est commune à tous les acides aminés. R (pour radical) représente une chaîne latérale spécifique, c'est elle qui détermine la nature de l'acide aminé. Il existe vingt acides aminés différents dans les protéines.

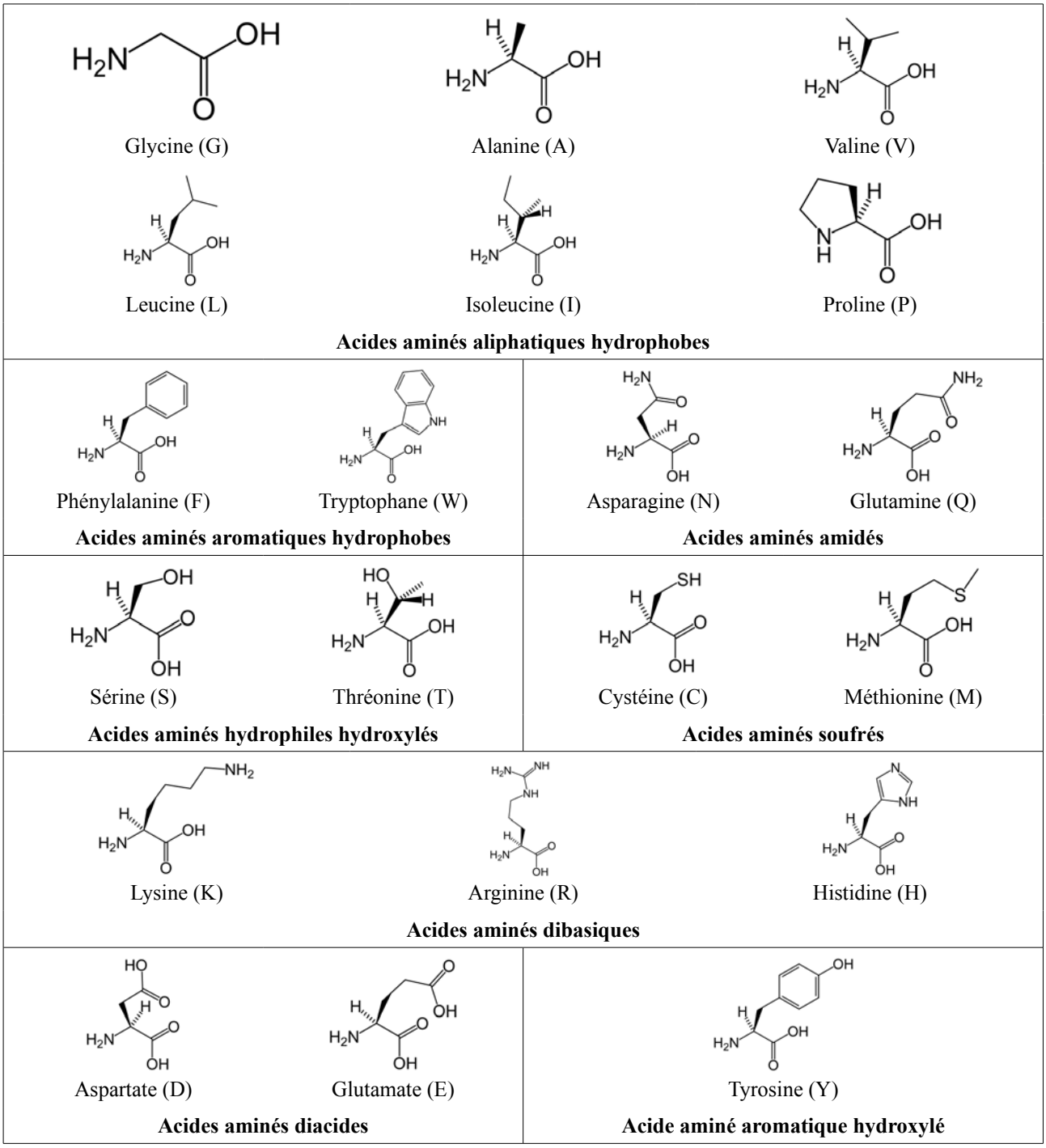


Figure 2.1.2 – Les différents acides aminés présents dans les protéines

Il existe deux stéréo-isomères de chaque acide aminé, suivant la position à droite ou à gauche du groupement -NH_2 en représentation de Fischer [Fischer1891] (type de représentation utilisé dans la figure 2.1.2). La forme gauche est la plus répandue dans la nature.

Les acides aminés sont regroupés par familles en fonction des propriétés chimiques du radical (Fig. 2.1.2). Plusieurs types de classement sont possibles puisque certains acides aminés peuvent

entrer dans plusieurs catégories. Le diagramme de Venn est utilisé [Venn1880] (Fig. 2.1.3) afin de représenter l'ensemble de leurs propriétés.

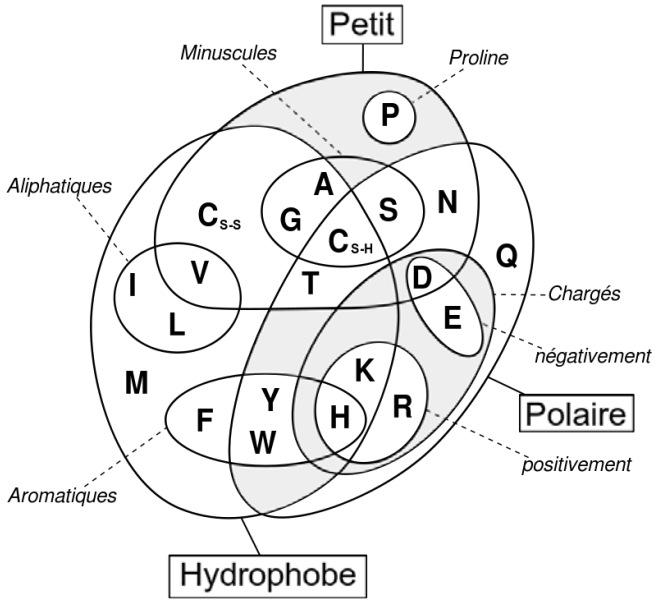


Figure 2.1.3 – Diagramme de Venn des propriétés des acides aminés. Les différentes catégories sont faites en fonction de la nature de la chaîne latérale R. Trois grandes catégories se chevauchent, les hydrophobes, les polaires et les petits. Il existe également plusieurs sous-catégories, les aliphatiques, les aromatiques, les neutres, les chargés positivement, les chargés négativement, les minuscules et la proline. Certains acides aminés, telle la glutamine, n'appartiennent qu'à une seule catégorie, a contrario*, d'autres telle la thréonine, qui fait partie des trois catégories majeures, partagent plusieurs appartenances*

Pour former une protéine, les acides aminés se lient entre eux par une liaison covalente plane qui résulte de l'association du groupement amine (NH_2) d'un résidu avec le groupement acide carboxylique (COOH) du résidu suivant. C'est une réaction de condensation qui libère d'une part le dipeptide formé par l'association des acides aminés et d'autre part une molécule d'eau (Fig. 2.1.4).

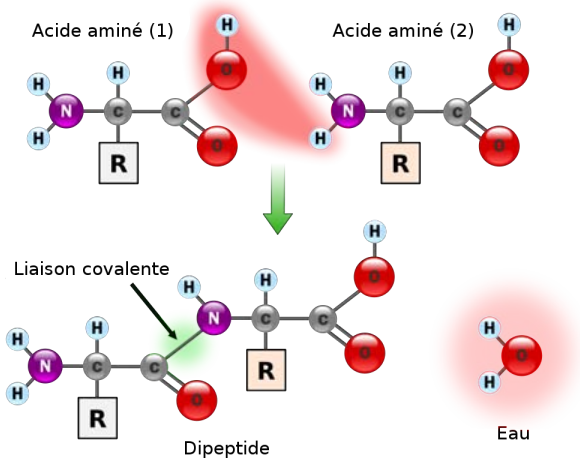


Figure 2.1.4 – Réaction de condensation lors de la formation d'un dipeptide

La liaison peptidique relie les atomes CONH, elle est dite plane parce que les centre des atomes CONH et les deux Cα des acides aminés sont situés dans un même plan. Les plans de deux liaisons peptidiques successives s'orientent l'un par rapport à l'autre par des rotations des liaisons C-N et C-C. C'est la disposition de ces plans qui va déterminer la structure tridimensionnelle de la protéine.

Les structures du Cα et de la liaison peptidique imposent une géométrie spatiale à l'enchaînement des acides aminés, représentée par trois angles dièdres φ_i ($C_{i-1}N_iC_\alpha C_i$), ψ_i ($N_iC_\alpha C_iN_{i+1}$) et ω_i ($C_\alpha C_iN_{i+1}C_{\alpha+1}$) (Fig. 2.1.5).

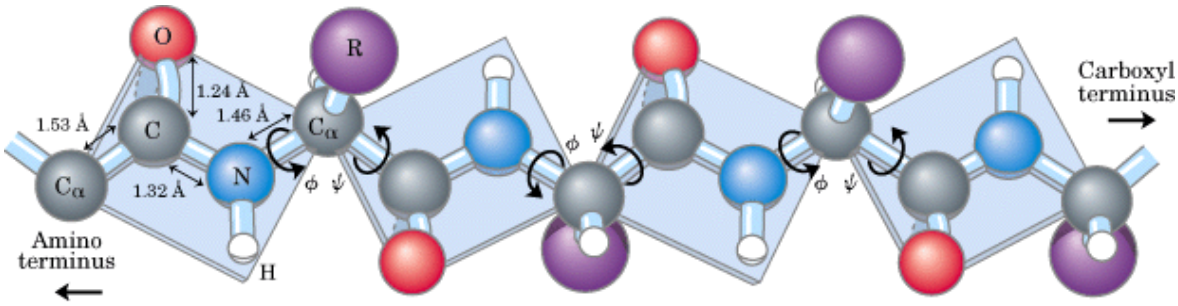


Figure 2.1.5 – Les angles dièdres de la liaison peptidique

L'enchaînement des acides aminés est appelé structure monodimensionnelle, structure primaire de la protéine ou séquence. L'étude de la séquence des acides aminés d'une protéine a été rendue possible par l'introduction de réactifs chimiques spécifiques des groupes des chaînes latérales, et en particulier de l'extrémité amino-terminale, c'est à dire à partir du premier acide aminé de la chaîne polypeptidique. Une avancée décisive fut apportée par Frederick Sanger qui développa

une technique utilisant un réactif spécifique, le 1-fluoro-2,4-dinitrobenzène pour l'identification des acides aminés N-terminaux. Sanger et ses collaborateurs utilisèrent cette méthode pour déterminer la séquence des deux chaînes de l'insuline, ce qui leur demanda dix années de travail. En 1953 Sanger [Sanger1953a; Sanger1953b] publie les 53 acides aminés de la structure primaire de l'insuline, il sera récompensé du prix Nobel de Chimie en 1958 pour ses travaux sur la structure des protéines, et notamment l'insuline.

Les années qui s'ensuivirent virent le nombre de séquences résolues augmenter de telle façon qu'en 1965 Margaret O. Dayhoff [Dayhoff1965] publie la première édition de l'atlas des séquences protéiques.

D'abord manuelle, les techniques de séquençage se sont peu à peu perfectionnées, puis automatisées. Les banques de données actuelles contiennent un grand nombre de séquences (de l'ordre de plusieurs dizaines de milliers) obtenues, soit par séquençage de la protéine, soit par séquençage du gène correspondant. Grâce au développement intense de la bioinformatique, des algorithmes permettant la recherche d'homologies, et l'alignement de séquences permirent de révéler l'existence de familles et superfamilles de protéines. C'est à dire de protéines contenant plus de 30 à 40 % d'identité de séquence. À ce jour, plus de 3 000 superfamilles ont été identifiées.

Nous connaissons alors un premier niveau de structure. Cependant les protéines ne sont pas des objets linéaires, la chaîne polypeptidique se replie pour acquérir sa structure dans l'espace à trois dimensions. La structure tridimensionnelle d'une protéine est la condition nécessaire à l'expression de ses propriétés fonctionnelles. Ainsi, il est possible d'observer les structures secondaires, les hélices α et les brins β s'associant en feuillet β, les coudes et les coils. Ces structures secondaires s'associent dans l'espace pour former la structure tertiaire des protéines. Enfin, dans certains cas, les macromolécules sont constituées de plusieurs chaînes, ou monomères, pour former les structures quaternaires. Seules les méthodes expérimentales de résolution des structures atomiques que sont la cristallographie aux rayons X et la résonance magnétique nucléaire permettent d'accéder à ces informations structurales.

2.2 Méthodes expérimentales de résolution structurale

2.2.1 Cristallographie aux rayons X

C'est la cristallographie aux rayons X qui fournira la technologie nécessaire à la résolution de la structure tridimensionnelle des protéines. La découverte de la diffraction des rayons X par les cristaux est due à Max von Laue, Paul Knipping et Walter Friedrich [Laue1912], mais la théorie précise fut établie par William et Lawrence Bragg en 1913 [Bragg1913a; Bragg1913b]. Le phénomène de diffraction repose sur le fait que la longueur d'onde des rayons X est plus petite que la distance entre deux atomes. Lawrence Bragg considéra le cristal comme un réseau tridimensionnel dans lequel les atomes sont arrangés périodiquement. Les rayons X sont réfléchis par les divers plans du réseau cristallin. Pour ces travaux William et Lawrence Bragg obtiennent le prix Nobel de physique en 1915 ; Lawrence a alors 25 ans, il est le plus jeune lauréat du Nobel à ce jour.

William Astbury [Astbury1931a; Astbury1931b; Astbury1933; Astbury1934; Astbury1935] caractérisa la diffraction de plusieurs protéines fibreuses, et il découvrit que toutes les protéines étudiées contenaient un faible nombre de diagrammes de diffraction, principalement deux qu'il nomma α et β. La forme β, avec une structure en zigzag, était plus étendue que la forme α, avec une structure hélicoïdale, et ces formes étaient capables de se transformer l'une dans l'autre. La transition de la forme α vers la forme β était obtenue par simple étirement ou en fonction de l'humidification de la protéine. La forme α caractérisait toute une classe de protéines appelées k-m-e-f pour kératine, myosine, épidermine et fibrinogène.

Par la suite, Linus Pauling, travailla sur l'analyse des clichés de diffraction de l'α- et de la β-kératine publiées par Astbury. En 1948, il montre que l'α-kératine présente une répétition de 5,1 Å, et que cela correspond à 3,6 résidus d'acides aminés par tour d'hélice. Seulement cette démonstration contredit apparemment les données d'Astbury, c'est pourquoi il ne publie pas sa découverte et qu'il n'en fait part qu'à sa femme, Ava Helen. Désireux de démontrer sa théorie il collabore avec Robert Corey, ils publient ensemble la structure de l'hélice α, du feuillet β et établissent les cinq règles principales qui gouvernent la structure des protéines [Pauling1951a; Pauling1951b; Pauling1951c; Pauling1951d; Pauling1951e; Pauling1951f; Pauling1951g].

Ces règles sont :

1. la longueur et les angles de liaison ont des valeurs standard ;
2. la liaison peptidique est planaire ;
3. les angles dièdres φ et ψ (Fig. 2.1.5) ont des valeurs restreintes à quelques orientations favorables ;
4. le maximum de liaisons hydrogène entre les groupes C=O...HN sont formées ;
5. la longueur et la déformation angulaire de la liaison hydrogène varient dans des limites étroites.

Les conditions (3) et (4) ne sont plus considérées comme essentielles aujourd'hui pour beaucoup de protéines.

Pauling et Corey décrivirent l'hélice α qui comporte 3,6 résidus d'acides aminés par tour d'hélice, le pas de l'hélice étant de 5,4 Å. Les angles φ et ψ ont été définis. L'hélice α peut-être droite ou gauche. L'autre type de structure est le feuillet β, structure dans laquelle la chaîne se replie en zigzag et les liaisons hydrogène s'établissent entre plusieurs chaînes disposées parallèlement ou de manière antiparallèle. Les structures élaborées par Pauling permettaient de reconstituer les digrammes de diffraction théoriques. Il reçut le prix Nobel de Chimie en 1954 pour l'ensemble de ses travaux.

Les travaux de Pauling sont à l'origine de l'orientation des recherches de G.N. Ramachandran [Ramachandran1963] qui posa les fondements de l'analyse conformationnelle des chaînes polypeptidiques. En 1963, il proposa une méthode analytique pour représenter les différents types de conformations protéiques basés sur les valeurs des angles dièdres, φ et ψ, qui gouvernent l'orientation d'un amino-acide. Cette représentation bidimensionnelle est connue sous le nom de diagramme de Ramachandran (Fig. 2.2.1.1), dans laquelle φ et ψ sont les coordonnées. Principalement deux zones sont explorées et permettent de décrire la majeure partie des conformations adoptées par les résidus dans les protéines : les hélices α (φ=-47°) et les feuillets β (φ=+113° ou +135°).

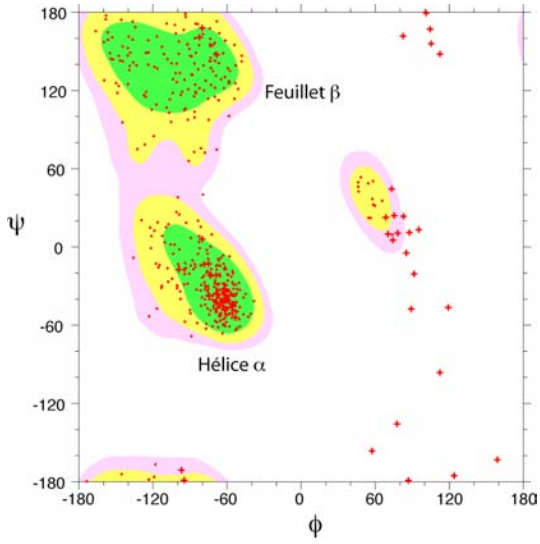


Figure 2.2.1.1 – Diagramme de Ramachandran

C'est Max Perutz qui, après seize années de recherche, mit au point la méthode permettant de déterminer la structure tridimensionnelle des protéines, et il lui fallut huit ans pour réussir à l'appliquer et publier, en 1960, la structure de l'hémoglobine [Perutz1960]. Cependant c'est John Kendrew, étudiant puis collègue de Perutz, qui en utilisant la méthode mise au point par ce dernier, sera le premier, en 1958, à résoudre la structure tridimensionnelle d'une protéine, la myoglobine [Kendrew1958]. Les cristallographes furent alors surpris de constater la complexité et l'absence de symétrie auxquelles ils s'attendaient. L'arrangement des acides aminés était bien plus compliqué que ne l'avait prédit aucune théorie de la structure des protéines.

Kendrew construisit une représentation physique de la myoglobine qui avait une échelle de 5 cm/Å. Ce modèle comprenait 2500 tiges de métal à l'intérieur d'un cube de deux mètres de côté. Des barrettes de couleur attachées aux tiges représentaient les densités électroniques. Mais la véritable forêt formée par les tiges métalliques rendait ce modèle difficile à interpréter, et sa taille posait des problèmes pour le déplacer.

Max Perutz et John Kendrew obtinrent le prix Nobel de Chimie en 1962.

2.2.2 Résonance Magnétique Nucléaire

L'utilisation de la cristallographie à rayons X peut poser des problèmes, car il faut pouvoir obtenir des cristaux de protéines qui diffractent correctement et cela peut s'avérer impossible. C'est pourquoi plusieurs groupes de chercheurs tentèrent d'appliquer la spectroscopie de Résonance Magnétique Nucléaire (RMN) à l'étude des protéines. Les premières tentatives datent des années 1960, mais les premiers spectres étaient très mal résolus. Les spectromètres de 220 MHz n'étaient

pas assez puissants, il faudra attendre les années 1980 que des appareils à plus haut champ existent et la mise au point de la technique de RMN bidimensionnelle par Kurt Wüthrich en 1981 [Wüthrich2001] pour réellement aborder l'étude structurale des protéines. Kurt Wüthrich obtient le prix Nobel de Chimie en 2006 pour ses travaux sur l'utilisation de la RMN multidimensionnelle pour l'étude de la structure des protéines.

À partir de 1980, l'introduction d'appareils de résonance magnétique nucléaire à haut champ offre un moyen de déterminer la structure de protéines de taille moyenne, en solution et non plus dans un cristal. Entre 1982 et 1983, les structures de plusieurs petites protéines sont ainsi déterminées.

La puissance des spectromètres RMN ne cessa de croître par la suite, en 1987 les premiers appareils RMN à 600 MHz font leur apparition, puis viennent les appareils à 800 MHz, 900 MHz, et actuellement le spectroscope RMN le plus puissant au monde a une fréquence de 1 000 MHz.

La technique de résolution a également évolué, de la RMN bidimensionnelle nous sommes passés à la RMN multidimensionnelle qui utilise non seulement la résonance du proton, mais également celle du carbone 13 et de l'azote 15.

Mais cette méthode présente également des limites car elle exige de très fortes concentrations de protéine, et pendant longtemps elle n'est restée applicable qu'à des molécules dont la masse moléculaire ne dépassait pas 25 000 Daltons. Actuellement la limite se situe vers des protéines dont la masse moléculaire est de 10^6. La RMN, tout comme la cristallographie à rayons X, est une méthode d'imagerie indirecte : en effet, afin de les exploiter, les spectrogrammes obtenus doivent être étudiés avec des méthodes théoriques associées, permettant, en utilisant des données expérimentales de type « effet nOe », d'obtenir un ensemble de structures possibles et donc de modèles envisageables de la molécules étudiée. Ces différents modèles, qui ne violent pas les données expérimentales, sont données dans un même fichier dans la base de données des structures macromoléculaires qu'est la PDB.

2.2.3 Protein Data Bank - *PDB*

La *Protein Data Bank* ou PDB [Bernstein1977; Berman2000] est une banque mondiale de données de structures tridimensionnelles de macromolécules biologiques, essentiellement des protéines et des acides nucléiques. Ces structures sont principalement déterminées par les deux méthodes précédemment décrites que sont la cristallographie à rayons X et la RMN. Les structuralistes du monde entier sont tenus d'y déposer leurs données expérimentales, de cette façon elles appartiennent au domaine publique. La consultation des structures est gratuite et peut se faire *via* l'internet [Berman2003].

La PDB est créée en 1971 par le laboratoire national de Brookhaven, elle sera transférée en 1998 au « *Research Collaboratory for Structural Bioinformatics* » composé de l'université de Rutgers, l'université du Winsconsin, du « *National Institute of Standards and Technology* » (NIST) et du « *San Diego Supercomputer Center »*. Son financement est assuré par plusieurs organismes, la « *National Science Foundation* », le « *Department of Energy* », la « *National Library of Medicine* » et le « *National Institute of General Medical Sciences* ».

En 2003, la « *Worldwide Protein Data Bank* » voit le jour et est composée de trois membres : RCSB (USA), PDBe (Europe) et PDBj (Japon).

À sa création en 1971, la PDB contient sept structures. À partir des années 1980 le nombre de structures augmente de façon considérable avec les avancées majeures réalisées dans les domaines de la cristallographie à rayons X et de la RMN (Fig. 2.2.3.1). Dans les années 1990, la NIST exige le dépôt de toutes les données structurales sur la PDB et un accès par l'internet est créé. De ce fait, l'utilisation de la PDB, jusqu'alors réservée à quelques experts, est ouverte à tous. Désormais ce sont des chercheurs en biologie, en chimie, en informatique, des enseignants et des étudiants qui forment les usagers de cette banque de données mondiale. Au 6 décembre 2009 la PDB contient 57 173 structures protéiques (Tableau 2.2.3.1).

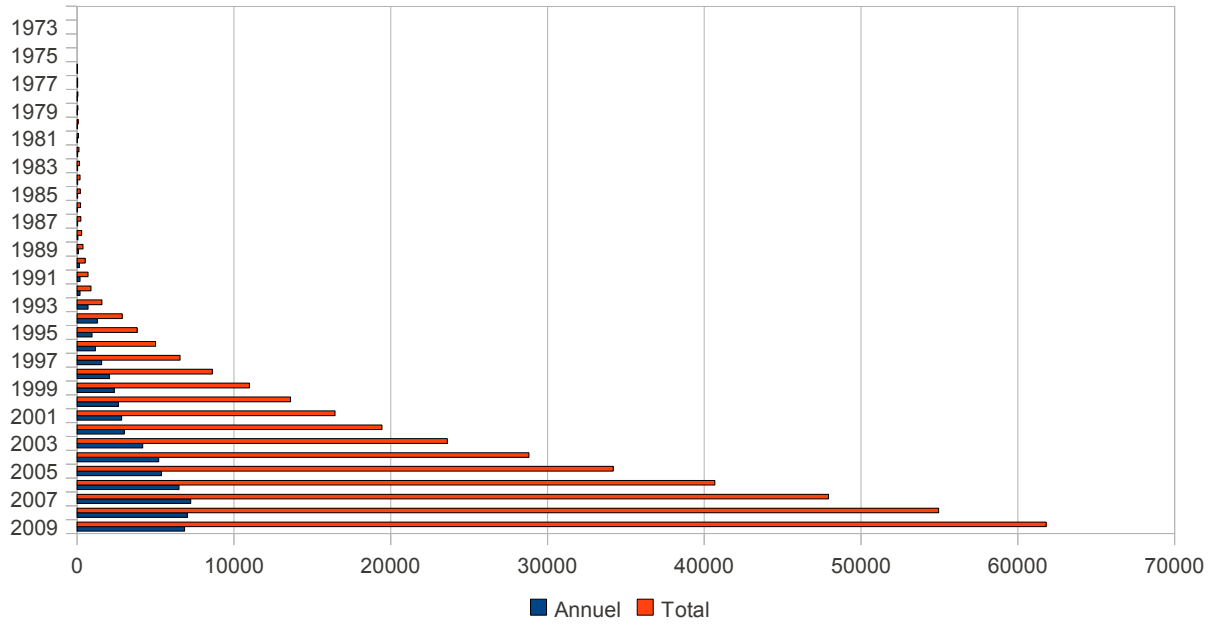


Figure 2.2.3.1 – Croissance annuelle du nombre total de structures de la PDB

Méthode	Protéines	Acides nucléiques	Complexes protéines/acides nucléiques	Autres	Total
Cristallographie à Rayons X	49761	1178	2295	17	86,2%
RMN	7102	882	150	7	13,2%
Microscopie électronique	176	16	66	0	0,4%
Hybride	18	1	1	1	0,03%
Autres	116	4	4	13	0,2%
Total	57173	2081	2516	38	61808

Tableau 2.2.3.1 – Statistiques de la PDB au 6 décembre 2009

2.3 Analyse structurale et règles topologiques

Malgré la grande diversité des protéines existant chez les êtres vivants, la connaissance d'un nombre croissant de structures a permis d'établir quelques règles topologiques qui président au repliement des chaînes polypeptidiques, et a montré l'existence de motifs structuraux en nombre limité qui se retrouvent dans les diverses protéines, ouvrant ainsi de nouveaux champs d'exploration et d'études statistiques en tout genre aux chercheurs. En effet, s'il existe un grand nombre de protéines différentes, dont beaucoup sont des enzymes, la diversité structurale est beaucoup plus réduite. Un organisme humain possède environ 5 millions de protéines différentes, et le nombre de protéines présentes dans toutes les espèces vivantes peut être évalué à environ 10^{11}. Il existe donc une très grande diversité d'espèces moléculaires. Nous pouvons réduire cette diversité de deux manières, soit par analogie fonctionnelle, soit par analogie structurale. Si les différences entre les

protéines assurant la même fonction dans des espèces ou des organes différents ne sont pas prises en compte, ce nombre peut être réduit à 10^5. Si nous considérons l'analogie structurale, ce nombre peut être réduit à un peu plus d'une centaine. Ces deux types de réduction correspondent à deux aspects différents de l'évolution. La réduction par analogie fonctionnelle correspond à l'évolution des espèces, et la réduction par analogie structurale correspond à l'évolution des protéines.

Comme nous l'avons déjà vu, la structure des protéines fait apparaître un ordre hiérarchique. La séquence des acides aminés, qui forment la chaîne polypeptidique, représente la structure primaire de la protéine. La structure secondaire est constituée d'éléments de structure réguliers ; la superstructure secondaire résulte de l'interaction des éléments de structures secondaires entre eux. La structure tertiaire désigne la conformation spatiale relativement compacte de la protéine. La structure quaternaire est formée par l'association non covalente de sous-unités identiques ou non. La connaissance d'un grand nombre de structures protéiques au niveau atomique a permis de découvrir l'existence d'invariants structuraux qui se retrouvent dans différentes protéines.

Les premières approches analytiques datent de la fin des années 1960, elles cherchaient à déterminer les règles qui président à la formation des structures secondaires par études statistiques. Il n'existe dans les protéines que deux types de structures régulières, l'hélice α et les brins β. Ces éléments structuraux sont reliés par des boucles plus ou moins grandes dans le repliement de la chaîne polypeptidique.

Trois types de méthodes furent développées, des méthodes empiriques utilisant les règles prédictives déduites de l'examen des structures connues, des méthodes basées sur théorie de la mécanique statistique et des méthodes qui utilisent la théorie de l'information.

Dans les méthodes empiriques, la fréquence d'occurrence de chaque acide aminé dans un état conformationnel défini est déduite de l'observation des structures aux rayons X de plusieurs protéines. En 1974, Chou et Fasman [Chou1974] développèrent une analyse basée sur l'examen de 15 protéines comportant 2 473 acides aminés. Cette méthode est encore utilisée dans la communauté scientifique.

Les méthodes basées sur la théorie de la mécanique statistique furent introduites par Lewis en 1970 [Lewis1970]. Elles admettent la prédominance des interactions à courte distance dans l'établissement des structures secondaires.

Le troisième type de méthode, qui est une application des théories de l'information pour déterminer les préférences conformationnelles des acides aminés, fut introduit par Robson et Pain

en 1971 [Robson1971]. Cette approche consiste à évaluer la tendance de chaque acide aminé à déterminer sa propre conformation et celle de ses voisins.

Ces différentes méthodes permettent, connaissant une séquence (structure primaire) d'effectuer une prédiction des structures secondaires, principalement les hélices α et les brins β.

2.4 Brins et feuillets β

Comme nous venons de le voir dans les paragraphes précédents, la structure des protéines est complexe et met en jeu, à différents niveaux, des éléments structuraux spécifiques : parmi ceux-ci nous trouvons les brins β qui peuvent s'associer en feuillet.

Les feuillets β sont la seconde forme régulière de structures secondaires derrière les hélices α, ces feuillets sont constitués de brins β connectés latéralement par des liaisons hydrogène. Les feuillets β sont dits « plissés » à cause de l'aspect en « dent de scie » des brins β comme nous pouvons le voir sur la figure 2.4.1.

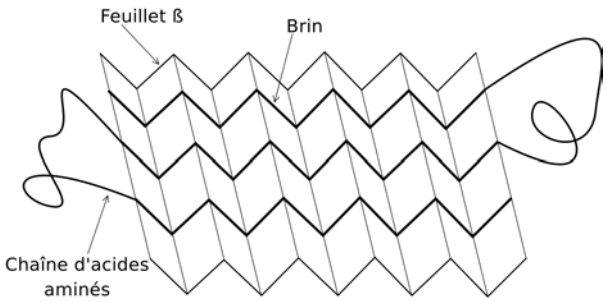
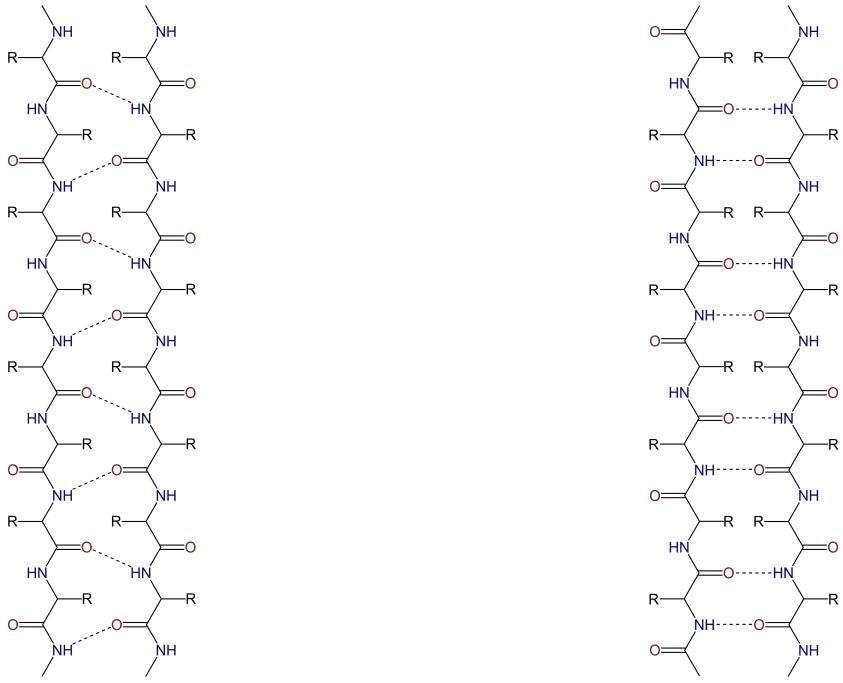


Figure 2.4.1 – Schématisation d'un feuillet β plissé

Les feuillets β sont constitués de chaînes d'acides aminés très étirées, la chaîne polypeptidique est plus étirée lorsqu'elle forme un feuillet β que lorsqu'elle forme une hélice α. Un seul brin β n'est pas stable, mais l'association de plusieurs brins par l'intermédiaire de liaisons hydrogène entre les groupes CO et NH des brins voisins, sont stables. Il existe deux types de feuillets β : les parallèles et les antiparallèles (Fig. 2.4.2).



(a) (b)

Figure 2.4.2 – Les deux types de feuillet β, (a) parallèle et (b) antiparallèle

L'arrangement antiparallèle possède la stabilité la plus forte, et ce parce que la liaison hydrogène entre le groupement carbonyl et le groupement amine est plane et plus directionnelle. Dans ce cas les angles dièdres φ et ψ valent respectivement -140° et 135°, ce qui fait que les résidus adjacents sont très proches et forment donc des liaisons hydrogène.

Dans le cas d'un arrangement parallèle, les brins β étant tous orientés dans le même sens, les liaisons hydrogène inter-chaînes ne sont pas planes et donc cet arrangement est beaucoup moins stable que l'antiparallèle. Les angles dièdres φ et ψ valent respectivement -120° et 115° si bien que deux résidus adjacents ne forment pas de liaisons hydrogène : le résidu *i* d'un brin forme une liaison avec le résidu *j*-1 et *j*+1du brin adjacent. Il est rare d'observer un feuillet β parallèle composé de moins de cinq brins, ce qui laisse penser qu'une telle conformation est trop instable pour subsister.

Un troisième type d'arrangement existe : un brin β peut être lié de manière parallèle d'un côté et antiparallèle de l'autre. Ce type d'arrangement est le moins courant à cause de sa nature instable.

Parfois les liaisons hydrogène d'un feuillet β ne sont pas parfaitement arrangées, ce phénomène est connu sous le nom de « *β bulge* ». Il s'agit d'un acide aminé d'un brin β qui ne forme aucune liaison hydrogène car ses angles dièdres correspondent à ceux d'un résidu d'une hélice α : φ et ψ valant -60° et -45°. L'importance des *β bulges* réside dans leur influence sur la fonction d'une protéine. Par exemple, il a été démontré leur influence sur le site actif de la Super-Oxide Dismutase.

Les différents types de connexions entre brins β dans les différentes protéines connues furent analysées par Jane Richardson en 1977 [Richardson1977]. La même année, Cyrus Chothia et ses collaborateurs [Chothia1977] dégagent un certain nombre de règles qui présidents à l'association des brins β pour former des feuillets et à l'interaction de ces feuillets entre eux. L'association des brins β donne naissance à différentes topologies. Les brins β possèdent une flexibilité intrinsèque qui leur permet de se tordre ou de se courber, et ainsi de s'associer de différentes manières. Les principaux motifs formés par les brins β sont l'épingle à cheveu (« *hairpin* »), le méandre β, la clé grecque et le « *jellyroll* ». Dans ces différents motifs les brins β sont antiparallèles. Ainsi, quatre brins β consécutifs dans la chaîne d'acides aminés peuvent être arrangés de douze manières différentes suivant les connexions pour former un feuillet β à quatre brins, comme le montre la figure 2.4.3.

Sur un même brin β, les chaînes latérales des acides aminés successifs pointeront alternativement en haut puis en bas. Nous remarquons également que les Cα des brins β adjacents sont alignés et que leurs chaînes latérales pointent dans la même direction.

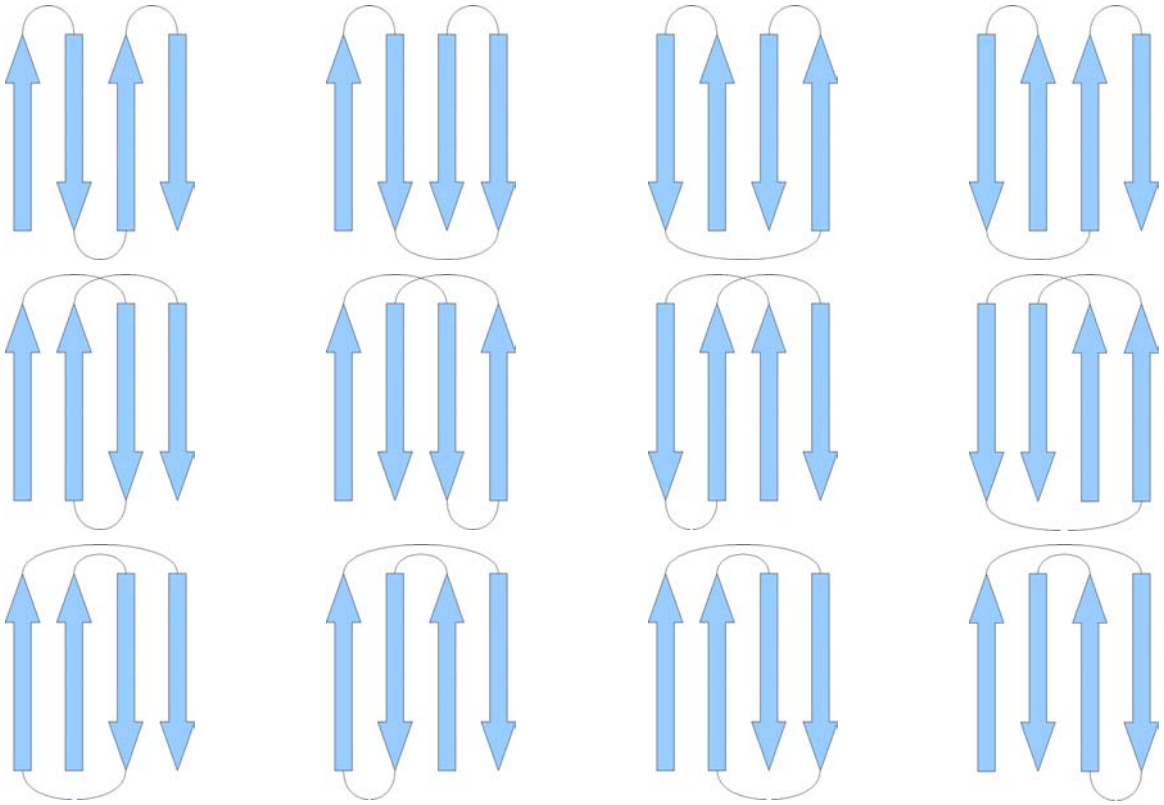


Figure 2.4.3 – Les douze arrangements possibles de quatre brins β consécutifs, symbolisés par des flèches

Il existe plusieurs sites internet permettant d'étudier ces structures secondaires, parmi lesquels le « *Beta Sheet's World* »[1] ainsi que des logiciels, comme par exemple TOPDRAW [Bond2003], permettant de tracer des diagrammes en deux dimensions du même type que ceux de la figure 2.4.3.

2.5 Classification des protéines

L'analyse des structures protéiques a révélé que la structure tridimensionnelle a été mieux conservée au cours de l'évolution que la structure primaire, ce qui signifie que deux protéines sans homologie de séquence importante peuvent avoir la même structure spatiale. Par exemple, le domaine de fixation du coenzyme dans quatre déshydrogénases, la lactate, la malate, la glycéraldéhyde-3-phosphate et l'alcool déshydrogénase, ont la même structure tertiaire alors qu'elles ne possèdent que quatre acides aminés en commun sur les 94 qui composent ce domaine.

L'analyse structurale montre également que les protéines de grande taille sont constituées de domaines structuraux distincts. Un domaine structural est une région bien définie de la protéine qui a toutes les caractéristiques d'une protéine globulaire, c'est à dire la compacité et la stabilité. Nous pouvons donc considérer ces domaines comme des régions indépendantes reliées les unes aux autres par le biais de liaisons covalentes au sein d'une même protéine. Les domaines structuraux peuvent donc être classés de façon indépendante. Le concept de domaine structural a été introduit en 1970 par Gerald Edelman [Edelman1970] suite à l'observation des immunoglobulines. Il reçu le prix Nobel de Médecine en 1972 pour ses travaux sur les anticorps.

2.5.1 Classes structurales

Une fois qu'un domaine structural a été identifié et isolé, c'est le type et la succession des structures secondaires qui le composent qui va déterminer sa classe d'appartenance. Quatre classes sont ainsi décrites : « tout α », « tout β », « α/β » et « α+β ». Nous considérons également une cinquième classe qui regroupe les domaines ne contenant que peu ou pas de structures secondaires, la classe « petit ou irrégulier ».

La classe tout α regroupe des domaines composés à plus de 90 % d'hélices α. Les hélices α de domaines de cette classe sont de manière générale longues, plus de vingt acides aminés. L'hémoglobine est une protéine de la classe tout α (Fig. 2.5.2.1).

La classe tout β est constituée de domaines contenant des brins β organisés majoritairement en feuillets β antiparallèles. Les immunoglobulines font partie de la classe tout β (Fig. 2.5.2.1).

1 http://www-lbit.iro.umontreal.ca/bSheet/index.html

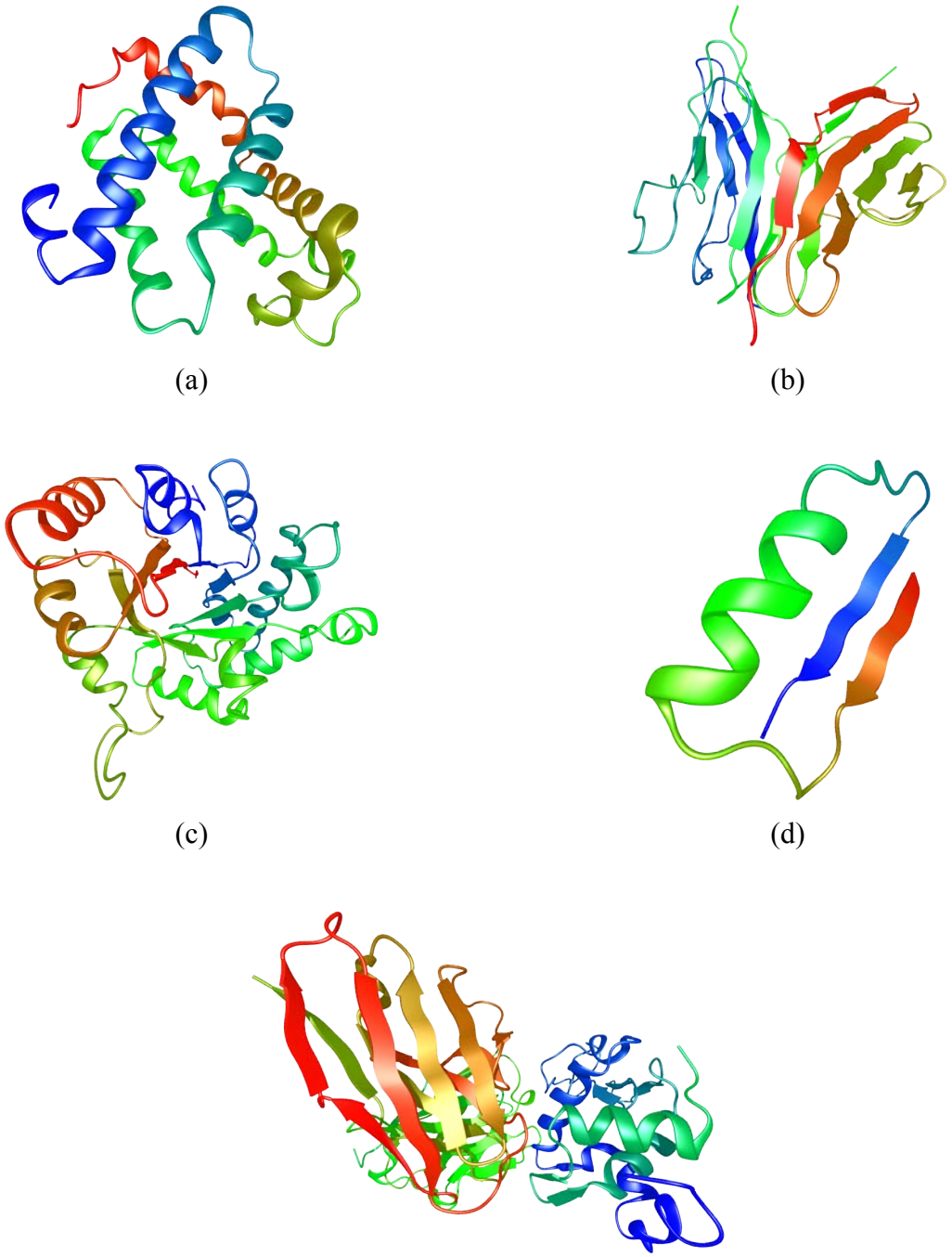
La classe α/β caractérise les domaines structuraux possédant une alternance régulière d'hélices α et de brins β. Nous y retrouvons beaucoup de feuillets β parallèles, mais comment deux brins β adjacents sont-ils connectés ? Si deux brins adjacents se suivent dans la séquence des acides aminés, les deux extrémités qui doivent être reliées se trouvent sur les bords opposés du feuillet. La chaîne polypeptidique doit traverser le feuillet β d'un bord à l'autre pour pouvoir relier le brin suivant. Dans ce cas, la jonction est souvent réalisée par une hélice α. La chaîne polypeptidique doit tourner deux fois en formant des boucles et le motif composé est donc : un brin β, une boucle, une hélice α, une boucle et un brin β. Ce motif, dénommé βαβ est présent dans quasiment toutes les structures possédant un feuillet β parallèle. Le plus souvent les brins β sont enfouis au cœur du domaine et recouverts par les hélices α. Les « *TIM barrels* » sont une parfaite illustration de cette classe (Fig. 2.5.2.1).

La classe α+β correspond à des domaines contenant des hélices α ainsi que des brins β répartis dans des régions distinctes. L'alternance des structures secondaires observable dans la classe α/β n'est pas présente. Le lysozyme appartient à cette classe (Fig. 2.5.2.1).

2.5.2 « Folds » et « superfolds »

Le nombre, le type, la connectivité, et l'arrangement des structures secondaires déterminent la structure tertiaire d'une protéine, son repliement ou « *fold* ». Lorsque des domaines structuraux contiennent les mêmes structures secondaires principales, connectées topologiquement de la même façon, ils seront classés dans le même *fold*. Les éléments périphériques de structure secondaire, qui peuvent représenter une part importante du domaine structural, peuvent être très différents. Dans un même *fold* les ressemblances ne proviennent pas d'une origine commune, mais de conditions physico-chimiques qui imposent la topologie.

Nous parlons de « *superfold* » pour désigner des domaines structuraux non apparentés, qui ne possèdent aucune similarité de séquence, qui ont des fonctions différentes, mais qui présentent un repliement semblable. Un *superfold* est définit à partir du moment où un même repliement a été observé dans au moins trois domaines sans homologie de séquence significative. Neuf types de repliements entrent dans cette catégorie, le pli globine, le trèfle, les hélices antiparallèles, le pli immunoglobuline, le sandwich αβ, le *jellyroll*, le double tour, le pli αβ et le *TIM barrel* (Fig. 2.5.2.2).



(a) (b) (c) (d) (e)

Figure 2.5.2.1 – Illustration des différentes classes structurales. (a) La classe tout α représentée par une hémoglobine [3A5A], (b) la classe tout β représentée par une immunoglobuline [2WP3], (c) la classe α/β représentée par un TIM barrel *[8TIM] en (d) on peut observer un motif βαβ composé par un brin β, une boucle, une hélice α, une boucle et un brin β, cet exemple est tiré d'une triose phosphate isomérase [1AMK], et en (e) la classe α+β représentée par un lysozyme [3A67]*

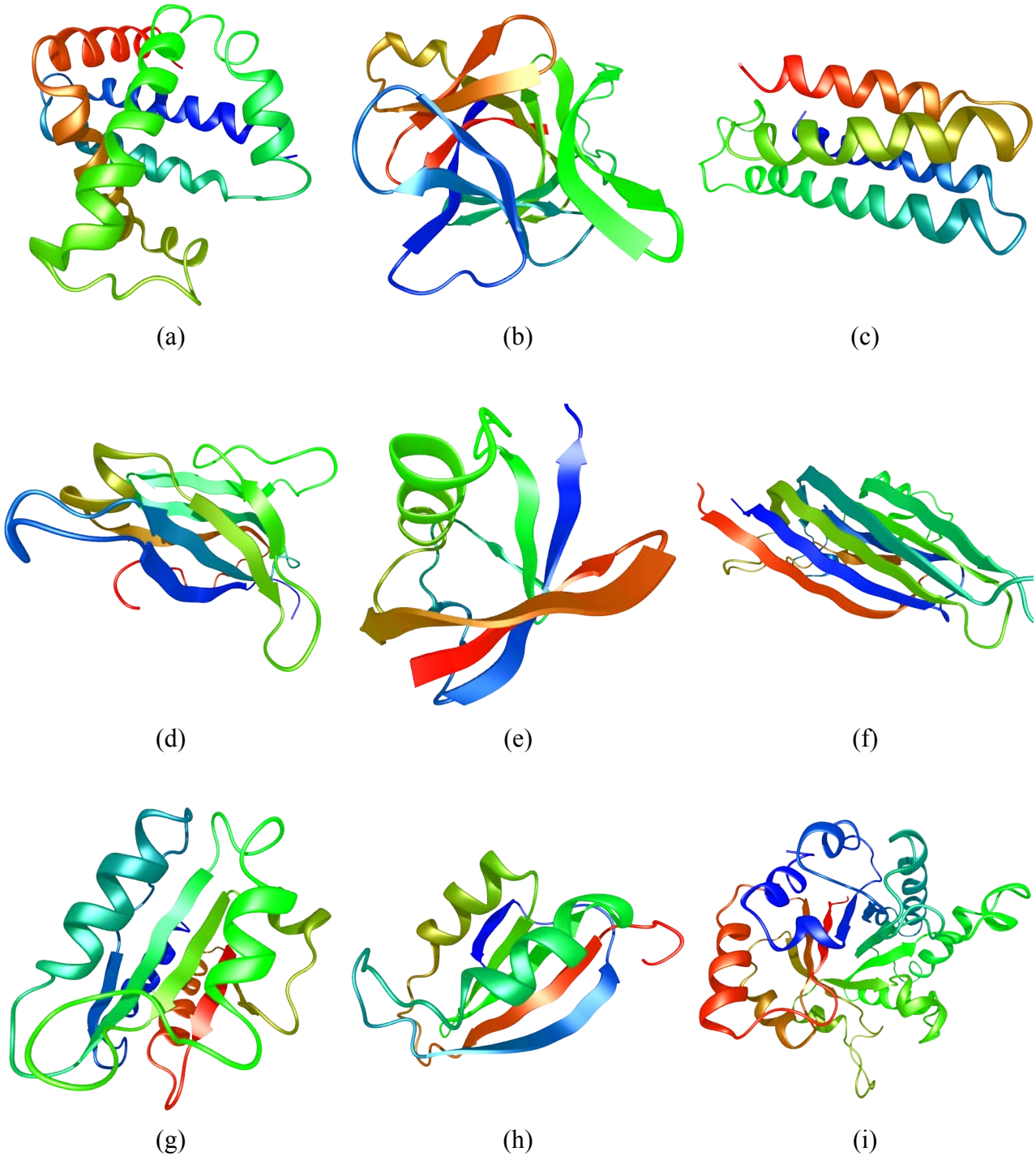


Figure 2.5.2.2 – Les neuf superfolds. *(a) Le pli globine [1THB], (b) le trèfle [1I1B], (c) les hélices antiparallèles [256B], (d) le pli immunoglobuline [2RHE], (e) le sandwich αβ [1UBQ], (f) le* jellyroll *[2BUK], (g) le double tour [2FOX], (h) le pli αβ [1APS] et (i) le* TIM barrel *[7TIM]*

2.5.3 Familles et superfamilles

Il est considéré que des domaines structuraux appartiennent à une même famille lorsqu'ils possèdent une identité de séquence supérieure à 30 %. Pour une homologie comprise entre 20 et 30 % la parenté entre les deux domaines est incertaine, il s'agit de la « *twilight zone* ». Pour une homologie de séquence inférieure à 20 % il est estimé qu'il n'y a aucune parenté entre les deux domaines, il s'agit de la « *midnight zone* ».

Les deux membres d'une même famille ont donc une très forte connexion évolutive. Cependant, dans certains cas deux membres d'une même famille ne peuvent avoir que 15 % de similarité ; dans ces cas précis les liens, évolutifs ont pu se faire sur la base d'arguments structuraux et fonctionnels, c'est la cas de la famille des globines.

Une famille ne peut contenir qu'un seul membre si aucune relation n'a pu être établie avec un autre domaine.

Une superfamille regroupe une ou plusieurs familles dont les homologies de séquence sont inférieures au seuil nécessaire à former une famille, mais dont la comparaison des structures et des fonctions suggère une origine commune.

2.5.4 Classifications structurales

Il existe plusieurs types de classification qui prennent en compte différents niveaux de hiérarchie. Les deux méthodes les plus utilisées sont CATH (« *Classification, Architecture, Topology, Homology* ») élaborée par Janet Thornton et ses collaborateurs en 1997 [Thornton1997], et SCOP (« *Structure Classification Of Proteins* ») introduite par Chothia en 1995 [Chothia1995].

CATH est une méthode de classification hiérarchique des structures protéiques présentes dans la PDB. Afin de les classifier, les domaines protéiques sont individualisés en utilisant des méthodes automatiques et manuelles. Si une chaîne protéique possède une grande homologie de séquence (homologie de séquence de l'ordre de 80 %), ainsi qu'une structure similaire avec une chaîne déjà individualisée, alors le découpage du nouveau domaine est fait automatiquement en produisant les découpes aux mêmes endroits de la séquence. Sinon, les découpages se font manuellement à partir de résultats produits par divers algorithmes, ainsi que d'informations présentes dans la littérature. Il y a deux types d'algorithmes, ceux qui s'appuient sur la structure (CATHEDRAL [Redfern2007], SSAP [Orengo1996], DETECTIVE [Swindells1995a; Swindells1995b], PPU [Holm1994], DOMAK [Siddiqui1995]), et ceux qui se basent sur la séquence (Profile HMMs [Krogh1994]).

Une fois les domaines individualisés, il existe deux procédures de classification pour CATH : les méthodes automatiques, et les méthodes qui combinent les traitements manuels et automatiques.

Les méthodes automatiques recherchent de grandes homologies de séquence et de structure (35 % d'homologie de séquence) avec des domaines faisant déjà partie de CATH. Dans ce cas, le domaine est automatiquement classé dans la même catégorie que celui déjà répertorié.

Les méthodes combinant traitements manuels et automatiques classent les domaines selon quatre niveaux (Fig. 2.5.4.1) :

1. C pour classe (C-level) : la classe est déterminée en fonction de la composition en structures secondaires de la structure. Il existe trois grandes classes : tout α, tout β et α-β. Cette dernière classe comprend les structures α/β et α+β.
2. A pour architecture (A-level) : ce niveau décrit la forme globale du domaine structural en utilisant les structures secondaires, mais sans prendre en considération leurs relations. C'est en fait une simple description de l'arrangement des structures secondaires.
3. T pour topologie (T-level) : les structures sont regroupées en fonction de leur topologie, ou de leur repliement dans le cœur du domaine.
4. H pour superfamilles homologues (H-level) : ce niveau regroupe les domaines qui peuvent être décrits comme homologues, et qui sont supposés avoir un ancêtre commun. Les similitudes sont identifiées soit par une haute homologie de séquence, soit par homologie de structure en utilisant l'algorithme SSAP.

SCOP, à l'instar de CATH, est une méthode de classification hiérarchique des structures protéiques. Sa grande différence avec CATH est que le classement se fait principalement de façon manuelle. Sinon, cette méthode comprend également quatre niveaux :

1. Classe : les structures sont classées en fonction de leur architecture structurelle globale.
2. Repliement : regroupe les structures qui possèdent les mêmes arrangements de structures secondaires, mais qui n'ont pas de parenté d'évolution.
3. Superfamilles : regroupe les homologies structurales et fonctionnelles suffisantes pour déduire des relations évolutives qui ne sont pas nécessairement visibles de par leurs séquences.
4. Familles : ce niveau ne concerne que les homologies de séquence.

Étant donné son classement manuel, le traitement des structures est plus long que pour CATH.

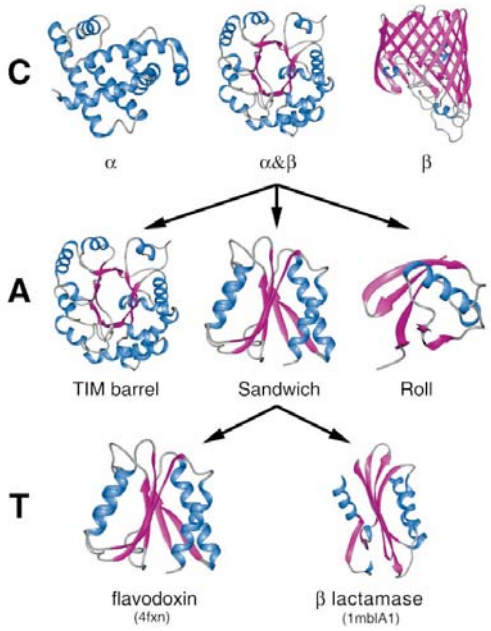


Figure 2.5.4.1 – Illustration des trois premiers niveaux de CATH, d'après [Thornton1997]

L'ensemble des protéines n'est pas encore classé, mais au fur et à mesure que de nouvelles structures sont résolues, il n'apparaît pas de repliement fondamentalement nouveau. Trente repliements différents sont recensés dans CATH, le nombre de repliements apparaît donc comme très petit par rapport au nombre de protéines et à leur diversité.

2.6 Relation structure-fonction des protéines

L'utilisation d'agents dénaturants sur une protéine aura pour conséquence la désorganisation de sa structure spatiale sans casser les liaisons peptidiques, seules les liaisons secondaires de type liaison hydrogène ou pont disulfure sont concernées. Nous observons alors un dépliement de la chaîne polypeptidique. Les agents dénaturants sont multiples et de différentes natures. Il existe des agents physiques tels que l'augmentation de la température qui provoque une agitation thermique des atomes et qui engendre un rupture des liaisons faibles, ce phénomène est facilement observable lors de la cuisson d'un blanc d'œuf. Un changement de pH entraîne une modification des charges portées par les groupements ionisables, les liaisons ioniques et hydrogène sont donc modifiées. Il existe également des agents dénaturants chimiques, comme l'urée qui fragilise les liaisons hydrogène, les détergents qui modifient les interactions avec le solvant, et les bases et les acides qui altèrent le pH.

Outre les modifications de structure subies par les protéines suite à des expositions à ces différents agents, il existe plusieurs conséquences importantes, dont la perte de l'activité biologique. C'est en effet la structure tridimensionnelle de la protéine qui lui confère sa fonction. Si une protéine est dénaturée, elle perd sa structure tridimensionnelle et donc sa fonction. Afin de comprendre le rôle d'une protéine, il est donc nécessaire d'en connaître la structure. C'est la relation structure-fonction des protéines.

En 1969, Cyrus Levinthal [Levinthal1968a] remarque qu'à cause du grand nombre de degrés de liberté présents dans une protéine (les angles φ et ψ), une macromolécule possède un nombre gigantesque de conformations possibles, il parle de 3^{300} ou 10^{143}. Considérant ces chiffres, si une protéine doit, pour atteindre sa conformation finale passer par l'ensemble des conformations possibles, il lui faudrait un temps supérieur à l'âge de l'univers. Pourtant, il est constaté qu'une protéine se replie en un temps qui va de la milliseconde à la minute. Il s'agit là du paradoxe de Levinthal [Zwanzig1992].

Une partie de la réponse à ce paradoxe a été trouvée (par anticipation) en 1954, lorsque Christian Anfinsen [Anfinsen1954] énonce que toute l'information nécessaire au repliement spatial d'une protéine dans son état final (actif) est présent dans sa structure primaire. En 1961, il démontre ce principe [Anfinsen1961] en prouvant que la ribonucléase peut revenir à sa conformation initiale après dénaturation, c'est à dire après la perte de sa structure tridimensionnelle normale, et récupérer son activité enzymatique. Il obtiendra le prix Nobel de Chimie en 1972 pour ses travaux.

Levinthal suggéra que le repliement est accéléré et guidé par la formation rapide d'interactions locales qui déterminent le futur repliement.

La connaissance du repliement des protéines et des relations structure-fonction sont fondamentales pour une bonne compréhension des phénomènes physiopathologiques qui existent. Ainsi, dans le cas qui nous intéresse dans cette étude, une détermination précise, associée à une compréhension topologique des feuillets β, de leur organisation dans le repliement protéique et le rôle qu'ils pourraient jouer dans les relations structure-fonction peuvent être importantes.

L'apparition relativement récente des techniques de graphisme moléculaire et de modélisation moléculaire peut nous aider à mieux appréhender tous ces phénomènes.

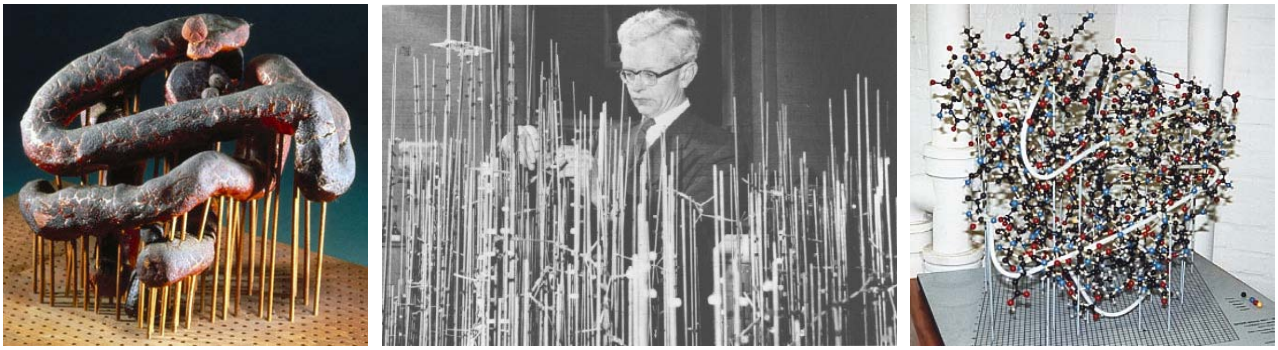
2.7 Historique de la modélisation moléculaire

De nos jours, lorsque nous évoquons la modélisation moléculaire, cela concerne des représentations tridimensionnelles sur un écran d'ordinateur ou dans un système de réalité virtuelle. Il est bien évident que suite aux premières résolutions de structures, il a été tenté de les représenter et que les ordinateurs de cette époque étaient alors bien loin des performances des machines actuelles. Il n'était même pas concevable de les manipuler sur de tels supports. La solution envisagée était de construire des modèles physiques afin de les représenter le plus fidèlement possible.

Nous allons passer en revue les deux grands modes de représentation : les physiques et les virtuels.

2.7.1 Modèles physiques

Le tout premier modèle représentant une protéine date de 1957, il s'agit d'un modèle en plasticine de myoglobine fait par John Kendrew (Fig. 2.7.1.1a) un an avant sa résolution par cristallographie à rayons X. Le premier modèle élaboré à partir d'une structure connue fut celui, un an plus tard, de la myoglobine par John Kendrew, décrit dans le paragraphe consacré à la cristallographie à rayons X (Fig. 2.7.1.1b). Peu de temps après sa publication de 1961, Kendrew reçut beaucoup de demandes pour obtenir des modèles comme le sien, c'est pourquoi en 1965 il demanda à A. A. Barker, employé à l'université de Cambridge, de concevoir des modèles de la myoglobine (Fig. 2.7.1.1c). Ces modèles étaient de types boules/bâtons, et étaient représentés à une échelle de 2,5 cm/Å. La production commença en mai 1966 et se termina en mars 1968, durant cette période 29 commandes furent honorées au rythme d'un modèle par mois.



(a) (b) (c)

Figure 2.7.1.1 – Les modèles de myoglobine de 1957 à 1966. (a) modèle en plasticine de 1957, (b) John Kendrew terminant le modèle de 1958 à base de tiges de métal et de barrettes colorées, (c) modèle réalisé par Barker en 1966

Le deuxième date de 1968 avec Frederic Richards qui après avoir résolu la structure de la ribonucléase [Wyckoff1967a; Wyckoff1967b] créa sa propre représentation, « *The Richards' Box* » [Richards1968], alors qu'il est en congé sabbatique de l'université d'Oxford. Les densités électroniques obtenues par cristallographie étaient imprimées sur papier, et les lignes de contour de ces densités étaient tracées en reliant les valeurs identiques. Ces lignes étaient ensuite reproduites sur des plaques transparentes montées verticalement et régulièrement espacées, créant ainsi une carte de densité électronique en 3D. Des miroirs sans tain étaient disposés de façon à surimposer la carte de densité électronique sur un modèle en fil de fer. L'échelle obtenue était de 1 cm par Å (Fig. 2.7.1.2a). Ce modèle était plus grand qu'un homme, et prenait beaucoup de place, il a été pourtant largement utilisé par les cristallographes du monde entier. D'ailleurs les premiers équivalents sur ordinateur furent baptisés « Electronic Richards' boxes ».

Ces premiers modèles, bien que très intéressants pour la communauté scientifique, étaient très imposants et, de ce fait, très difficiles à transporter. Ce problème sera en partie résolu par Byron Rubin, un cristallographe travaillant avec Jane Richardson. Dans les années 1970, Byron inventa le « *Byron's Bender* » [Rubin1972], une machine capable de plier un fil de fer pour lui faire suivre le squelette d'une protéine (Fig. 2.7.1.2b). Ces modèles étaient petits et légers ce qui les rendaient aisément transportables. Un exemple de leur utilité est survenu dans le milieu des années 1970 lors d'une conférence. À cette époque, seule une douzaine de structures protéiques avaient été résolues. David Davies avait apporté un modèle « *Bender* » du fragment Fab de l'immunoglobuline, et Jane et David Richardson un modèle *Bender* de la superoxide dismutase. En comparant leurs modèles, ils se rendirent compte que les deux protéines avaient un repliement similaire, alors qu'ils n'ont que

9 % d'homologie de séquence. Cet événement est la première reconnaissance de ce qui est connu maintenant comme étant la superfamille des immunoglobulines.

En 1973, Rubin construisit un modèle du squelette de la rubredoxine d'une hauteur de 150 cm. Ce modèle remporta, en tant que sculpture, la compétition Chandler à l'université de Caroline du Nord. Byron a également crée un modèle de la collagénase neutrophile humaine qui est en exposition permanente au Smithonian Institute de Washington.

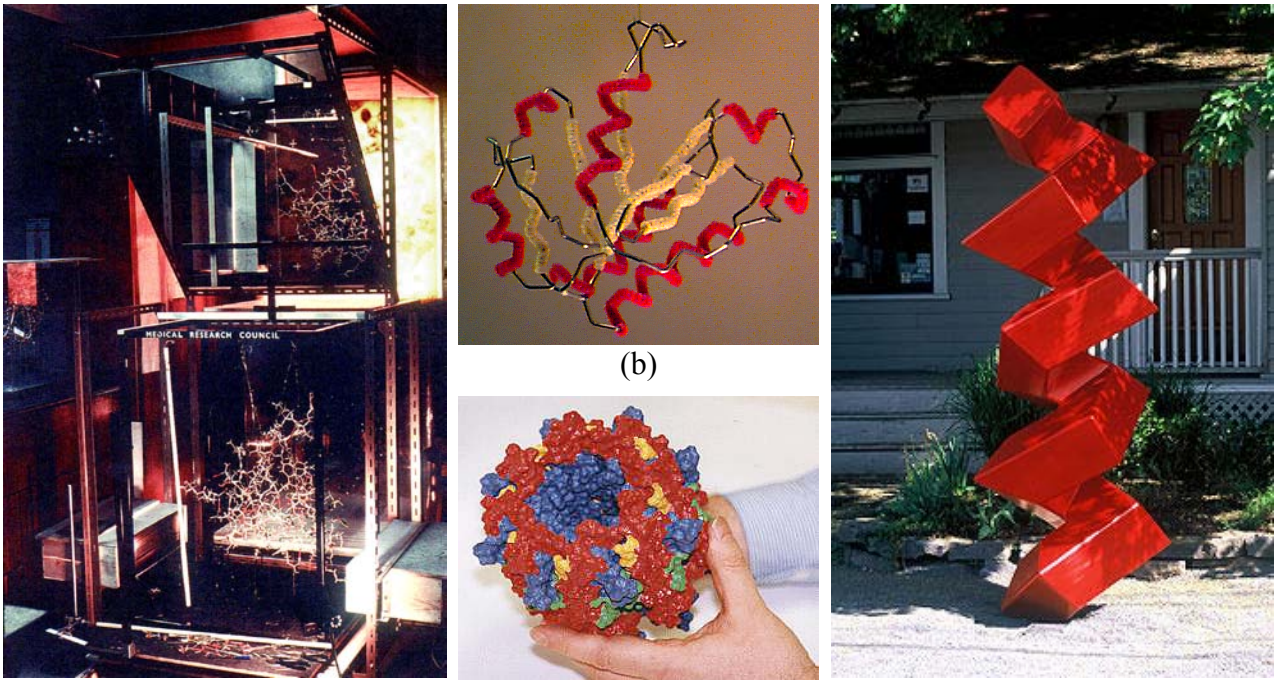


Figure 2.7.1.2 – Évolution des représentations physiques. (a) The Richards' Box *de 1968, (b) un modèle de Byron de 1972, (c) modèle de* rapid prototyping *de Micheal Bailey de la fin des années 90, (d) sculpture de Julian Voss-Andreae de 2004 en mémoire de Linus Pauling*

De par l'aspect relativement artistique de la structure des protéines, beaucoup d'artistes se sont intéressés à leur représentation. Le sculpteur Bathsheba Grossman fut le premier à les représenter dans des blocs de verre à l'aide d'un laser. Edgar Meyer, après sa carrière de cristallographe, fabriqua des représentations en bois de diverses structures avec l'aide de son logiciel SCULPT.

À la fin des années 90, Michael Bailey (SD Supercomputer Center) travailla avec Tim Herman (école d'ingénieur du Milwaukee) sur une technologie permettant d'obtenir rapidement un modèle physique d'une protéine donnée (Fig. 2.7.1.2c). Des modèles commencèrent à être disponibles en 2000 *via* la société 3D Molecular Designs.

Pour terminer ce paragraphe sur les modèles physiques nous pouvons également évoquer l'hélice α de Julian Voss-Andreae (Fig. 2.7.1.2d). Cette œuvre de 3 mètres de haut datant de 2004 a été créée en mémoire de Linus Pauling qui a découvert la structure de l'hélice α, elle est située devant la maison de Pauling à Portland (Oregon).

2.7.2 Modèles virtuels

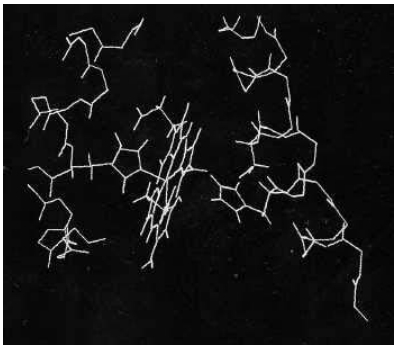
Le premier système pour l'affichage interactif de structures moléculaires a été conçu au MIT en 1964. Cyrus Levinthal et son équipe [Levinthal1966; Levinthal1968b] ont conçu un programme pour travailler sur les protéines. Ce programme permettait l'étude des interactions à courte distance entre les atomes. L'affichage se faisait sur un écran d'oscilloscope monochrome (surnommé Kluge) montrant les structures dans un rendu de type « fil de fer » (Fig. 2.7.2.1a et 2.7.2.1b). L'effet tridimensionnel était rendu en faisant tourner la structure en permanence à l'écran. La vitesse de rotation pouvait être contrôlée à l'aide d'un dispositif de forme hémisphérique, disposé devant l'écran, sur lequel l'utilisateur laissait reposer sa main. L'ancêtre de la souris en quelque sorte.

C'est au milieu des années 1970 que pour la première fois une structure protéique a été résolue et modélisée uniquement sur ordinateur, sans passer par un modèle physique [Beem1977]. Pour ce faire, David et Jane Richardson utilisèrent un système appelé GRIP [Tainer1982] de l'université de Caroline du Nord.

À la fin des années 1970, de plus en plus de cristallographes construisent les structures nouvellement résolues uniquement sur ordinateur (« *Electronic Richards' boxes* ») et non plus des modèles physiques. Un des avantages majeurs de la modélisation par ordinateur est qu'il conserve les coordonnées atomiques, alors qu'avec un modèle physique il faut les mesurer manuellement, atome par atome.

En 1980, le TAMS [Feldmann1980] (« *Teaching Aids for Macromolecular Structures* ») fait son apparition. C'est un système peu onéreux qui consiste à regarder des images stéréoscopiques à l'aide de lunettes dans lesquelles deux diapositives sont enfichées (Fig. 2.7.2.1c). Le TAMS contenait des diapositives concernant les liaisons peptidiques, les hélices α, les feuillets β, la structure tertiaire, la structure quaternaire, les groupes prosthétiques et les sites actifs. Ce système comprenait 116 paires de diapositives couleurs et les lunettes servant à visualiser la stéréoscopie (Taylor Merchant). Chaque image était accompagnée d'un paragraphe de description et d'une question.

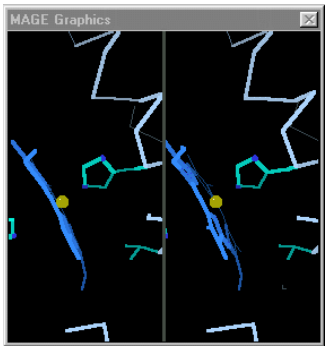
(a)



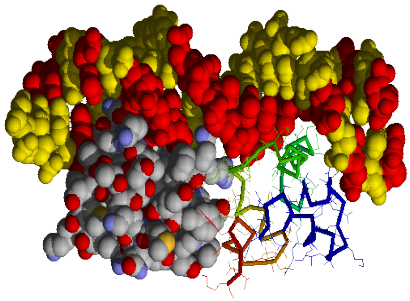
(b)

(c)

(d)



(e)



(f)

Figure 2.7.2.1 – Évolution des représentations virtuelles. (a) le système de Cyrus Levinthal de 1964, nous pouvons voir l'écran de l'oscilloscope ainsi que le dispositif de contrôle, (b) une représentation en fil de fer de la myoglobine avec le système de Levinthal, (c) le dispositif de visualisation de Taylor Merchant utilisé pour le TAMS en 1980, nous voyons une paire d'images stéréoscopiques enfichées dans le dispositif, (d) une machine Evans & Sutherland PS390 de la fin des années 1980, écran 19", processeur cadencé à 10 MHz, 512 Kb de mémoire pour le système et 2 Mb de mémoire pour l'utilisateur, nous ne pouvons observer ici que l'écran ainsi que des dispositifs d'interactions, l'unité centrale de la taille d'un petit réfrigérateur n'est pas visible, (e) un exemple de kinemage *qu'il est possible d'obtenir, (f) illustration de RasMol utilisant plusieurs modes de visualisation avec projection d'ombres*

Dans les années 80 les ordinateurs de prédilection des cristallographes sont ceux fabriqués par Evans & Sutherland (Fig. 2.7.2.1d). Chaque ordinateur coûtait 250 000 $, il était capable d'afficher une carte de densité électronique, et l'utilisateur pouvait faire correspondre manuellement une chaîne d'acide aminé à l'intérieur de la carte. L'écran couleur affichait une chaîne d'acides aminés, avec un rendu fil de fer, qui pouvait tourner en temps réel. Ces systèmes utilisaient des graphismes vectoriels, les matrices de rotations étaient calculées par des processeurs dédiés, chaque dimension possédait son processeur (x, y et z). Le logiciel utilisé sur ces machines était FRODO [Jones1978].

Durant cette période, David et Jane Richardson, furent parmi les premiers à développer des programmes de représentations graphiques sur ordinateur de structures moléculaires. À la fin des années 80 cela donna naissance au programme CHAOS [Richardson1989] écrit dans le langage de l'ordinateur Evans & Sutherland PS300. En France dans les années 1980, Jean-Paul Mornon et son équipe développèrent sur ces machines un logiciel dédié : MANOSK [Thomas1990].

En 1992 les Richardson développèrent le « *kinemage* » [Richardson1992] (pour *kinetic image*), basé sur leurs programmes MAGE et PREKIN (Fig. 2.7.2.1). Du fait de son implémentation sur Macintosh, il s'agit du premier programme de visualisation moléculaire accessible au plus grand nombre. Ce système a été décrit dans l'article principal du premier numéro du journal « *Protein Science* », le programme était fourni sur une disquette avec chaque numéro. Toutes les instructions pour utiliser les programmes PREKIN et MAGE étaient détaillées dans l'article pour pouvoir créer de nouveaux *kinemages*. Dans les cinq années qui suivirent, plus de 1000 *kinemages* avaient été crées pour illustrer des articles de *Protein Science*.

Le grand avantage des *kinemages* est que cela représente uniquement la sélection et le point de vue recherchés. Cela représente exactement ce que l'auteur veut montrer. Ce grand avantage est aussi un grand défaut, car l'utilisateur souhaiterait également pouvoir explorer la structure par lui même, sans *a priori*. Ce manque sera comblé par l'arrivée de RasMol en 1993 [Sayle1995].

L'histoire de RasMol commence en 1989, alors que Roger Sayle étudie l'informatique à l'Imperial College. Il s'intéresse particulièrement au problème de perception de la profondeur pour les représentations sur ordinateur. Il écrit alors un algorithme de « *raytracing* » suffisamment rapide pour permettre de représenter des rotations d'images projetant des ombres. Il en résulte le second algorithme le plus rapide de *raytracing* de sphères. Cependant, son algorithme ne sait produire que des sphères et nécessite un processeur parallèle pour fonctionner. En 1990, Roger Sayle intègre l'université d'Edimbourg, ou il continue le développement de son programme sous la direction du cristallographe Andrew Coulson. Plusieurs améliorations en terme de rapidité de l'algorithme font que la limite de transfert des processeurs parallèles ne suffit plus, il implémente alors son algorithme sur des machines mono-processeur, tout d'abord sur des machines Unix, et plus tard Windows et Macintosh. Le programme devint un outil de visualisation moléculaire complet, et dès 1993 il est utilisé pour l'enseignement et pour la création d'images de publications. Le programme devient libre de droit en juin 1993 après la soutenance de thèse de Roger Sayle (Fig. 2.7.2.1f).

Le nom RasMol vient de « *raster* » qui signifie « tableau de pixels d'un écran », et « molecules ».

Le fait que Roger Sayle publie le code source de son programme a permis son adaptation sur de nombreux systèmes, et son inclusion dans de nombreux programmes dérivés.

2.7.3 Importance de l' « open source »

Le terme « *open source* » qualifie un logiciel dont le code source est diffusé librement, mais cela implique également qu'il ne peut être vendu, que sa distribution sous forme compilée est autorisée et que n'importe qui peut modifier le code selon ses convenances et le distribuer.

L' open source a une influence considérable dans le monde de la recherche, et notamment en modélisation moléculaire. Il est certain que si Roger Sayle n'avait pas distribué le code source de RasMol, ses développements auraient été nettement plus limités. Le fait que n'importe quel chercheur puisse apporter sa pierre à l'édifice en développant de nouvelles fonctionnalités, et qu'il puisse ensuite distribuer librement sa version du logiciel permettent une large diffusion des innovations.

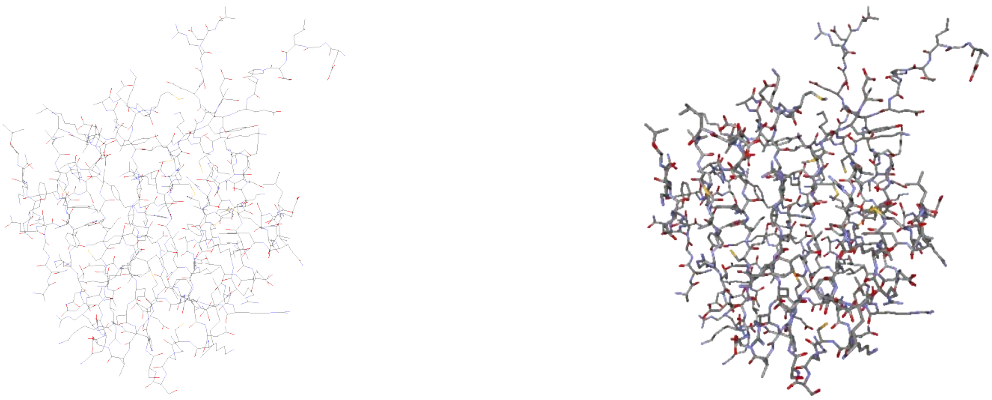
Depuis RasMol de nombreux logiciels de modélisation moléculaire sont distribués en open source, tels que VMD [Humphrey1996] (Visual Molecular Dynamics), NAMD [Phillips2005] (NAnoscale Molecular Dynamics), PyMol [DeLano2008] (Python Molecule), Swiss-PDBViewer [Guex1997], Jmol [Gezelter] (Java Molecule), CHIME [Rzepa1994; Casher1995] (CHemical mIME), BALLView [Kohlbacher2000; Moll2006] (Biochemical Algorithms Library View). Ces outils sont parmi les plus utilisés dans la communauté des chimistes, des biologistes et des biochimistes.

2.8 Modes classiques de visualisation

Les macromolécules sont composées de quelques milliers à plusieurs dizaines de milliers d'atomes. Ce sont des objets très complexes, il est donc difficile de tirer des enseignements de leur simple observation. C'est pourquoi il existe différents modes de visualisation, utilisés afin de simplifier la complexité des macromolécules en fonction de ce que nous souhaitons observer.

Si nous souhaitons représenter la structure globale d'une protéine, il est possible d'utiliser le rendu fil de fer, il s'agit du mode historique de visualisation (Fig. 2.7.2.1a et 2.7.2.1b). Ce type de rendu représente les atomes et les liaisons inter-atome par des lignes. Chaque atome a son code couleur qu'il est possible de représenter sur les lignes. Nous constatons sur la figure 2.8.1a que ce

mode de visualisation bien qu'en 3D ne rend pas d'impression de profondeur, c'est dû en partie au fait que les lignes ont toujours la même épaisseur quelques soit leur distance du point de vue, et qu'elles ne rendent pas d'effet d'éclairage.



(a) (b)

Figure 2.8.1 – Modes de visualisation en fil de fer (a) et en bâtons (b) [1914]

Le mode souvent associé au rendu fil de fer est le rendu en bâtons, dans lequel des cylindres sont utilisés au lieu des lignes. Les cylindres sont des objets surfaciques qui peuvent donc interagir avec la lumière (Fig. 2.8.1b).

Le type visualisation le plus connu est certainement le « boule-bâtons ». Ce mode ci permet de localiser précisément les atomes par une représentation sous forme de sphères (Fig. 2.8.2).

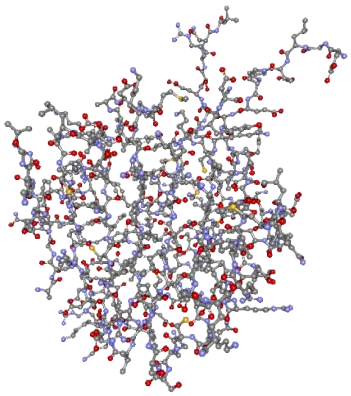


Figure 2.8.2 – Visualisation boules-bâtons [1914]

Afin de représenter la forme globale d'une protéine, nous utilisons des modes de visualisation tels que van der Waals, qui représente les rayons de van der Waals, ou encore des modes tels que le « *Solvent Accessible Surface* » ou SAS et le « *Solvent Excluded Surface* » ou SES [Lee1971]. La

surface de van der Waals est représentée par la frontière de l'ensemble des boules formées par des sphères ayant pour centre un atome et pour rayon, le rayon de van der Waals correspondant à l'atome (Fig. 2.8.3a). La surface accessible au solvant est l'ensemble des positions possibles du centre d'une sonde sphérique représentant le solvant, qui roulerait sur la surface de van der Waals (Fig. 2.8.3b), et la surface exclue au solvant prend en compte le recouvrement des creux par la sonde (Fig. 2.8.3c), dans ce type de représentation il est possible de modifier le rayon de la sonde.

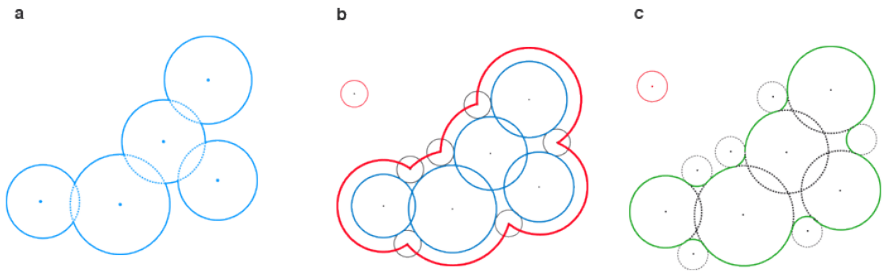
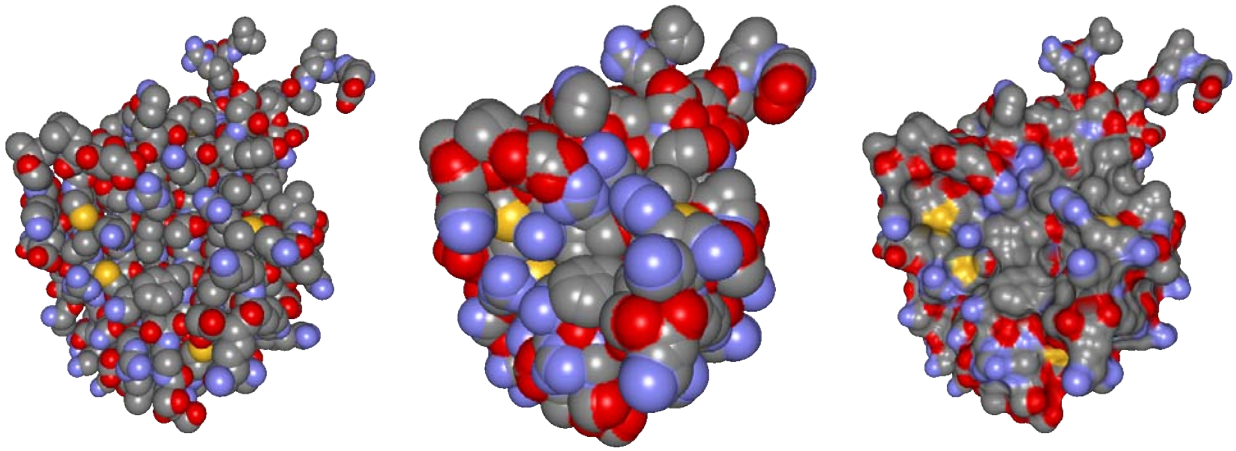


Figure 2.8.3 – Méthodes de calcul des surfaces de van der Waals, accessibles et exclues au solvant, pour van der Waals la surface des boules de van der Waals est conservée (a), pour la surface accessible au solvant, une sonde d'un rayon déterminé passe le long de la surface de van der Waals et seul l'ensemble des positions possibles du centre de cette sphère est représenté (b), et pour la surface exclue au solvant, n'est pris en compte que le recouvrement des creux par cette même sonde (c)

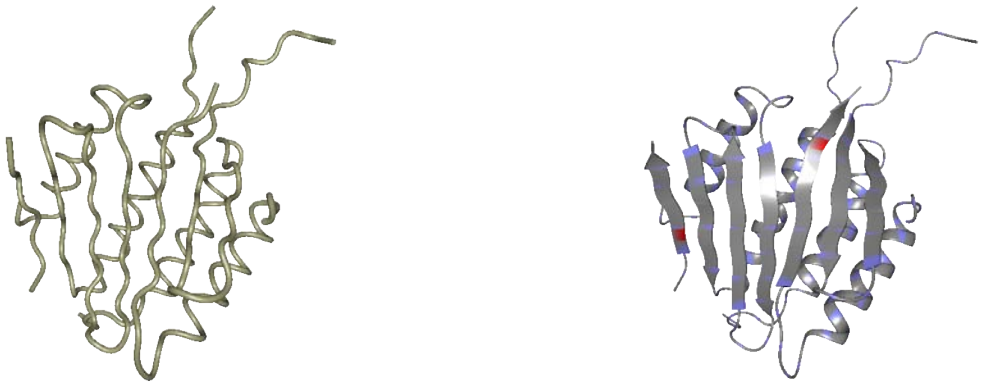


(a) (b) (c)

Figure 2.8.4 – Représentation des surfaces de van der Waals (a), accessible au solvant (b) et exclue au solvant (c) [1914]

Si nous souhaitons ne représenter que certaines structures d'une protéine, nous utilisons des rendus simplifiés ou symboliques. Un rendu de type « tube » ou « squelette » est utilisé pour ne

représenter que le squelette de la protéine (Fig. 2.8.5). Parler de rendu squelette est abusif dans ce cas, car seuls les Cα sont utilisés pour la représentation et non l'ensemble NCαCO constituant le véritable squelette de la protéine.



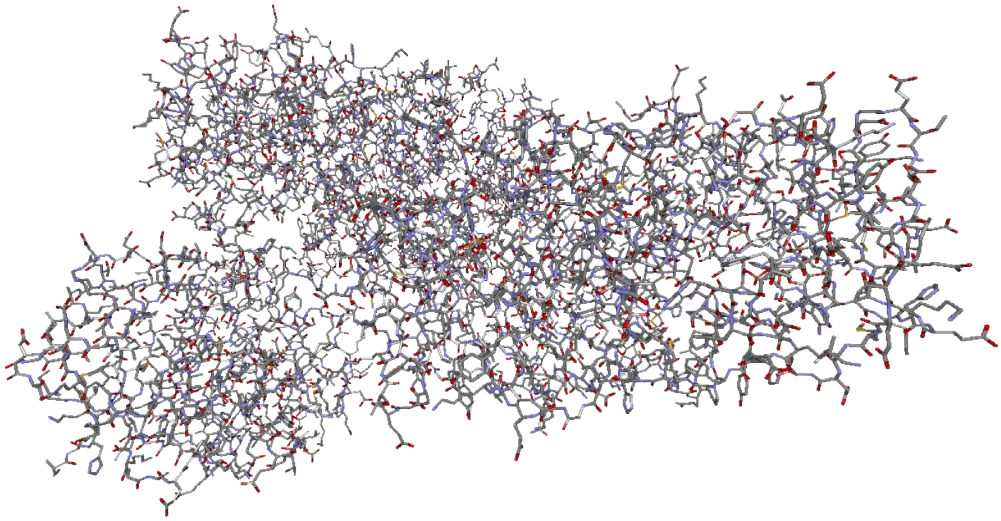
(a) (b)

Figure 2.8.5 – Rendu de type squelette (a), seuls les Cα sont utilisés pour cette représentation. Rendu de type cartoon *(b), les hélices α sont représentées sous forme de serpentins et les brins β sous forme de flèches, nous constatons ici que c'est l'association de brins β qui forme un feuillet β [1914]*

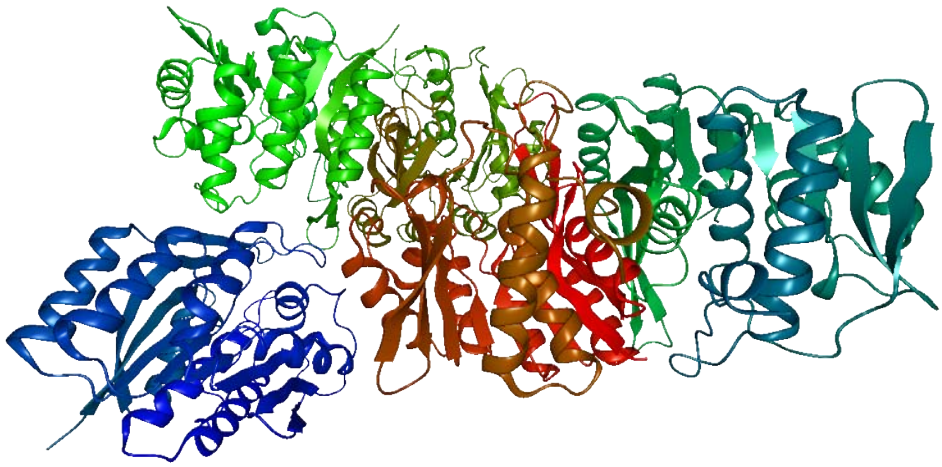
Le rendu symbolique le plus utilisé est le rendu « *cartoon* » (Fig. 2.8.5). Celui-ci a été rendu populaire grâce aux possibilités de tracé bidimensionnel qu'à offert le logiciel Molscript [Kraulis1991] développé Per Kraulis dans les années 1990. Dans ce type de rendu, nous cherchons à représenter les structures secondaires, les hélices α sont sous formes de serpentins, dans certains logiciels elles sont représentées sous forme de cylindres et les brins β, constituants de base des feuillets β, sont sous formes de flèches. Nous pouvons constater sur la figure 2.8.5b que les flèches suivent l'orientation du plan des liaisons peptidiques, cela n'est pas systématiquement le cas, les flèches peuvent également être représentées droites.

Comme cela a été présenté dans « *Nature Methods* » en mars 2010 [O'Donoghue2010], il existe d'autres modes de visualisation, mais ceux précédemment cités sont les plus importants et les plus utilisés. Parmi ces derniers, force est de constater que certains manques subsistent. Si nous observons une protéine en mode bâtons (Fig. 2.8.6), il est impossible d'observer les structures secondaires à cause de la complexité de l'objet. Le mode *cartoon* paraît être le plus approprié, seulement nous ne pouvons distinguer avec clarté que les hélices α (Fig. 2.8.6), il est aisé de les repérer même dans une structure de grande taille. La réelle difficulté provient des feuillets β car ils

ne sont pas représentés, seuls les brins β le sont. Sur la figure 2.8.6 nous pouvons voir quelques brins β mais il est impossible de localiser, visualiser, dénombrer ou caractériser les feuillets correspondants. Il n'existe aucun mode de visualisation capable de représenter un feuillet β dans sa globalité.



(a)



(b)

Figure 2.8.6 – (a) Représentation en bâtons, il est impossible d'observer les structures secondaires. (b) Représentation cartoon*, il est aisé d'observer les hélices α, nous pouvons voir quelques brins β, mais il est impossible de localiser, visualiser, dénombrer ou caractériser les feuillets β [2I5B]*

2.9 Matériels et méthodes

Dans cette partie, les choix de langages de programmation et de bibliothèques, que nous avons utilisés afin de réaliser ce travail, seront détaillés. Ensuite, nous étudierons les splines de Catmull-Rom, ainsi que les courbes et surfaces de Bézier dont nous nous servirons afin de représenter les feuillets β. Nous verrons également la structure, et les données contenues dans un fichier PDB, puis le principe de l'attribution des structures secondaires dont nous aurons besoin par la suite. Nous terminerons par une introduction aux principes de la simulation de dynamique moléculaire.

2.9.1 C++, OpenGL et Qt

L'ensemble des développements réalisés durant ce travail de thèse l'ont été en utilisant le langage de programmation orienté objet C++. Ce langage est l'un des plus utilisés et il n'appartient à personne, par conséquent tout le monde peut l'utiliser librement. Nous verrons plus avant que ce langage a été utilisé en fonction du logiciel open source choisi comme support à ce travail.

Pour l'affichage, la bibliothèque graphique utilisée est OpenGL (Open Graphics Library) qui, comme son nom l'indique, est libre d'utilisation. La bibliothèque OpenGL est très largement utilisée dans la communauté d'informatique graphique pour représenter des scènes en trois dimensions et en temps réel.

De manière générale les interactions avec un programme se font *via* une interface graphique. Dans ce travail la bibliothèque d'interface utilisée est Qt. Qt appartient depuis janvier 2008 à Nokia, qui en janvier 2009 prend la décision de passer sous une licence LGPL (ou GNU LGPL pour GNU Lesser General Public License, licence publique générale limitée GNU en français). Cette nouvelle licence permet le développement de logiciels propriétaires sans achat de licence commerciale.

2.9.2 Splines de Catmull-Rom

Une spline est une fonction définie par morceaux par des polynômes. Cette définition basique est incomplète, car nous pourrions en déduire que les courbes de Bézier sont des courbes splines ce qui n'est pas le cas. En effet, les courbes splines permettent de lisser un polygone au moyen d'une courbe paramétrique telle que la modification de la position d'un seul point de contrôle ne modifie pas l'intégralité de la courbe. Seule la partie modifiée aura a être recalculée.

Les splines de Catmull-Rom sont des splines interpolantes cubiques dans lesquelles les tangentes en chaque point sont calculées en utilisant le point de contrôle précédant et suivant.

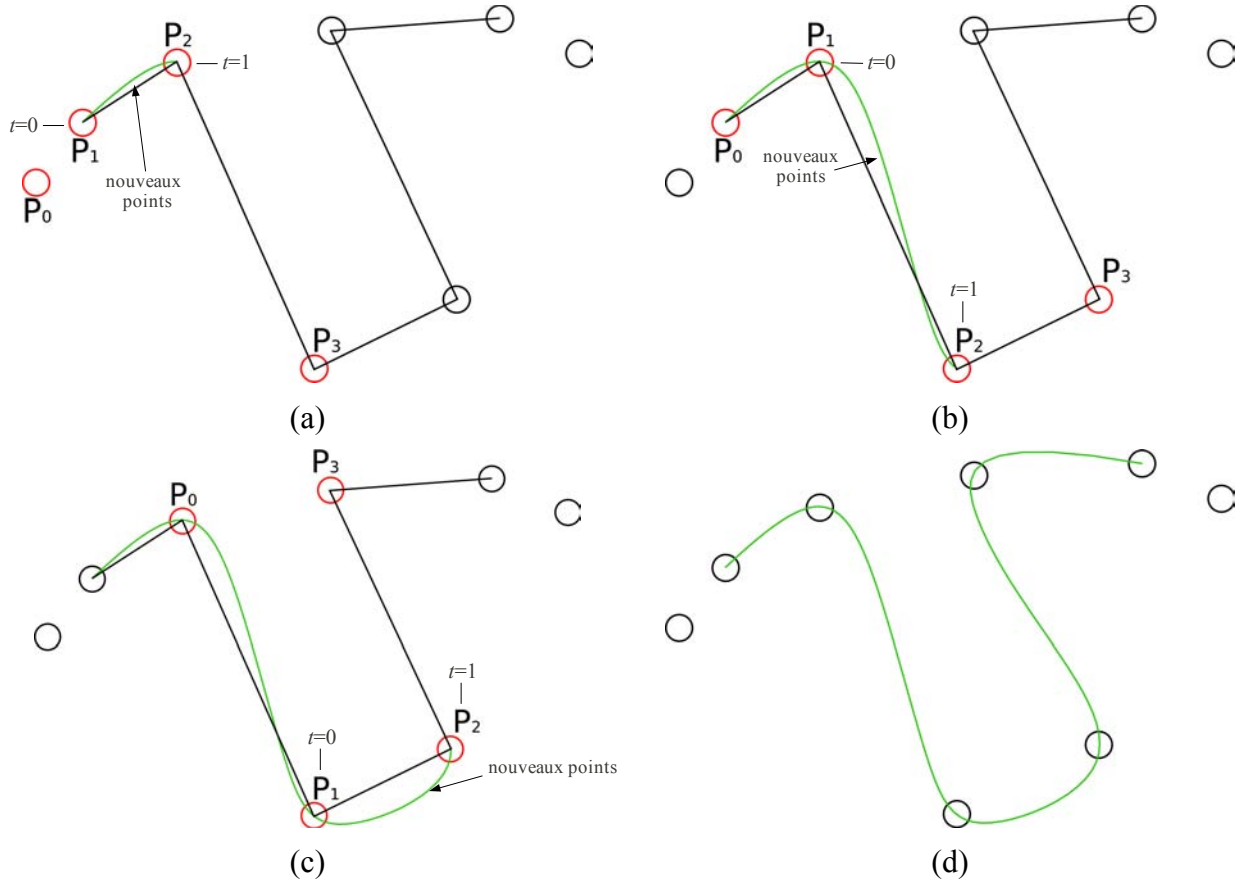


Figure 2.9.2.1 – Les différentes étapes de la construction d'une spline de Catmull-Rom. (a) Les nouveaux points sont calculés entre P_1 et P_2 en faisant varier la valeur de t, (b) on change les points de contrôle pour pouvoir calculer la suite de la spline, les splines de Catmull-Rom étant continues en C_1 les deux morceaux de courbe sont parfaitement jointifs, (c) on change les points de contrôle jusqu'à obtenir la totalité de la courbe spline (d). Nous constatons que la spline passe bien par l'ensemble des points de contrôle à l'exception du premier et du dernier

Nous pouvons les définir de cette façon :

$$q(t)=\frac{1}{2}(t^3\ t^2\ t\ 1)\cdot\begin{pmatrix}-1 & 3 & -3 & 1\\ 2 & -5 & 4 & -1\\ -1 & 0 & 1 & 0\\ 0 & 2 & 0 & 0\end{pmatrix}\cdot\begin{pmatrix}P_0\\ P_1\\ P_2\\ P_3\end{pmatrix}$$

Que nous pouvons également écrire :

$$q(t)=\frac{1}{2}((-P_0+3P_1-3P_2+P_3)t^3+(2P_0-5P_1+4P_2-P_3)t^2+(-P_0+P_2)t+2P_1)$$

Pour calculer un nouveau point, deux points de contrôle de chaque côté sont nécessaires. Il n'y aura donc aucun nouveau point calculé entre P_0 et P_1 et entre P_2 et P_3, la spline passera donc par l'ensemble des points de contrôle, à l'exception du premier et du dernier. La position du nouveau point correspond à la valeur de *t*, qui représente la distance entre les deux points de contrôle les plus proches. La variable *t* vaut 0 sur P_1 et *t* vaut 1 sur P_2. Les points entre P_1 et P_2 ont une valeur de *t* comprise entre 0 et 1 (Fig. 2.9.2.1).

2.9.3 Courbes et surfaces de Bézier

Les courbes de Bézier sont des courbes polynomiales paramétriques inventées en 1962 par Pierre Bézier ingénieur en mécanique et électricité chez Renault. Ces courbes seront utilisées pour concevoir des pièces automobiles à l'aide d'ordinateurs. En 1971 viendra la naissance d'Unisurf, pionnier en DAO (Dessin Assisté par Ordinateur) et CFAO (Conception et Fabrication Assistées par Ordinateur). Désormais les courbes de Bézier ont de nombreuses applications dans le domaine de la synthèse d'images et le rendu de polices de caractères.

La théorie énonce que pour n+1 points de contrôle (P_0, …, P_n), une courbe de Bézier est définie par l'ensemble des points $\sum_{i=0}^{n} B_i^n(t)\,\boldsymbol{P}_i, t\in[0,1]$ et où les B_i^n sont les polynômes de Bernstein. La suite des points P_0, …, P_n forme le polygone de contrôle de Bézier.

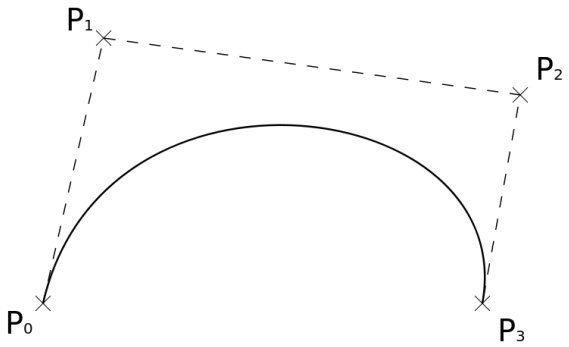


Figure 2.9.3.1 – Une courbe de Bézier définie par les points de contrôle P_0, P_1, P_2 et P_3. La courbe commence au point P_0 et se termine au point P_3, mais ne passe pas a priori par les autres points de contrôle qui participent cependant à l'allure générale de la courbe

Comme nous pouvons le constater sur la figure 2.9.3.1, la courbe de Bézier est à l'intérieur de l'enveloppe convexe des points de contrôle. La courbe commence au point P_0 et se termine au point P_n, sans passer *a priori* par les autres points de contrôle. Le contrôle d'une telle courbe est global,

modifier un point de contrôle va entraîner la modification de toute la courbe et pas seulement le voisinage du point de contrôle.

La forme cubique (de degré 3) s'écrit $\boldsymbol{P}(t)=\boldsymbol{P}_0(1-t)^3+3\boldsymbol{P}_1 t(1-t)^2+3\boldsymbol{P}_2 t^2(1-t)+\boldsymbol{P}_3 t^3$, t étant compris entre 0 et 1. Cette forme est la plus utilisée car elle permet la continuité en tangence et en courbure de deux courbes raccordées.

Une courbe P(t) peut être décomposée en deux courbes P_L et P_R (Fig. 2.9.3.2), dont les points de contrôle sont respectivement (L_1, L_2, L_3, L_4) et (R_1, R_2, R_3, R_4) avec :

$$\begin{pmatrix} L_1' \\ L_2' \\ L_3' \\ L_4' \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{4} & \frac{2}{4} & \frac{1}{4} & 0 \\ \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{pmatrix} \cdot \begin{pmatrix} A \\ B \\ C \\ D \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} R_1' \\ R_2' \\ R_3' \\ R_4' \end{pmatrix} = \begin{pmatrix} \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \\ 0 & \frac{1}{4} & \frac{2}{4} & \frac{1}{4} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} A \\ B \\ C \\ D \end{pmatrix}$$

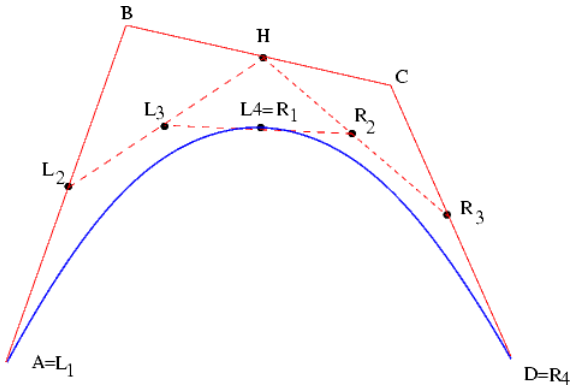


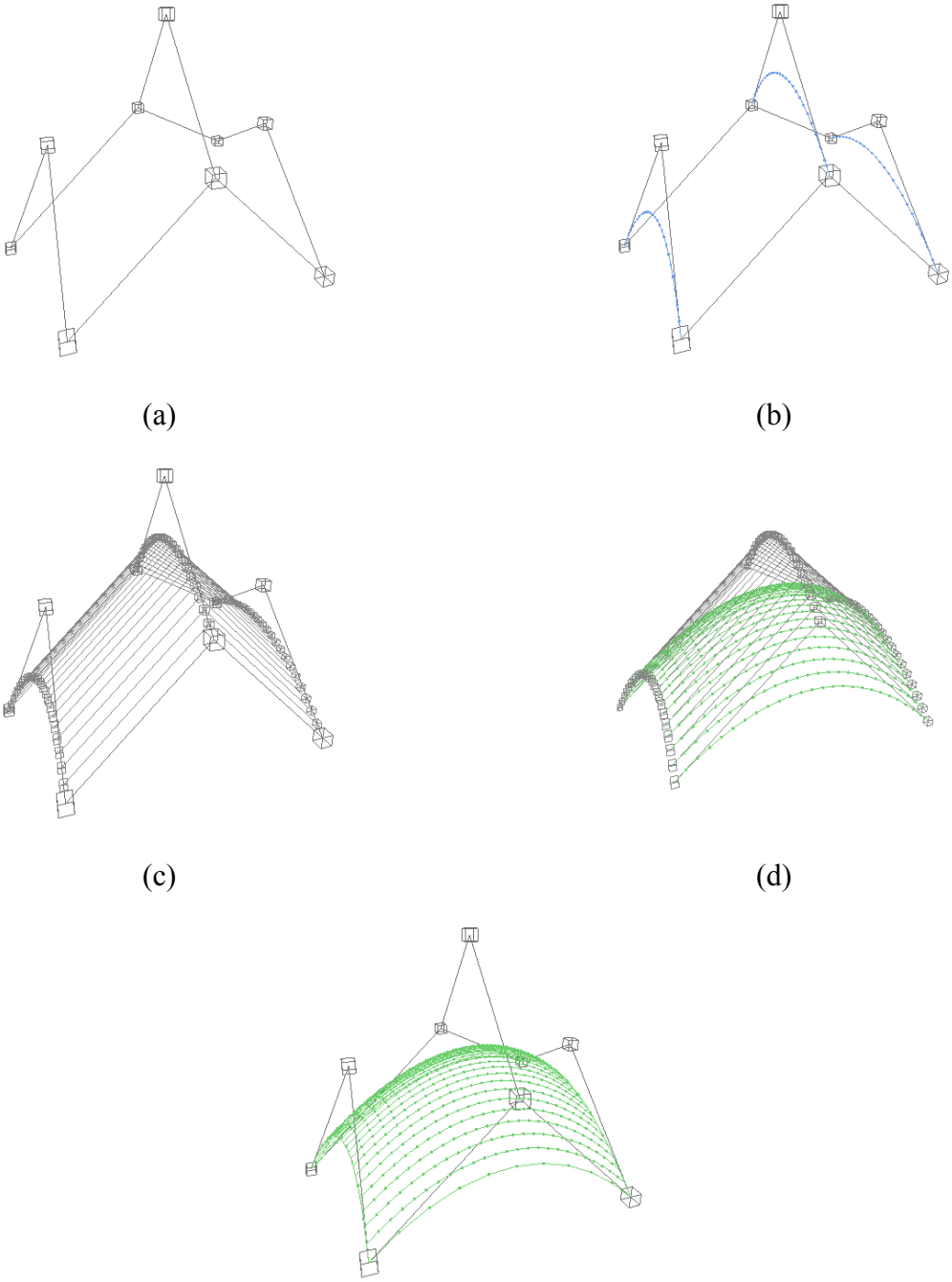
Figure 2.9.3.2 – Décomposition d'une courbe de Bézier P(t) en deux courbes P_L et P_R avec leurs points de contrôle respectifs (L_1, L_2, L_3, L_4) et (R_1, R_2, R_3, R_4)

Nous pouvons également définir des surfaces grâces aux courbes de Bézier, ce sont les surfaces ou carreaux de Bézier.

Étant donnée une matrice [M] de points dans l'espace $A_{i,j}$, la surface de Bézier correspondante est l'ensemble des points M générés par les valeurs comprises entre 0 et 1 des variables u et v du polynôme :

$$\overrightarrow{OM}=\sum_{i=0}^{n}\sum_{j=0}^{m} C_n^i v^i (1-v)^{(n-i)} C_m^j u^j (1-u)^{(m-j)} \overrightarrow{OA_{ij}}$$

Nous constatons alors qu'il s'agit de deux courbes de Bézier imbriquées, d'où les paramètres u et v. Fort de cette constatation il est aisé de construire une surface en utilisant uniquement les calculs sur des courbes de Bézier comme nous pouvons le voir sur la figure 2.9.3.3.



(a) (b)

(c) (d)

(e)

Figure 2.9.3.3 – Construction d'une surface de Bézier. Cet exemple définit une matrice de neuf points de contrôle (a), à partir de laquelle les courbes de Bézier sont calculées, en considérant uniquement les colonnes (ou les lignes) de la matrice (b). Les courbes ainsi obtenues vont servir comme courbes de points de contrôle, il suffit de discrétiser les courbes avec le même nombre de points et chaque point correspond à un nouveau point de contrôle (c), les courbes de Bézier sont alors calculées pour les nouveaux points de contrôle (d) et nous obtenons une surface de Bézier (e)

Nous pouvons également écrire :

$$\boldsymbol{P}(u,v)=[u^3\,u^2\,u\,1]\cdot\left[\boldsymbol{B}\cdot\boldsymbol{P}\cdot\boldsymbol{B}^T\right]\cdot\begin{bmatrix}v^3\\v^2\\v\\1\end{bmatrix} \text{ avec } \boldsymbol{B}=\begin{bmatrix}-1 & 3 & -3 & 1\\ 3 & -6 & 3 & 0\\ -3 & 3 & 0 & 0\\ 1 & 0 & 0 & 0\end{bmatrix} \text{ la matrice de Bézier.}$$

2.9.4 Structure et données d'un fichier PDB

Pour tracer notre modèle de feuillet β, nous utiliserons les données provenant de l'expérimental, et contenues dans les fichiers de la PDB, ou des données qui seraient produites par l'utilisateur et écrites dans un format PDB. L'activité principale de la PDB est de stocker et distribuer des fichiers de coordonnées de molécules biologiques. Ces fichiers répertorient les atomes de chaque protéine et leur position dans l'espace. Ces fichiers sont disponibles dans plusieurs formats (PDB, mmCIF, XML...). En général un fichier au format PDB possède une large section en-tête de texte qui résume la protéine, des informations et des détails sur la structure, s'ensuit la séquence des acides aminés et la liste de l'ensemble des atomes et leurs coordonnées. Ces fichiers contiennent également les observations expérimentales utilisées pour déterminer les coordonnées atomiques.

Champ	Description
HEADER	première ligne, contient le code PDB, la date de dépôt et la classification de la protéine
TITLE	description de la macromolécule
COMPND	description du contenu macromoléculaire
SOURCE	source biologique de la macromolécule
KEYWDS	liste de mots clés décrivant la macromolécule
EXPDTA	technique expérimentale utilisée pour déterminer la structure
AUTHOR	liste des auteurs
REVDAT	date de révision
JRNL	citation dans la littérature
REMARK	remarques générales, leur structuration est libre
SEQRES	séquence primaire de la macromolécule
HELIX	liste des acides aminés en conformation hélice α
SHEET	liste des acides aminés en conformation feuillet β
SSBOND	liste des couples de cystéines formant des ponts disulfures
ATOM	coordonnées atomiques
TER	représente une fin de chaîne polypeptidique
END	dernière ligne, signale la fin du fichier

Tableau 2.9.4.1 – Champs les plus importants présents dans un fichier PDB et leurs descriptions

Les lignes de chaque fichier PDB sont numérotées, et chaque ligne possède 80 colonnes. Ce format est issu des programmes FORTRAN des années 1970. Les six premières colonnes contiennent le nom du champ, les informations des champs sont séparées par des espaces. Ils existent beaucoup de types de champs, seuls les plus importants pour la compréhension d'un fichier sont détaillés dans le tableau 2.9.4.1.

Pour ce travail SHEET et ATOM sont les champs les plus importants. Le champ SHEET est utilisé pour identifier la position des feuillets β dans la protéine, ils sont à la fois nommés et numérotés. Chaque brin d'un feuillet possède le nom et le numéro du feuillet correspondant ainsi qu'un numéro qui lui est propre afin de pouvoir l'identifier dans le feuillet. Pour chaque brin nous connaissons le numéro du premier et du dernier résidu.

```
         1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
SHEET    1   A 5 THR A 107  ARG A 110  0
SHEET    2   A 5 ILE A  96  THR A  99 -1  N  LYS A  98   O  THR A 107
SHEET    3   A 5 ARG A  87  SER A  91 -1  N  LEU A  89   O  TYR A  97
SHEET    4   A 5 TRP A  71  ASP A  75 -1  N  ALA A  74   O  ILE A  88
SHEET    5   A 5 GLY A  52  PHE A  56 -1  N  PHE A  56   O  TRP A  71
SHEET    1   B 5 THR B 107  ARG B 110  0
SHEET    2   B 5 ILE B  96  THR B  99 -1  N  LYS B  98   O  THR B 107
SHEET    3   B 5 ARG B  87  SER B  91 -1  N  LEU B  89   O  TYR B  97
SHEET    4   B 5 TRP B  71  ASP B  75 -1  N  ALA B  74   O  ILE B  88
SHEET    5   B 5 GLY B  52  ILE B  55 -1  N  ASP B  54   O  GLU B  73
```

Figure 2.9.4.1 – Extrait du champ SHEET d'un fichier PDB. Le codage du fichier étant par colonne, les numéros de la première ligne représentent les dizaines et ceux de la deuxième les unités

Sur la figure 2.9.4.1 nous pouvons observer un exemple de description de feuillet β dans un fichier PDB. Comme il est mentionné plus haut, les six premières colonnes correspondent au nom du champ, nous trouvons ensuite le numéro du brin, puis le nom du feuillet auquel il appartient et derrière le nombre de brins qui composent le feuillet. Dans notre exemple il y a deux feuillets, A et B, composés de cinq brins chacun. Viennent ensuite le nom du premier acide aminé du brin, suivi par l'identifiant de sa chaîne polypeptidique et de son numéro dans la séquence primaire. Nous retrouvons ensuite les mêmes informations pour le dernier acide aminé du brin.

Sur les colonnes 39 et 40 se trouve une information très importante qui est le sens du brin. S'il s'agit du premier brin du feuillet β alors l'information est à 0 car il est le premier point de repère. Si l'information est à 1 alors le brin est parallèle par rapport à son précédent, sinon il est à -1 et antiparallèle par rapport à son précédent. Dans l'exemple nos deux feuillets β sont antiparallèles car

tous leurs brins sont antiparallèles les uns par rapport aux autres.

La description d'un feuillet β se fait toujours d'un bord à un autre, c'est à dire que le premier ainsi que le dernier brin déclarés seront obligatoirement un bord du feuillet.

```
         1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
ATOM     32  N   ARG A   3      11.281  86.699  94.383  0.50 35.88           N
ATOM     33  CA  ARG A   3      12.353  85.696  94.456  0.50 36.67           C
ATOM     34  C   ARG A   3      13.559  86.257  95.222  0.50 37.37           C
ATOM     35  O   ARG A   3      13.753  87.471  95.270  0.50 37.74           O
ATOM     36  CB  ARG A   3      12.774  85.306  93.039  0.50 37.25           C
ATOM     37  CG  ARG A   3      11.754  84.432  92.321  0.50 38.44           C
ATOM     38  CD  ARG A   3      11.698  84.678  90.815  0.50 38.51           C
ATOM     39  NE  ARG A   3      12.984  84.447  90.163  0.50 39.94           N
ATOM     40  CZ  ARG A   3      13.202  84.534  88.850  0.50 40.03           C
ATOM     41  NH1 ARG A   3      12.218  84.840  88.007  0.50 40.76           N
ATOM     42  NH2 ARG A   3      14.421  84.308  88.373  0.50 40.45           N
```

Figure 2.9.4.2 – Extrait du champ ATOM d'un fichier PDB. Le codage du fichier étant par colonne, les numéros de la première ligne représentent les dizaines et ceux de la deuxième les unités

Sur la figure 2.9.4.2 nous pouvons observer la déclaration des coordonnées atomiques d'un acide aminé. Les six premières colonnes nous confirment que nous sommes dans un champ de type ATOM, à chaque ligne correspond donc un atome. La première information disponible est le numéro de série de l'atome, chaque atome a son propre numéro dans un fichier PDB. S'ensuit son nom, le nom de l'acide aminé auquel il appartient, le nom de sa chaîne polypeptidique, et le numéro de son acide aminé. Dans cet exemple, nous avons la description complète d'une arginine appartenant à la chaîne A et possédant le numéro 3.

Nous trouvons ensuite les coordonnées spatiales de l'atome en x, y et z. Puis les facteurs d'occupation et de température. Le facteur d'occupation correspond à la représentation de la quantité de chaque conformation dans un cristal de protéine. En effet, les cristaux sont composés de la même protéine en grande quantité, parfois il se peut qu'il y ait des différences géométriques entre ces protéines cristallisées et c'est pourquoi nous utilisons le facteur d'occupation. Le facteur de température (également appelé « facteur B ») est en réalité un coefficient d'agitation thermique. Les protéines et les atomes qui les composent ne sont pas des objets fixes, le coefficient d'agitation thermique correspond à leur propension à vibrer.

2.9.5 Prédiction de structures secondaires

La prédiction de structures secondaires consiste à trouver les éléments de structures secondaires d'une protéine à partir de sa séquence primaire et/ou d'en effectuer son attribution à partir de sa structure tridimensionnelle. Il existe plusieurs algorithmes de prédiction/attribution de structures secondaires, tels STRIDE [Heinig2004], DEFINE [Richards1988], et, la méthode utilisée dans ce travail, DSSP [Kabsch1983] (Dictionnaire de Structure Secondaire de Protéines). DSSP est la méthode utilisée pour l'attribution de structures secondaires des fichiers de la PDB.

DSSP a été créé par Kabsch et Sander en 1983. Cette méthode détermine la structure secondaire sur la base de l'arrangement des liaisons hydrogène selon le schéma proposé par Corey et Pauling en 1951. Il y a huit types de structures secondaires définies dans DSSP :

- G : hélice 3_{10}. Le carbonyle du résidu *i* forme une liaison hydrogène avec l'amide du résidu *i*+3, sa longueur minimale est de trois résidus.
- H : hélice α. Le carbonyle du résidu *i* forme une liaison hydrogène avec l'amide du résidu *i*+4, sa longueur minimale est de quatre résidus.
- I : hélice π. Le carbonyle du résidu *i* forme une liaison hydrogène avec l'amide du résidu *i*+5, sa longueur minimale est de cinq résidus.
- T : coude fermé par une liaison hydrogène, sa longueur est de 3, 4 ou 5 résidus.
- E : brin β étendu au sein d'un feuillet parallèle ou antiparallèle, sa longueur minimale est de deux résidus.
- B : résidu isolé dans un pont β (paire formant une liaison hydrogène de type feuillet β).
- S : coude sans liaison hydrogène.

Les résidus qui ne sont dans aucune des conformations ci-dessus sont classés dans « pelote », la huitième catégorie. La plupart des méthodes d'attribution de structure secondaire n'utilisent que trois types de conformation : les hélices qui regroupent les types G, H et I, les brins qui regroupent les types E et B, et les boucles qui regroupent les types T et S.

La structure secondaire est définie par l'arrangement des liaisons hydrogène, il est donc très important de bien les définir. DSSP attribue des charges partielles q_1 de +0,42e et -0,42e respectivement sur le carbone et l'oxygène du groupement carbonyle et q_2 de +0,20e et -0,20e respectivement sur l'hydrogène et l'azote du groupement amide. L'énergie électrostatique est définie

par $E = q_1 q_2 \left(\frac{1}{r_{ON}} + \frac{1}{r_{CH}} + \frac{1}{r_{OH}} + \frac{1}{r_{CN}} \right) .332\, kcal/mol$ ou r_{AB} représente la distance interatomique entre A et B. DSSP considère qu'une liaison hydrogène existe si et seulement si E est inférieur à -0,5 kcal/mol. Bien que ce calcul ne soit qu'une approximation, cette règle est acceptée pour la détermination de la structure secondaire des protéines.

2.9.6 Dynamique moléculaire

La dynamique moléculaire est la méthode de prédilection pour l'étude des mouvements de faible amplitude sur une échelle de temps de l'ordre de quelques nanosecondes. Pour effectuer une simulation de dynamique moléculaire (DM), il est nécessaire d'avoir ce que nous appelons un champ de forces, c'est à dire l'ensemble des paramètres dédiés aux atomes impliqués dans des groupes chimiques pour une forme de fonction énergie potentielle. Toute molécule isolée dans une conformation stable aura une énergie potentielle minimum. Dans une simulation de DM, nous effectuons une intégration des équations du mouvement issues de la loi de Newton : $\sum \vec{F}_i = m\vec{a}$. La force dérivant du potentiel EP, il est possible de remonter à l'ensemble des paramètres étant donné que tout système isolé a une énergie totale constate, c'est à dire la somme de l'énergie potentielle et de l'énergie cinétique. Cette dernière est liée à la vitesse et, par la physique statistique, à la température. En conséquence, lorsque nous faisons varier les paramètres de température, nous influençons la distribution des vitesses pour chacun des atomes, et donc leurs positions et leurs accélérations. Une simulation de DM est donc l'étude, pas après pas, d'une trajectoire qui correspond à un échantillonnage temporel de l'hypersurface potentielle. La résolution analytique n'étant pas possible pour les macromolécules, très souvent, la fonction énergie potentielle utilisée, et son champ de force associé, sont définis pour des interactions liées (liaisons, angles, torsions) ou des interactions non liées (contributions de van der Waals, et contributions de type Coulomb). L'évaluation de cette énergie se fait pas à pas, la structure atomique associée étant sauvegardée dans un fichier dit « trajectoire ». C'est la structure tridimensionnelle initiale et le fichier trajectoire qui seront utilisés dans BALLView pour mettre en évidence les aspects dynamiques d'un feuillet β sous la forme d'un « tapis volant ».

Chapitre 3
Modélisation et visualisation scientifique

« I want to share something with you: The three little sentences that will get you through life. Number 1: Cover for me. Number 2: Oh, good idea, Boss! Number 3: It was like that when I got here. »

Homer J. SIMPSON

3.1 Problématique

Le but de ce travail est de développer un mode de représentation des feuillets β sous la forme de surfaces dont la caractéristique majeure est de ressembler à une feuille de papier sur laquelle différentes propriétés, structurales ou physico-chimiques par exemple, peuvent être symbolisées. La problématique de ce travail est qu'il faut pouvoir représenter ces feuillets non pas sur une protéine particulière ou sur un type de protéines particulier, mais sur l'ensemble des fichiers présents dans la PDB ou écrits au format PDB. Il faut que la solution soit la plus générique possible afin que cela puisse fonctionner sur tous les fichiers des utilisateurs potentiels. Pour cela il nous faut nous appuyer sur les données présentes obligatoirement dans un fichier PDB. Certains critères doivent être impérativement présents pour qu'un fichier soit accepté dans la PDB.

Pour représenter des feuillets β, deux informations sont essentielles : quels acides aminés sont en conformation β et quelles sont les coordonnées spatiales de ces résidus.

Pour qu'un fichier soit soumis à acceptation dans la banque de données, il faut que sa structure secondaire soit connue, déterminée (au moyen de DSSP [Kabsch1983]) et inscrite dans le fichier à l'aide des champs prévus à cet effet qui ont été détaillés dans la section 2.9.4 vue précédemment. Une fois que les résidus en conformation β ont été identifiés dans le champ SHEET d'un fichier PDB, il suffit de croiser les informations des numéros de résidus et du nom de leur chaîne avec les données des champs ATOM afin de connaître les coordonnées spatiales des atomes qui composent ces résidus. À l'aide des champs SHEET et ATOM, nous savons donc quels résidus sont en conformation β et quelles sont leurs coordonnées spatiales. Nous pouvons alors commencer à représenter les feuillets β sous forme de surfaces et non plus de simples associations de flèches.

3.2 BALLView

Afin de réaliser ce travail nous avions deux possibilités : soit partir de zéro et écrire en intégralité un logiciel de modélisation moléculaire, soit reprendre un logiciel existant *open source* et le compléter pour nos besoins. Nous avons opté pour la seconde possibilité car le fait d'écrire complètement un logiciel aurait pris beaucoup de temps et ce de façon inutile car le sujet de cette thèse n'est pas la conception d'un nouveau programme de modélisation.

Il existe de très nombreux logiciels de modélisation moléculaire, suite à l'impulsion donnée par Roger Sayle avec son logiciel RasMol beaucoup d'entre eux sont distribués en *open source*. Pour trouver le logiciel qui nous a servi de base pour ce travail, nous ne nous sommes bien sur intéressés qu'à cette catégorie.

Logiciel	**Langage**	**Graphismes**	**Interface utilisateur**	**Dernière version**	**Site Internet**
VMD	C	OpenGL	Python	08/01/09	www.ks.uiuc.edu/research/vmd
PyMOL	C	OpenGL	Python	05/10/09	www.pymol.org
RasMol	C	RasMol	GTK/Xlib GUI	23/07/09	www.rasmol.org
BALLView	C++	OpenGL	Qt	12/02/10	www.ballview.org

Tableau 3.2.1 – Logiciels de modélisation moléculaire open source *parmi les plus utilisés et leurs caractéristiques*

Durant les prospections pour trouver le logiciel qui a servi de base aux travaux présentés ici, de nombreux logiciels ont été testés. Nous n'allons bien sur pas tous les détailler, notre attention a été retenue plus particulièrement par ceux dont les caractéristiques sont présentées dans le tableau 3.2.1.

Le choix a été fait en fonction d'un certain nombre de critères, le premier d'entre eux est que l'ensemble des modes de visualisation classiques soit présent (voir le paragraphe 2.8). Il faut ensuite que le code du logiciel soit régulièrement mis à jour, de façon à ne pas s'enfermer dans un logiciel qui n'évoluera plus et qui risque de ne plus être compatible sur de futures plateformes. Un autre critère important est que ce logiciel utilise des bibliothèques logicielles qui elles aussi sont maintenues à jour régulièrement. Sur le tableau 3.2.1 nous constatons que les dates des dernières versions des quatre logiciels en lice sont récentes, la première de nos conditions est donc respectée. En ce qui concerne les bibliothèques graphiques utilisées, nous constatons que RasMol développe sa propre bibliothèque alors que les autres utilisent OpenGL, standard reconnu pour ses performances et très régulièrement mis à jour. Le fait de ne pas utiliser une bibliothèque comme OpenGL laisse supposer que nous ne bénéficierons pas des derniers progrès réalisés en matière de représentation graphique, RasMol est donc écarté. En ce qui concerne les bibliothèques d'interface utilisateur, parmi les logiciels restants, il y a Python et Qt. Nous constatons que la bibliothèque Python est très utilisée pour les logiciels de modélisation moléculaire, cependant Qt est très largement utilisée dans la communauté informatique car c'est une bibliothèque très complète et régulièrement mise à jour. Le choix s'arrête donc sur BALLView qui est le seul logiciel du tableau 3.2.1 à utiliser Qt. BALLView est de plus codé en C++, qui est un langage objet, ce qui permet une meilleure organisation du code, ainsi qu'une grande flexibilité.

3.3 Des carbones α à une surface β

Il a été établi dans le paragraphe 3.1 qu'un fichier PDB contient l'ensemble des données nécessaires pour représenter un feuillet β. Nous disposons en effet des coordonnées de l'ensemble des atomes de chaque acide aminé en conformation β, seulement nous ne cherchons pas à représenter l'ensemble des atomes de chaque acide aminé, mais une surface. Il nous faut donc déterminer ce que nous devons représenter pour chaque résidu. Étant donné que nous cherchons à visualiser des feuillets β également appelés feuillets plissés β, l'aspect plissé est en conséquence très important. C'est l'alternance haut-bas de la position des carbones α consécutifs sur un même brin qui produit cet effet plissé des feuillets (cf. Fig. 2.4.1 et 2.4.2). Nous utiliserons donc les carbones α pour représenter les divers acides aminés.

3.3.1 Première tentative « chaotique »

La première tentative de représentation d'un feuillet β par l'intermédiaire des carbones α consistait à relier les acides aminés des brins voisins. Dans un fichier PDB chaque feuillet β possède un nom, comme il est spécifié dans le paragraphe 2.9.4 qui traite de la description du champ SHEET. Ce nom est généralement composé du nom de la chaîne polypeptidique et d'un numéro correspondant au numéro du feuillet β, dans le cas où plusieurs feuillets existent sur une même chaîne.

Ainsi, si une chaîne « A » comprend deux feuillets β, ils s'appelleront « A1 » et « A2 ». De plus chaque brin au sein d'un feuillet possède un numéro qui lui est propre; ainsi, des brins consécutifs auront des numéros qui se suivent. L'association du nom du feuillet avec le numéro du brin fait qu'il est impossible de confondre deux brins d'une même protéine.

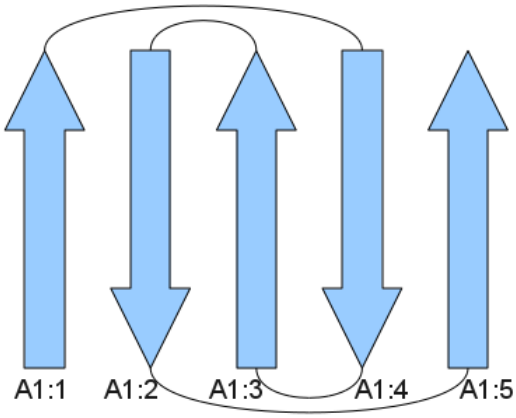


Figure 3.3.1.1 – Représentation d'un feuillet β composé de cinq brins β numérotés séquentiellement, pour chaque brin le nom de son feuillet est renseigné suivi du numéro du brin séparé par le signe de ponctuation « : »

Comme nous pouvons le voir sur la figure 3.3.1.1 chacun des brins qui compose le feuillet porte le nom du feuillet ainsi qu'un numéro qui correspond au numéro du brin dans ce feuillet. Cet exemple montre que les numéros sont attribués en fonction de la place du brin dans le feuillet et non en fonction de sa place dans la chaîne. Deux brins adjacents auront bien des numéros qui se suivent alors qu'ils ne suivent pas dans la chaîne. Ainsi les brins « A1:1 » et « A1:2 » sont voisins bien que, si nous considérons leurs places dans la chaîne, c'est à dire le long de la séquence, ils sont les plus éloignés. C'est sur cette information que nous allons nous baser pour savoir quels acides aminés doivent être reliés.

Une fois les informations récoltées et filtrées, nous disposons d'une liste de tous les carbones α dont le résidu est en conformation β, et nous savons pour chacun à quel brin et à quel feuillet il appartient. Il faut maintenant construire la surface des feuillets.

Il existe beaucoup d'algorithmes pour créer des maillages de surfaces. Mais nous sommes face à un problème très spécifique dans la mesure où nous souhaitons passer d'une structure globalement linéique (la chaîne de la structure primaire) à une structure localement surfacique (le feuillet β). C'est pourquoi nous avons eu besoin de développer notre propre algorithme de maillage. Cet algorithme crée un maillage incrémental par couple de brins consécutifs d'un même feuillet. Il est basé sur de simples tests de distance afin de déterminer quel triangle doit être crée, les brins n'étant pas nécessairement les uns faces aux autres, et ne comportant pas toujours le même nombre de résidus.

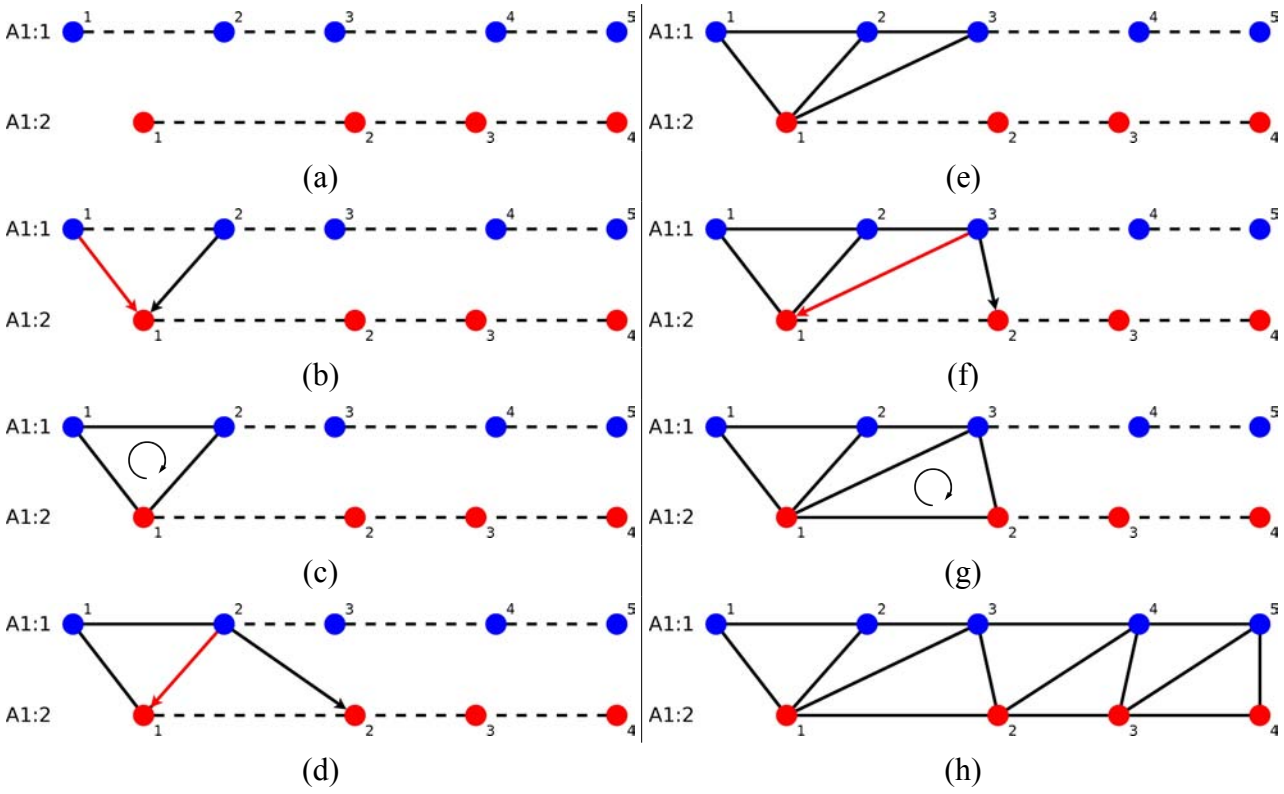


Figure 3.3.1.2 – Illustration du déroulement de l'algorithme de maillage

Sur la figure 3.3.1.2 nous pouvons observer le déroulement de l'algorithme de maillage ; (a) nous disposons du couple de brins consécutifs A1:1 et A1:2, les résidus de ces brins possèdent un numéro représentant leur position ; (b) on compare $d1 = \|\overrightarrow{A1:1_1 A1:2_1}\|$ et $d2 = \|\overrightarrow{A1:1_2 A1:2_1}\|$ si d2 > d1, comme c'est le cas dans notre exemple, nous créons le triangle $t(A1:1_1, A1:1_2, A1:2_1)$

sinon nous créons le triangle $t(A1:1_1, A1:2_2, A1:2_1)$. Une fois ce triangle résolu, nous continuons le long des brins et, en (d), nous comparons $d1 = \|\overrightarrow{A1:1_2 A1:2_1}\|$ et $d2 = \|\overrightarrow{A1:1_2 A1:2_2}\|$, nous créons le triangle $t(A1:2_1, A1:1_2, A1:2_3)$ en conséquence, comme nous le constatons en (e). Nous poursuivons ainsi jusqu'à atteindre les derniers résidus des brins.

Une fois le maillage entre deux brins consécutifs terminé, dont le résultat obtenu est visible sur la figure 3.3.1.2h, nous passons aux deux brins suivants. Si nous reprenons notre exemple nous allons donc passer des brins A1:1 et A1:2 aux brins A1:2 et A1:3 et nous relançons notre algorithme jusqu'à avoir atteint le dernier couple de brins du feuillet.

Sur l'exemple donné, nous constatons que la numérotation des résidus des deux brins se fait dans la même direction, de la gauche vers la droite, ce qui signifie que les brins sont parallèles. Hors, les feuillets β sont en grande majorité antiparallèles ; dans ce cas l'algorithme développé renverra un résultat aberrant en maillant les extrémités opposées des brins consécutifs. Nous avons vu dans la section 2.9.4 que le champ SHEET nous renseigne sur le sens des brins par rapport au brin précédent. Dans le cas où des brins consécutifs sont antiparallèles, il suffit de numéroter les résidus dans l'autre sens. Comme nous pouvons le voir sur la figure 3.3.1.3a, les brins A1:1 et A1:2 sont antiparallèles : il faut donc changer la numérotation du brin A1:2 tel que présenté sur la figure 3.3.1.3b. Avant d'exécuter l'algorithme de maillage, il faut donc faire une première passe pour modifier la numérotation des résidus si besoin est.

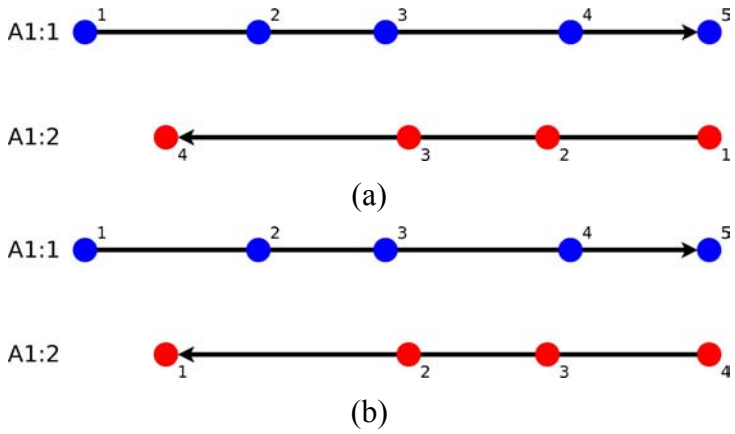


Figure 3.3.1.3 – Changement de l'ordre de la numérotation des résidus dans le cas où deux brins consécutifs sont antiparallèles

L'ordre dans lequel les résidus d'un triangle sont organisés n'est pas anodin, comme nous pouvons le constater sur la figure 3.3.1.2c et 3.3.1.2g, les sommets tournent, pour les deux constructions possibles de triangle, dans le sens horaire. Cela est nécessaire pour le calcul des

normales en chaque sommet du maillage. Les normales servent pour l'éclairage des surfaces, c'est la normale d'une surface qui va déterminer la façon dont elle renvoie la lumière. Afin que les normales soient toutes orientées dans le même sens par rapport à la surface que nous allons générer, il faut absolument que les triangles construits aient le même sens de rotation. La normale d'un triangle (A, B, C) est obtenue en normalisant le vecteur calculé avec le résultat du produit vectoriel suivant : $\vec{N} = \overrightarrow{AB} \wedge \overrightarrow{AC}$. Pour l'éclairage de la surface chaque sommet doit avoir sa normale, nous allons donc attribuer aux sommets A, B et C la normale qui vient d'être calculée. Seulement, un même sommet peut participer jusqu'à six triangles différents, cela représente donc six normales différentes pour ce même sommet. Nous calculons la moyenne de ces vecteurs et l'attribuons au sommet concerné, il est fait de même pour l'ensemble des sommets composant la surface.

Si les sommets des triangles ne tournaient pas tous dans le même sens, l'éclairage de la surface ne serait pas uniforme et certains triangles seraient sombres tandis que ses voisins seraient éclairés normalement.

La figure 3.3.1.4 illustre le comportement de l'algorithme sur un fichier PDB : (a) seuls les carbones α du feuillet β sont représentés et il a été attribué une couleur différente à chaque brin afin de les différencier ; (b) on peut observer le maillage obtenu grâce à l'algorithme développé ; (c) la surface correspondante au maillage avec l'éclairage rendu grâce aux normales calculées.

Ce modèle présente plusieurs avantages, tout d'abord il est compatible avec l'ensemble des fichiers présents dans la PDB car il s'appuie sur le champ SHEET qui est obligatoirement renseigné, et il est également très rapide à calculer et très peu gourmand en terme de ressources informatiques. Il permet également de visualiser de façon immédiate un feuillet β au sein d'une protéine, et ce de façon bien plus efficace qu'avec la méthode *cartoon* qui représente les feuillets β uniquement par le biais de flèches représentant les brins β les constituants. Cependant, la surface ainsi représentée est très chaotique à cause du caractère plissé des brins β.

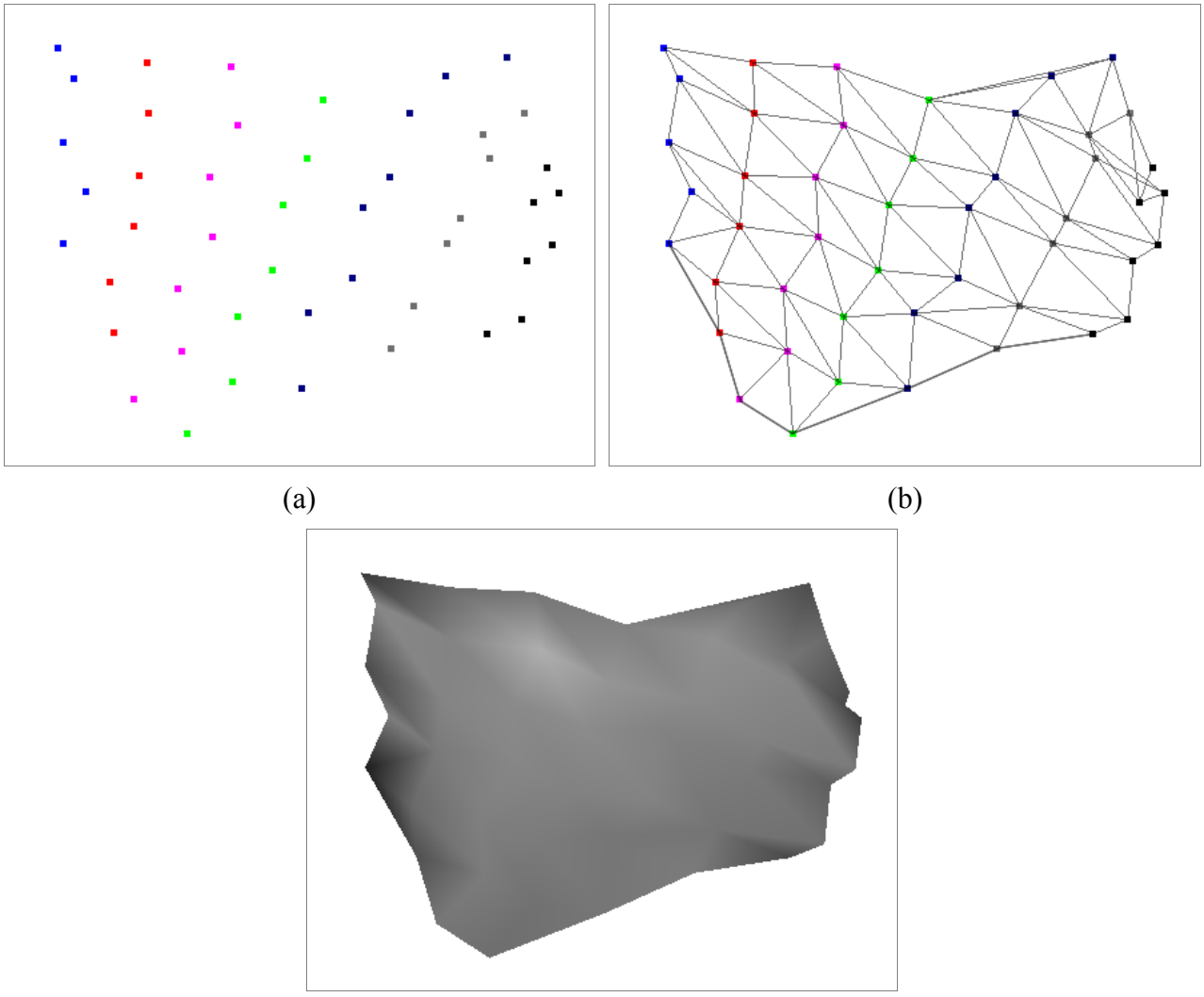


Figure 3.3.1.4 – Illustration du résultat fourni par l'algorithme développé sur le fichier avec le code PDB 1914 [1914]

3.3.2 Modèle de Catmull-Rom

3.3.2.1 Interpolation par brin β

Afin d'obtenir une surface plus lisse et sur laquelle l'aspect plissé des feuillets β serait plus visible, nous avons opté pour l'utilisation de courbes splines. Le type de courbe spline utilisé sera celui développé par Edwin Catmull et Raphael Rom en 1974. Les splines de Catmull-Rom ont été choisies car elles sont efficaces, et ont pour particularité d'être interpolantes, et par conséquent de passer par les positions exactes de leurs points de contrôle. Dans notre cas, les points de contrôle sont les positions des carbones α et nous souhaitons que notre surface passe par l'ensemble des carbones α afin de préserver l'aspect plissé des feuillets β.

Cependant ce type de spline présente un inconvénient pour l'utilisation que nous souhaitons en faire car, comme cela a été décrit dans la section 2.9.2, il n'y a pas d'interpolation entre le premier et le deuxième point de contrôle, ainsi qu'entre l'avant-dernier et le dernier point de contrôle. La solution mise en place est d'utiliser la position du carbone α précédant le brin β et du carbone α suivant ce même brin. Les étapes (a) et (b) de la figure 3.3.2.1.1 illustrent la récupération de la position de ces points de contrôle.

Une fois ces points récupérés nous pouvons calculer nos splines de Catmull-Rom, à chaque brin β correspond une spline. Comme décrit dans la section 2.9.2 nous utilisons un facteur *t* pour ce calcul qui représente la distance entre deux points successifs. Si, par exemple, à chaque étape du calcul de la spline de Catmull-Rom nous incrémentons la valeur de *dt* de 0,25, trois points seront interpolés comme nous pouvons le voir sur la figure 3.3.2.1.1c. La dernière étape de ce calcul consiste en la suppression des points supplémentaires récupérés car ils ne font pas partie des brins (Fig. 3.3.2.1.1d).

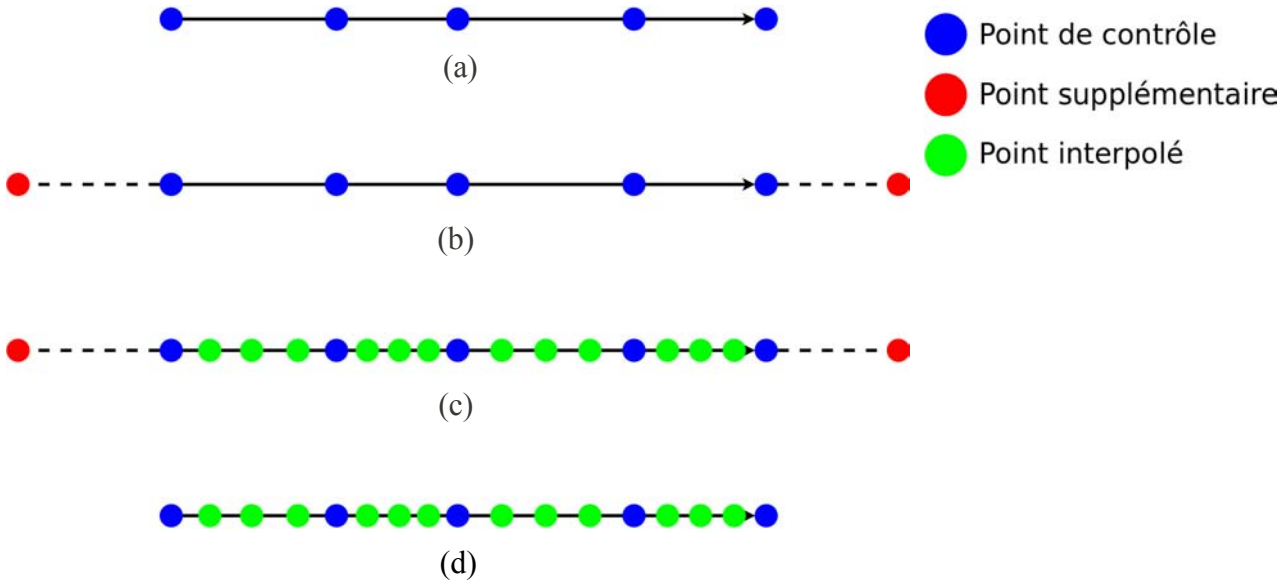


Figure 3.3.2.1.1 – Étapes nécessaires à l'utilisation des splines de Catmull-Rom : (a) la position de chaque carbone α du brin sert de point de contrôle ; (b) on récupère les positions des carbones α précédant et suivant les brins β nécessaires au calcul de la spline étant donné que le premier et le dernier point de contrôle n'y sont pas inclus ; (c) on calcule les positions des points interpolés en utilisant Catmull-Rom avec un incrément dt de 0,25 dans cet exemple ; (d) on ne conserve pas les points supplémentaires pour la suite

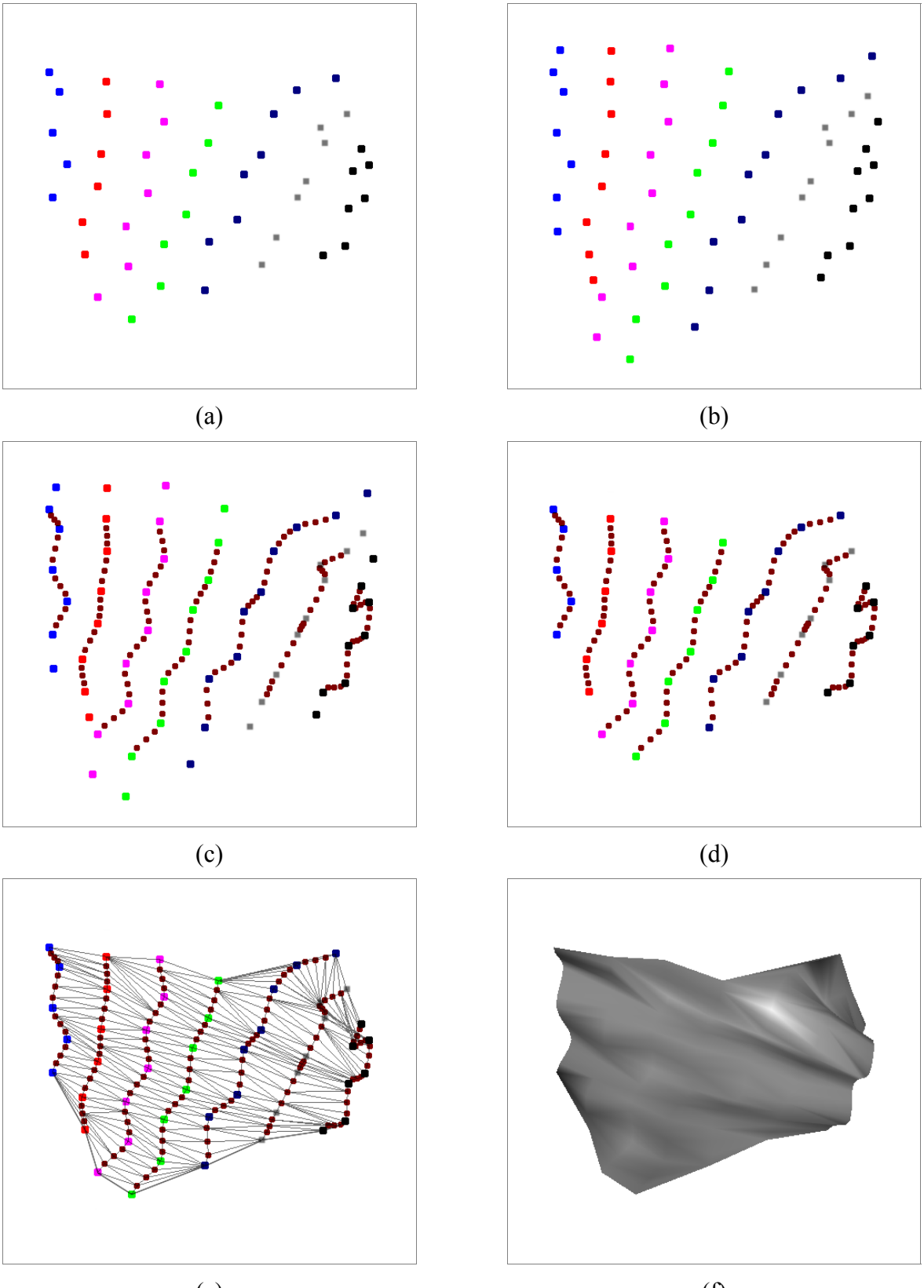


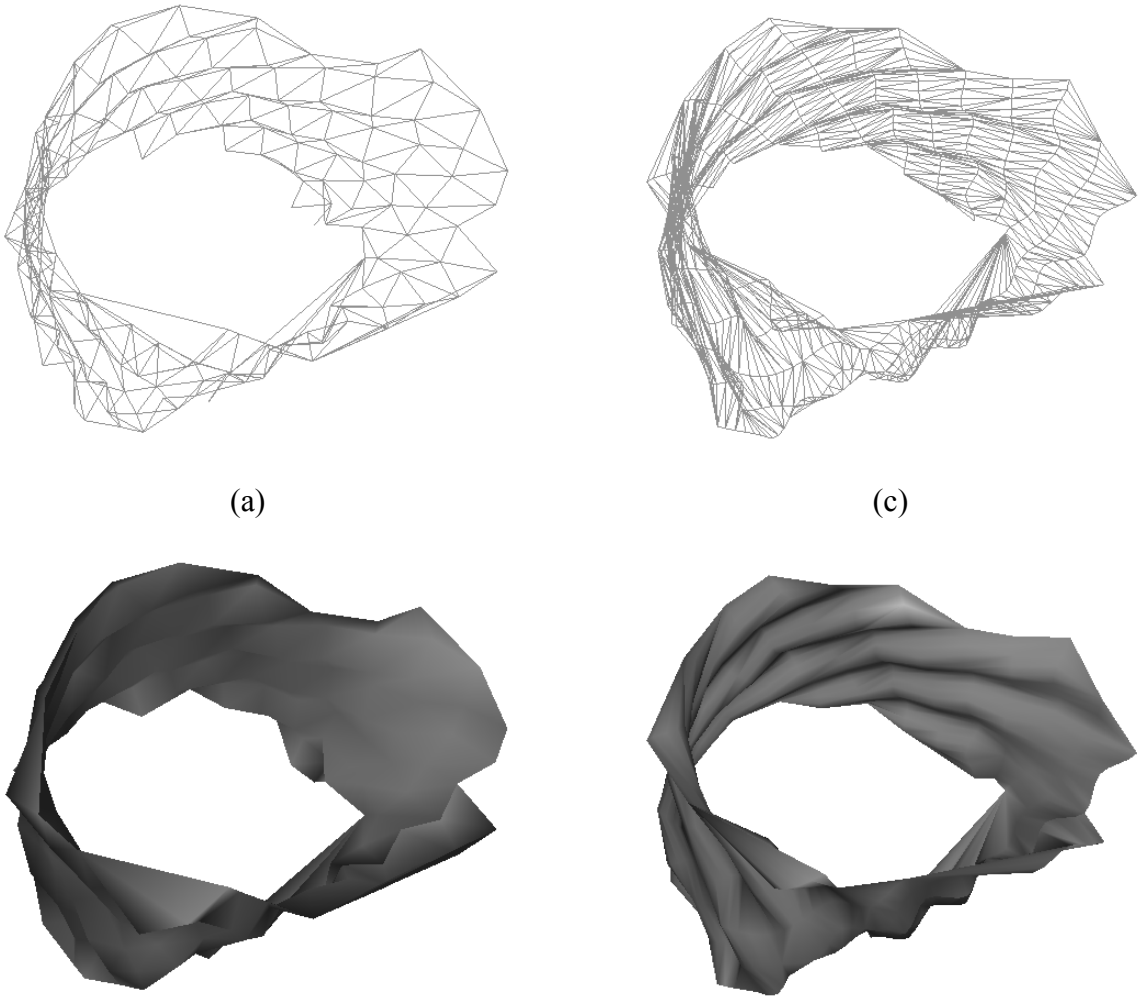
Figure 3.3.2.1.2 – Illustration du déroulement de l'algorithme utilisant les splines de Catmull-Rom pour représenter la surface d'un feuillet β [1914]

Une fois les splines calculées pour chacun des brins β, le même algorithme de maillage, que celui utilisé dans le paragraphe précédent, basé sur des tests de distance, est utilisé.

Sur la figure 3.3.2.1.2 nous pouvons suivre le déroulement de l'algorithme mis en place en utilisant les splines de Catmull-Rom. L'illustration (a) nous montre les points de contrôle correspondants aux carbones α des différents brins du feuillet, une couleur différente a été attribuée à chaque brin afin de pouvoir les différencier ; sur l'illustration (b) nous utilisons les positions des carbones α précédant et suivant des brins ; (c) nous montre le résultat des calculs des splines de Catmull-Rom, nous constatons bien qu'il n'y a aucun point interpolé entre le premier et le deuxième point de contrôle, ainsi qu'entre le dernier et l'avant-dernier pour chaque brin ; en (d) nous ne conservons pas les points supplémentaires car ils n'appartiennent pas aux divers brins, nous ne cherche donc pas à les représenter ; en (e) nous pouvons voir le résultat du maillage obtenu par l'algorithme détaillé dans le paragraphe précédent (cf. paragraphe 3.3.1) et en (f) il est possible d'observer la surface correspondante au maillage.

La figure 3.3.2.1.3 illustre les différences entre les deux algorithmes que nous venons de détailler, (a) et (c) nous montrent les différences au niveau du maillage et (b) et (d) les différences au niveau de la surface générée par le maillage. Nous constatons immédiatement que la surface obtenue sur (d) est nettement moins chaotique et beaucoup plus lisse que sur (c), l'aspect plissé de (d) démontre l'utilité du choix des carbones α comme points de contrôle.

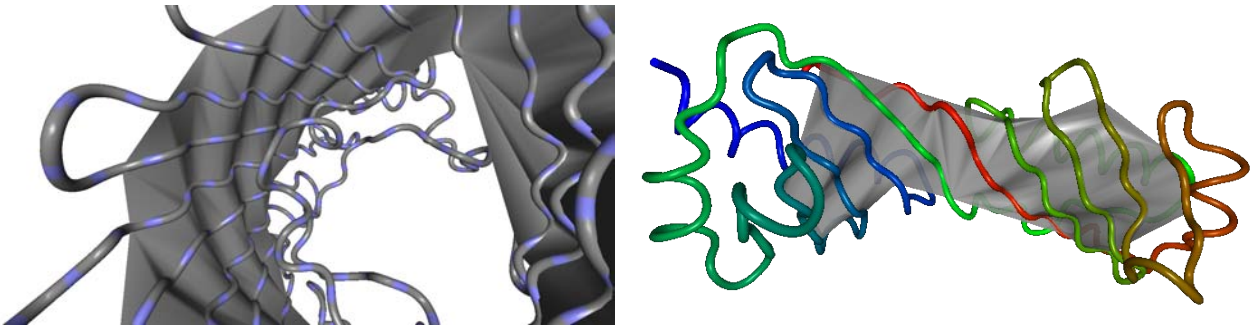
Les exemples montrés sur la figure 3.3.2.1.4 illustrent les qualités de notre modèle, sur (a) nous constatons que notre surface épouse parfaitement le squelette de la protéine et sur (b) nous constatons que la visualisation du feuillet ainsi que sa taille et sa forme est immédiate, un effet de transparence a été ajouté sur ce feuillet β de façon à ce qu'il soit possible d'observer les structures cachées par notre surface.



(a) (c)

(b) (d)

Figure 3.3.2.1.3 – Illustration des différences existantes entre les algorithmes de maillage avec et sans l'utilisation des splines de Catmull-Rom [1PRN]



(a) (b)

Figure 3.3.2.1.4 – Exemples d'utilisations de notre modèle couplé avec un rendu de type tube représentant le squelette de la protéine. (a)[1PRN] et (b)[1YTB]

3.3.2.2 Interpolation bidimensionnelle

Malgré les qualités évidentes de notre modèle, ce dernier souffre de défauts : le maillage est très irrégulier dans la mesure où l'échantillonnage au niveau des brins β est beaucoup plus important que le nombre de brins ce qui a pour conséquence de créer des triangles très étirés. La surface est bien lissée le long des brins, mais pas entre les brins. Afin de résoudre ce problème nous avons développé une méthode d'interpolation bidimensionnelle à base de splines de Catmull-Rom.

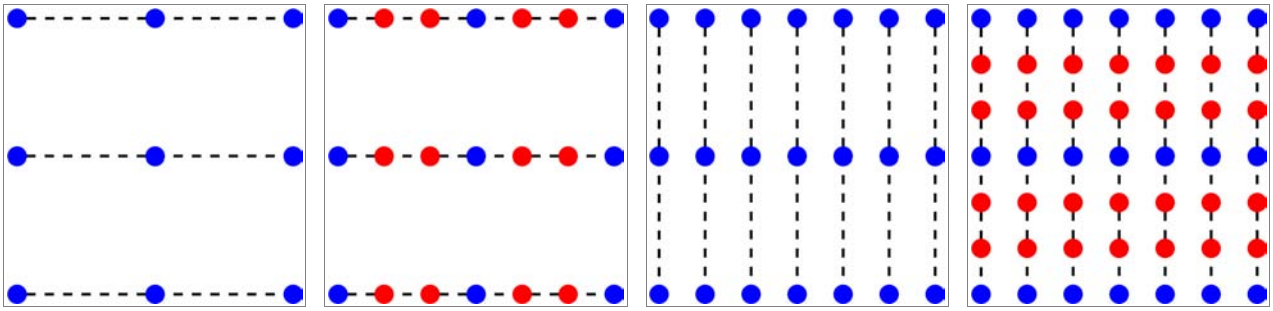


Figure 3.3.2.2.1 – Concept de l'interpolation bidimensionnelle utilisant les splines de Catmull-Rom

La figure 3.3.2.2.1 représente le concept de cette interpolation bidimensionnelle. Dans un premier temps, en (a), nous utilisons les carbones α comme points de contrôle de la même façon que pour les autres algorithmes puis, en (b), nous calculons les nouveaux points (les points supplémentaires nécessaires aux calculs ne sont pas illustrés sur ces figures). À partir de (c), nous ne considérons plus les splines dans le sens des brins β, nous passons de la structure primaire de la protéine à sa structure tertiaire locale. Ce sont les points interpolés qui serviront de support aux splines. Voici en (d) le résultat escompté d'une telle méthode : l'obtention d'un maillage régulier tout au long du feuillet β et non plus uniquement le long des brins β.

Sur la figure 3.3.2.2.1, l'algorithme est simplifié dans la mesure où, en (a), nous constatons qu'il y a exactement le même nombre de points de contrôle par spline, signifiant qu'il y a exactement le même nombre d'acides aminés par brin β. Dans la réalité, il y a rarement le même nombre d'acides aminés par brin β sur un même feuillet. La difficulté de cet algorithme réside dans le fait qu'il faut s'assurer de la présence du même nombre de points sur chaque spline après la première interpolation. De plus, afin d'obtenir un maillage régulier, l'espace entre chaque point interpolé devra être régulier tout au long d'une spline.

Afin d'obtenir le même nombre de points sur chaque spline, nous allons définir un pas d'interpolation, qui correspondra à l'espace entre deux points interpolés, spécifique à chaque spline.

Pour cela il nous faut connaître la distance séparant chaque couple de carbones α consécutifs, ainsi que la longueur totale de chaque brin β qui correspond à l'addition des distances entre les carbones α consécutifs, il ne s'agit pas de la distance projetée sur l'axe du brin β mais bien de la distance inter-atomique réelle. La longueur de chaque brin est normalisée : nous bornons entre 0 et 1 la longueur de nos brins, de façon à ce que les distances entre les carbones α consécutifs représentent un pourcentage de la longueur totale. Le pas d'interpolation correspond donc à $p = \frac{1}{(n-1) * nb_{brins}}$ où p représente le pas d'interpolation, n le nombre de points souhaités par spline et nb_{brins} le nombre de brins β présents dans le feuillet β. Le pas d'interpolation est corrélé au nombre de brins composants le feuillet : de cette façon, un feuillet qui aura deux fois plus de brins qu'un autre, aura deux fois plus de points interpolés. Le facteur t de la spline se calcule donc : $t = \frac{p}{\|\overrightarrow{P_1 P_2}\|}$ où $\|\overrightarrow{P_1 P_2}\|$ représente la distance entre le point de contrôle P_1 et le point de contrôle P_2, c'est à dire la distance entre les deux carbones α entre lesquels les points sont interpolés.

Les points interpolés sont calculés entre P_1 et P_2. Lorsque la position de ces points atteint la position de P_2 il faut procéder au changement des points de contrôle. À cette fin nous incrémentons une variable de la valeur de t à chaque calcul de point et dès que la longueur $\|\overrightarrow{P_1 P_2}\|$ additionnée aux longueurs précédentes, est dépassée, nous changeons de points de contrôle.

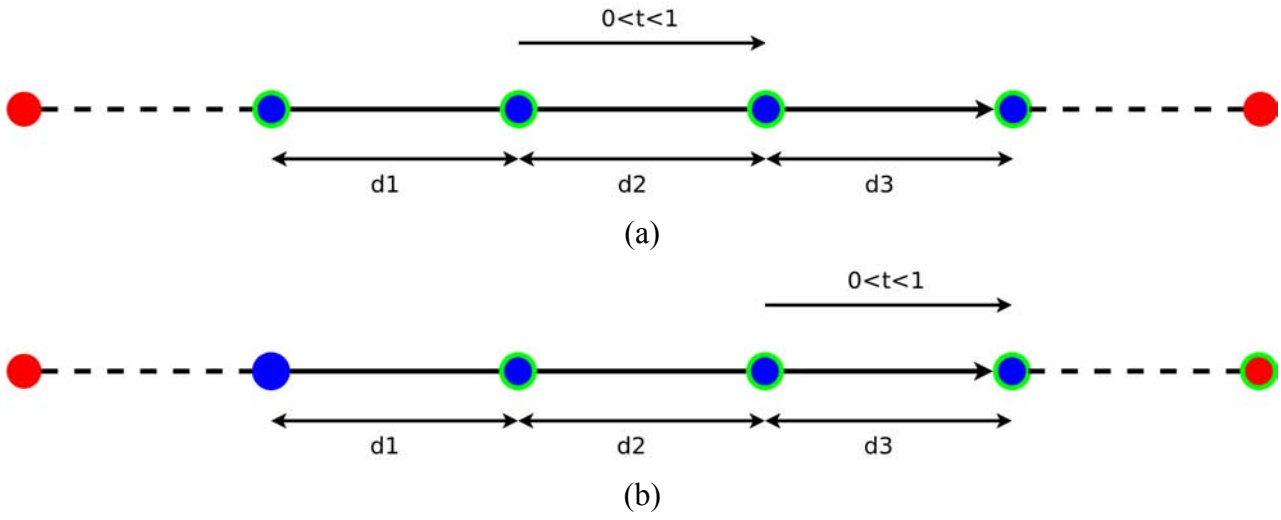


Figure 3.3.2.2.2 – Illustration du changement de points de contrôle en fonction des distances calculées, les points interpolés ne sont pas représentés

Sur la figure 3.3.2.2.2a, les points de contrôle actifs sont ceux qui sont détourés de vert. À chaque étape du calcul des points, nous incrémentons la valeur *t* ainsi qu'une variable nommée *v* de la valeur du pas d'interpolation *p*. Dès que la valeur de *v* est supérieure à la distance entre les points de contrôle P_1 et P_2 additionnée aux distances inter-carbones α précédentes, nous changeons de points de contrôle. Pour reprendre la figure 3.3.2.2.2a, lorsque la valeur de *v* sera supérieure à *d1+d2,* alors nous changeons de points de contrôle tel qu'il est décrit dans (b). Le calcul de la spline s'arrête lorsque *v* contient une valeur supérieure ou égale à 1.

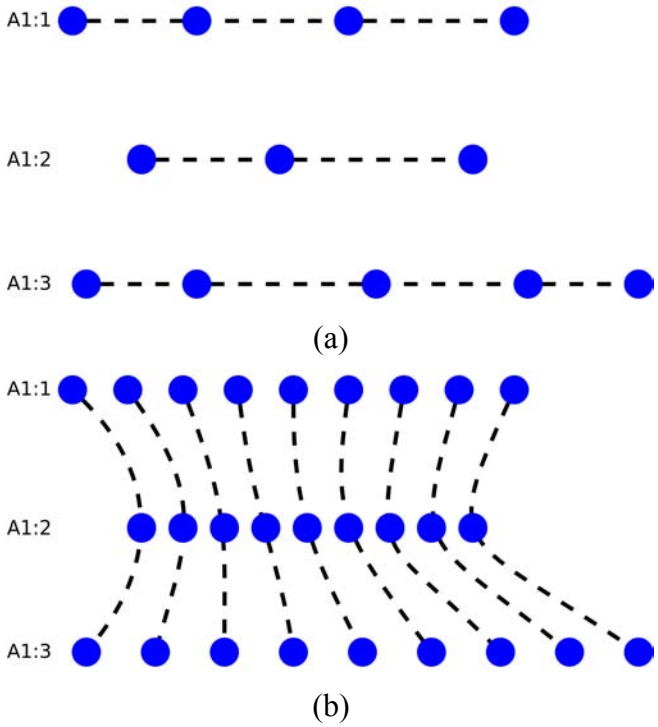


Figure 3.3.2.2.3 – (a) Avant interpolation, chaque spline possède un nombre différent de points de contrôle ; (b) après interpolation, chaque spline possède le même nombre de points de contrôle nous pouvons alors définir de nouvelles splines

Une fois cette étape achevée, toutes les splines auront le même nombre de points de contrôle. Cette étape est illustrée dans la figure 3.3.2.2.3 : en (a), nous constatons que le nombre de points de contrôle varie d'une spline à l'autre, cela correspond au fait que les brins β ont un nombre d'acides aminés différents ; en (b), après avoir calculé le pas d'interpolation et appliqué notre algorithme, nous constatons que nous avons le même nombre de points par splines et qu'il est possible de définir les splines pour la suite de l'algorithme tel que nous l'avons conceptualisé dans la figure 3.3.2.2.1.

Pour les nouvelles splines que nous devons interpoler, il n'y a pas de nécessité de calculer un pas d'interpolation car ces splines ont le même nombre de points de contrôle. Dans un soucis

d'unicité nous calculons *t* de façon à avoir le même nombre de points interpolés dans les deux dimensions de la surface. Cependant, étant donné que pour une spline de Catmull-Rom il n'y a pas d'interpolation entre le premier et le deuxième point de contrôle, ainsi qu'entre le dernier et l'avant-dernier, nous devons donc créer de nouveaux points de contrôle aux extrémités de nos splines. Pour ce faire, nous devons calculer les points symétriques du deuxième point de contrôle par rapport au premier, ainsi que de l'avant-dernier point de contrôle par rapport au dernier. Une fois l'interpolation réalisée, nous ne conservons pas les points extrêmes et nous pouvons mailler notre surface. Cette fois il n'est pas nécessaire d'utiliser l'algorithme de maillage basé sur les tests de distance étant donné que nous avons le même nombre de points d'une spline à l'autre, il suffit de créer nos triangles à l'aide de la position occupée par les points dans leurs splines. Les triangles sont créés deux par deux : $t_1(S_{n:i}, S_{n:i+1}, S_{n+1:i})$ et $t_2(S_{n:i+1}, S_{n+1:i+1}, S_{n+1:i})$, où S_n représente la spline numéro *n* et *i* représente l'indice d'un point d'intérêt dans une spline. Comme dans le cas du premier algorithme développé, il est important que l'ordre des sommets des triangles soit respecté afin que la surface soit bien orientée et que les calculs de normales soient justes. La normale à un sommet est calculée comme étant la moyenne des normales des triangles auxquels ce sommet participe.

Sur la figure 3.3.2.2.4 nous pouvons observer le déroulement de cet algorithme. Les étapes qui vont de (a) à (c) sont identiques aux étapes de l'algorithme précédent. En (d) les points extrêmes ont été calculés par symétrie afin de pouvoir calculer nos splines de Catmull-Rom ; (e) nous montre le résultat obtenu après interpolation et nous ne conservons pas les points extrêmes qui n'appartiennent pas au feuillet β comme nous pouvons le voir sur (f). Nous constatons à cette étape que l'échantillonnage est parfaitement régulier, l'interpolation bidimensionnelle permet de résoudre le problème de l'algorithme précédent qui créait des distances importantes entre les divers points, si bien que le maillage générait des triangles très étirés. L'étape (g) montre le résultat du maillage qui se base uniquement sur l'ordre des points dans les splines, et (h) présente la surface correspondante.

La surface que nous obtenons par interpolation bidimensionnelle n'est plus chaotique, elle présente l'aspect plissé caractéristique des feuillets β.

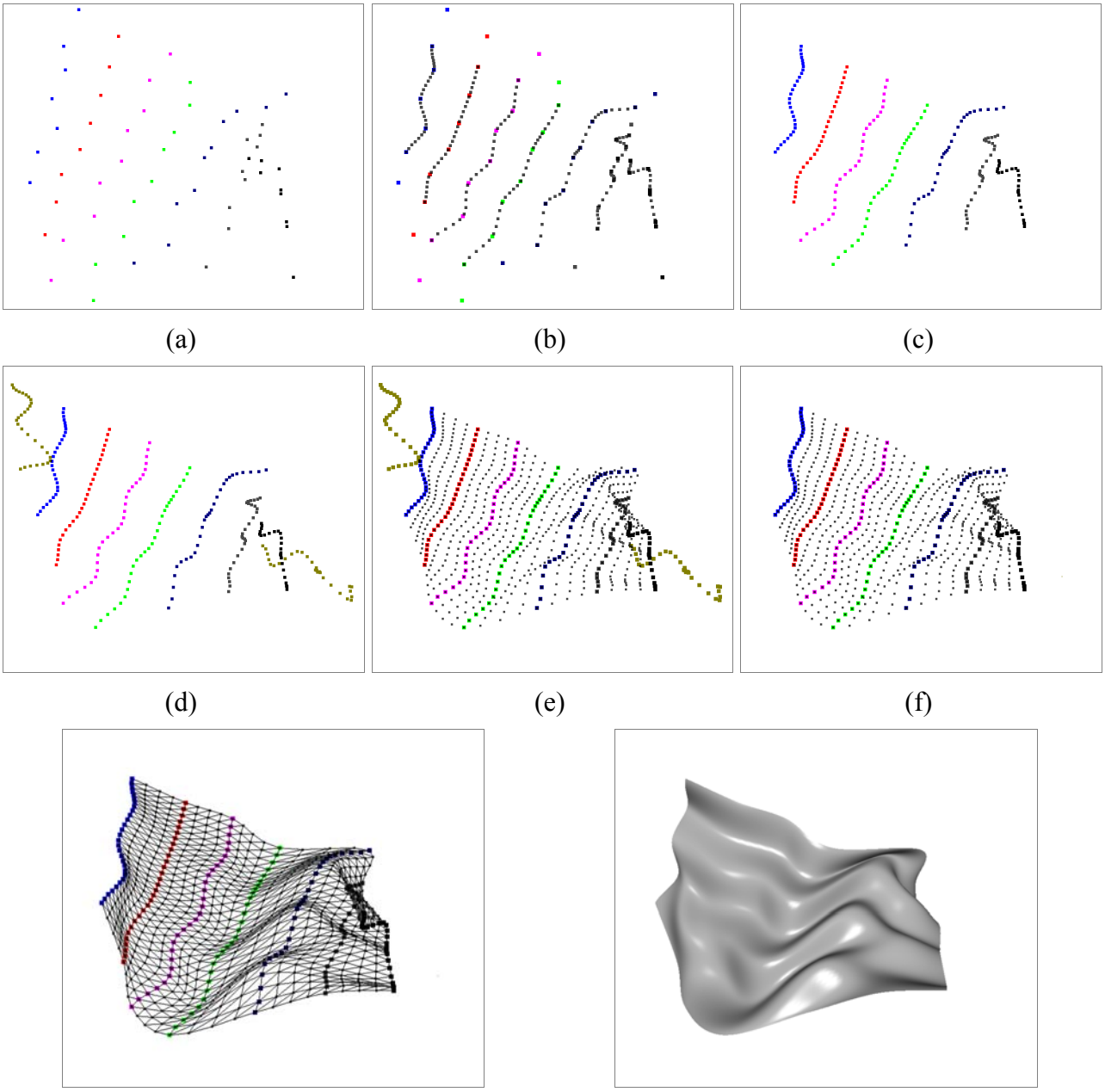
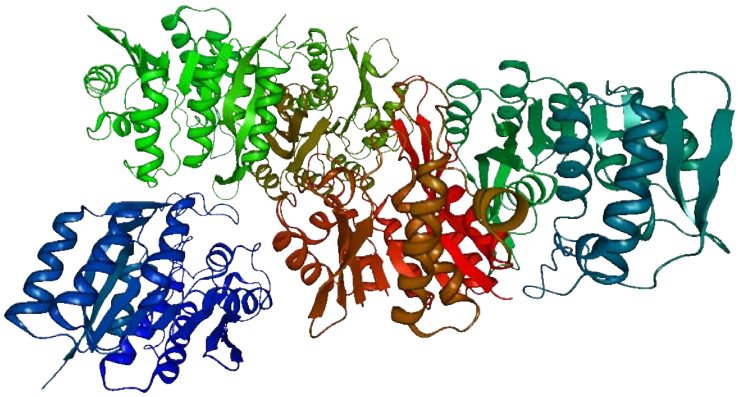


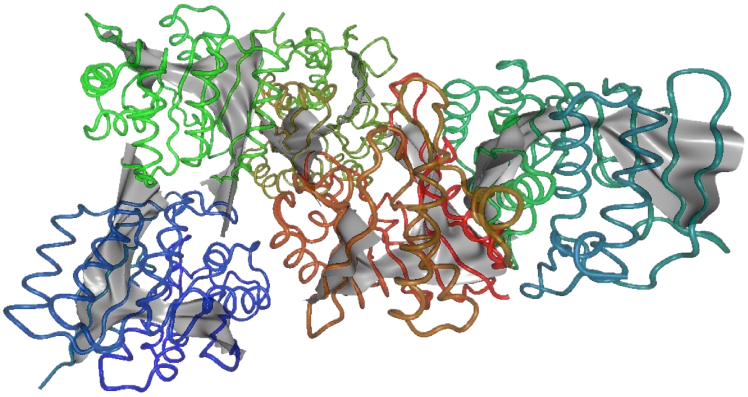
Figure 3.3.2.2.4 – Illustration des étapes de l'algorithme de maillage utilisant une interpolation bidimensionnelle basée sur les splines de Catmull-Rom [1914]

Nous pouvons dorénavant exploiter cette surface comme le montre la figure 3.3.2.2.5. Cette figure reprend l'exemple de la section 2.8. Sur la partie (a), il est impossible de distinguer avec précision les structures secondaires, nous apercevons les hélices α ainsi que quelques brins β mais il est impossible de localiser précisément les feuillets β présents, de les dénombrer ou de décrire leur forme globale. La partie (b) illustre la même protéine avec un rendu tube couplé à notre mode de visualisation des feuillets β utilisant l'interpolation bidimensionnelle basée sur les splines de

Catmull-Rom ; il est aisé de visualiser ces feuillets β, il n'est pas nécessaire d'être un expert du domaine pour constater que cette protéine comporte cinq feuillets β dont les formes sont sensiblement identiques.



(a)



(b)

Figure 3.3.2.2.5 – Intérêt de la représentation des feuillets β dans leur intégralité [2I5B]

Malgré ce résultat probant, un défaut majeur subsiste. En effet le maillage sera systématiquement calculé tout au long de deux brins β consécutifs, alors que dans certains cas de figure deux brins d'un même feuillet peuvent s'éloigner sur une partie de leur longueur et être alors trop éloignés pour former des liaisons hydrogène, comme nous pouvons le voir sur la figure 3.3.2.2.6. Comme expliqué dans la section 2.4, ce sont les liaisons hydrogène qui lient les acides aminés inter-brins les uns aux autres ; si deux brins sont trop éloignés alors les acides aminés ne peuvent pas se lier et il n'y a donc pas de raison de représenter la surface du feuillet à ces endroits. Nous avons alors décidé d'utiliser un autre modèle, celui de Bézier pour représenter ces feuillets.

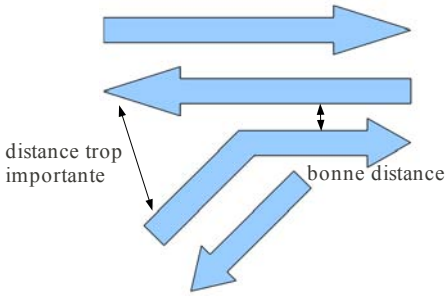


Figure 3.3.2.2.6 – Dans un même feuillet β deux brins consécutifs peuvent ne pas être liés sur toute leur longueur s'ils sont trop éloignés l'un de l'autre

3.3.3 Modèle de Bézier

3.3.3.1 Principe

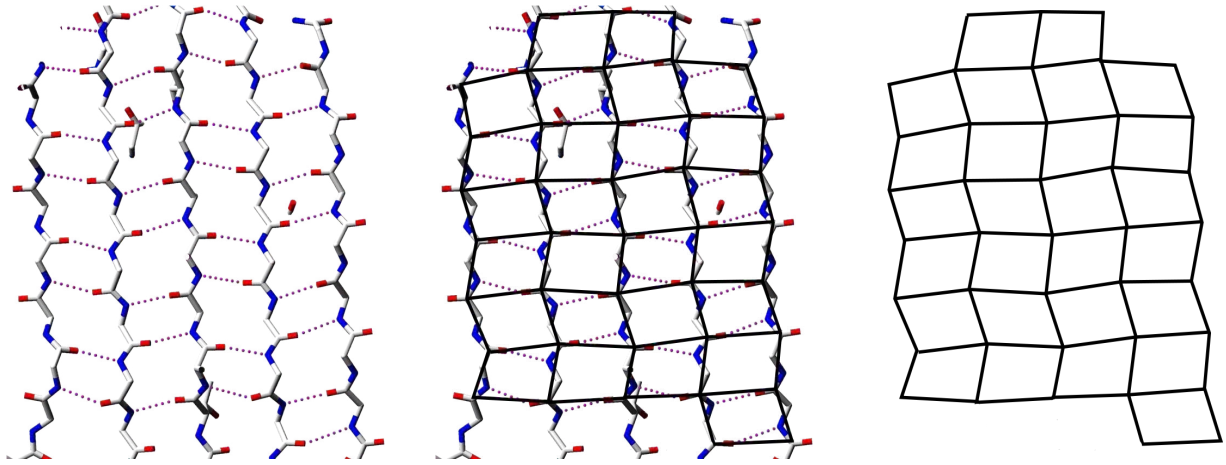
Bien que le modèle créé à partir d'une interpolation bidimensionnelle de splines de Catmull-Rom soit utilisable, il peut, dans certains cas, représenter une surface là où il n'y en a pas. Afin de résoudre ce problème, l'algorithme a été complètement repensé et nous sommes partis sur de nouvelles bases. La donnée essentielle qu'il faut considérer est de savoir si deux résidus inter-brins sont liés ou non, deux résidus étant liés s'il y a présence de liaisons hydrogène comme nous pouvons le voir dans la section 2.4. Ce sont les liaisons hydrogène qui constituent les associations de brins β entre eux : s'il y a liaison hydrogène entre deux acides aminés, nous pouvons alors mailler sans risque de se tromper. Seulement, les coordonnées des liaisons hydrogène n'apparaissent pas obligatoirement dans les fichiers présents dans la PDB. Pour cela, il nous faut passer par l'utilisation d'un algorithme d'affectation de la structure secondaire des protéines (cf. section 2.9.5). BALLView utilise l'algorithme DSSP de Kabsh et Sander [Kabsch1983], qui est également utilisé pour la validation des fichiers PDB. Nous utiliserons l'implémentation présente dans BALLView pour nos calculs. Cet algorithme se base uniquement sur les coordonnées atomiques présentes dans les fichiers PDB, cela signifie que l'algorithme développé utilisera uniquement le champ ATOM des fichiers PDB (cf. section 2.9.4) et non plus les données du champ SHEET. De cette façon, nous nous affranchissons partiellement du format PDB et ce modèle sera non seulement compatible avec l'ensemble des fichiers présents dans la PDB, mais également avec les fichiers créés par des chercheurs n'ayant pas déposés leurs données.

3.3.3.2 Algorithme de prédiction de structures secondaires

Afin que notre algorithme récupère les données nécessaires à son fonctionnement, il faut faire appel à l'algorithme d'attribution de structures secondaires présent dans BALLView et en récupérer les données qui nous intéressent, à savoir les acides aminés liés par des liaisons hydrogène. Seulement les liaisons hydrogène ne sont pas présentes uniquement dans les feuillets β, elles contribuent principalement à la structure tertiaire de la protéine. Étant donné qu'après exécution de l'algorithme nous connaissons les diverses conformations des acides aminés et que nous avons récupéré une liste des liaisons hydrogène, il faut corréler ces informations et lorsque deux acides aminés liés par une liaison hydrogène sont en conformation β, alors ils font partie d'un même feuillet β et leurs brins sont liés.

3.3.3.3 Carreaux de Bézier

Pour créer notre surface nous allons utiliser des carreaux de Bézier, ce sont des surfaces de Bézier (cf. section 2.9.3) qui nécessitent seize points de contrôle pour être bicubiques. Nous allons dans un premier temps définir des quadrilatères de quatre points de contrôle, qui correspondent aux carbones α de quatre acides aminés liés au sein d'un même feuillet β. Un quadrilatère sera composé de deux carbones α successifs en conformation β, appartenant donc à un même brin β, ainsi que des deux carbones α successifs en conformation β, et donc du brin β voisin, liés aux premiers par liaisons hydrogène (Fig. 3.3.3.3.1).



(a) (b) (c)

Figure 3.3.3.3.1 – Illustration de la définition des quadrilatères qui serviront à calculer les carreaux de Bézier. (a) Brins β adjacents d'un feuillet parallèle ; (b) quadrilatères reliant les Cα en conformation β quatre par quatre ; (c) représentation des quadrilatères obtenus

3.3.3.4 Calcul et uniformisation des normales

De la même façon que pour les algorithmes développés précédemment, il faut s'assurer de pouvoir calculer correctement les normales nécessaires au bon éclairage de nos surfaces. Un quadrilatère peut être décomposé en deux triangles de deux façons différentes et il faut que les triangles de tous les quadrilatères formés aient des sommets qui respectent le même ordre. Mais, au moment de la création des quadrilatères, il n'y a aucun moyen de savoir s'ils sont orientés dans le même sens, car les numéros des résidus (leurs numéros dans la chaîne polypeptidique) ne reflètent pas leur position dans le feuillet comme nous pouvons le voir sur la figure 3.3.1.1. Afin que toutes les normales soient orientées de la même manière nous utilisons un algorithme de propagation tel qu'il est illustré sur la figure 3.3.3.4.1. Le principe de cet algorithme est de considérer un quadrilatère qui va nous servir de référence pour l'orientation de la normale : sur notre figure, en (a), nous avons une représentation d'une surface composée de plusieurs quadrilatères ; pour chaque quadrilatère, le vecteur normal est représenté sous la forme d'une flèche. Nous constatons que ces vecteurs ne sont pas tous orientés dans le même sens. Nous définissons un quadrilatère de référence en (b), c'est son orientation qui définira celle de tous les quadrilatères composants cette surface. Dans ce but, nous allons propager l'orientation de référence sur les quadrilatères voisins par 4-connexité, un quadrilatère ayant au maximum quatre voisins puisqu'il ne possède que quatre arêtes. Sur l'illustration (b), les flèches rouges montrent les voisins du quadrilatère de référence, ceux qui seront traités en premier par notre algorithme. À l'aide d'un simple produit scalaire nous allons déterminer si le vecteur normal de référence est colinéaire avec chacun de ses voisins. Si tel n'est pas le cas alors nous réordonnons les sommets du quadrilatère de façon à ce que son orientation concorde avec celle du quadrilatère de référence. Ainsi, si le vecteur normal de Q_1 n'est pas colinéaire avec le vecteur normal de référence nous passons de $Q_1(P_1, P_2, P_3, P_4)$ à $Q_1(P_4, P_3, P_2, P_1)$; de cette façon le vecteur normal sera orienté dans l'autre direction comme nous pouvons le voir sur la figure en (c).

Une fois les quadrilatères voisins traités, ils ne doivent plus être modifiables par notre algorithme pour cela nous leur donnons une propriété indiquant qu'ils ont été vérifiés et qu'il ne faut pas les traiter à nouveau. Nous pouvons alors continuer l'algorithme en recherchant les voisins des quadrilatères venant d'être traités : une fois les voisins identifiés nous pouvons à nouveau faire nos calculs de produit scalaire, et ce jusqu'à ce qu'il n'y ait plus de quadrilatère n'ayant pas été vérifié. Nous obtenons le résultat présenté en (g), l'ensemble des vecteur normaux de la surface sont orientés dans le même sens. Les vecteurs normaux des sommets d'un quadrilatère sont obtenus en

calculant les normales des quatre triangles le composant. Au sein d'un même quadrilatère le vecteur normal d'un sommet est donc la moyenne des vecteurs normaux des trois triangles possédant ce sommet. De plus, un même sommet peut participer à quatre quadrilatères, il faut également prendre en compte les vecteurs normaux de ces quadrilatères.

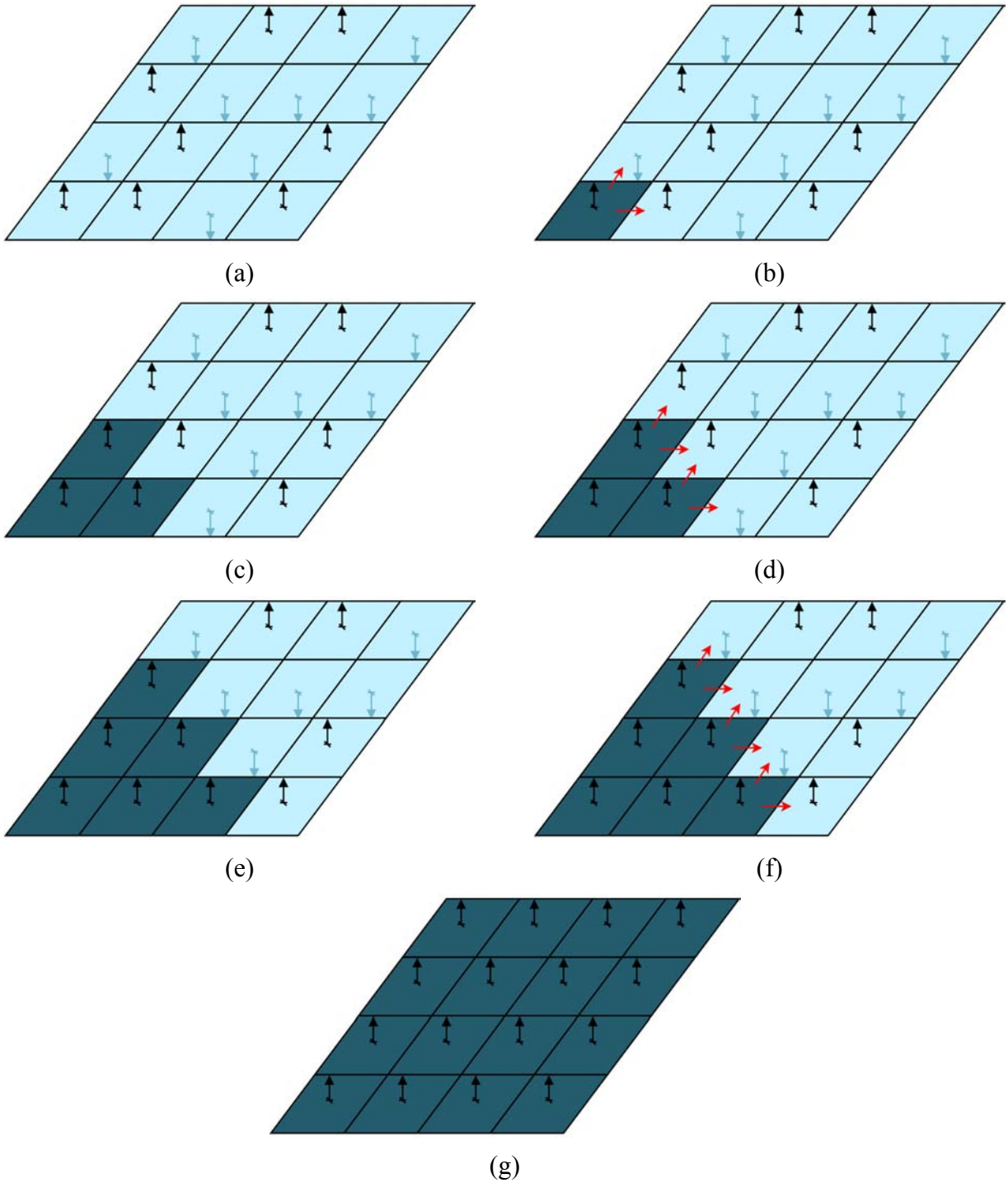


Figure 3.3.3.4.1 – Étapes de l'algorithme de propagation utilisé pour l'uniformisation des normales

3.3.3.5 Calcul des points de contrôle

Une fois l'ensemble des normales unifiées, avant de pouvoir calculer nos surfaces de Bézier il nous faut calculer les points de contrôle manquants. Comme il a été spécifié plus haut il faut seize points de contrôle pour calculer une surface de Bézier bicubique, hors nous n'en disposons que de quatre (cf. figure 3.3.3.5.1).

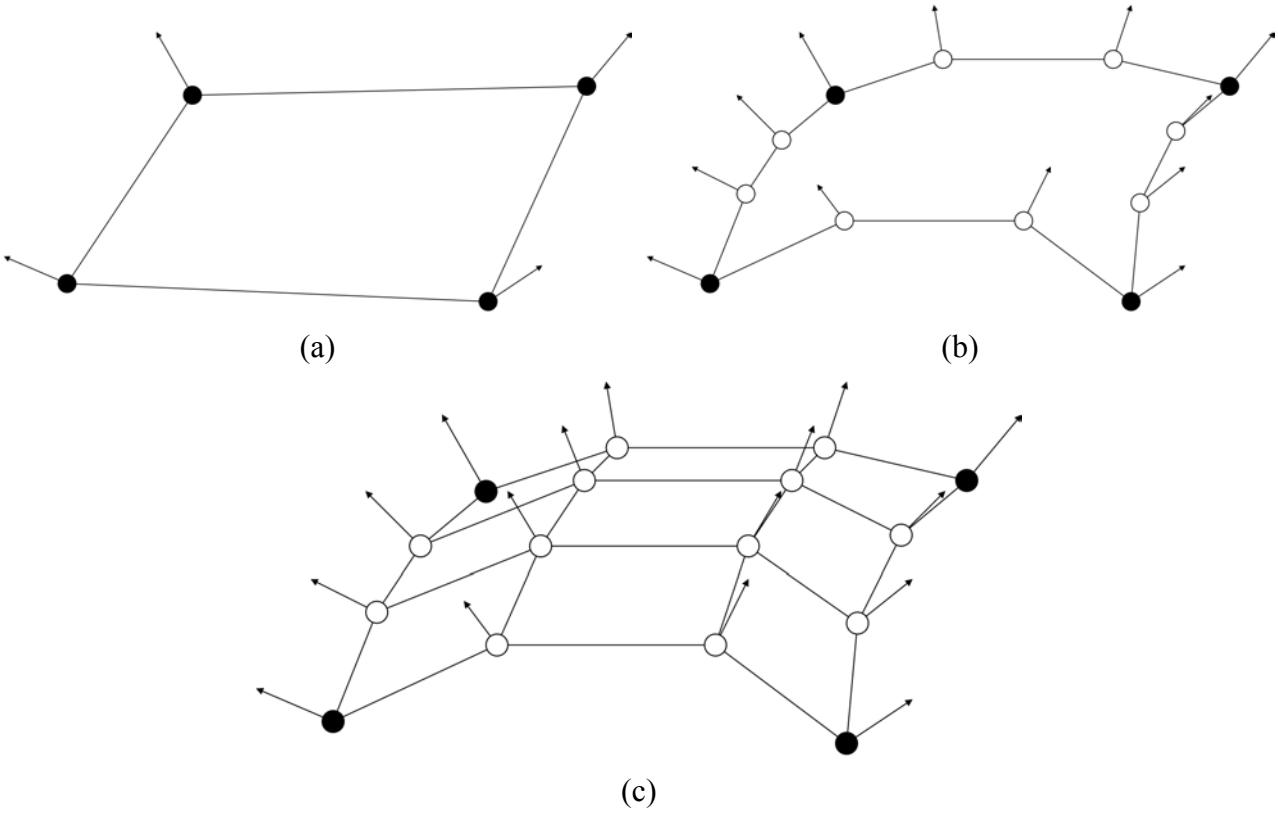


Figure 3.3.3.5.1 – Étapes de calcul des seize points de contrôle à partir des quatre du départ

Il faut donc construire les points manquants à partir des informations dont nous disposons. Pour un même quadrilatère, nous connaissons les positions de ses points dans l'espace ainsi que des vecteurs normaux à chaque sommet. Pour calculer les points de contrôle nécessaires, nous allons utiliser les vecteurs tangents aux normales, et nous allons commencer par calculer l'ensemble des points extérieurs pour obtenir le résultat visible sur la figure 3.3.3.5.1b. Pour calculer le vecteur tangent à la normale de l'un des points de contrôle, nous effectuons deux produits vectoriels successifs. Sur la figure 3.3.3.5.2, nous ne représentons qu'une des arêtes d'un quadrilatère, ainsi comme nous pouvons le voir en (a) il n'y a que deux points de contrôle, $\vec{u}$ correspond au vecteur normal du point et $\vec{v}$ correspond au vecteur qui va du point P_1 au point P_2. Le vecteur $\vec{w}$ est le résultat du produit vectoriel : $\vec{u} \wedge \vec{v}$. Une fois ce vecteur obtenu, nous calculons son produit vectoriel avec la normale afin d'obtenir le vecteur tangent à la normale que nous pouvons observer

en (b). C'est sur ce vecteur tangent que nous allons positionner notre point de contrôle. Étant donné qu'il faut deux points supplémentaires par arête, nous positionnons le premier sur le vecteur tangent à un tiers de la distance séparant les deux points extrêmes comme nous pouvons le voir en (c).

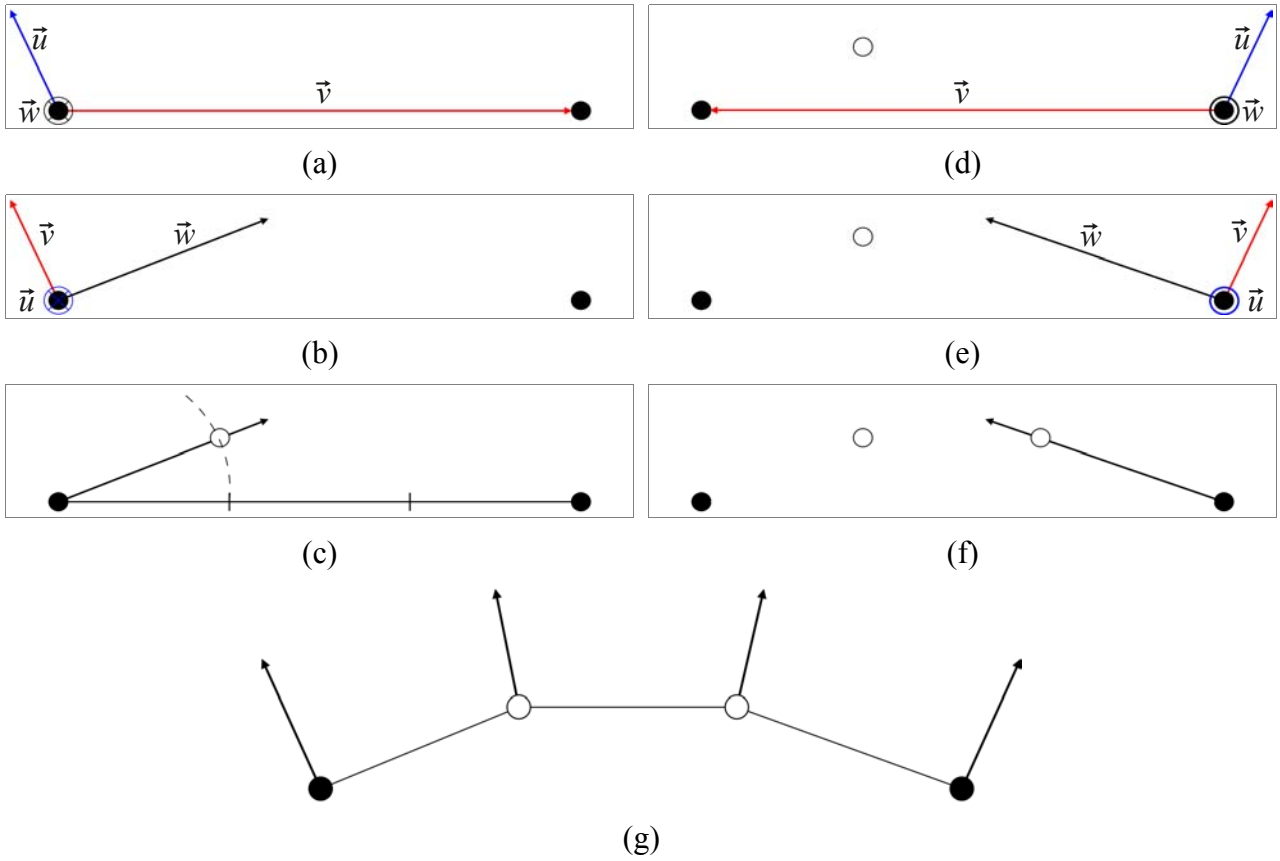


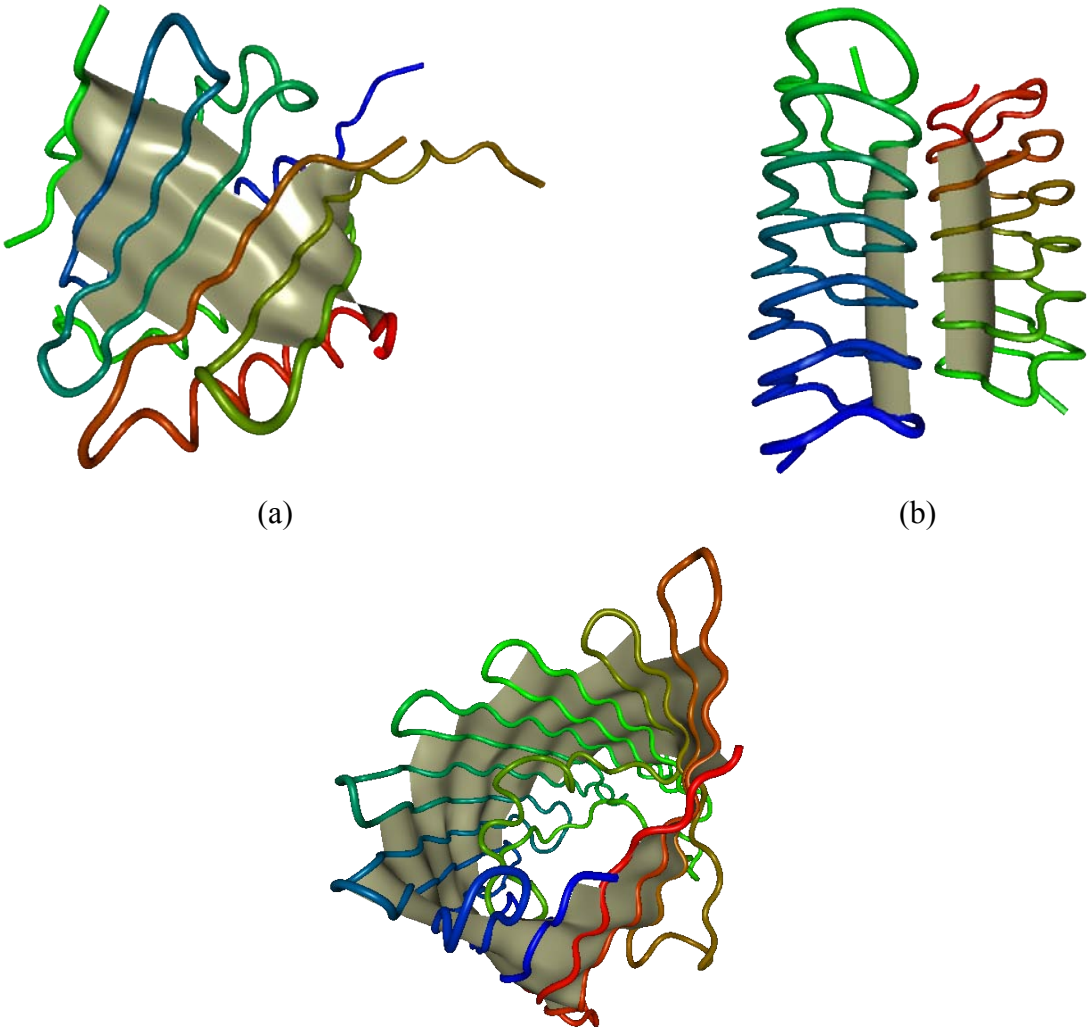
Figure 3.3.3.5.2 – Calcul des points de contrôle manquants sur une arête du quadrilatère

Maintenant que nous avons calculé le premier point de contrôle manquant, nous calculons le second et pour cela nous changeons les vecteurs de calcul, comme nous pouvons le voir sur la figure en (d). Nous remarquons cependant que le produit vectoriel $\vec{u} \wedge \vec{v}$ produit un vecteur $\vec{w}$ dont l'orientation est à l'opposé du vecteur $\vec{w}$ calculé en (a), car les vecteurs de calculs ne sont pas dans la même direction.

Une fois nos points calculés, nous obtenons le résultat visible en (g) : les normales des nouveaux points sont calculées par une simple interpolation des normales extrêmes, ainsi la normale d'un point nouveau représente les deux tiers de la normale du point extrême le plus proche, et un tiers du point extrême le plus éloigné. Le résultat obtenu sur l'ensemble d'un quadrilatère est représenté sur la figure 3.3.3.5.1b. Il ne manque plus que les points internes qui sont obtenus en nous basant sur les points de contrôle nouvellement calculés, et nous obtenons les seize points de contrôles nécessaires comme nous pouvons le constater sur la figure 3.3.3.5.1c.

3.3.3.6 Calcul de la surface de Bézier

Nous disposons de l'ensemble des points de contrôle nécessaires aux calculs de la surface de Bézier. Chaque carreau de Bézier est calculé séparément et étant donné le calcul des points de contrôle fait à partir des normales aux sommets, l'ensemble des carreaux de Bézier va former une surface continue. Nous réalisons le calcul tel qu'il a été énoncé dans la section 2.9.3 et nous obtenons les résultats présentés dans la figure 3.3.3.6.1.



(a) (b)

(c)

Figure 3.3.3.6.1 – Résultats obtenus avec les carreaux de Bézier, ces résultats sont couplés avec le mode de visualisation squelette. (a) [1914], (b) [1EZG] et (c) [1PRN]

Les résultats présentés sur ces figures montrent que les surfaces générées épousent parfaitement le squelette de la protéine, et que nous ne constatons pas de discontinuité entre les carreaux de Bézier. En utilisant ce mode de visualisation, nous sommes certains de ne calculer notre maillage qu'aux endroits nécessaires.

Ces modèles, de Catmull-Rom et de Bézier, évoquent un tapis volant de par leur aspect plissé. C'est pourquoi nous proposons de les baptiser SheHeRASADe pour « *Sheets Helper for RepresentAtion of SurfAce Descriptors* ».

3.4 Représentations

Une fois ces modèles de représentation des feuillets β développés, il était important de penser des outils autour de ces modèles, afin que ces derniers soient pleinement exploitables. Nous nous sommes donc intéressés à l'intégration et à la représentation d'informations importantes concernant les feuillets β.

3.4.1 Intégration dans BALLView

L'interface de BALLView a été modifiée afin d'accueillir les modes de représentation des feuillets β qui ont été développés. Deux items ont été ajoutés dans la liste déroulante des modes de visualisation disponibles : « *Catmull-Rom Beta Sheet* » pour le modèle utilisant les splines de Catmull-Rom et « *Bezier Beta Sheet* » pour le modèle utilisant les carreaux de Bézier. L'utilisation de ces modes est donc complètement transparente pour l'utilisateur du logiciel. Lorsque nous accèdons aux options de ces modèles, la fenêtre qui gère les paramètres est composées d'autant d'onglets qu'il y a de feuillets β dans la protéine que représentée. Chacun des onglets porte le nom d'un feuillet β, et pour chacun des feuillets nous connaissons sa surface en Å². De plus, un bouton présent sur chaque onglet permet de choisir si nous désirons représenter ou non le feuillet concerné. Ainsi sur des systèmes très complexes, présentant de nombreux feuillets, il est possible de choisir spécifiquement les feuillets que nous souhaitons.

3.4.2 Textures

Afin de représenter des informations importantes directement sur la surfaces des feuillets β, nous avons opté pour l'utilisation de textures symboliques. L'avantage de nos modèles par rapport au mode *cartoon* est que nous représentons les feuillets β dans leur globalité, nous donnant ainsi accès à leur forme globale et facilitant leur visualisation. Cependant le mode *cartoon* offre une information de première importance qui est le sens des brins β. Il est en effet difficile d'appréhender le sens de ces brins uniquement à l'aide des surfaces que nous générons. C'est pourquoi nous avons choisi de représenter le sens des brins sous forme de textures, appliquées sur la surface.

Afin de pouvoir représenter des textures sur les surfaces de nos feuillets, il faut calculer les coordonnées de textures, c'est à dire quelles coordonnées de l'image de la texture en deux

dimensions (coordonnées en x et y) seront attachées à quel point de la surface (coordonnées en x, y et z). Ces calculs de coordonnées de texture sont sensiblement différents pour nos deux modèles, c'est pourquoi ces cas seront traités séparément.

3.4.2.1 Modèle de Catmull-Rom

La texture utilisée pour ce modèle est l'image d'une flèche (cf. Fig 3.4.2.1.1) qui servira à la fois à représenter localement le sens du brin, mais également à matérialiser chaque acide aminé. Ainsi, à l'emplacement des brins présents dans notre surface, nous allons représenter les acides aminés sous forme de flèches orientées dans le sens du brin.

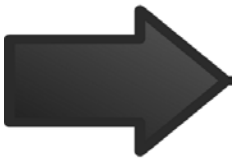
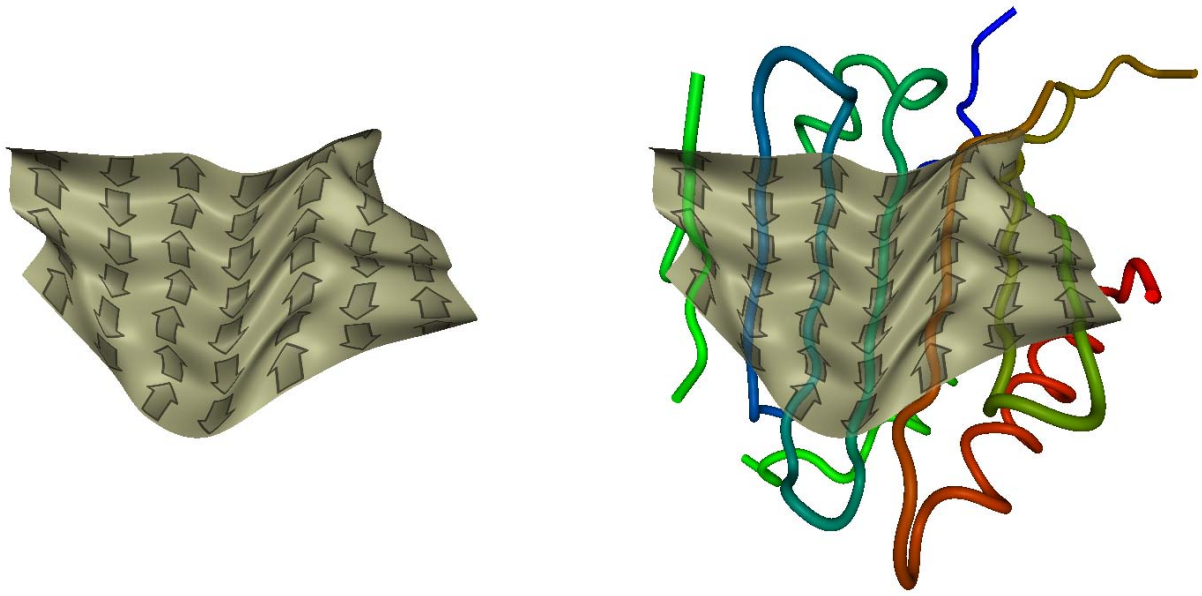


Figure 3.4.2.1.1 – Image représentant une flèche utilisée pour texturer la surface de Catmull-Rom

Dans le cas du modèle de Catmull-Rom, nous disposons d'un maillage global qui représente l'ensemble d'un feuillet β, pour appliquer notre texture il nous faut connaître le nombre de points qui représentent un acide aminé. Ce calcul est facilité par le fait que l'échantillonnage de notre surface est régulier, de cette façon chaque acide aminé est représenté par le même nombre de points. Pour calculer les coordonnées des textures que nous allons appliquer, il nous faut connaître le nombre de brins β présents, le nombre d'acides aminés que contiennent chacun de ces brins β et enfin le nombre de points interpolés. Le nombre de points interpolés sur chaque spline est calculé en fonction du nombre de brins présents dans le feuillet β, ce nombre est toujours impair de cette façon il y a toujours un nombre impair de points entre deux brins, cela rend le calcul des coordonnées de texture plus simple. Pour connaître le nombre de points utilisés sur l'axe *x* pour représenter un acide aminé, il faut diviser le nombre de points interpolés pour un brin, par le nombre d'acides aminés présents. Et en *y*, nous divisons le nombre de points interpolés le long du feuillet par le nombre de brins présents. Nous connaissons alors les points sur lesquels nous allons plaquer notre texture pour représenter un acide aminé. De plus, il nous faut connaître le sens des brins pour savoir dans quel sens plaquer la texture. Une fois ces calculs achevés nous obtenons les résultats visibles sur la figure 3.4.2.1.2.

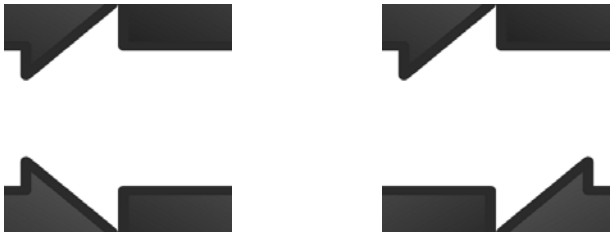


(a) (b)

Figure 3.4.2.1.2 – Résultat obtenu sur le modèle de Catmull-Rom en texturant la surface du feuillet β de la protéine [1914]. (a) représente le feuillet texturé par des flèches, chaque acide aminé est représenté par une flèche et les flèches pointent dans la direction des brins. (b) représente le même feuillet couplé à un rendu de type squelette

3.4.2.2 Modèle de Bézier

Pour le modèle de Bézier le plaquage de texture est très différent, chaque carreau de Bézier va être traité séparément. Pour ce modèle, chacun des coins d'un carreau représente le carbone α d'un acide aminé, chaque carreau représente donc un quart de quatre acides aminés différents. L'image de notre texture doit alors représenter un quart de quatre flèches. De plus, il y a deux arrangements possibles pour ces flèches, car les carreau représentent deux acides aminés consécutifs d'un même brin en relation avec deux acides aminés consécutifs du brin voisin et comme nous l'avons vu dans la section 2.4 il existe deux arrangements pour les brins : ils sont soit parallèles, soit antiparallèles. Nous devons donc utiliser deux textures différentes, une pour les brins parallèles et une autre pour les brins antiparallèles comme nous pouvons le voir sur la figure 3.4.2.2.1.

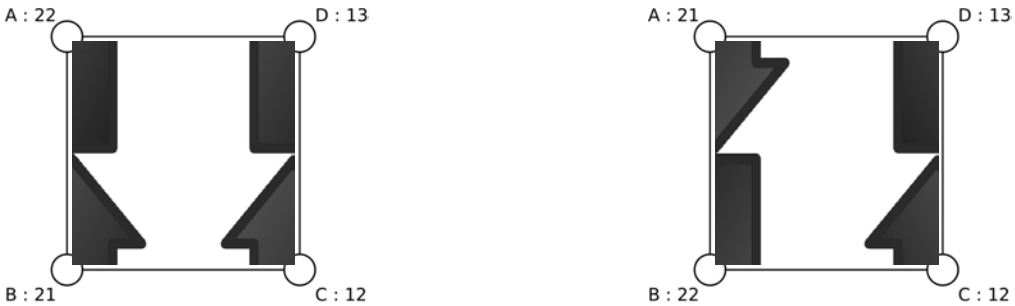


(a) (b)

Figure 3.4.2.2.1 – Textures utilisées pour le modèle de Bézier, (a) dans le cas parallèle, et (b) dans le cas antiparallèle

Nous constatons sur cette figure que l'image (a) représente quatre morceaux de flèches dont l'association formera des flèches parallèles les unes aux autres, et que l'image (b) représente des morceaux de flèches dont l'association formera des flèches antiparallèles les unes aux autres.

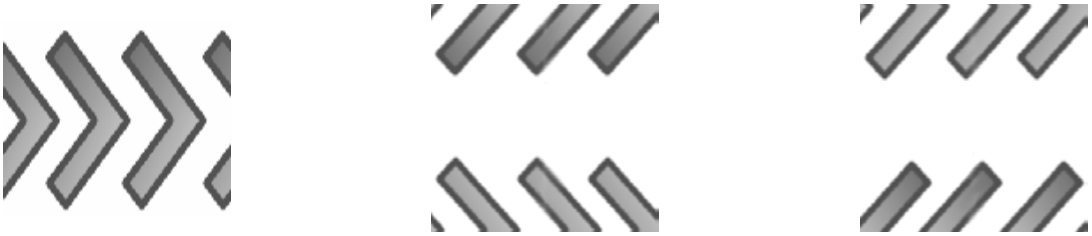
Pour savoir quelle texture appliquer à quel carreau il faut consulter les numéros portés par les acides aminés le composant. Considérons un carreau dont les coins sont (A, B, C, D), A et B étant les acides aminés consécutifs du premier brin et C et D ceux du brins voisin. Si le numéro de A et supérieur à celui de B et que le numéro de C est supérieur à celui de A, alors nous appliquons la texture parallèle avec pour origine de l'image le point B, cela permet d'orienter la texture dans le sens des brins. Par contre si le numéro de A est inférieur à celui de B, et que celui de D est supérieur à celui de C, alors nous appliquons la texture antiparallèle avec pour origine le point A. Cet exemple est repris dans la figure 3.4.2.2.2.



(a) (b)

Figure 3.4.2.2.2 – Illustration du choix de la texture et du choix du point d'origine en fonction des numéros des acides aminés

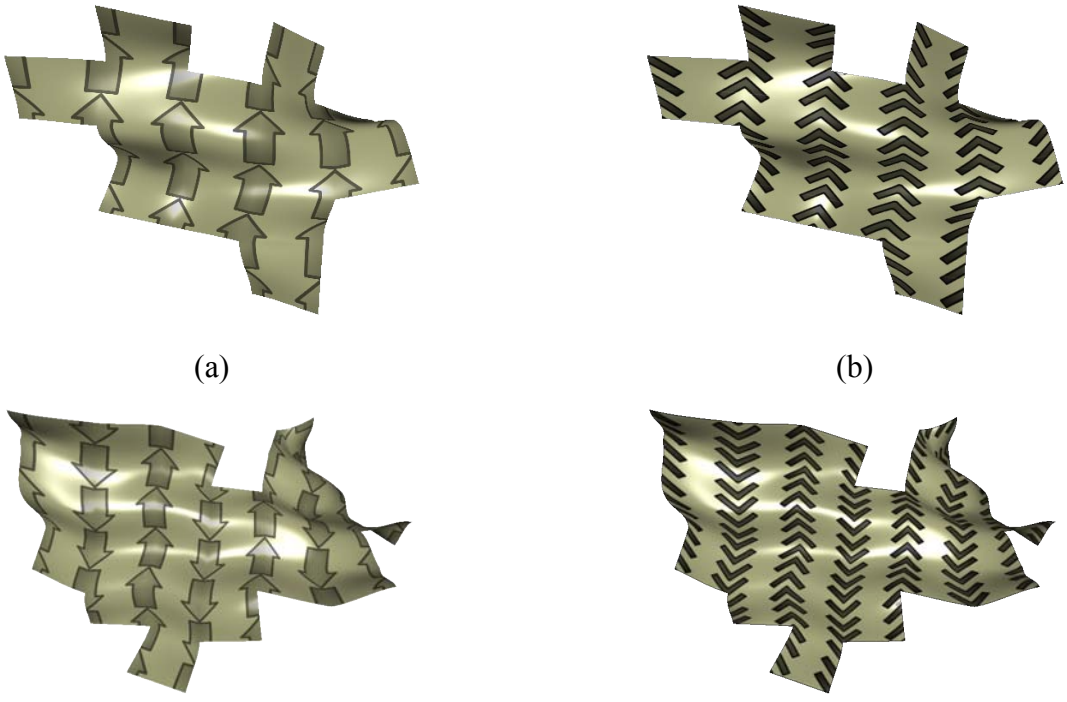
Nous utilisons également des textures représentant des chevrons, afin de représenter de manière très claire le sens des brins β. Le principe du plaquage de la texture est identique, à celui utilisé pour les textures représentant des flèches. La figure 3.4.2.2.3 illustre les images utilisées dans ce cas.



(a) (b) (c)

Figure 3.4.2.2.3 – (a) Texture représentant des chevrons, utilisée pour le modèle de Catmull-Rom ; (b) et (c) textures représentant des chevrons, utilisée pour le modèle de Bézier, dans le cas parallèle en (b) et antiparallèle en (c)

La figure 3.4.2.2.4 présente des résultats de surfaces obtenues avec le modèle de Bézier texturé. Nous constatons sur ces exemples que l'information représentée est très claire : le sens de chaque brin apparaît de façon évidente. Nous pouvons en déduire que le feuillet de la protéine [1E20] est parallèle, alors que celui de la protéine [1914] est antiparallèle. En utilisant la texture en flèche, deux informations importantes sont directement visibles sur la surface : le sens des brins et la position des acides aminés au sein des brins, chaque acide aminé est matérialisé par une flèche.

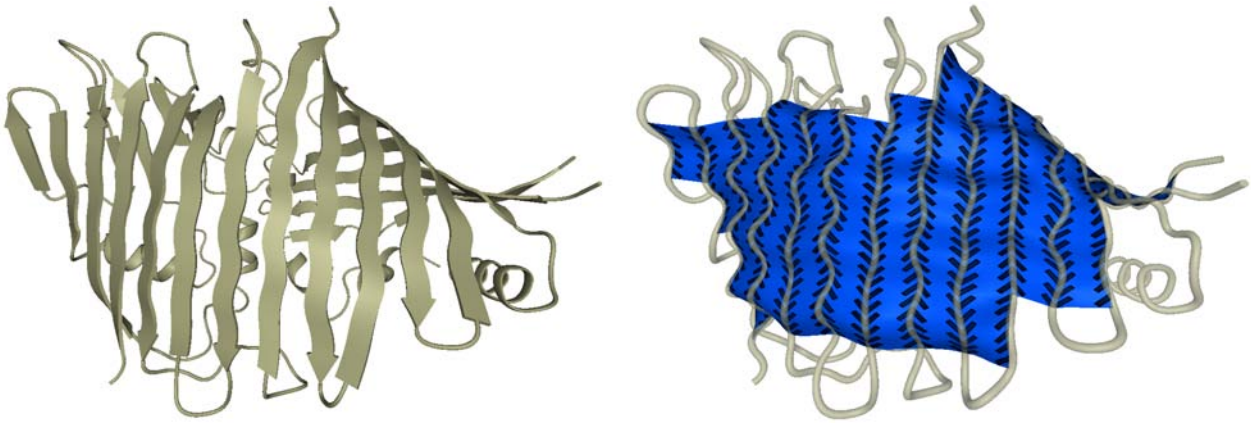


(a) (b)

(c) (d)

Figure 3.4.2.2.4 – Résultats du modèle de Bézier texturé. (a) et (b) représentent le feuillet β parallèle de la protéine [1E20]. (c) et (d) représentent le feuillet β antiparallèle de la protéine [1914]. Nous avons utilisé des textures en flèches, et en chevrons.

Parmi les modes classiques de visualisation le seul représentant le sens des brins est le mode *cartoon*, si nous comparons ce dernier avec notre modèle de Bézier texturé, force est de constater que notre représentation est la plus claire, comme nous pouvons le voir sur la figure 3.4.2.2.5. En effet, si nous souhaitons observer le sens d'un brin β sur un modèle *cartoon* il faut se référer à une extrémité du brin ce qui n'est pas toujours aisé si le brin est de longueur importante. Sur notre modèle le sens du brin est répété tout au long de ce dernier ce qui rend l'information plus accessible.



(a) (b)

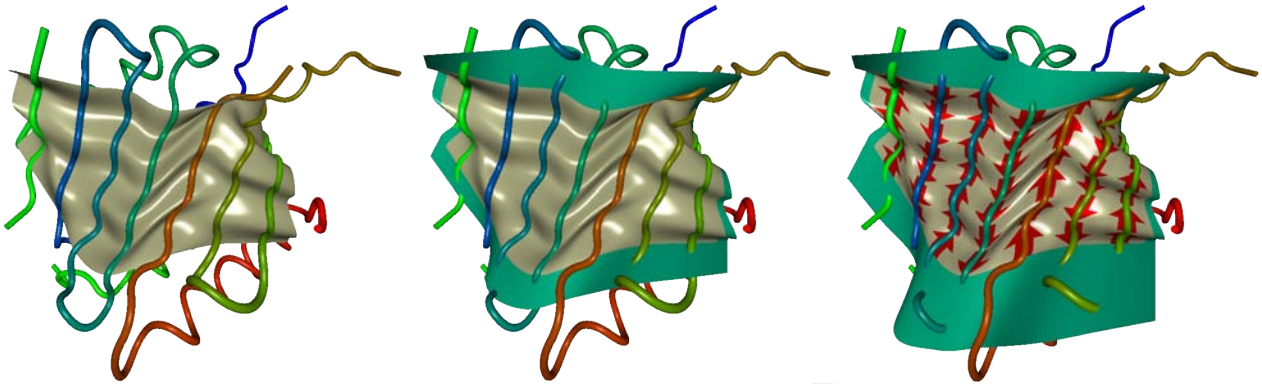
Figure 3.4.2.2.5 – Représentation du domaine 4bclA00 de la classe tout β de CATH. (a) Représentation de type cartoon, (b) représentation de Bézier texturée par des chevrons et couplée à un rendu de type squelette transparent à 35 %

Que ce soit sur le modèle de Catmull-Rom ou de Bézier, le plaquage de texture permet de représenter des informations importantes telles que le sens des brins et la position des acides aminés directement sur les surfaces que nous générons. Ces représentations sont très intuitives et sont utiles aussi bien dans la recherche que dans l'enseignement, l'utilisateur n'ayant pas de connaissance approfondie des feuillets β comprendra aisément ces informations et cela lui permettra de mieux appréhender ce type de structure secondaire.

3.4.3 Extension des feuillets β

De manière générale, nous observons qu'un feuillet β est rarement seul au sein d'une protéine : il est souvent accompagné d'autres structures secondaires, que ce soient d'autres feuillets β ou bien d'hélices α. Nous pouvons alors considérer que ces structures se stabilisent mutuellement, et qu'ils ne s'agit pas simplement de structures secondaires les unes à côté des autres, mais de superstructures secondaires. Afin de tenter l'observation de l'influence des feuillets β sur ces superstructures, nous avons souhaité être capable de les étendre dans une direction voulue.

Afin d'étendre nos représentations de feuillets β, nous avons créé des points symétriques à la surface dans la direction souhaitée. Dans la fenêtre de préférence du modèle, il y a possibilité de choisir une ou plusieurs directions et de les étendre comme nous le souhaitons. Nous obtenons les résultats montrés par la figure 3.4.3.1. L'extension dans une direction donnée, n'est calculée que par rapport aux derniers points existants dans cette direction. Cela peut entraîner des représentations bizarres, dues au repliement, possiblement important, du feuillet. Nous n'utilisons pas l'équation du plan du feuillet car lorsque celui-ci se replie sur lui même, l'équation de son plan est inutilisable.



(a) (b) (c)

Figure 3.4.3.1 – Exemples de l'extension d'un feuillet β. (a) représente le feuillet d'origine, sans extension, (b) représente le même feuillet avec une extension différente de chaque côté, et (c) représente ce feuillet avec une extension plus importante sur le côté inférieur. Sur l'ensemble de ces exemples, le modèle est couplé à une représentation squelette, en (c) la surface est texturée [1914]

3.4.4 Modes de coloration

Les modes de coloration classiques, implémentés dans BALLView, sont bien sûr utilisables avec nos modèles. Cependant, nous avons voulu développer quelques modes de coloration, supplémentaires, intéressants à coupler avec nos modèles.

3.4.4.1 Mode personnalisable

Dans ce mode, nous pouvons attribuer la couleur que nous souhaitons à un acide aminé. Ce mode existe également dans tous les logiciels de modélisation moléculaire. Il est souvent long à paramétrer si nous voulons représenter plusieurs acides aminés de la même couleur, car le choix est effectué séparément pour chaque acide aminé. Le but de ce mode de coloration est de rendre beaucoup plus simple et rapide l'utilisation de la coloration par nom d'acide aminé. Pour ce faire, nous avons développé une interface présentant une liste de tous les acides aminés, en face de

laquelle se trouvent six colonnes et à chaque colonne correspond une couleur éditable selon le désir de l'utilisateur. Ces colonnes sont composées de boutons que l'utilisateur sélectionne s'il souhaite que tel acide aminé soit représenté de telle couleur. Cette interface est représentée dans la figure 3.4.4.1.1a. Il est bien plus aisé et rapide de paramétrer ce mode pour colorer des groupes d'acides aminés que de paramétrer le mode classique qui nécessite de donner à chaque acide aminé une couleur de façon individuelle.

À ce mode a été ajouté une liste déroulante, comme nous pouvons le constater sur la figure 3.4.4.1.1a, qui contient des configurations prédéfinies. Ces préférences sont issues du diagramme de Venn de la nature des acides aminés (cf. Fig 2.1.3), elles sont au nombre sept :

- Small_o : représente les petits acides aminés, et les autres,
- Polar_apolar : représente les acides aminés polaires, et les apolaires,
- Polar_N_I : représente les acides aminés polaires neutres, les acides aminés polaires chargés et les autres,
- Polar_pos_neg : représente les acides aminés polaires chargés positivement, chargés négativement et les autres,
- Polar_N_pos_neg : représente les acides aminés polaires neutres, chargés positivement, négativement et les autres,
- Hydrophobic_o : représente les acides aminés hydrophobes et les autres,
- Aromatic_o : représente les acides aminés aromatiques et les autres.

Quelques exemples de ces préférences sont donnés dans la figure 3.4.4.1.1. Lorsque nous sélectionnons une préférence, les couleurs des acides aminés sont changées de manière automatique.

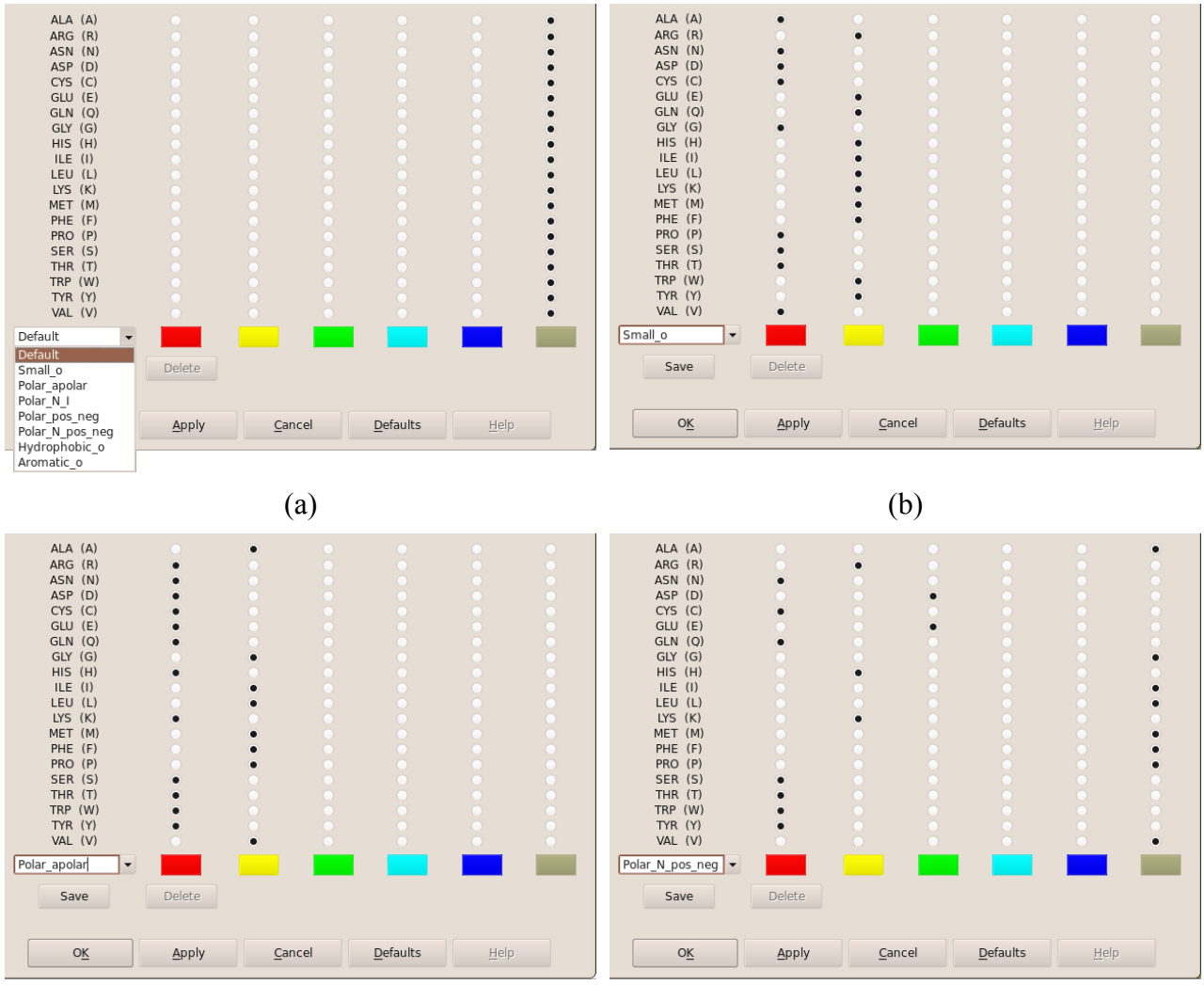


Figure 3.4.4.1.1 – Fenêtre de configuration de notre mode de coloration personnalisable

Comme il a été déclaré plus haut, il est possible de créer ce que nous pouvons appeler des classes de couleur personnalisées. Une fois nos classes créées, il est possible de leur donner un nom et de les sauvegarder de manière à pouvoir les réutiliser sans avoir à les redéfinir à chaque utilisation. La liste déroulante est éditable de façon à pouvoir donner un nom à nos classes, et un bouton « *Save* » est présent. La sauvegarde se fait dans un fichier XML contenant l'ensemble des classes que utilisables. L'utilisation d'un fichier XML permet à un utilisateur souhaitant faire profiter de ses classes de couleur à un autre utilisateur de lui envoyer ce fichier, et qu'il puisse l'utiliser et le modifier à son tour. Il suffit d'intégrer ce fichier au répertoire contenant l'exécutable de BALLView. La figure 3.4.4.1.2 présente un exemple de ce fichier XML, les différentes classes de couleurs sont définies dans la balise « *ColorModels* », chaque balise « *model* » a son propre nom et

pour chaque acide aminé, représenté par son code à une lettre, correspond le numéro de la colonne qui lui est associée, et donc sa couleur.

```
<!DOCTYPE predefColorModels>
<ColorModels>
 <model name="Default"          R="5" P="5" V="5" Q="5" C="5" S="5" A="5" E="5" F="5" G="5"
                                T="5" K="5" N="5" D="5" I="5" W="5" M="5" Y="5" L="5" H="5" />
 <model name="Small_o"          R="1" P="0" V="0" Q="1" C="0" S="0" A="0" E="1" F="1" G="0"
                                T="0" K="1" N="0" D="0" I="1" W="1" M="1" Y="1" L="1" H="1" />
 <model name="Polar_apolar"     R="0" P="1" V="1" Q="0" C="0" S="0" A="1" E="0" F="1" G="1"
                                T="0" K="0" N="0" D="0" I="1" W="0" M="1" Y="0" L="1" H="0" />
 <model name="Polar_N_I"        R="1" P="5" V="5" Q="0" C="0" S="0" A="5" E="1" F="5" G="5"
                                T="0" K="1" N="0" D="1" I="5" W="0" M="5" Y="0" L="5" H="1" />
 <model name="Polar_pos_neg"    R="0" P="5" V="5" Q="5" C="5" S="5" A="5" E="1" F="5" G="5"
                                T="5" K="0" N="5" D="1" I="5" W="5" M="5" Y="5" L="5" H="0" />
 <model name="Polar_N_pos_neg" R="1" P="5" V="5" Q="0" C="0" S="0" A="5" E="2" F="5" G="5"
                                T="0" K="1" N="0" D="2" I="5" W="0" M="5" Y="0" L="5" H="1" />
 <model name="Hydrophobic_o"    R="1" P="1" V="0" Q="1" C="0" S="1" A="0" E="1" F="0" G="0"
                                T="0" K="0" N="1" D="1" I="0" W="0" M="0" Y="0" L="0" H="0" />
 <model name="Aromatic_o"       R="1" P="1" V="1" Q="1" C="1" S="1" A="1" E="1" F="0" G="1"
                                T="1" K="1" N="1" D="1" I="1" W="0" M="1" Y="0" L="1" H="0" />
</ColorModels>
```

Figure 3.4.4.1.2 – Exemple du contenu du fichier XML de préférences de coloration

3.4.4.2 Coloration de type « Hydrophobic Cluster Analysis » *- HCA*

La méthode HCA (*Hydrophobic Cluster Analysis*) [Gaboriaud1987] permet de montrer la présence d'amas locaux d'acides aminés hydrophobes associés aux faces internes des structures secondaires tels que les feuillets β. Cette méthode prend en compte la proximité spatiale des acides aminés distants dans la séquence protéique, il est alors possible de souligner la présence d'amas hydrophobes qui correspondent majoritairement aux faces internes des structures secondaires. Cette méthode peut également être utilisée pour prédire la structure secondaire, les amas hydrophobes étant souvent caractéristiques d'un type de repliement pour les protéines globulaires.

Pour notre mode de coloration, nous n'allons pas analyser la structure tertiaire de nos protéines, mais uniquement colorer les acides aminés suivant la méthode HCA. La méthode HCA repose sur une dichotomie simple entre les acides aminés hydrophobes et les autres. Les hydrophobes sont plus généralement constitutifs du cœur de la protéine, tandis que les hydrophiles et les neutres sont maintenus à la surface de la protéine, au contact du solvant. Les résidus hydrophobes utilisés par HCA pour constituer les amas hydrophobes sont les sept acides aminés : valine, isoleucine, leucine, phénylalanine, méthionine, tyrosine et le tryptophane. Ce sont les sept acides aminés qui se caractérisent par une propension plus importante à former des structures secondaires de type hélice α ou feuillet β plutôt que des « *coils* », qui sont des structures non périodiques localisées sur trois ou quatre acides aminés consécutifs [Callebaut1997]. Il faut également considérer que dans certains cas la cystéine, l'alanine ou la thréonine peuvent intégrer ce

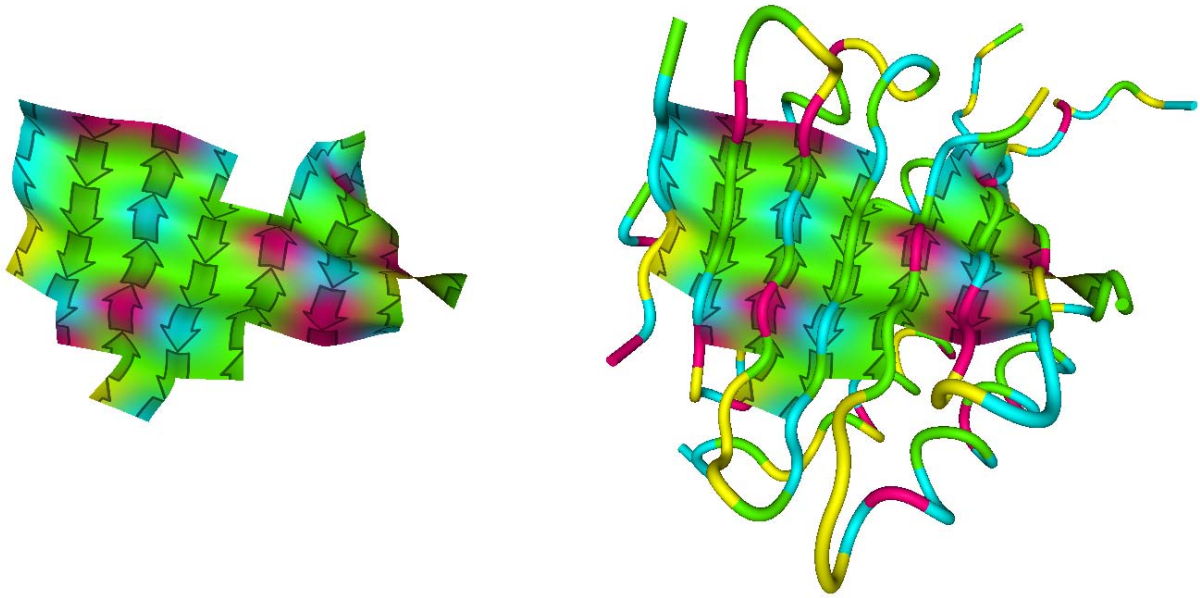
groupe d'acides aminés. Après avoir défini un alphabet hydrophobe, nous définissons un alphabet *coil* composé par des résidus ayant une propension nettement plus importante à former des *coils* que des hélices α ou des feuillets β. Cet alphabet est composé de la proline, la glycine, l'acide aspartique, l'asparagine et la sérine [Callebaut1997].

Pour notre mode de coloration, les acides aminés de ces alphabets seront représentés avec chacun une couleur distincte que l'utilisateur sera libre de modifier à sa convenance. Il y aura quatre couleurs différentes comme nous pouvons le constater dans le tableau 3.4.4.2.1, une couleur pour l'alphabet hydrophobe, une pour l'alphabet *coil*, une pour les acides aminés pouvant intégrer l'alphabet hydrophobe et une couleur pour les acides aminés n'appartenant à aucune section.

Acides aminés	**Alphabet hydrophobe**	**Alphabet *coil***	**Hydrophobes occasionnels**	**Autres**
Alanine (A)			×	
Arginine (R)				×
Asparagine (N)		×		
Acide aspartique (D)		×		
Cystéine (C)			×	
Acide glutamique (E)				×
Glutamine (Q)				×
Glycine (G)		×		
Histidine (H)				×
Isoleucine (I)	×			
Leucine (L)	×			
Lysine (K)				×
Méthionine (M)	×			
Phénylalanine (F)	×			
Proline (P)		×		
Sérine (S)		×		
Thréonine (T)			×	
Tryptophane (W)	×			
Tyrosine (Y)	×			
Valine (V)	×			

Tableau 3.4.4.2.1 – Détails des catégories utilisées pour le mode de coloration HCA

En utilisant ce mode de coloration, nous obtenons le type de résultat observable sur la figure 3.4.4.2.1. Sur cette figure, les résidus hydrophobes sont en vert, les *coils* en jaune, les hydrophobes fuchsia en jaune et les autres en cyan. Nous constatons que le feuillet β est majoritairement composé de résidus hydrophobes en vert et en fuchsia ; de plus, les flèches permettent de bien visualiser l'emplacement des acides aminés. Ce mode de coloration est bien sûr applicable aux autres modes de visualisation comme il est visible en (b) avec un rendu squelette utilisant ce mode de coloration.



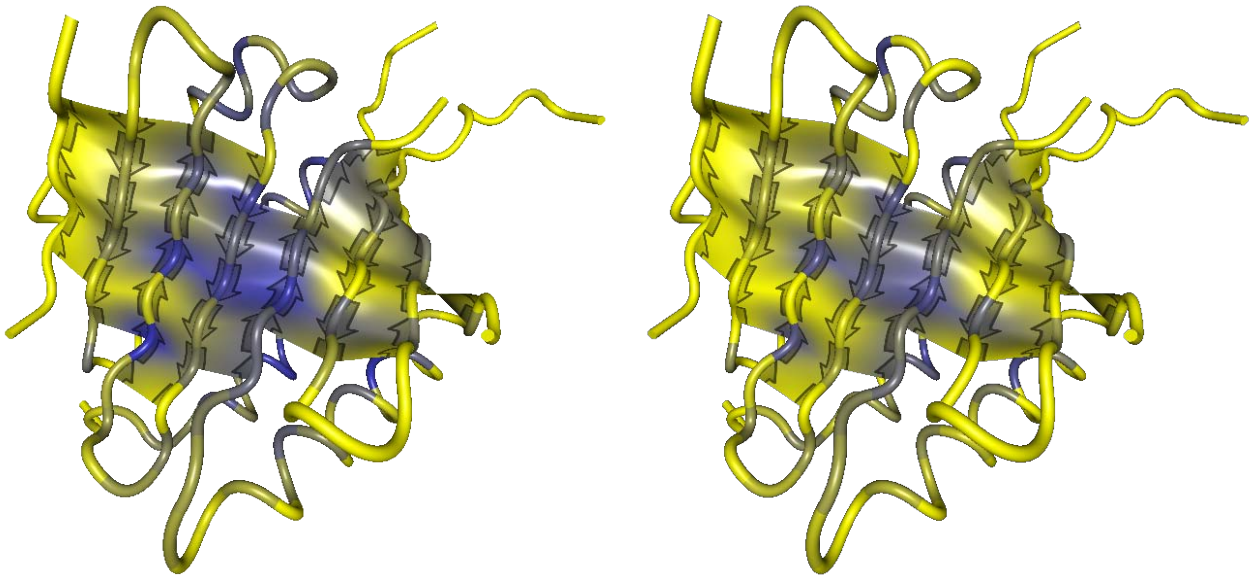
(a) feuillet en HCA (b) feuillet et squelette en HCA

Figure 3.4.4.2.1 – Résultats obtenus avec le mode de coloration HCA sur [1914]. (a) représente un feuillet β texturé à l'aide de flèches représentant les acides aminés et leur direction, (b) représente le même feuillet β couplé à un rendu de type squelette coloré avec la méthode HCA également. Les hydrophobes sont en vert, les coils *en jaune, les occasionnels en fuchsia et les autres en cyan*

3.4.4.3 Facteur de température

Le mode de coloration représentant le facteur de température, ou coefficient d'agitation thermique, est un mode classique présent dans l'ensemble des logiciels de modélisation moléculaire. L'innovation ici n'est pas dans la représentation du facteur, mais dans la façon de le calculer. Le mode classique va représenter une couleur plus ou moins intense en fonction de la valeur du facteur de température, et ce pour chaque atome. Chaque atome aura donc une couleur dont l'intensité correspond à son facteur de température.

Lorsqu'il s'agit de colorer une surface telle qu'une surface accessible au solvant, alors, la valeur représentée à un point de la surface correspond à une moyenne des valeurs des atomes les plus proches de ce point. C'est ce dernier point qui fait que nous avons développé notre propre mode de coloration du facteur de température, car nos surfaces sont « accrochées » aux carbones α des acides aminés composant les feuillets β, or les valeurs que nous avons pour ces atomes ne sont pas représentatives des acides aminés concernés. C'est pourquoi notre mode de coloration peut représenter soit les valeurs moyennes des facteurs de température des atomes constituant les chaînes latérales des acides aminés, soit les valeurs des atomes ayant le facteur de température le plus important des chaînes latérales. Dans ce mode l'intensité de la couleur dépend également de la valeur du facteur de température : plus le facteur de température est élevé, plus la couleur est intense. Il est possible de fixer une borne supérieure.



(a) moyenne (b) maximum

Figure 3.4.4.3.1 – Résultats obtenus avec notre mode de coloration du facteur de température. (a) représente le feuillet β et le squelette de la protéine [1914] colorés avec ce mode dont la valeur maximale du facteur de température a été fixée à 50 et ou chaque acide aminé est représenté par la moyenne des valeurs de sa chaîne latérale, (b) représente la même protéine avec le même mode de coloration, mais chaque acide aminé est représenté par la valeur maximale de sa chaîne latérale. Les valeurs les plus faibles sont en bleu, et les plus importantes en jaune. La protéine 1914 n'est peut-être pas la plus parlante pour illustrer ce mode de coloration mais nous nous avons choisi cette protéine depuis le début comme modèle

La figure 3.4.4.3.1 montre des résultats obtenus avec ce mode de coloration : (a) nous représentons la valeur moyenne du facteur de température des chaînes latérales des acides aminés, et en (b) nous représentons les valeurs maximales trouvées dans les chaînes latérales de ces mêmes acides aminés. Pour les deux résultats, la valeur maximale du facteur de température représenté est de 50, les valeurs les plus faibles sont en bleus et les plus importantes en jaune. Ces options sont modifiables *via* les fenêtres d'interface qui ont été créées pour ce modèle.

3.4.4.4 Coloration de type « Molecular Hydrophobicity Potential *» - MHP*

Afin de pouvoir représenter les potentiels hydrophobes d'une protéine nous utilisons les données calculées par la méthode MHP (*Molecular Hydrophobicity Potential*) [Efremov1993; Efremov2007]. MHP est une méthode empirique permettant d'évaluer et de visualiser les propriétés hydrophobes et hydrophiles de molécules pour chacun des atomes les constituants. Pour représenter les valeurs de ces potentiels, nous utilisons une méthode de coloration comparable à celle utilisée pour la représentation des facteurs de température : nous calculons pour chaque carbone α la valeur MHP moyenne de la chaîne latérale correspondante.

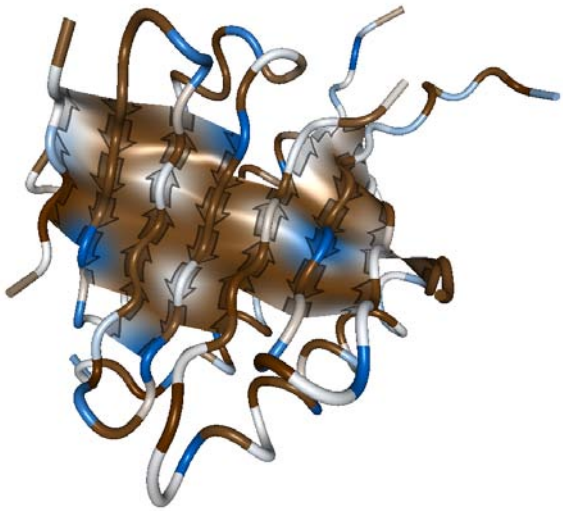


Figure 3.4.4.4.1 – Résultats obtenus avec le mode de coloration MHP sur le feuillet β et le squelette de la protéine [1914]. L'hydrophobie est en marron, l'hydrophilie en bleu et le neutre en blanc

Dans ce mode, trois couleurs sont utilisées : une couleur hydrophobe, une couleur hydrophile et une couleur neutre. L'intensité de la couleur dépend de la valeur MHP : plus la valeur MHP est hydrophobe, plus la couleur hydrophobe est intense et plus la valeur MHP est hydrophile, plus la couleur hydrophile est intense. Dans les deux cas, plus la valeur est neutre, plus la couleur représentée s'approche de la couleur neutre.

La figure 3.4.4.4.1 montre le résultat obtenu avec ce mode de coloration sur la protéine [1914]. Les valeurs hydrophobes sont représentées en marron, les hydrophiles en bleu et les neutres en blanc.

3.4.4.5 Zones de stabilité d'un feuillet β sur le modèle de Bézier

Ce mode, ne fonctionnant qu'avec le modèle de Bézier, a pour but de visualiser les zones de stabilité d'un feuillet β. Une zone est considérée comme stable lorsque les brins β adjacents sont proches, et comme instable lorsque les brins β adjacents sont éloignés. Étant donné que la distance entre deux carbones α consécutifs varie très peu (environ 3,8 Å), seules les distances entre les carbones α des brins β adjacents peuvent faire varier de manière significative l'aire d'un carreau de Bézier. C'est sur ce paramètre que se base ce mode de coloration. L'aire de chaque carreau de Bézier est calculée, ces valeurs sont utilisées pour déterminer les couleurs et leurs intensités. Trois couleurs sont utilisées : pour les zones les plus stables, pour les zones instables, et pour les zones à stabilité moyenne. Les couleurs sont appliquées indépendamment sur chaque carreau de Bézier.

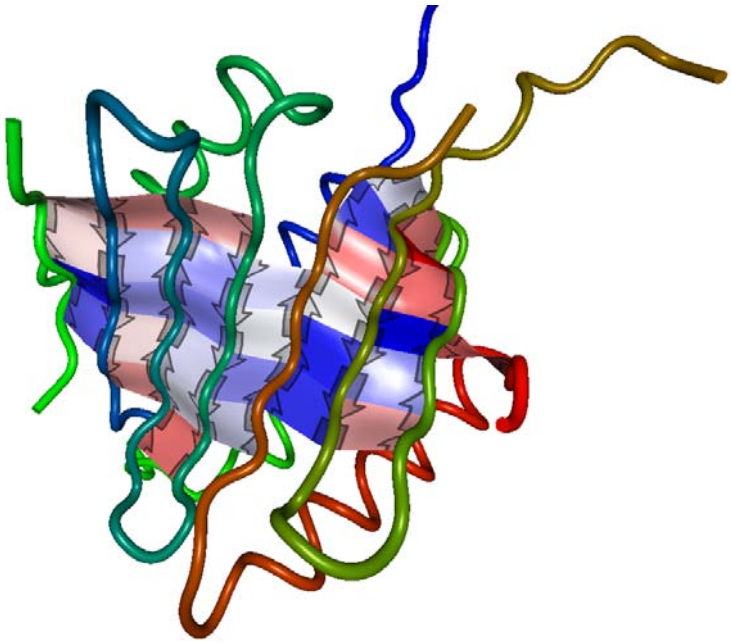


Figure 3.4.4.5.1 – Résultat obtenu avec le mode de représentation des zones de stabilité d'un feuillet β sur la protéine [1914]. Les zones stables sont en bleu, les instables en rouge et les zones et celles à stabilité moyenne en blanc

La figure 3.4.4.5.1 montre l'utilisation de ce mode de coloration sur la protéine [1914]. Sur cette figure les zones les plus stables sont en bleu, les zones instables en rouge et les zones moyennes en blanc.

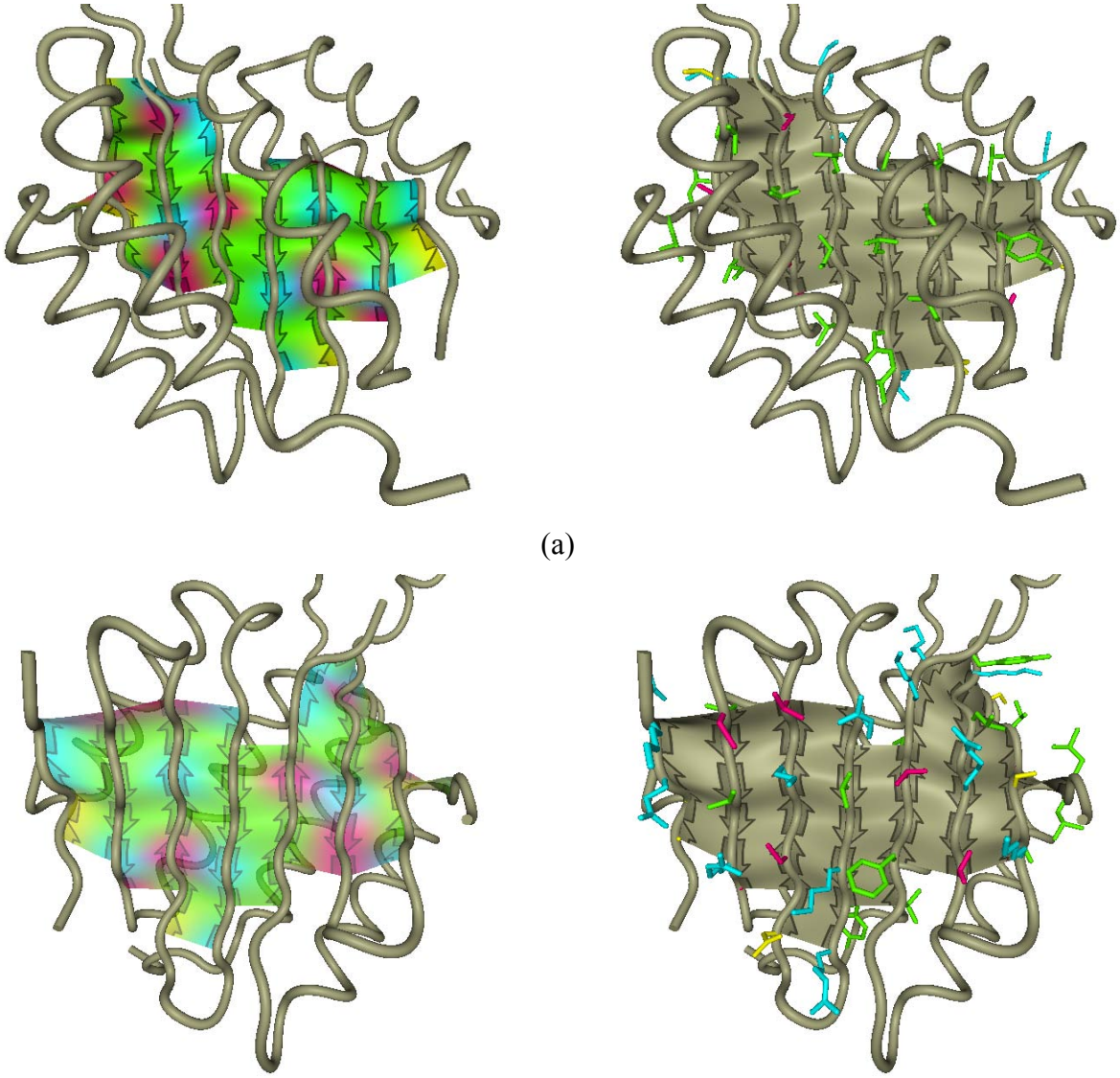
Sur cette figure, il est possible de voir que les zones les plus « fragiles » en terme de distances s'observent aux extrémités de chaque brin, ainsi qu'entre les 2 brins à droite.

3.4.5 Chaînes latérales

Les modes de visualisation que nous avons développés représentent uniquement les surfaces des feuillets β. Cependant, il peut être intéressant de représenter également les chaînes latérales des acides aminés appartenant à ces feuillets. Pour ce faire, il a été ajouté un mode qui permet de ne visualiser que les chaînes latérales des acides aminés en conformation β. Il est alors intéressant de coupler nos modes feuillets β avec ce mode chaînes latérales β.

Si nous couplons notre modèle de feuillet β avec la représentation des chaînes latérales des acides aminés en conformation β en utilisant le mode de coloration HCA tel que présenté sur la figure 3.4.5.1, nous constatons que les résidus hydrophobes de couleur verte sont majoritairement du côté représenté en (a). Nous pouvons en déduire qu'il s'agit du côté hydrophobe, ce côté étant celui où se situent les hélices α. Cela nous permet une interprétation supplémentaire. En effet, pour une protéine globulaire il est nécessaire que de grandes zones hydrophobes ne se trouvent pas en contact avec le solvant : en conséquence elles sont soit enfouies à l'intérieur de la protéine, soit elles correspondront à des zones qui devront être en interaction avec d'autres partenaires présentant une complémentarité de forme et de propriétés physico-chimiques.

Grâce à ces visualisations il est très aisé et rapide de repérer la face interne des feuillets β, et ce même pour un néophyte.



(a)

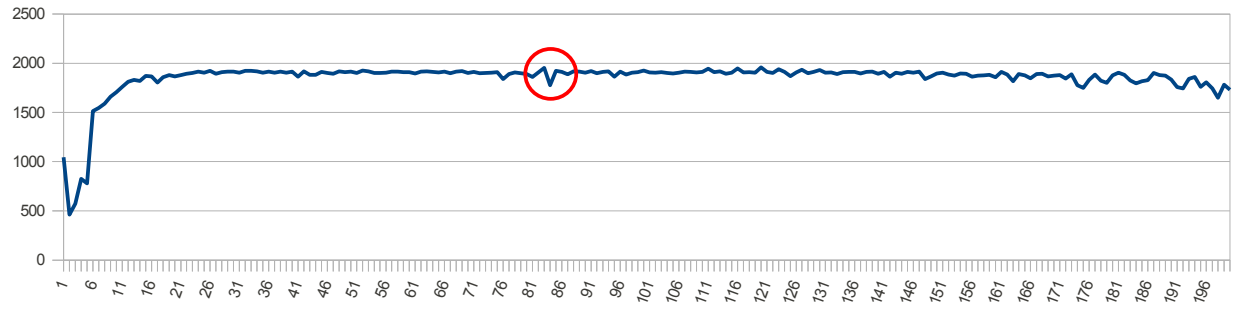
(b)

Figure 3.4.5.1 – Visualisation du feuillet β de la protéine [1914] couplée au mode de visualisation des chaînes latérales des acides aminés en conformation β coloré par HCA. Nous remarquons que la face du feuillet β représentée en (a) est plus hydrophobe que celle en (b) car il y a plus de chaînes colorées en vert. La face hydrophobe correspond à celle des hélices α de la protéine. Le feuillet en (b), à gauche, est transparent afin que nous puissions voir l'arrangement de ces hélices du côté hydrophobe

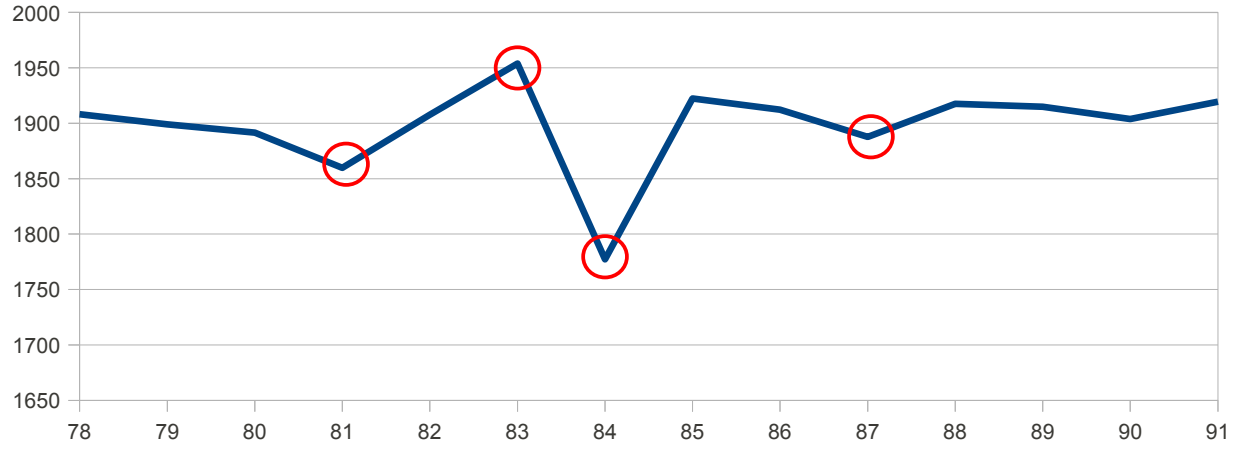
3.5 Visualisation dynamique

3.5.1 Dynamique moléculaire

Lors de la visualisation d'une simulation de dynamique moléculaire, les aires des feuillets β d'une protéine sont stockées dans des fichiers. Chaque fichier porte le nom de la protéine, ainsi que le numéro du feuillet β dont les aires sont conservées. Si nous prenons l'exemple du fichier PDB [1YTB] dont la protéine contient deux feuillets β, deux fichiers seront donc crées : 1YTB_1.txt et 1YTB_2.txt. Ces données peuvent servir à établir des graphiques représentants l'aire d'un feuillet β au cours d'une simulation dynamique.



(a)



(b)

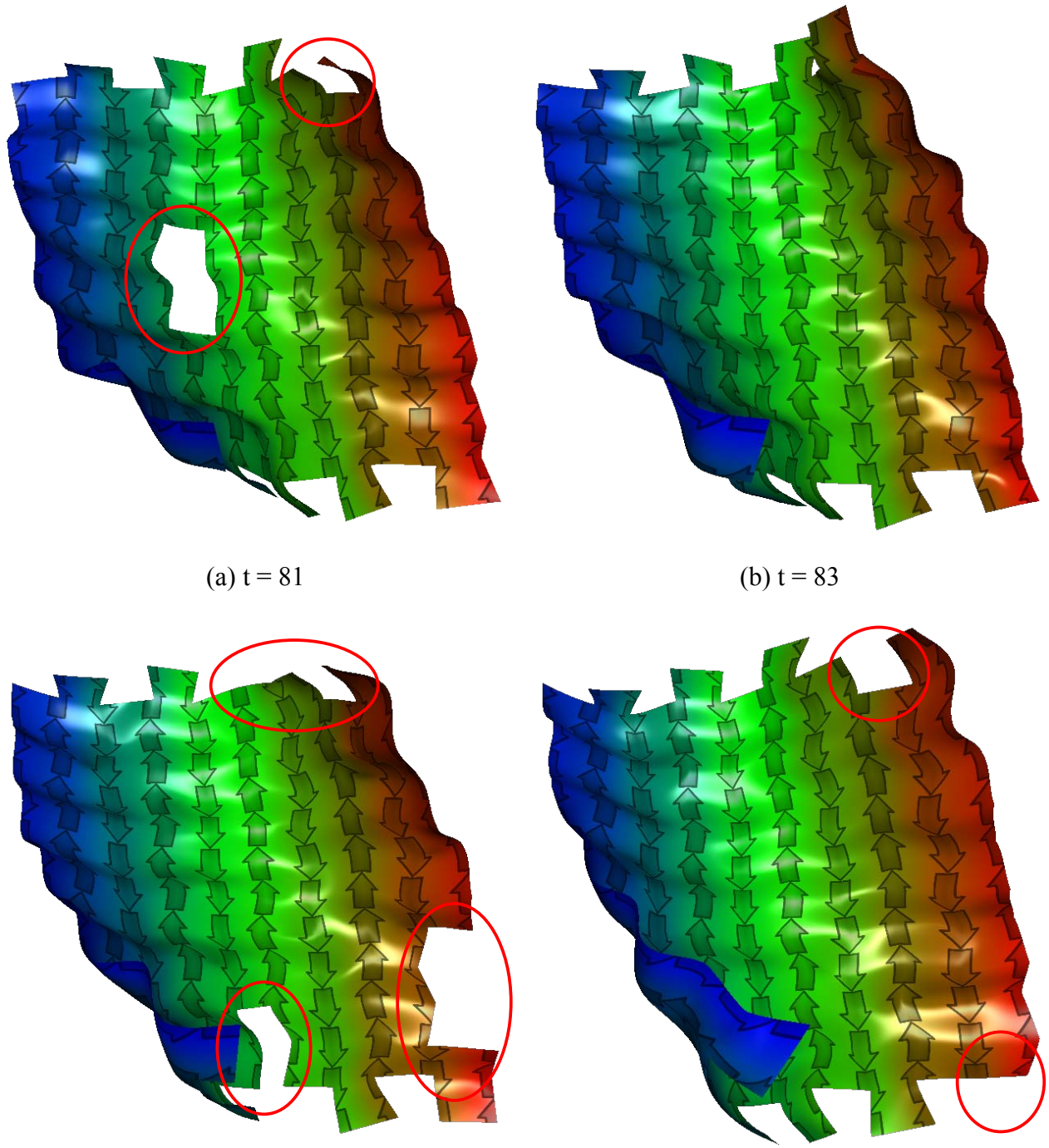
Figure 3.5.1.1 – Graphiques représentants l'évolution de la surface, en Å², d'un feuillet β au cours d'une simulation de dynamique moléculaire. La figure (a) représente la simulation complète, et (b) la portion de la simulation correspondant aux pas de simulation 78 à 91 entourés sur (a). Les pas de simulation entourés sur (b) sont reportés dans la figure 3.5.1.2

Cette méthodologie a été utilisée sur un modèle, construit par Nicolas Belloy, que nous utilisons dans notre laboratoire, le SiRMa. Il s'agit d'un feuillet β plan qui s'enroule sur lui même au cours de la simulation et qui finit par se refermer. La figure 3.5.1.1a présente le graphique obtenu, il représente l'aire du feuillet β au cours de la simulation. L'abscisse représente l'aire en Å² et l'ordonnée le temps de simulation.

Nous constatons en (a) qu'au début de la simulation le feuillet β a une aire d'environ 1000 Å² pour s'effondrer jusqu'à environ 500 Å² et enfin se stabiliser à une valeur inférieure à 2000 Å². Cet effondrement initial s'explique par le fait que le feuillet se scinde en trois entités distinctes, pour ensuite se reformer. En observant ce graphique nous pouvons déceler des zones de variations significatives au cours de la simulation, telle celle détourée de rouge sur la figure 3.5.1.1a. Cette zone correspond aux pas de simulation allant de 78 à 91, cette zone est représentée sur le graphique de la figure 3.5.1.1b. Nous constatons que la surface du feuillet β varie beaucoup sur cette période de la simulation. Le feuillet β correspondant aux pas 81, 84, 85 et 87 (entourés sur la figure 3.5.1.1b) est représenté sur la figure 3.5.1.2.

Cette figure présente le feuillet aux étapes 81, 83, 84 et 87. Le feuillet (a) qui correspond à l'étape 81 présente un trou, ainsi qu'une légère invagination en haut à droite de l'illustration, ce qui explique que sur la figure 3.5.1.1b le graphique montre une diminution de la surface. L'étape 83, observable en (b), présente un feuillet d'une surface plus importante. L'étape 84 correspond au feuillet dont la surface est la plus faible, sur la figure 3.5.1.2c nous constatons que ce feuillet présente plusieurs invaginations ce qui explique la diminution visible sur le graphique. La figure 3.5.1.2d, correspond à l'étape 87 dont le feuillet a une surface légèrement inférieure à celui de l'étape 83. Les cercles rouges sur la figure 3.5.1.2 correspondent aux zones manquantes par rapport à l'étape 83, (b) sur la figure.

Ce type de graphique permet donc d'identifier, très facilement, des séries d'étapes durant lesquelles d'importantes variations de surfaces sont observables. Avec ce graphique il est également simple de déterminer lorsqu'un feuillet β se scinde en plusieurs entités et qu'il se reforme, comme c'est le cas durant les premières étapes du graphique de la figure 3.5.1.1a.



(a) t = 81 (b) t = 83

(c) t = 84 (d) t = 87

Figure 3.5.1.2 – Illustrations du feuillet β aux pas de simulation 81 (a), 83 (b), 84 (c) et 87 (d). Les informations données par le graphique de la figure 3.5.1.1 (a) sont cohérentes puisqu'aux pas de simulation où les surfaces sont les moins importantes, 81 (a) : 1860 Å² et 84 (c) : 1880 Å², nous constatons l'existence de trous ainsi que d'invaginations. L'étape 83 (b) présente la plus grande surface : 1960 Å², et l'étape 87 (d) présente une surface moindre de 1890 Å²

3.5.2 Autres aspects dynamiques

À partir des outils présentés dans le paragraphe ci-dessus, nous avons pu visualiser des fichiers PDB issue de la Résonance Magnétique Nucléaire (RMN). Ces fichiers contiennent jusqu'à plusieurs dizaines de modèles de la même molécule, issus de l'expérimental et dont les contraintes de résolution ne violent pas les données expérimentales. Lorsqu'un feuillet β est impliqué dans les variations locales observées par RMN, nous pouvons tout à fait obtenir un effet similaire à celui observé par simulations de dynamique moléculaire.

Nous avons par ailleurs testé notre application sur des résultats d'Analyse harmonique des Modes Normaux (NMA). Là encore le mode de visualisation se prête totalement à l'observation des différents modes mettant en jeu un feuillet β.

Chapitre 4
Intérêts de SheHeRASADe et applications

« Son, if you really want something in this life, you have to work for it. Now quiet! They're about to announce the lottery numbers. »

Homer J. SIMPSON

Tout au long du développement de nos différents modèles de visualisation, nous avons testé et surtout amélioré, en fonction des problèmes rencontrés, chacune de nos représentations. Nous avons testé environ 1750 fichiers issus de la PDB, avec tous les types de modèles et de définitions de couleurs possibles. Nous ne pourrons bien évidemment pas présenter de façon exhaustive les possibilités de nos représentations. En conséquence nous avons effectué certains choix dans les applications que nous allons présenter. De plus, dans ce manuscrit, nous avons pris le parti de présenter nos feuillets dans une couleur unique et de ne pas entrer dans une description complexe avec les différents modes de coloration. Il est évident que seul l'usage du logiciel et de ces représentations permet d'appréhender les nombreuses possibilités scientifiques que procure « SheHeRASADe ».

Dans un premier temps, nous présenterons les différents éléments de structures secondaires, tertiaires et quaternaires sur diverses protéines. Dans le dernier cas, nous nous intéresserons à des homodimères et des hétérodimères, afin d'observer la formation de feuillets β résultant de

l'interaction de deux macromolécules. L'étude des fichiers issus de la PDB, considérant les différents types de repliements structuraux, nous a amené à tester nos modèles sur les bases de données de classification structurale que sont CATH [Thornton1997] et SCOP [Chothia1995]. Ici nous ne présenterons que le niveau le plus élevé de CATH, en choisissant les exemples les plus représentatifs de chaque classe structurale. Ensuite, nous avons appliqué nos modèles à la superfamille des immunoglobulines, dont les protéines ont pour particularité d'avoir un type de repliement spécifique : le pli immmunoglobuline, un des neuf *superfolds*. Ensuite, nous avons utilisé les travaux de Nicolas Prudhomme [Prudhomme2009], qui présente une base de données sur le repliement de type immunoglobuline, afin d'évaluer les possibilités de nos modèles.

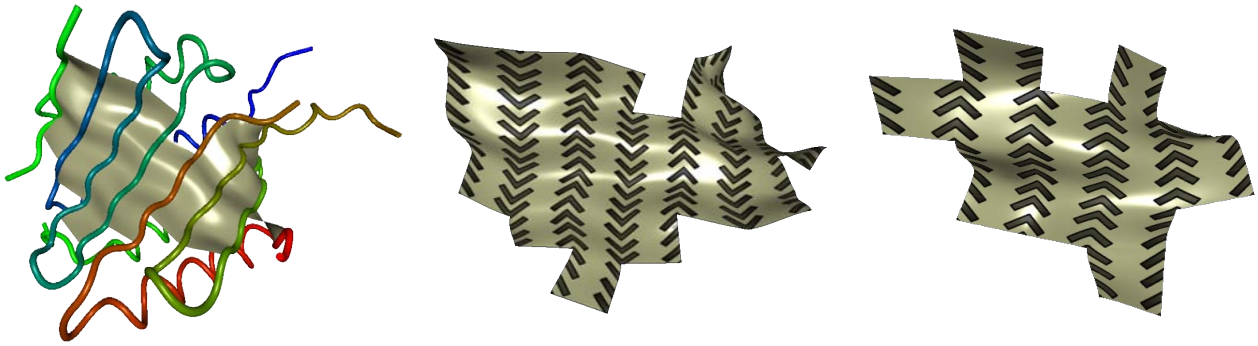
Il est évident que le feuillet β prend toute son importance dans un domaine de recherche qui est actuellement au centre des problématiques de santé publique : les fibres amyloïdes impliquées dans les pathologies amyloïdogéniques. Nous avons utilisé nos modèles pour constater la plus-value apportée sur les différents types de solénoïdes β, qui apparaissent souvent comme motifs structuraux de base pour le déclenchement de structures amyloïdogéniques. Comme précédemment, nous avons tenté de cibler l'intérêt de notre travail sur la base de données des protéines présentant des caractéristiques d'amyloïdes : AmyPDB [Pawlicki2008]. La taille importante de cette base de données nous a amené à ne pas la représenter de manière exhaustive, c'est pourquoi nous avons choisi de ne l'illustrer que par quelques exemples. Enfin, sur une fibre amyloïde hypothétique de 27 886 amino-acides, nous avons appliqué nos différentes textures et colorations.

4.1 Intérêts de SheHeRASADe sur les différents niveaux de structures

A la vue des feuillets β présentés dans les précédents chapitres de cette thèse, il apparaît que le mode de représentation que nous avons proposé est tout à fait approprié pour mettre en évidence, sous la forme d'une « feuille » de papier, complète ou non, déchirée ou non, la structure secondaire régulière associée. Durant le développement de ces travaux, les possibilités de ces représentations et la plus-value apportée par ce type de visualisation ont été manifestes tant sur les aspects topologiques ou topographiques, que sur l'application de textures ou de modes de coloration présentant une dimension sémantique supplémentaire. Toutefois, ces derniers aspects ne seront pas discutés dans les choix applicatifs ci-après.

4.1.1 Représentation des structures secondaires

Sur la figure 4.1.1.1 il est évident que nous pouvons aisément appréhender la structure complète de chaque feuillet β et il apparaît que l'aspect des structures secondaires de type β est clairement défini. Par exemple, la représentation utilisant des chevrons plaqués sur la surface nous donne immédiatement le sens des brins β constituant le feuillet que ce soit dans le cas d'un feuillet parallèle (figure 4.1.1.1b) ou antiparallèle(figure 4.1.1.1c). Cette dimension sémantique supplémentaire est fondamentale lorsque nous cherchons à focaliser notre intérêt sur un amino-acide donné. Ceci s'avère être très utile dans le cas de très grands feuillets dont la vue d'ensemble est très difficile à appréhender.



(a) (b) (c)

Figure 4.1.1.1 – (a) Représentation du feuillet β de type Bézier, couplé à un rendu de type squelette sur la protéine [1914] ; (b) représentation du feuillet antiparallèle de la protéine [1914] texturé avec des chevrons ; (c) représentation du feuillet parallèle la protéine [1E20] texturé avec des chevrons

4.1.2 Représentations des structures tertiaires et quaternaires

Les représentations de type Catmull-Rom ou Bézier prennent toute leur signification lorsque nous considérons la problématique de visualisation à un niveau plus élevé de complexité. Ainsi lorsque plusieurs feuillets sont présents dans les structures tertiaires, nous avons très rapidement la possibilité de les visualiser, ensembles ou séparément, d'évaluer leurs surfaces respectives, d'envisager leur positionnement, les uns par rapport aux autres, mais également au sein de structures tertiaire ou quaternaire. Nous avons choisi quelques modèles représentatifs de protéines permettant d'illustrer la plus-value de ce type de visualisation. Les protéines choisies pour décrire ce niveau des structures tertiaires et quaternaires sont présentées ci-dessous.

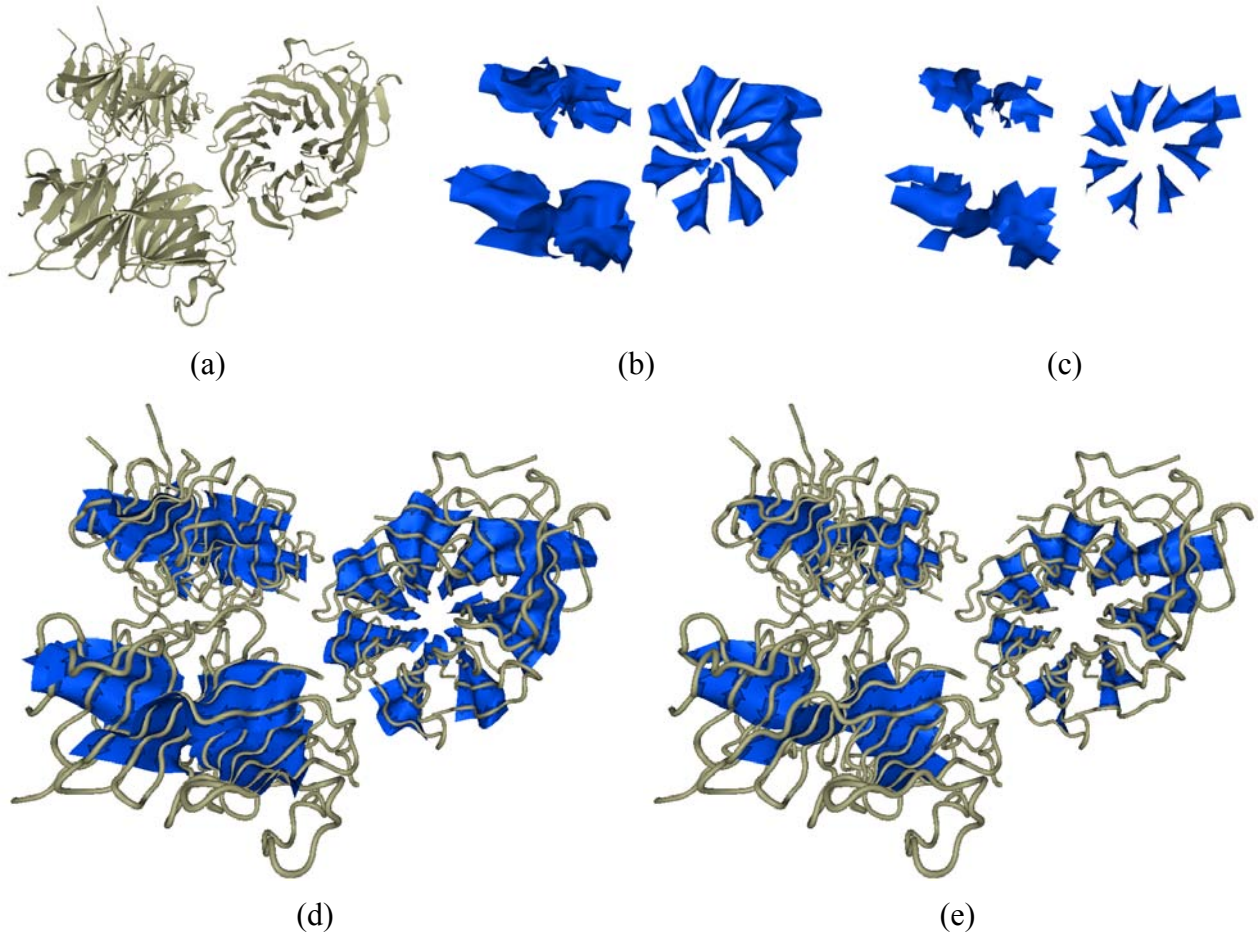


Figure 4.1.2.1 – Domaine C-terminal WD40 de TUP1 [1ERJ]. (a) Représentation de type cartoon*, (b) et (c) représentations respectivement de type Catmull-Rom et Bézier, (d) et (e) représentations respectivement de Catmull-Rom et de Bézier, texturées par des flèches, couplées à des représentations squelette*

Le domaine C-terminal WD40 de TUP1 est un inhibiteur de la transcription. Cette protéine, visible sur la figure 4.1.2.1, est arrangée en trimère, dont chaque monomère contient un « *β-propeller* ». Un *β-propeller* est une association de feuillets β, arrangés autour d'un axe central de manière torique. Pour les modèles de Catmull-Rom et de Bézier, visibles sur les figures 4.1.2.1b et 4.1.2.1c, nous dénombrons aisément les sept feuillets β et leur arrangement en *β-propeller* dont l'organisation autour d'un axe apparaît de façon évidente. Nous constatons que les feuillets sont *twistés*, c'est à dire que le premier et le dernier brin, d'un même feuillet, tendent à devenir perpendiculaires. Lorsque nous couplons ces représentations avec un mode de visualisation de type squelette, tel qu'en 4.1.2.1d et 4.1.2.1e, nous constatons que la protéine n'est quasiment composée que de feuillets. De plus, les textures nous indiquent qu'ils sont antiparallèles.

Nos modèles permettent d'identifier immédiatement le type d'arrangement des feuillets β, leur nombre, leur caractère antiparallèle, ainsi que leur topologie *twistée*. Par ailleurs, lorsque nous découpons le fichier de coordonnées en trois fichiers « indépendants » contenant chacun un *β propeller*, nous pouvons alors, par méthode de « *Root Mean Square Deviation* » (RMSD), superposer les modèles et comparer les feuillets β ainsi définis. Que ce soit d'un point de vue surfacique, topologique ou topographique (place des résidus au sein de chaque feuillet). Dans le cas présent, les éléments sont identiques et ne sont pas présentés. Cette approche de séparation de chacun des éléments peut aussi se faire pour chaque pale du *β propeller* et permettre ainsi d'identifier les différentes zones de chacune des pales, à la fois sur la forme, mais également, lors de l'utilisation de colorations de type HCA, sur le rôle d'une pale par rapport aux autres (données non présentées).

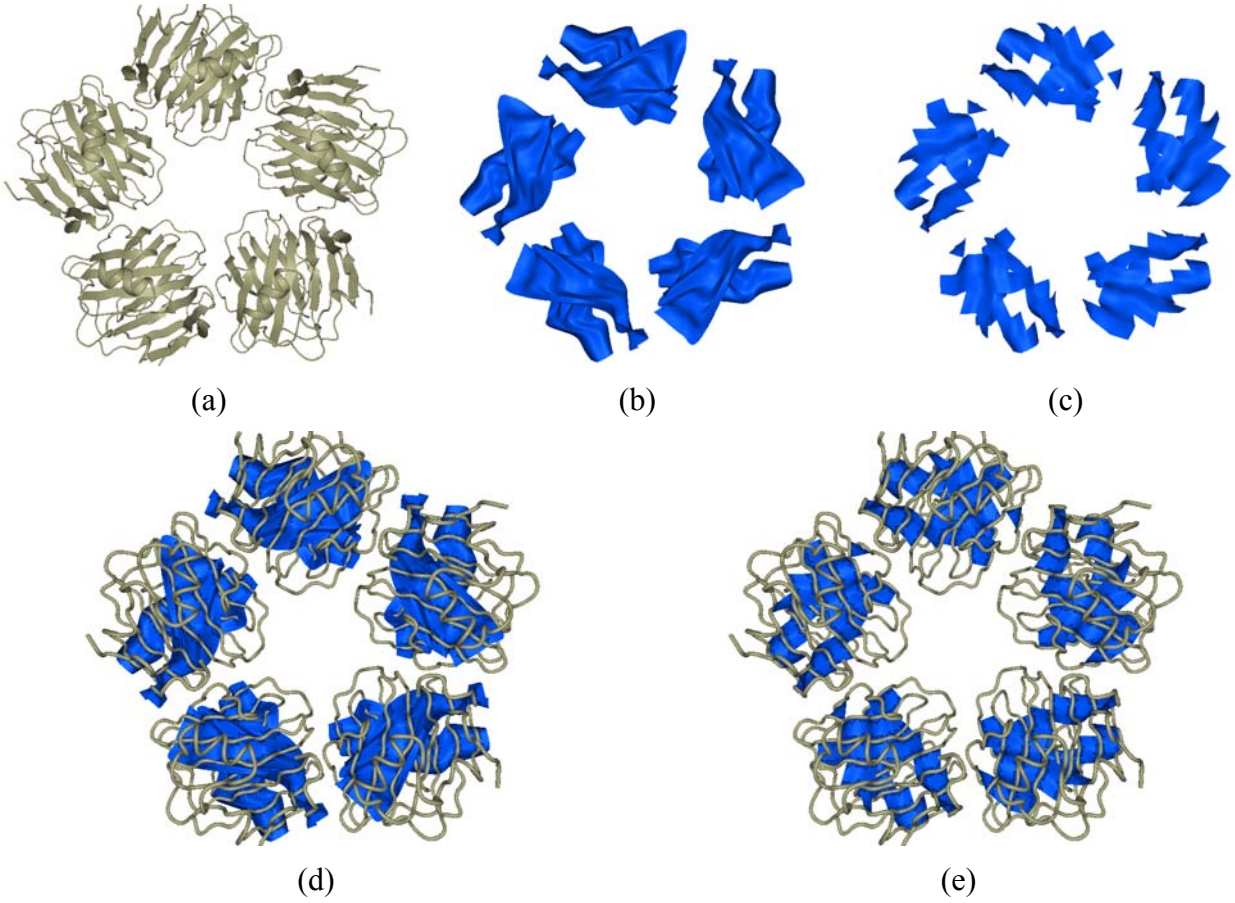


Figure 4.1.2.2 – Composant-P amyloïde [1LGN]. (a) Représentation de type cartoon*, (b) et (c) représentations respectivement de type Catmull-Rom et Bézier, (d) et (e) représentations respectivement de Catmull-Rom et de Bézier, texturées par des flèches, couplées à des représentations squelette*

La figure 4.1.2.2 représente le composant-P amyloïde (SAP pour « *Serum Amyloid P component* »), une glycoprotéine à symétrie radiale de cinq monomères formant un anneau.

Sur le modèle de Catmull-Rom, en (b), nous distinguons clairement que chacun des cinq monomères est composé de deux feuillets β disposés en *jellyroll*, un des neuf *superfolds*. Nous constatons également que ces cinq *jellyrolls* ont exactement la même topologie, ce qui est très difficile à estimer en observant le modèle *cartoon* visible en (a). Le modèle de Bézier, en (c), présente des feuillets nettement moins structurés. La différence majeure entre nos deux modèles étant que le modèle de Catmull-Rom se base sur les informations présentes dans le fichier PDB, alors que le modèle de Bézier utilise l'attribution issue de DSSP. Le *superfold jellyroll* est cependant bien conservé : il y a un couple de feuillets β par monomère, couples quasi-identiques sur l'ensemble du pentamère.

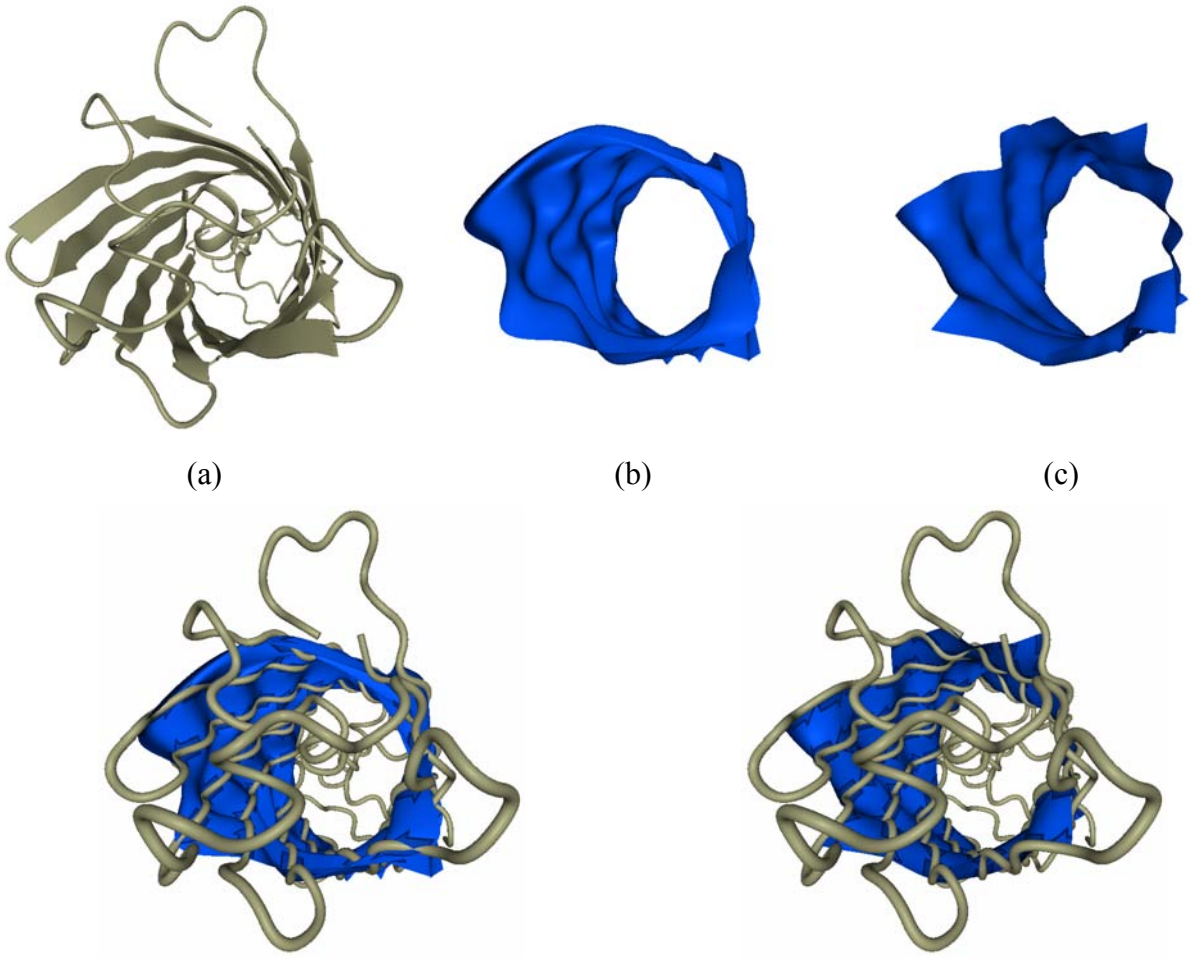
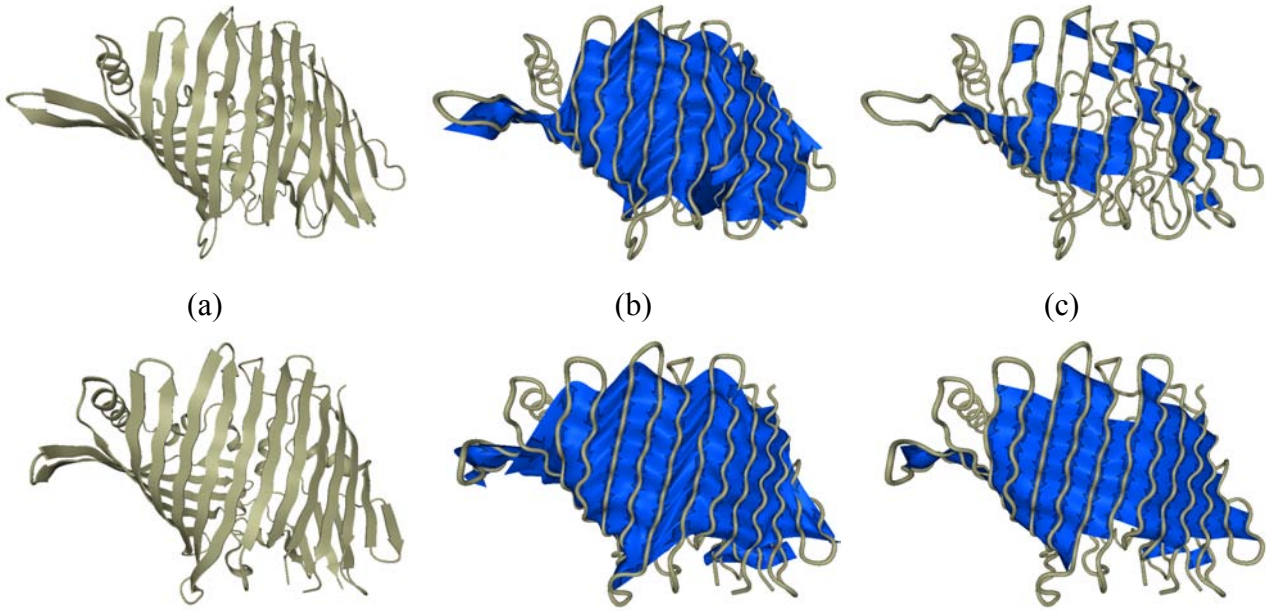


Figure 4.1.2.3 – GFP ou « Green Fluorescent Protein *» [1EMA]. (a) Représentation de type* cartoon*, (b) et (c) représentations respectivement de type Catmull-Rom et Bézier, (d) et (e) représentations respectivement de Catmull-Rom et de Bézier, texturées par des flèches, couplées à des représentations squelette*

Le modèle choisi considère la « *Green Fluorescent Protein* ». La GFP est constituée d'un *« β barrel »*, composé de onze brins β, dont la structure a été initialement décrite comme ayant la forme d'une cannette de soda (*« soda can shape »*).

Les modèles de Catmull-Rom et de Bézier appliqués à la GFP (figure 4.1.2.3a et 4.1.2.3b) présentent la forme en cannette de soda du *β barrel*. Dans ce cas, les deux modèles que nous avons développés sont très similaires, et l'aspect plissé des feuillets β est bien mis en relief. Il est important de noter que nos représentations conduisent à un modèle qui se referme sur lui même présentant la structure caractéristique citée ci-dessus. Une future projection de paramètres physico-chimiques du ligand sur la surface pourrait certainement apporter une information supplémentaire.



(a) (b) (c)

(d) (e) (f)

Figure 4.1.2.4 – Chaîne A de la protéine bactériochlorophylle A [1KSA] et [3ENI]. En (a), (b) et (c) protéine : [1KSA], avec respectivement un rendu de type cartoon*, un rendu de Catmull-Rom, texturé par des flèches, couplé à un rendu squelette et un rendu de Bézier, texturé par des flèches, couplé à un rendu squelette. En (d), (e) et (f) : protéine [3ENI], avec respectivement un rendu de type* cartoon*, un rendu de Catmull-Rom, texturé par des flèches, couplé à un rendu squelette et un rendu de Bézier, texturé par des flèches, couplé à un rendu squelette*

La figure 4.1.2.4a présente la chaîne A de la protéine [1KSA] en *cartoon*, nous constatons que cette chaîne est majoritairement constituée d'un grand feuillet β. Ce feuillet étant bien défini dans le fichier PDB, notre modèle de Catmull-Rom, en (b), le représente dans sa globalité. Mais le modèle de Bézier, visible en (c), présente un feuillet β très déstructuré. Le feuillet β, bien que parfaitement décrit dans le fichier, n'est pas du tout stable. Notre modèle permet donc de visualiser de manière très claire l'instabilité d'un feuillet.

La chaîne A de la protéine [3ENI], dont la représentation en mode *cartoon* est visible en (d), est le modèle raffiné de [1KSA]. Cette fois, les modèles de Catmull-Rom et de Bézier sont parfaitement en accord, et représentent la totalité du feuillet. Sur ces deux exemples la texture et le sens des brins prennent toute leur importance pour bien appréhender l'environnement de chaque partie de la macromolécule.

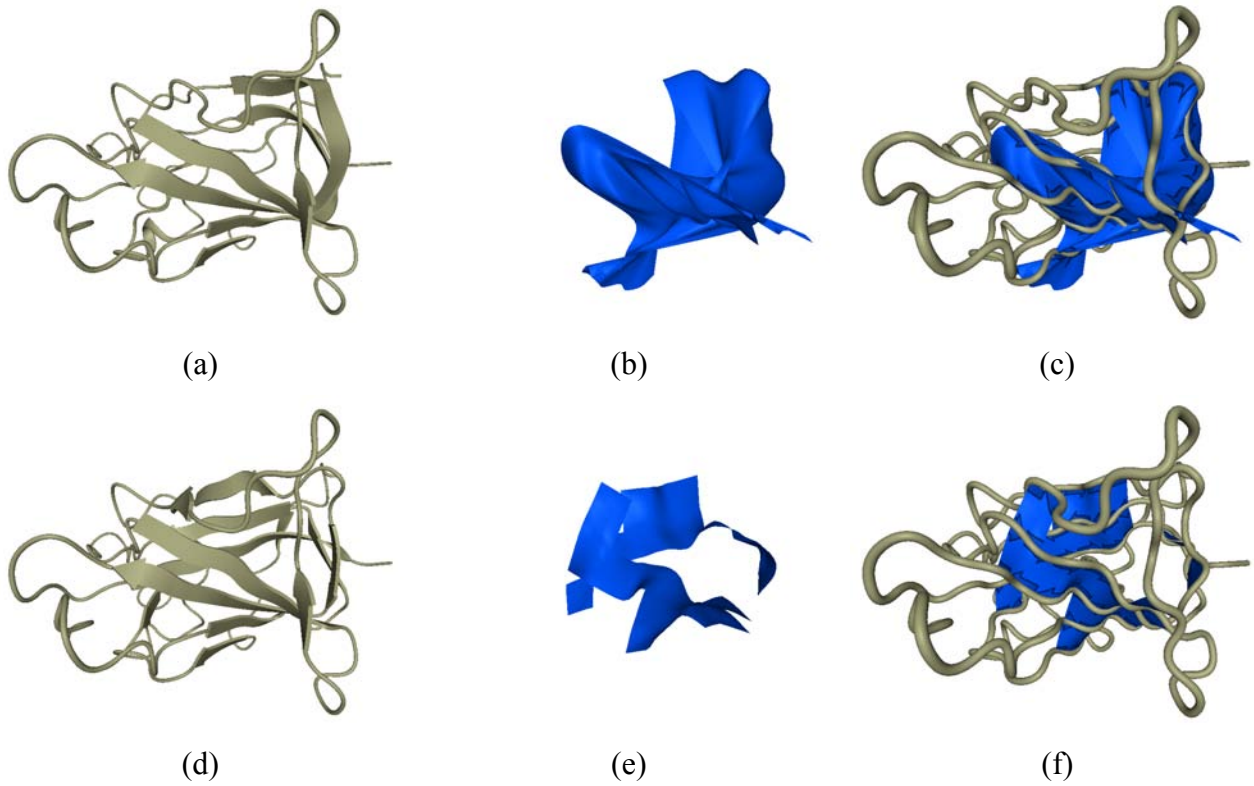
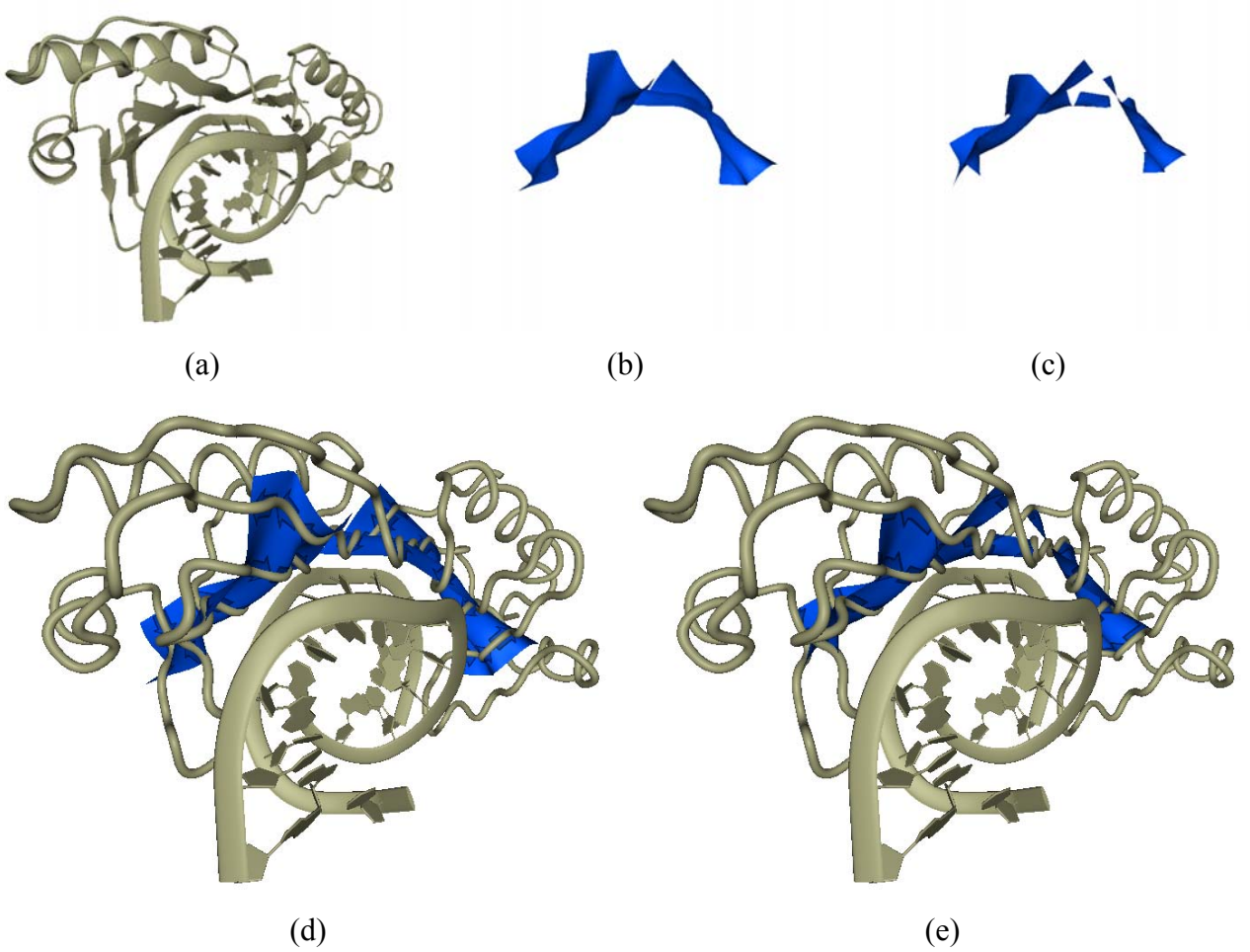


Figure 4.1.2.5 – DDR2 ou « Discoidin Domain Receptor 2 *» [2WUH]. (a) Représentation de type* cartoon*, (b) et (e) représentations respectivement de type Catmull-Rom et Bézier, (d) représentation de type* cartoon *après le calcul des structures secondaires par DSSP, (c) et (f) sont des représentations respectivement de Catmull-Rom et de Bézier, texturées par des flèches, couplées à des représentations squelette.*

La DDR2 a été choisie pour illustrer la détection aisée de définitions spécifiques, de feuillets β, voulues par les auteurs du dépôt de la structure dans la PDB. Ainsi, la représentation en mode *cartoon* de DDR2 (figure 4.1.2.5a) montre des brins β qui forment un coude très prononcé, ce qui est impossible puisque ce sont des structures étirées qui ne se replient pas. Lorsque nous utilisons le modèle de Catmull-Rom sur un tel feuillet, nous constatons, en (b), que le résultat obtenu n'est pas exploitable. Le fait que les brins soient définis ainsi est une volonté de l'auteur de ce dépôt PDB de les montrer comme tels, même si cela n'est pas structuralement possible. Notre modèle de Catmull-Rom peut être utilisé afin de détecter aisément ce genre d'aberrations.

Le modèle de Bézier, qui n'utilise pas les définitions des feuillets présents dans le fichier PDB, mais qui les recalcule, représente deux feuillets, et non plus un seul. La représentation en mode *cartoon*, en (d), après le calcul de DSSP, montre que les grands brins β ont été redéfinis en trois brins de tailles inférieures.



(a) (b) (c)

(d) (e)

Figure 4.1.2.6 – TBP, ou « TATA Binding Protein *», liée à la* « TATA box » *d'une séquence ADN [1TGH]. (a) Représentation de type* cartoon*, (b) et (c) représentations respectivement de type Catmull-Rom et Bézier, (d) et (e) sont des représentations respectivement de Catmull-Rom et de Bézier, texturées par des flèches, couplées à des représentations squelette*

Nous venons de montrer l'aptitude de nos différentes représentations dans la structuration tertiaire et ce qu'elles peuvent apporter. Dans l'exemple illustré sur la figure 4.1.2.6, nous avons souhaité évaluer l'intérêt d'une telle visualisation lors de l'interaction d'une protéine, présentant un feuillet β, avec d'autres partenaires macromoléculaires. Notre choix s'est naturellement porté sur une interaction protéine/ADN, la protéine choisie étant la « *TATA Binding Protein* ». La TBP se lie à l'ADN sur une séquence particulière : la *« TATA box »*. Nos modèles de Catmull-Rom et de Bézier, couplés au rendu de type squelette (figure 4.1.2.6d et 4.1.2.6e), permettent de véritablement visualiser la façon dont le feuillet β de la TBP vient s'enrouler sur la *TATA box*. Là encore, la possibilité de plaquer sur les surfaces une coloration qui serait issue des caractéristiques du partenaire (ici l'ADN) serait d'une grande nécessité.

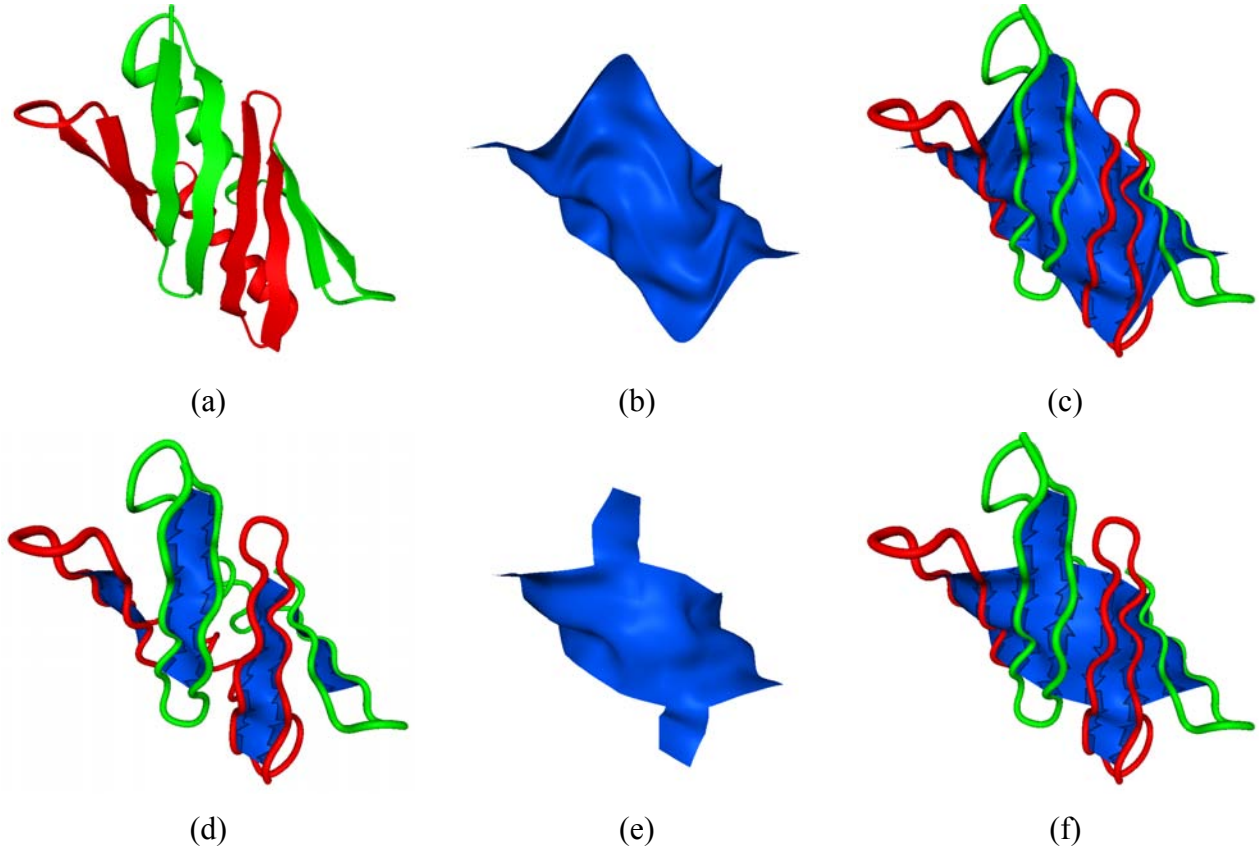


Figure 4.1.2.7 – Immunoglobulin G binding protein G [1Q10]. (a) Représentation de type cartoon*, (b) et (e) représentations respectivement de type Catmull-Rom et Bézier ; (c), (d) et (f) sont des représentations de Catmull-Rom et de Bézier, texturées par des flèches, couplées à des représentations squelette. Sur cette figure, les rendu de type* cartoon *et squelette sont coloré par chaîne*

Le dernier exemple que nous avons choisi de présenter concerne un fragment des immunoglobulines. Cette protéine représentée sur la figure 4.1.2.7 est composée de deux chaînes polypeptidiques, identiques, imbriquées l'une dans l'autre pour former un feuillet β à huit brins alternés deux à deux. En 4.1.2.7a nous avons représenté cette protéine en mode *cartoon*, avec une coloration par chaîne : la première chaîne est rouge, la seconde est verte. Nous constatons alors l'arrangement « deux par deux » des brins β.

Le modèle de Catmull-Rom représente parfaitement ce type de feuillet (figure 4.1.2.7b et 4.1.2.7c) étant donné que sa description dans le fichier PDB spécifie qu'il s'établit sur plusieurs chaînes. Par contre l'algorithme DSSP, tel qu'il est implanté dans BALLView, ne recherche pas d'éventuelles liaisons hydrogènes inter-chaînes. C'est pourquoi notre modèle de Bézier, au lieu de représenter un seul feuillet, représentait quatre feuillets de deux brins chacun comme nous pouvons

le voir en 4.1.2.7d. Nous avons alors modifié l'implémentation de DSSP afin de pouvoir appréhender des feuillets β établis sur la structure quaternaire des protéines, c'est à dire, sur l'association de chaînes polypeptidiques distinctes. Les résultats sont visibles en 4.1.2.7e et 4.1.2.7f.

Par ailleurs, pour valider notre approche, nous avons découpé dans deux fichiers distincts chacune des deux chaînes, celles-ci étant chargées dans Ballview comme deux molécules différentes. Nous avons pu alors retrouver le feuillet complet tel qu'il est présenté sur les figures ci-dessus. Cela est très important à plusieurs niveaux : cela signifie en premier lieu que notre modèle est capable de reconstruire des feuillets β que nous pourrions appeler quaternaires, soit au travers d'une interface structurale, soit, comme dans le cas présent, en raison d'une structuration tridimensionnelle fonctionnelle. Nous avons, par conséquent, testé la robustesse de notre approche dans le cadre de la prédiction d'interactions protéine-protéine qui se réalisent au travers de structurations β. Dans le cadre de l'étude, par exemple, de l'élafine, nous avons la présence d'un petit feuillet β à deux brins qui « gouverne » la boucle inhibitrice. Celle-ci s'avère non structurée lorsque l'inhibiteur est seul. Cependant dans le cadre du complexe formé avec l'élastase pancréatique de porc, nous observons la présence d'une structuration de la boucle inhibitrice en brin β qui vient se stabiliser avec le feuillet β qui gère la structuration du site actif de l'élastase. Notre modèle s'avère tout à fait pertinent pour représenter cet état de fait.

4.1.2.1 Application CATH

Nous nous sommes intéressés aux classes structurales de CATH dont les domaines contiennent des feuillets β : la classe tout β, la classe α/β et la classe α+β. Il existe vingt architectures différentes dans la classe tout β, et quatorze dans les classes α/β et α+β. Nous avons utilisé notre modèle sur l'ensemble de ces architectures, mais avons choisi de n'en représenter que quelques unes, observables sur les figures 4.1.2.1.1 et 4.1.2.1.2.

Sur ces figures, seul le modèle de Bézier est utilisé car les fichiers présents dans la base de données de CATH ne contiennent pas les informations sur les feuillets β nécessaires au calcul du modèle de Catmull-Rom.

Dans la première planche d'exemples, nous avons différents domaines de la classe tout β. Ainsi nous avons successivement les motifs « *solenoid* » , « *propellor* », « *aligned prism* », « *clam* » et enfin « *distorted sandwich* ». Décrire chacun d'entre eux ne présenterait certainement pas un grand intérêt. Cependant, il est évident, à la vision des modèles obtenus, que nous pouvons d'ores et déjà observer la nature de chacun des feuillets mis en place dans chacune des macromolécules. Ainsi pour la protéine décrite en figure 4.1.2.1.1a nous voyons sur la représentation en mode *cartoon* que les deux feuillets se font face. Notre modèle nous permet dans le cas présent d'obtenir deux surfaces dont la topologie spatiale conduit à deux morceaux d'un demi-cylindre. Nous pouvons constater sur les figures des architectures suivantes que notre modèle de Bézier est très dépendant de l'algorithme d'affectation de structures secondaires. La nécessité de pouvoir disposer d'un autre algorithme d'attribution dans BALLView semble impératif. Plusieurs algorithmes permettraient par ailleurs des comparaisons entre méthodes, qui pourraient s'avérer de première importance pour donner une signification aux zones de « fragilité » observées dans les feuillets.

Dans la seconde planche (figure 4.1.2.1.2), nous avons des exemples des classes α/β et α+β de CATH. En premier le motif « *2 layer sandwich* », le motif « *4 layer sandwich* », le motif « *αβ horseshoe* » et enfin le motif « *box* ». Au vue de ce qui a été présenté précédemment, il est normal que nous observions les feuillets en l'état sur les deux premières protéines. Sur la troisième figure, nous voyons nettement la forme en fer à cheval décrivant cette architecture. La dernière organisation spatiale du repliement, bien que complexe, est aisément mise en évidence grâce à notre modèle.

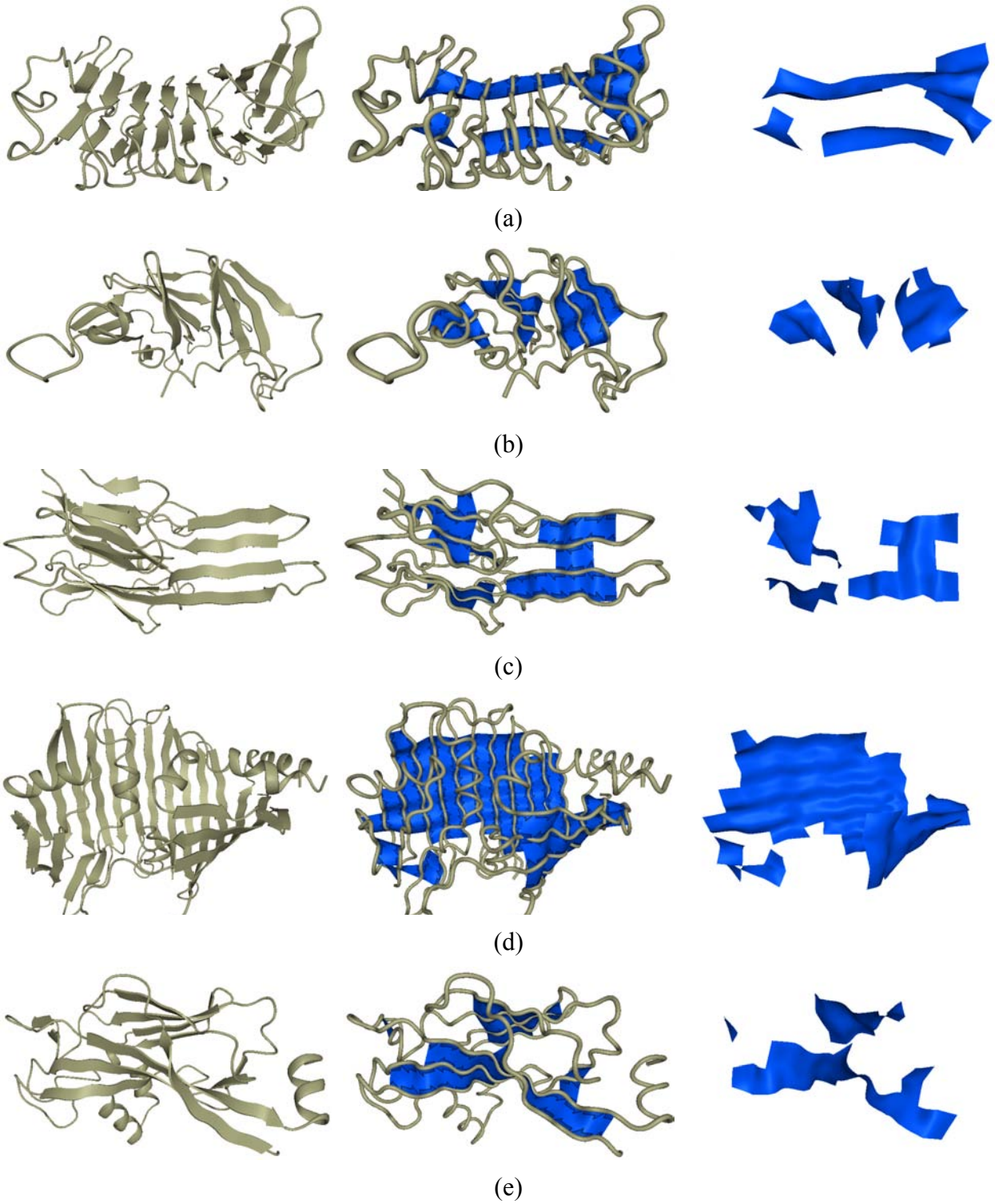


Figure 4.1.2.1.1 – Exemples de domaines de la classe tout β de CATH. (a) motif « 2 solenoid *» : 1k7iA01 ; (b) motif «* 3 propellor *» : 1n7vA01 ; (c) motif «* aligned prism *» : 1i5pA03 ; (d) motif «* clam *» : 4bclA00 ; (e) motif «* distorted sandwich *» : 1m3yA01. Chaque domaine est représenté par le modèle* cartoon*, le modèle de Bézier texturé, couplé à un modèle squelette, et le modèle de Bézier seul*

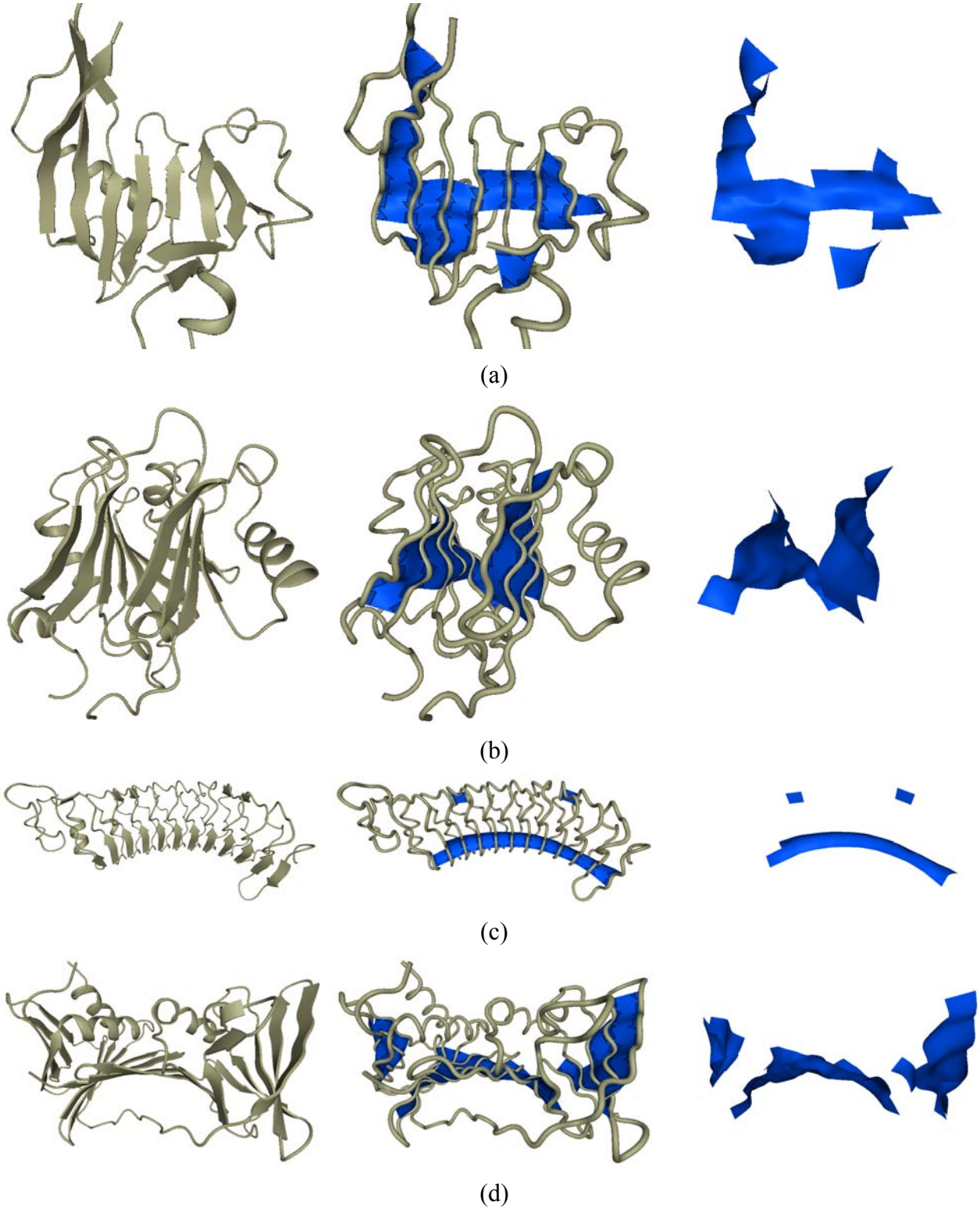
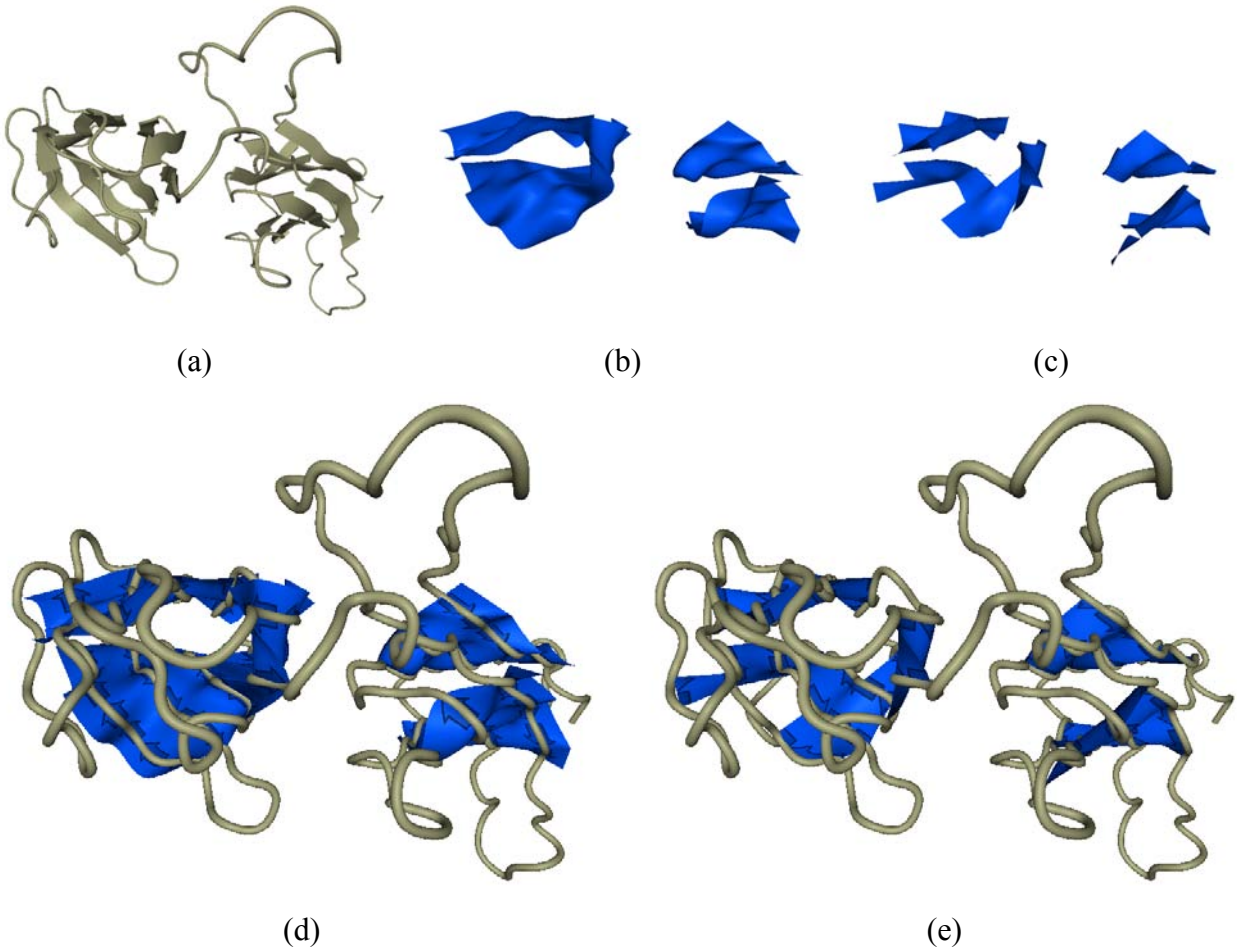


Figure 4.1.2.1.2 – Exemples de domaines des classes α/β et α+β de CATH. (a) motif « 2 layer sandwich » : 1c0pA02 ; (b) motif « 4 layer sandwich » : 1b25A01 ; (c) motif « αβ horseshoe » : 1oznA00 ; (d) motif « box » : 1t6lA00. Chaque domaine est représenté par le modèle cartoon, le modèle de Bézier texturé, couplé à un modèle squelette, et le modèle de Bézier seul

4.1.2.2 Application à la superfamille des immunoglobulines

Pour tester l'intérêt de nos modèles, nous avons recherché un modèle caractéristique des feuillets β. Pour cela, il nous fallait suffisamment de protéines présentant un motif relativement simple topologiquement et topographiquement. Nous avons alors pensé utiliser la banque de données décrite dans la thèse de Nicolas Prudhomme [Prudhomme2009] sur les domaines immunoglobulines. Elle présente cinquante-deux domaines de la banque d'Halaby et Mornon [Halaby1999] et vingt-cinq domaines de celle de Gerstein et Altman [Gerstein1995]. Les protéines qui présentent une identité de séquence supérieure à 70 % avec une autre ont été retirées de la banque. L'intérêt réside dans le fait que le taux d'identité étant faible, la structuration en feuillet β est prépondérante, les structures évoluant moins vites que les séquences. Un alignement multiple du domaine β a été effectué à l'aide du programme MUSTANG et est présenté dans la thèse de Nicolas Pruhomme sur la figure 1.12.3. Nous avons repris les informations contenues dans cette thèse pour visualiser les modifications structurales et séquentielles de chacun des brins.

Cependant, nous nous sommes heurtés à une faiblesse de notre mode de représentation. En effet, la superposition structurale des cinquante-six motifs immunoglobulines est tout à fait visible lorsque nous ne représentons que la trace des carbones alpha mis en jeu dans chacun des brins. Lorsque nous superposons les feuillets β correspondants, que ce soit avec une représentation de Catmull-Rom ou de Bézier, nous perdons toute information liée à l'alignement structural. Les solutions qui pourraient être apportées pour résoudre ce problème sont complexes. Nous pouvons penser qu'une visualisation successive de chaque feuillet, sous forme de film est envisageable. Si cette approche peut s'avérer intéressante pour distinguer des zones par coloration, elle sera cependant moins efficace. Le cas présent contenant cinquante-six protéines, il est quasiment impossible d'appréhender la totalité de ces modèles pour en tirer une information pertinente. Une autre approche pourrait être d'écrire, dans un fichier, les coordonnées du feuillet β que nous définissons pour chacune des protéines de l'alignement structural. Il faudrait ensuite, au moyen de méthodes graphiques, représenter le « feuillet β boudin », comme cela peut se visualiser pour la chaîne principale obtenue dans des données de RMN, pour représenter les zones de modifications structurales importantes. Là encore, l'information séquentielle portée par chacune des protéines ne sera pas d'une grande utilité et s'avérera inefficace à terme. Nous présentons cependant dans ce paragraphe quelques unes des structures immunoglobulines que nous avons retenues.



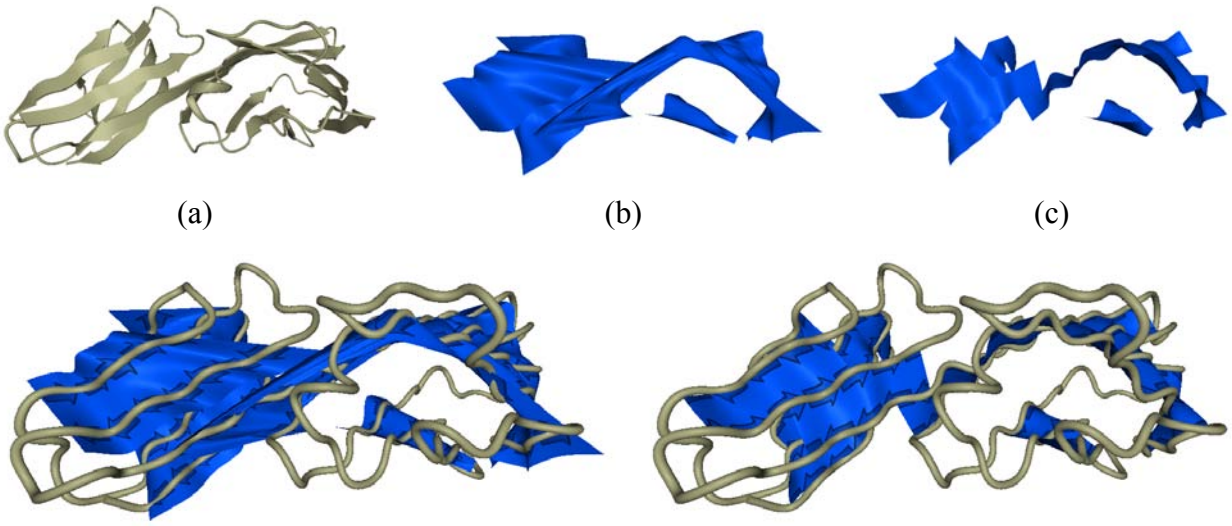
(a) (b) (c)

(d) (e)

Figure 4.1.2.2.1 – TCR ou « T Cell Receptor *» [1BEC]. (a) Représentation de type* cartoon*, (b) et (c) représentations respectivement de type Catmull-Rom et Bézier, (d) et (e) représentations respectivement de Catmull-Rom et de Bézier, texturées par des flèches, couplées à des représentations squelette*

Pour cette protéine, sur nos modèles de Catmull-Rom et de Bézier, nous distinguons très clairement la présence de deux couples de feuillets β se faisant face, ce sont deux plis immunoglobuline. Ce pli, caractéristique des immunoglobulines, fait partie des neuf *superfolds*.

En observant nos résultats nous pouvons donc affirmer que cette protéine fait partie de la classe structurale tout β, et que ses feuillets confirment son appartenance à la superfamille des immunoglobulines.



(a) (b) (c)

(d) (e)

Figure 4.1.2.2.2 – Le CD4, pour cluster *de différenciation 4 [1CID]. (a) Représentation de type* cartoon*, (b) et (c) représentations respectivement de type Catmull-Rom et Bézier, (d) et (e) représentations respectivement de Catmull-Rom et de Bézier, texturées par des flèches, couplées à des représentations squelette*

Le CD4 est une glycoprotéine qui s'exprime principalement à la surface des lymphocytes T CD4+. La description du fichier PDB fait état de quatre feuillets β. Les deux premiers, que nous pouvons observer à gauche sur nos illustrations, forment un pli immunoglobuline. Les deux derniers présentent un agencement étrange composé d'un assez petit feuillet de trois brins β, et d'un feuillet plus important qui se *twiste*.

Les représentations de ces feuillets sont fidèlement rendues sur le modèle de Catmull-Rom, mais le modèle de Bézier, lui, ne propose que trois feuillets β. Deux feuillets viennent s'associer pour n'en former qu'un seul, et chacune des faces de ce grand feuillet « regarde » un feuillet de moindre taille. Cela ressemble à la fusion, due à leur grande proximité, de deux plis immunoglobuline.

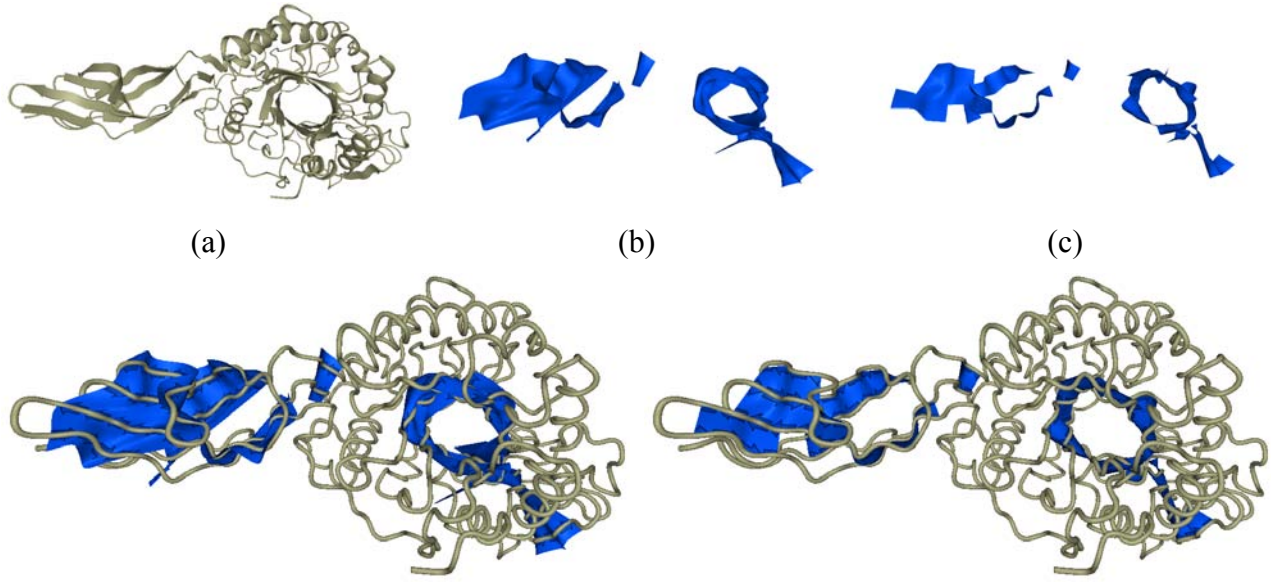


Figure 4.1.2.2.3 – Chitinase bactérienne [1CTN]. (a) Représentation de type cartoon*, (b) et (c) représentations respectivement de type Catmull-Rom et Bézier, (d) et (e) représentations respectivement de Catmull-Rom et de Bézier, texturées par des flèches, couplées à des représentations squelette*

Si nous observons le mode *cartoon*, en (a) sur notre figure, nous identifions un domaine composé de brins β sur la gauche de l'illustration, ainsi qu'un *TIM barrel* sur la droite. Cependant, l'utilisation de nos modèles démontre la présence d'un troisième domaine aux abords du *TIM barrel* : un pli αβ. Ce pli est difficilement visible sur la représentation *cartoon*, alors que nous le distinguons immédiatement sur nos représentations.

Nous préférerons l'utilisation de SheHeRASADe, aux types de représentations classiques, afin de détecter l'existence de domaines structuraux, car leur présence apparaît de façon claire et évidente.

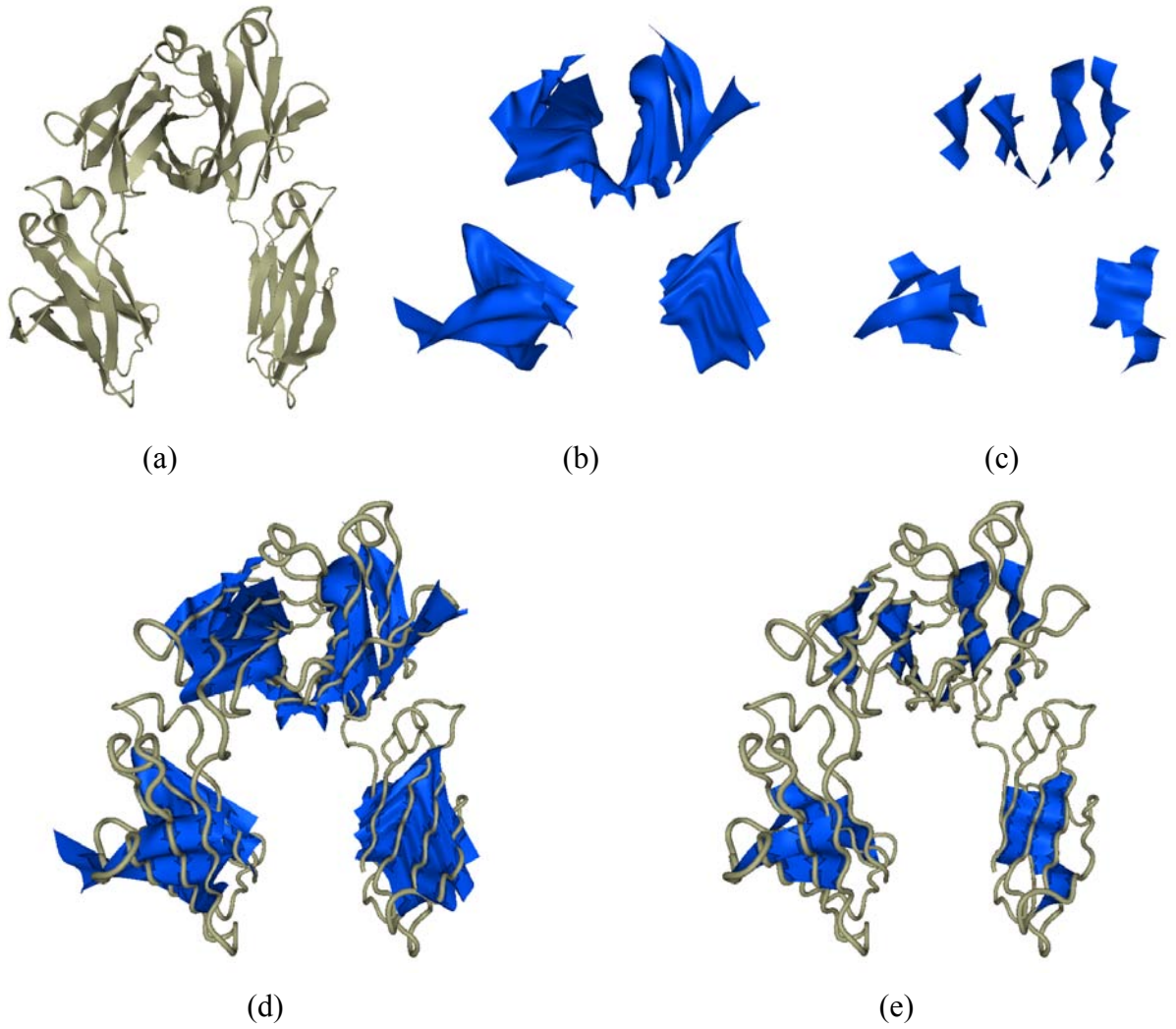


Figure 4.1.2.2.4 – Fragment Fc humain [1FC1]. (a) Représentation de type cartoon*, (b) et (c) représentations respectivement de type Catmull-Rom et Bézier, (d) et (e) représentations respectivement de Catmull-Rom et de Bézier, texturées par des flèches, couplées à des représentations squelette*

Le fragment Fc humain est un homodimère composé de quatre plis immunoglobuline. Les monomères s'associent en formant un complexe, très hydrophobe, de deux plis immunoglobuline. Dans le cas présent la structuration en fer à cheval se manifeste clairement au travers du dimère. Les feuillets se font face et jouent en regard les uns aux autres. Au delà du domaine pli β, il est à noter que ces structures tridimensionnelles du fragment Fc des immunoglobulines sont stabilisées par des glycannes multiples et complexes, dont les points de glycosylation sont en vis à vis des différents feuillets. Pouvoir projeter des paramètres géométriques ou physico-chimiques issus des glycannes, sur les feuillets pourraient certainement s'avérer très intéressant.

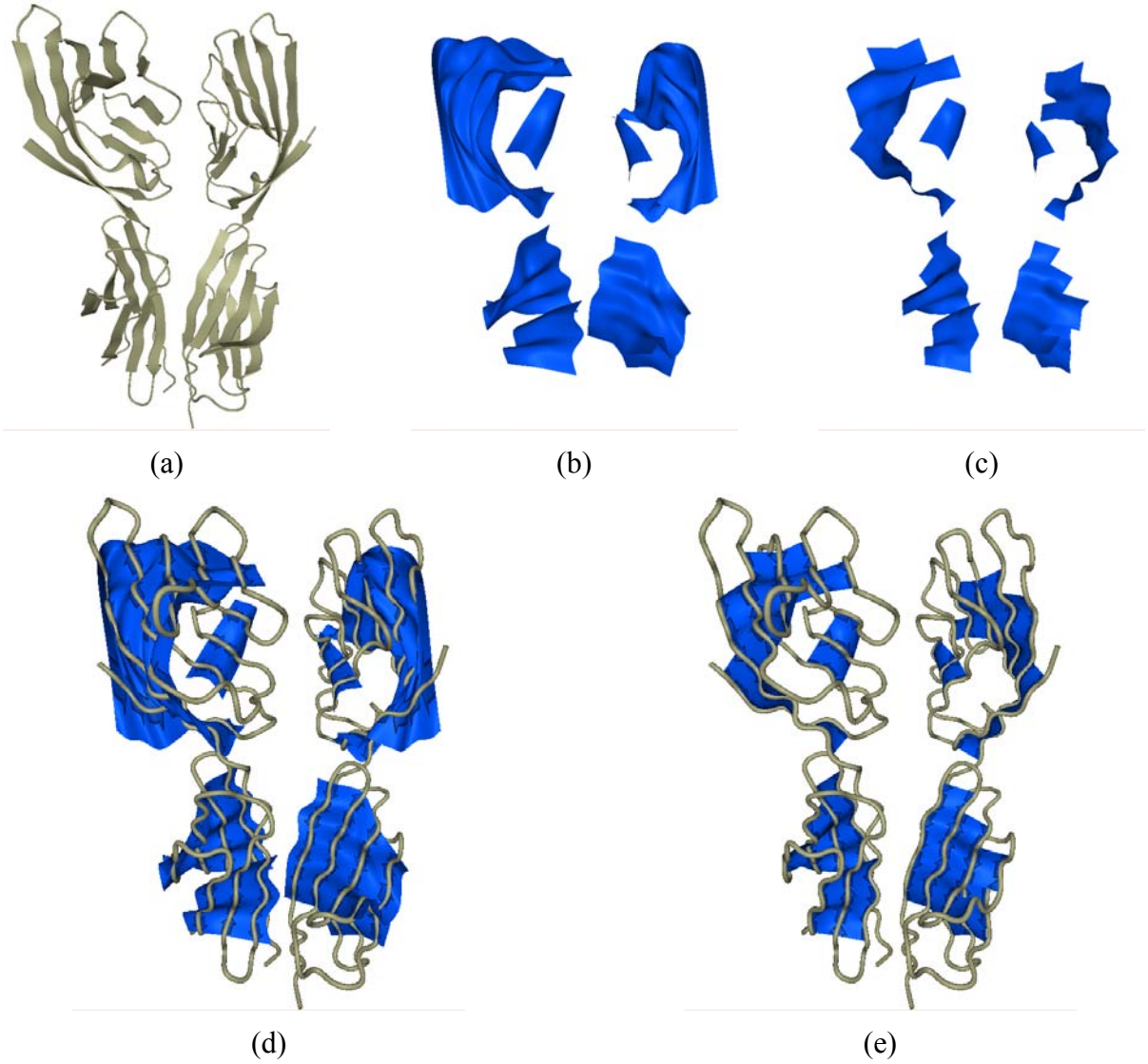
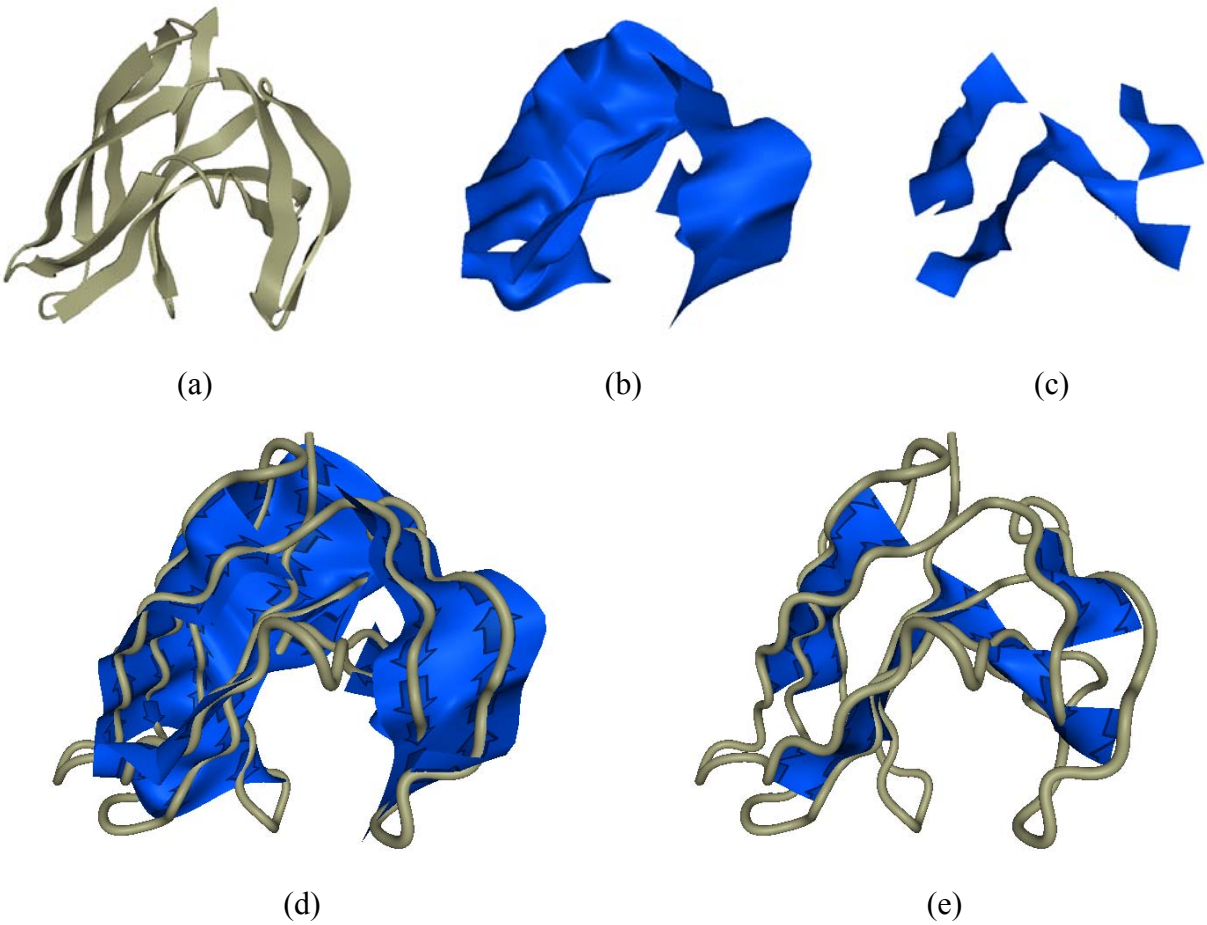


Figure 4.1.2.2.5 – Le CD2, pour cluster *de différenciation 2 [1HNG]. (a) Représentation de type* cartoon*, (b) et (c) représentations respectivement de type Catmull-Rom et Bézier, (d) et (e) représentations respectivement de Catmull-Rom et de Bézier, texturées par des flèches, couplées à des représentations squelette*

Le CD2 est une molécule d'adhésion cellulaire (CAM, ou « *Cell Adhesion Molecule* ») qui s'exprime à la surface des lymphocytes T et des cellules NK, ou « *Natural Killer* ». L'appartenance du CD2 à la superfamille des immunoglobulines, est due à la présence de deux plis immunoglobuline dans la partie extracellulaire de la protéine. Nos modèles (figure 4.1.2.2.5) permettent d'identifier clairement la partie extracellulaire comme étant la partie basse de nos illustrations, positions des plis caractéristiques de cette superfamille.



(a) (b) (c)

(d) (e)

Figure 4.1.2.2.6 – La macromycine [2MCM]. (a) Représentation de type cartoon*, (b) et (c) représentations respectivement de type Catmull-Rom et Bézier, (d) et (e) représentations respectivement de Catmull-Rom et de Bézier, texturées par des flèches, couplées à des représentations squelette*

La macromycine est composée d'un *β barrel* aplati, qui se situe sur la partie gauche des illustrations de la figure 4.1.2.6, et de deux feuillets β sur la droite. Cette structure, en forme de voûte, est un site de liaison au chromophore.

Sur le modèle de Catmull-Rom, en (b) et (d), l'observation de la voûte est rendue difficile par la description du *β barrel*. L'utilisation du modèle de Bézier, en (c) et (e), montre que le *β barrel* s'est scindé en deux feuillets, dont un a fusionné avec le feuillet central. Ce feuillet semble renforcé par les deux feuillets extérieurs. La voûte apparaît alors de façon très claire, et nous sommes à même de percevoir l'accessibilité de cette voûte, ce qui en fait un site de liaison privilégié.

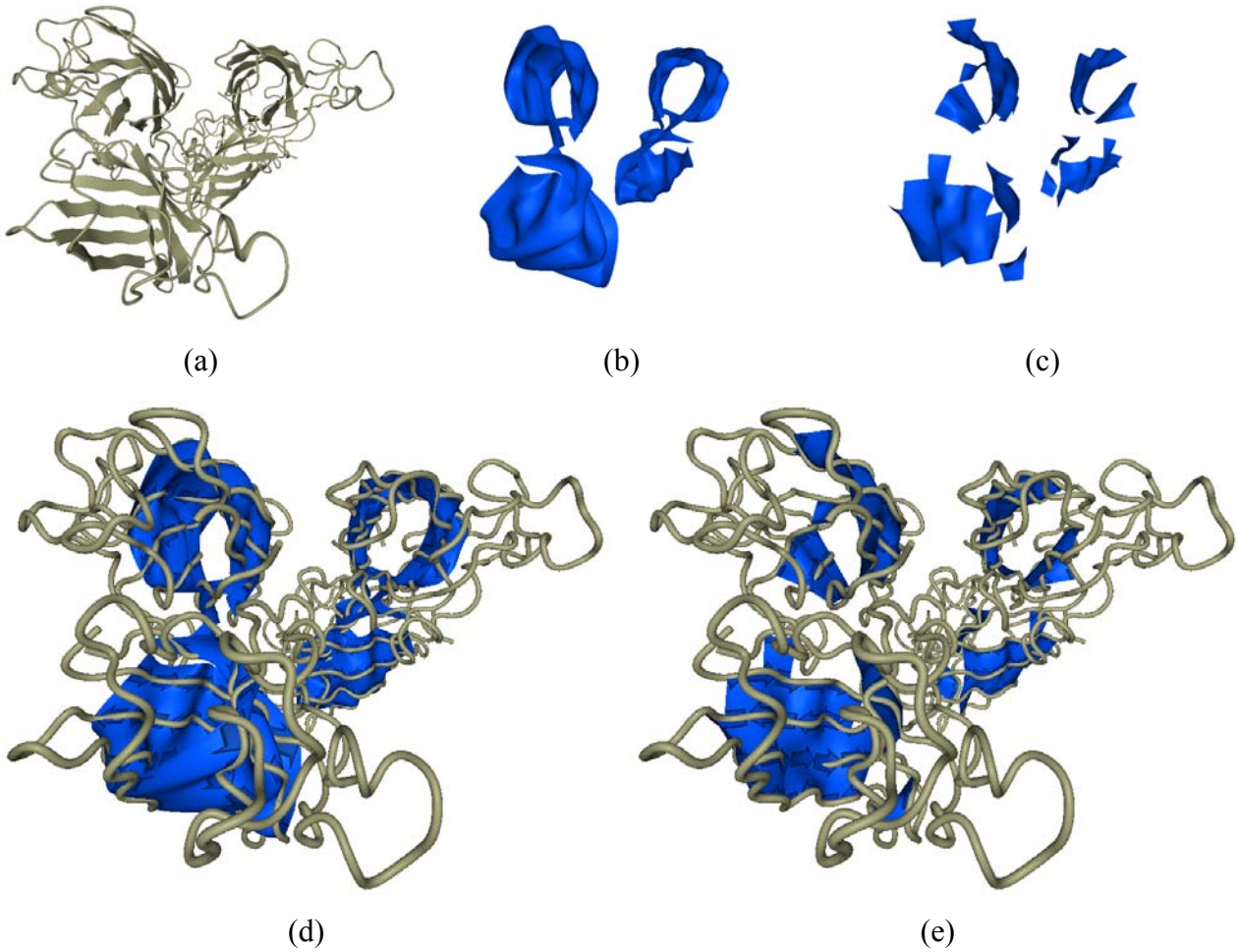


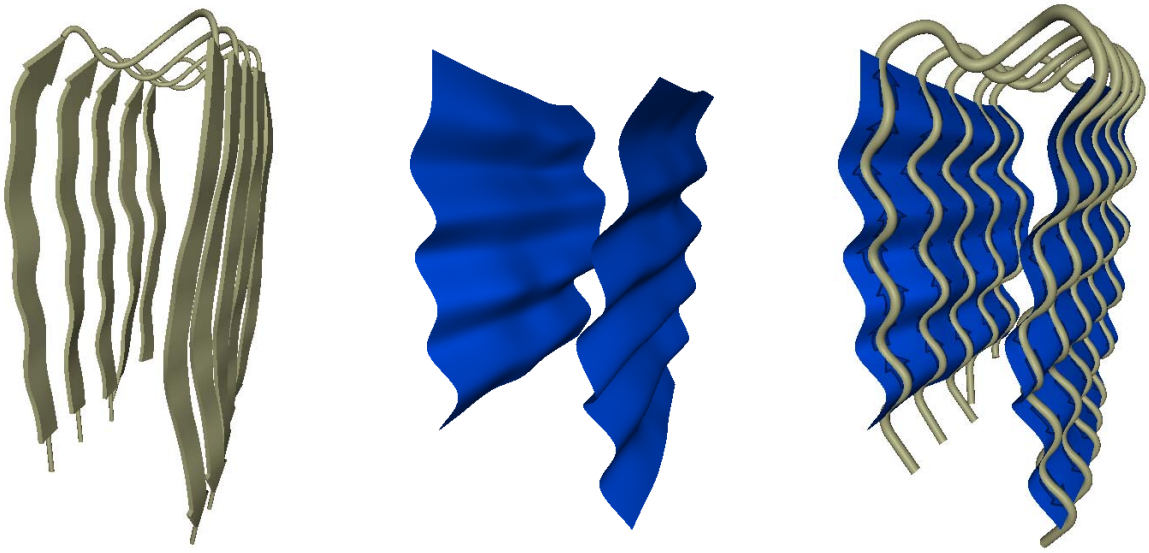
Figure 4.1.2.2.7 – La superoxyde dismutase [2SOD]. (a) Représentation de type cartoon*, (b) et (c) représentations respectivement de type Catmull-Rom et Bézier, (d) et (e) représentations respectivement de Catmull-Rom et de Bézier, texturées par des flèches, couplées à des représentations squelette*

La superoxyde dismutase est composée de quatre *β barrel*, ces structures sont visibles sur les illustrations du modèle de Catmull-Rom, en (b) et (d). L'utilisation du modèle de Bézier présente des feuillets β déstructurés, la structure de ces domaines serait donc instable.

4.2 Application sur les protéines amyloïdes

Les amyloïdoses sont des maladies à désordre conformationnel protéique. Si nous considérons une définition strictement médicale, les amyloïdoses sont un groupe de maladies caractérisées par un dépôt extracellulaire sous forme fibrillaire de matériels protéiques. Ces structures protéiques sont donc devenues insolubles pour des études expérimentales classiques, comme la cristallographie à rayons X et la RMN. En conséquence, les différents types de modélisation moléculaire et les simulations numériques, restent les méthodes privilégiées pour étudier ces objets. Nous verrons dans les quelques exemples suivants, que les modèles structuraux proposés reposent sur la répétition, par application de translation, ou d'opérations de symétrie, d'un même motif peptidique modèle. L'utilisation des modes de coloration s'avère alors d'une utilité limitée, et ne sera pas retenue. Cependant l'importance de ces pathologies dans les problématiques de santé publique, comme par exemple la protéine impliquée dans la maladie de Creutzfeld-Jakob ou le peptide A β responsable de la maladie d'Alzheimer, est de nos jours au centre de nombreux projets médicaux tendant à mieux appréhender les interactions moléculaires fines mises en jeu dans ces processus.

4.2.1 Motif minimum du peptide β 1-42 des amyloïdes



(a) (b) (c)

Figure 4.2.1.1 – Protéine amyloïde β A4 [2BEG]. (a) Représentation de type cartoon*, (b) représentations de type Catmull-Rom, (c) représentation de Catmull-Rom, texturée par des flèches, couplée à une représentation squelette*

La protéine amyloïde β A4, impliquée dans la maladie d'Alzheimer, présente deux feuillets β parallèles extrêmement réguliers qui ne se *twistent* pas. Sur la figure 4.1.2.2.7b, nous pouvons observer le résultat obtenu avec le modèle de Catmull-Rom : les deux feuillets sont parfaitement représentés et leur aspect plissé est très nettement visible.

Le modèle de Bézier n'est pas représenté car le résultat produit par DSSP ne fait état d'aucune structure secondaire, et *a fortiori*, d'aucun feuillet β. C'est une des limitations de cet algorithme.

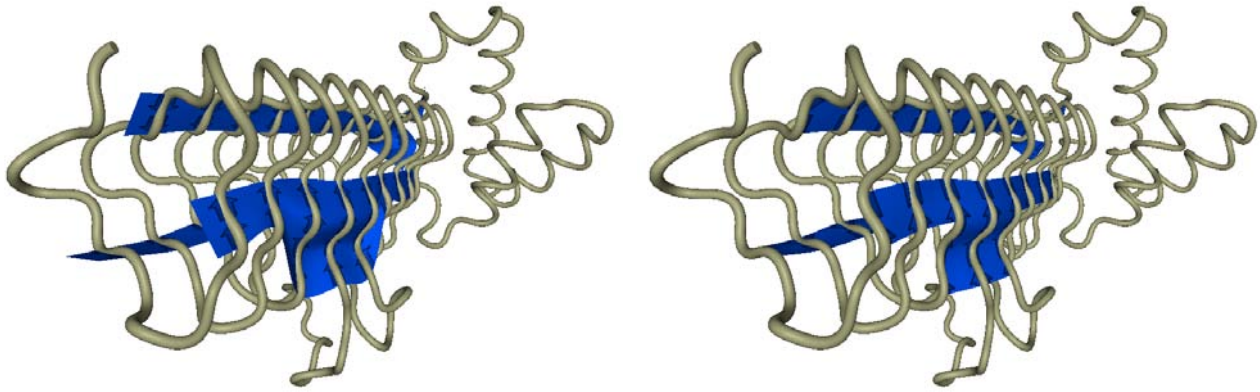
4.2.2 Les solenoïdes β

Récemment, Kajava et Steven [Kajava2006] ont publié une revue exhaustive sur les repliements β et l'aptitude de ces structures à produire des fibres amyloïdes. Parmi tous ces repliements, les solénoïdes β ont été particulièrement étudiés puisqu'ils sont exclusivement composés de brins β. L'agencement tridimensionnel de ces brins peut conduire à deux types d'enroulement solénoïdes qui sont soit à enroulement droit, soit à enroulement gauche.

Dans les deux planches d'exemples (figure 4.1.2.2.6 et 4.1.2.2.4), nous avons extrait, des nombreuses protéines avec un repliement solénoïde, les principales structures respectivement avec un enroulement gauche et un enroulement droit. De façon générale, les protéines à solénoïdes présentent des zones allongées, constituées de motifs répétés, vraiment différentes des feuillets observables dans des protéines globulaires. Ils sont constitués de segments de brins β avec des arcs β qui leurs succèdent, voire, dans certains cas, non pas des arcs mais de larges boucles qui donnent la spécificité du solénoïde. Il est également possible de constater la formation de liaisons hydrogène entre des brins β de parties différentes. De plus, les solénoïdes se différencient par des interactions inter-*coils* qui stabilisent les structures tridimensionnelles. Ces différentes interactions peuvent conduire à des formes spatiales caractéristiques de ces protéines.

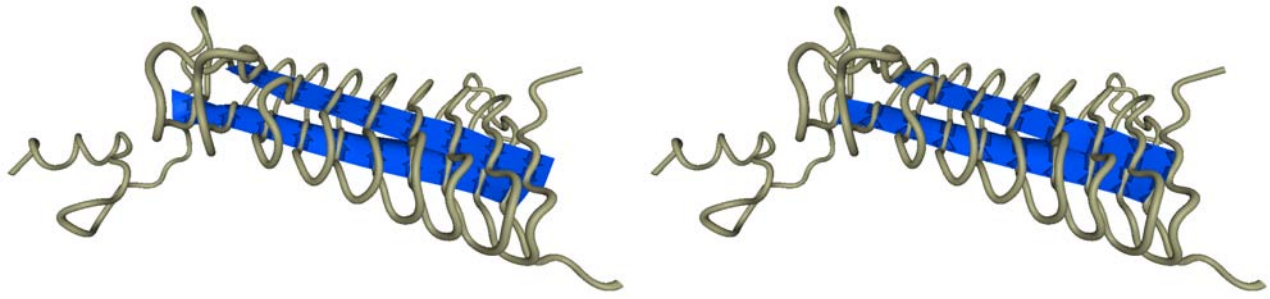
Ces caractéristiques, propres aux solénoïdes, se matérialisent avec nos modèles que ce soit sur la UDP-N-acetylglucosamine transférase [1J2Z], la protéine YadA [1P9H], pour les types gauches ou bien sur une pectate lyase C [1AIR], et sur la protéine SufD transporteur à fer [1VH4] pour les types droits. Dans chacun des cas, nous pouvons observer les interactions qui se mettent en place entre les feuillets, ainsi que les orientations relatives des feuillets entre eux. Notre modèle fait apparaître la complexité structurale nécessaire à l'établissement d'interactions entre plusieurs domaines. Cette étude nous prouve l'intérêt qu'il pourrait y avoir à projeter sur un feuillet β les informations séquentielles, structurales, ou autres, dans un environnement immédiat ou encore la projection mutuelle des paramètres sur des feuillets associés.

(a)



(b)

(c)



(d)

Figure 4.2.2.1 – Solénoïdes à enroulement gauche : (a) et (b) UDP-N-acetylglucosamine transférase [1J2Z] ; (c) et (d) la protéine YadA [1P9H]. Pour (a) et (c), l'ordre est le suivant : rendu de type cartoon*, rendu de Catmull-Rom, rendu de Bézier. Pour (b) et (d), l'ordre est le suivant : rendu de Catmull-Rom, texturé par des flèches, couplé à un rendu squelette, et rendu de Bézier, texturé par des flèches, couplé à un rendu squelette*

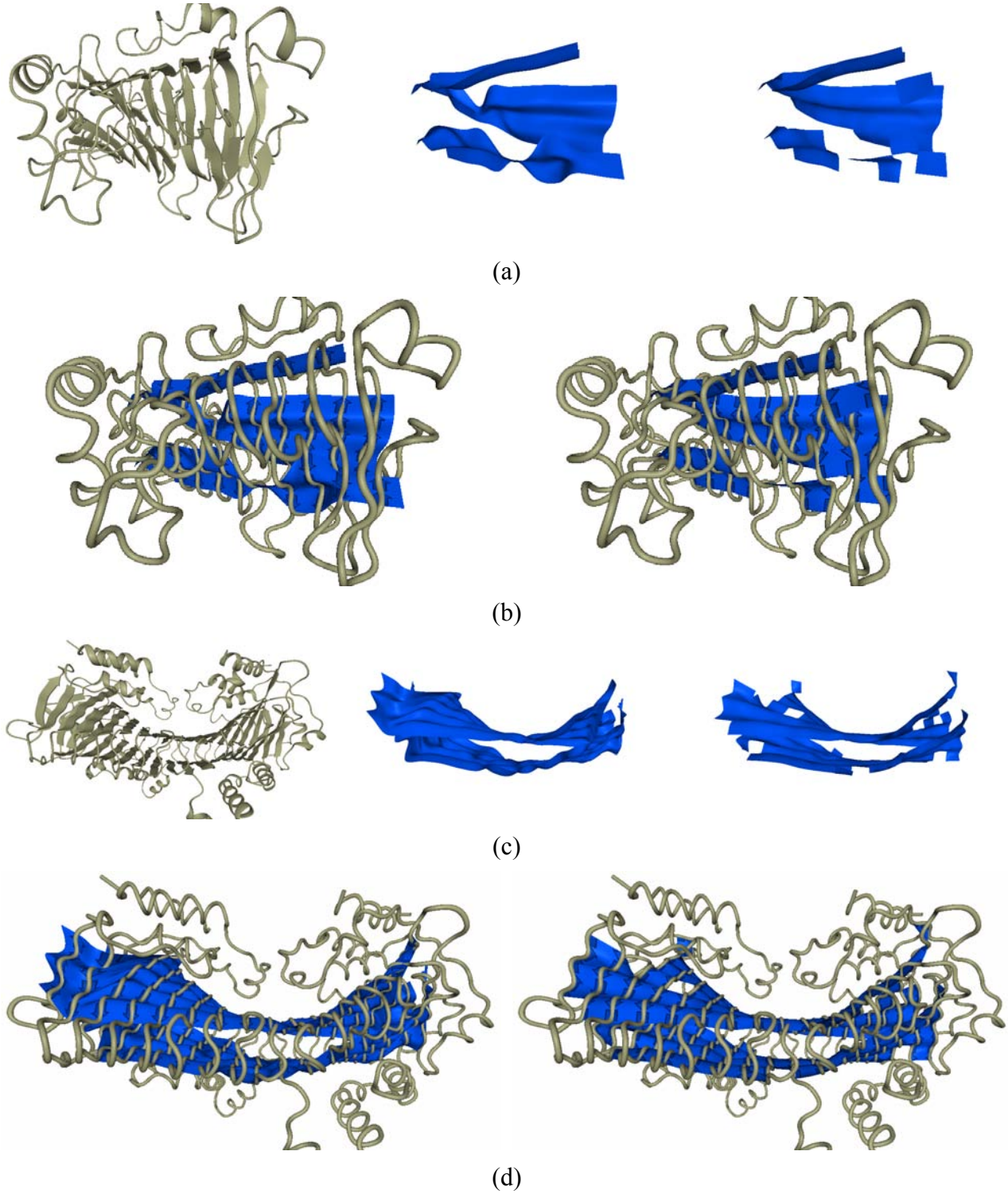
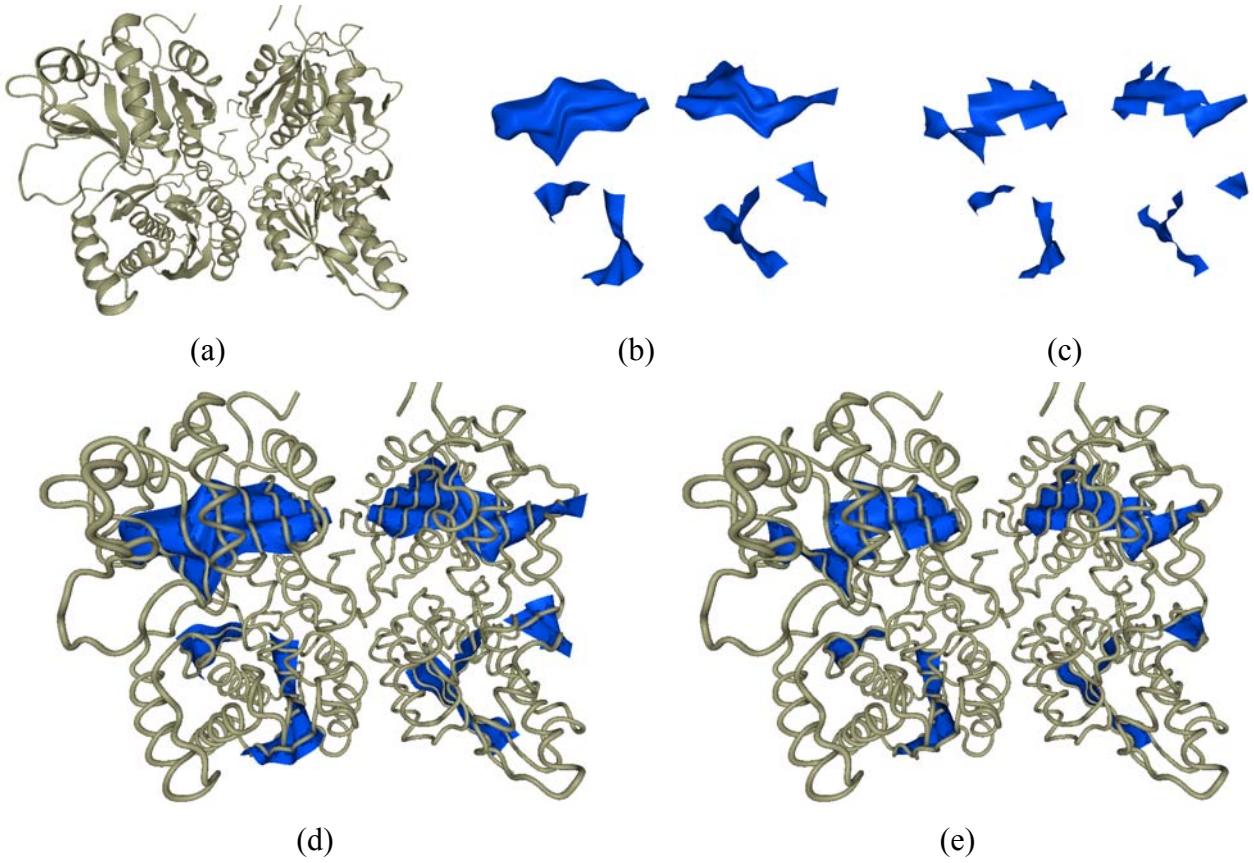


Figure 4.2.2.2 – Solénoïdes à enroulement droit : (a) et (b) représentations d'une pectate lyase C [1AIR] ; (c) et (d) représentations de la protéine SufD transporteur à fer [1VH4]. Pour (a) et (c), l'ordre est le suivant : rendu de type cartoon*, rendu de Catmull-Rom, rendu de Bézier. Pour (b) et (d), l'ordre est le suivant : rendu de Catmull-Rom, texturé par des flèches, couplé à un rendu squelette, et rendu de Bézier, texturé par des flèches, couplé à un rendu squelette*

4.2.3 AmyPDB

Pour approfondir l'intérêt des modes de visualisation que nous proposons, nous avons, après les quarante-cinq modèles de solénoïdes différents qui ont été identifiés, étudié les différentes structures tridimensionnelles recensées dans la banque de données AmyPDB maintenue à l'Université de Rennes. Nous avons observé plus d'une centaine de ces protéines, et nous présentons ici quelques exemples afin de démontrer les potentialités de nos modèles.



(a) (b) (c)

(d) (e)

Figure 4.2.3.1 – Famille amyloïde des ANF (« Atrial Natriuretic Factor *») [1T34]. (a) Représentation de type* cartoon*, (b) et (c) représentations respectivement de type Catmull-Rom et Bézier, (d) et (e) représentations respectivement de Catmull-Rom et de Bézier, texturées par des flèches, couplées à des représentations squelette*

Cette famille de protéines présente l'originalité d'une organisation quaternaire quasiment symétrique, conduisant à la visualisation de plusieurs feuillets β qui s'organisent deux par deux, et qui, pour des questions de structuration spatiale, sont orientés perpendiculairement. Il apparaît en observant la figure 4.1.2.2.3 que les feuillets perpendiculaires à la représentation proposée sont relativement différents, et présentent une « déstructuration » β locale. Il semble, en lisant la bibliographie associée, que cette zone soit mise en jeu dans la restructuration de la protéine.

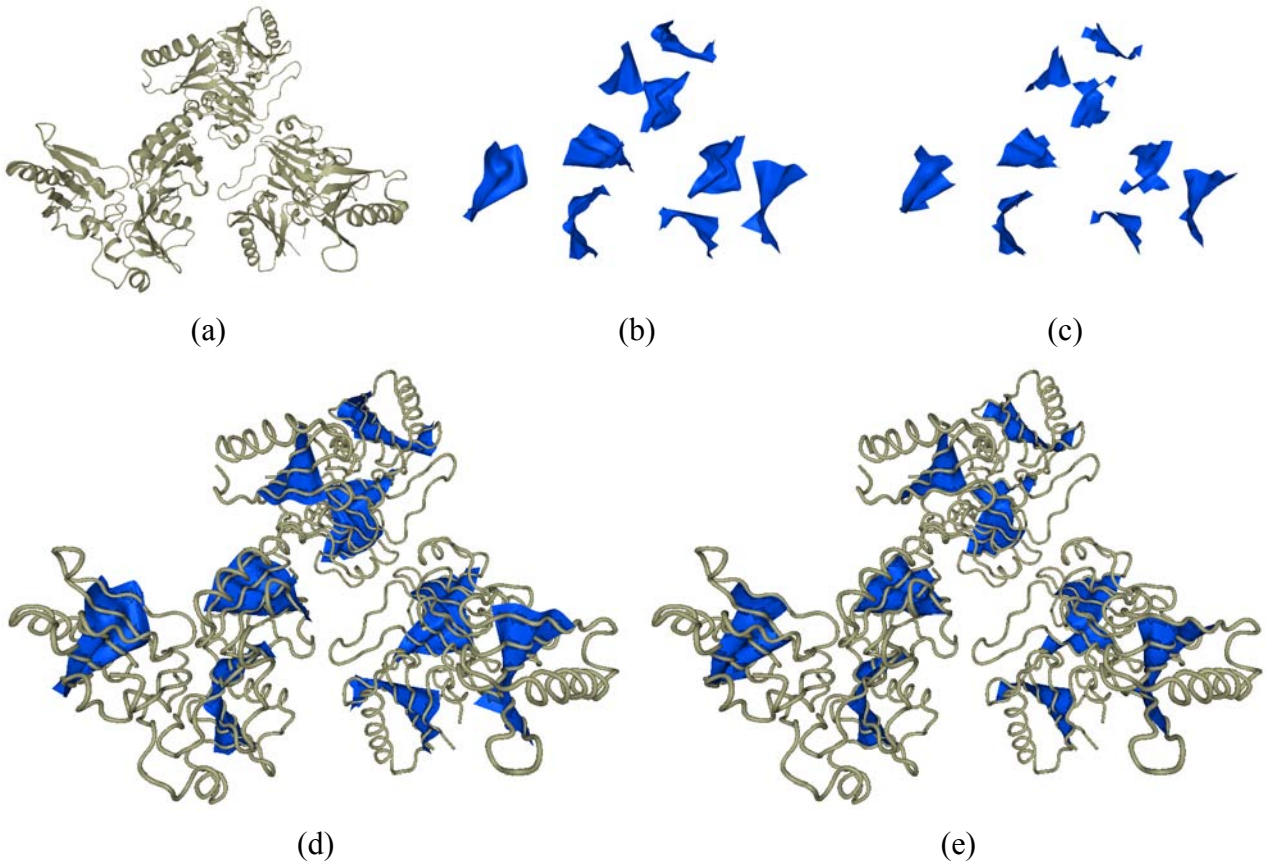
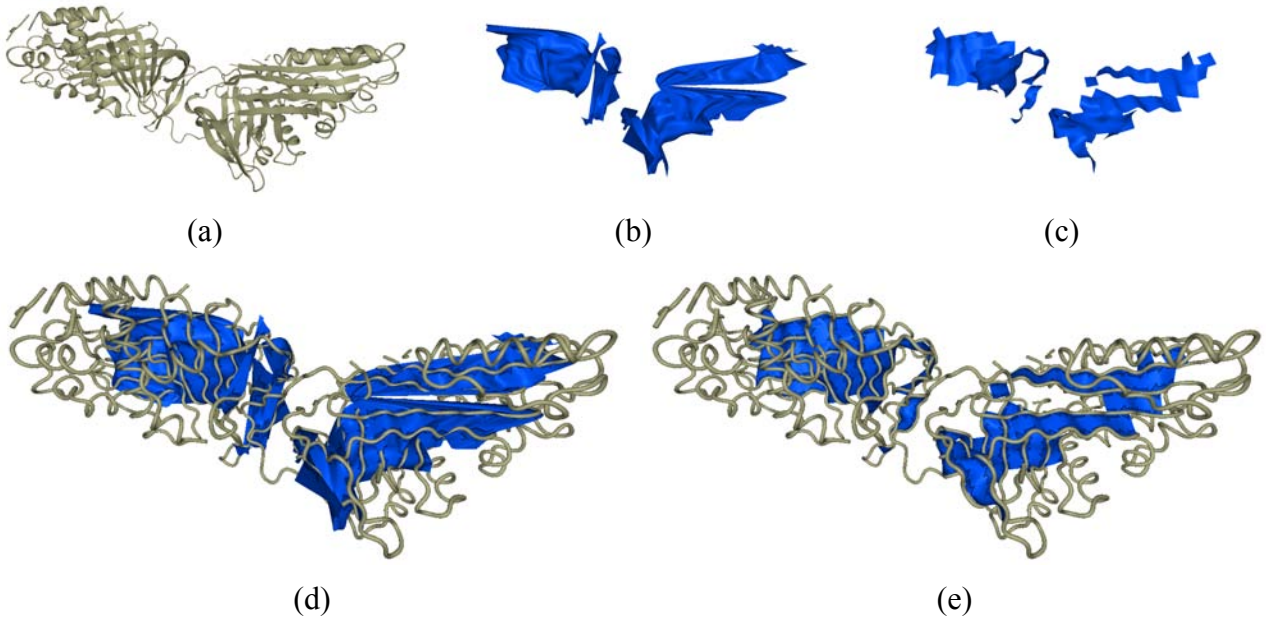


Figure 4.2.3.2 – Famille amyloïde des gelsolines [1P8X]. (a) Représentation de type cartoon*, (b) et (c) représentations respectivement de type Catmull-Rom et Bézier, (d) et (e) représentations respectivement de Catmull-Rom et de Bézier, texturées par des flèches, couplées à des représentations squelette*

Dans cette protéine nous pouvons observer l'organisation trimérique. Il s'avère que cette macromolécule s'organise spatialement avec les trois domaines nécessaires à son fonctionnement dans la capture de F-actine. Cette représentation montre à la fois les caractéristiques associées à la symétrie, mais également les spécificités locales que nous pouvons ponctuellement observer de l'une à l'autre des parties. De plus, la vision globale de l'ensemble des feuillets permet de mieux envisager le mode de capture qui pourrait s'effectuer si nous imaginons la modulation structurale de la macromolécule.



(a) (b) (c)

(d) (e)

Figure 4.2.3.3 – L'antithrombine humaine III de la famille amyloïde des serpines [1ATH]. (a) Représentation de type cartoon*, (b) et (c) représentations respectivement de type Catmull-Rom et Bézier, (d) et (e) représentations respectivement de Catmull-Rom et de Bézier, texturées par des flèches, couplées à des représentations squelettes.*

L'antithrombine est composée de deux grands domaines comprenant des feuillets β, nous remarquons sur le modèle de Catmull-Rom, sur la figure 4.1.2.2.2b et 4.1.2.2.2d, une profonde invagination du feuillet situé sur la droite de l'illustration. Cette invagination est due à la présence d'un petit brin β de seulement deux acides aminés. Sur le modèle *cartoon* sa présence est difficilement détectable, contrairement au modèle de Catmull-Rom qui montre clairement la particularité de ce brin. Le modèle de Bézier, quand à lui, révèle que la présence de ce brin est à l'origine d'une déchirure importante du feuillet, comme nous pouvons le constater en (c) et (e).

Ce brin d'une importance toute relative au regard du modèle *cartoon*, se révèle être d'une importance capitale pour la structure, et donc la fonction de cette protéine, avec l'utilisation de SheHeRASADe.

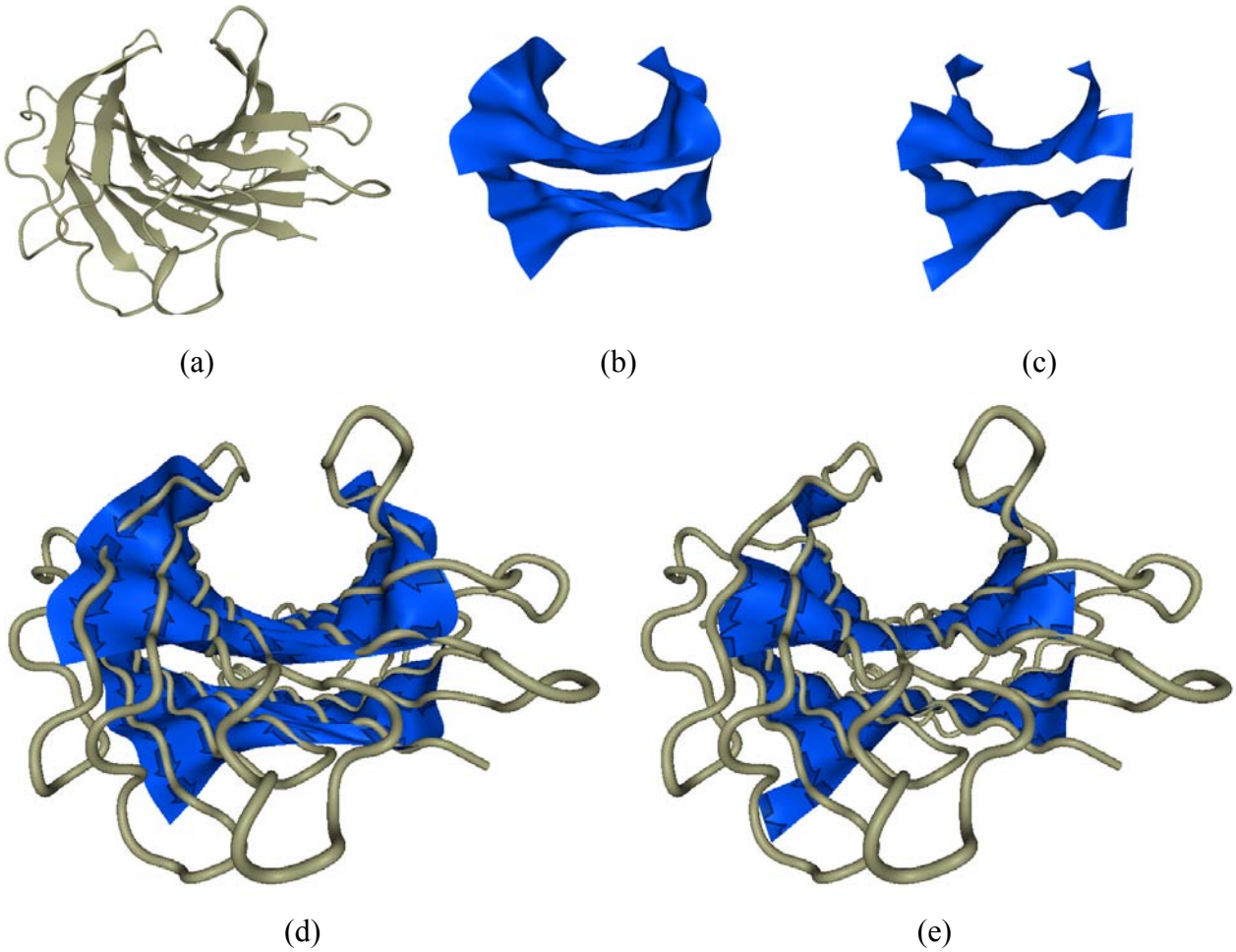
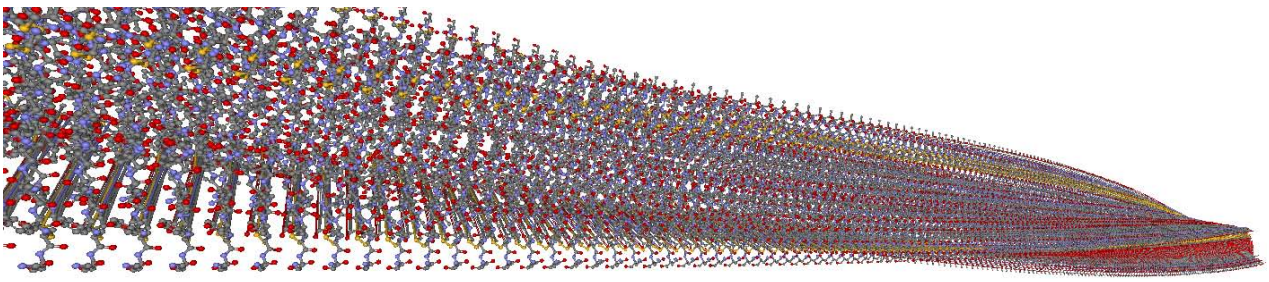


Figure 4.2.3.4 – Famille amyloïde des transthyrétines [1TFP]. (a) Représentation de type cartoon, *(b) et (c) représentations respectivement de type Catmull-Rom et Bézier, (d) et (e) représentations respectivement de Catmull-Rom et de Bézier, texturées par des flèches, couplées à des représentations squelette*

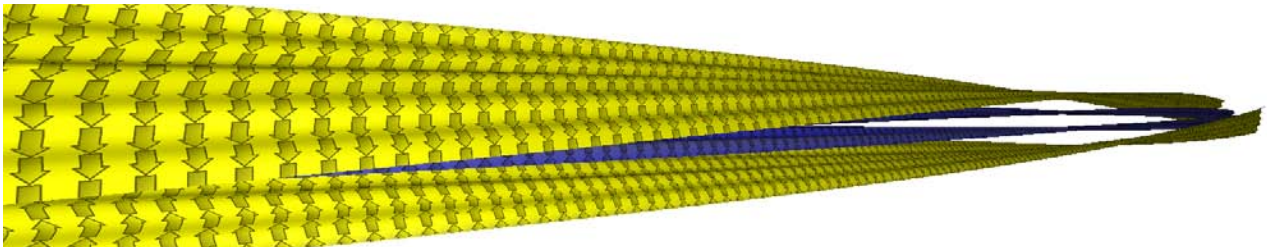
La transthyrétine est une protéine tout β qui est le principal transporteur de la thyroxine et de la vitamine A. Nous devinons aisément sa fonction de transporteur à la forme du feuillet β qui forme un demi-tube. L'utilisation de texture prend tout son sens et permet de mieux comprendre l'importance du feuillet β dans la structure tridimensionnelle complète.

4.2.4 Fibres amyloïdes

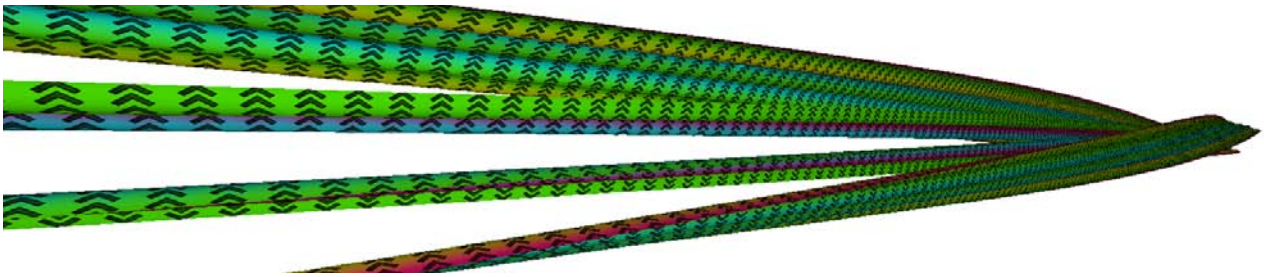
Pour finaliser l'intérêt de nos modes de visualisation sur les protéines amyloïdes, nous avons utilisé un modèle hypothétique, organisé en fibres amyloïdes, qui présente un millier de répétitions d'un même motif du peptide Aβ de l'amyloïde. Dans la figure 4.1.2.2.1a nous avons représenté le modèle tout atome de cette fibre. S'il est évident que la fibre présente une structure *twistée*, il est cependant difficile d'appréhender sa structuration spatiale. Les figures 4.1.2.2.1b et 4.1.2.2.1c présentent le modèle de Bézier, sur lequel figurent respectivement les flèches attachés à chaque amino-acide, et la texture en chevrons. Il est évident dans les deux cas que la nature de la fibre est tout à fait caractéristique et visible. La nature « simplificatrice » du feuillet permet d'augmenter la compréhension structurale précisément en visualisant l'organisation de chacun des « planchers β » les uns par rapport aux autres et leurs orientations relatives.



(a)



(b)



(c)

Figure 4.2.4.1 – Modèle hypothétique d'une fibre amyloïde obtenues par applications de paramètres de symétrie et de translation sur un double peptide amyloïde Aβ. (a) Représentation en mode boules-bâtons ; (b) représentation de type Bézier texturé par des flèches ; (c) représentation de Bézier texturée par des chevrons

Dans ce chapitre, nous avons évalué les différentes potentialités de SheHeRASADe tel que nous l'avons implémenté dans BALLView. Dans la première partie, nous avons étudié l'importance de pouvoir visualiser un feuillet β sous forme de « plancher » dans des protéines globulaires manifestant des structures secondaires, tertiaires ou quaternaires différentes. Les potentialités du modèle et la plus-value apportée, nous ont conduit à l'exploration des repliements majeurs de la banque de données CATH avec les classe tout β, α/β et α+β. Ensuite, nous avons focalisé notre attention sur un motif tout β, motif structural particulier observé dans le domaine des immunoglobulines.

La deuxième partie de ce chapitre applicatif s'est centrée sur des protéines amyloïdogéniques dont les rôles dans de nombreuses pathologies en font un sujet de préoccupation important, avec des conséquence majeures en santé publique. Après avoir étudié le peptide amyloïde Aβ, nous avons étudié les différents modèles de protéines à solénoïdes, puis la base de données AmyPDB pour enfin terminer par la représentation d'une fibre amyloïde hypothétique présentant un millier de motifs d'un peptide Aβ.

Au total, nous avons étudié environ 1750 protéines pour valider nos modèles de Catmull-Rom et de Bézier. Nous avons pris le parti de ne présenter que certains d'entre eux, sans rentrer dans la plus-value des différents modes de coloration qui auraient pu s'appliquer dans chacun des cas.

Chapitre 5
Conclusion et Perspectives

« Kids, you tried your best and you failed miserably. The lesson is: never try! »

Homer J. SIMPSON

5.1 Bilan et conclusion

L'objectif de ces travaux concernait la mise en place de nouveaux modes de visualisation d'éléments de la structure secondaire des protéines que sont les feuillets β. Jusqu'à présent, les méthodes consacrées se bornaient à représenter uniquement leurs éléments constitutifs, les brins β, qui par leur succession et leur orientation définissent les feuillets associés. C'est pourquoi nous avons mis en place une représentation surfacique modélisant les feuillets dans leur ensemble et non plus uniquement leurs brins. L'intérêt de cette approche a consisté en une visualisation à la fois simplifiée et permettant d'appréhender bien d'autres données comme la topologie et la topographie du feuillet dans son environnement structural. Dans ce but, deux approches différentes ont été expérimentées.

La première se fonde uniquement sur la définition des feuillets β présents dans les fichiers issus de la PDB, et utilise une interpolation bidimensionnelle basées sur les splines de Catmull-Rom. Ce modèle a été le premier, à notre connaissance, à représenter un feuillet β dans sa globalité. Contrairement aux modèles classiques listés très récemment par O'Donoghue et al. [O'Donoghue2010], son utilisation permet immédiatement de visualiser, de dénombrer, d'apprécier les formes ainsi que les topologies et topographies des feuillets β présents dans une protéine.

De plus, l'aspect plissé des feuillets β est alors clairement visible. Cependant, étant donné que le calcul de ces surfaces se fait à partir des données présentes dans les fichier issus de la PDB, et que ces données ne mentionnent que des informations sur la séquence de la protéine, nous ne sommes pas en mesure de représenter d'éventuelles déchirures, des bords crénelés ou des trous au sein d'un feuillet β.

La seconde approche utilise un algorithme d'attribution de structures secondaires afin de déterminer les liaisons hydrogène, l'algorithme utilisé jusqu'à présent est celui implémenté dans BALLView. De cette façon, notre modèle ne nécessite pas, pour représenter une protéine, que le fichier de données soit issu de la PDB mais uniquement qu'il utilise le formalisme PDB. Les résidus en conformation β impliqués dans des liaisons hydrogène sont utilisés pour créer des carreaux de Bézier. C'est l'association de ces carreaux qui forme la surface de notre représentation des feuillets. Ce modèle permet de matérialiser une surface uniquement entre les résidus effectivement liés, et ainsi de visualiser les déchirures, les bords plus ou moins lisses, les trous ainsi que les invaginations.

Plusieurs outils ont été développés autour de ces modèles afin de compléter leur intérêt scientifique. Parmi ces outils, plusieurs modes de coloration ont été mis au point. Ces modes ont été créés pour être utilisés avec nos modèles de feuillets β, mais ils sont toutefois exploitables pour tous les autres modèles disponibles dans BALLView (à l'exception du mode représentant la stabilité par zone d'un feuillet) et de plus, il est possible de les combiner pour augmenter l'information pertinente pour l'étude (par exemple coloration des résidus aromatiques sur le squelette, couplée à une coloration en surface par paramètre de facteur B issus de la PDB ou de paramètres physico-chimiques atomiques externes). L'utilité de ces modes, comme HCA, paramètres de facteur B, paramètres MHP ou tout autre type de paramètres que l'utilisateur souhaiterait entrer, personnalisables à volonté a été démontrée à travers divers exemples.

Nous sommes également à même de texturer nos surfaces afin de représenter des informations telles que les positions des acides-aminés, avec une flèche par amino-acide, ainsi que le sens des brins β au moyen de chevrons. Les nombreux exemples donnés tout au long de ce manuscrit permettent de constater l'intérêt évident de telles représentations. En effet, le sens d'un brin n'est jamais aussi évident que lorsqu'il est signifié tout au long de celui-ci, et cela permet localement de voir directement sur les feuillets s'ils sont parallèles ou anti-parallèles.

Afin de vérifier le rôle d'un ou de plusieurs feuillets β au sein d'un groupe formant une super structure secondaire, nous avons offert la possibilité d'étendre nos surfaces dans une direction souhaitée. Cette option peut se montrer très intéressante dans la compréhension de l'organisation spatiale des feuillets β au sein des super-structures secondaires mais aussi, par exemple, pour observer l'orientation des *turns* ou des *coils* au bout d'un feuillet (leur position en fonction de leur composition).

Nos modèles de visualisation des feuillets β, ainsi que les outils développés ont été intégrés au logiciel de modélisation moléculaire BALLView et sont pleinement utilisables *via* l'interface de ce dernier. Nos modèles sont opérationnels aussi bien pour une visualisation statique que dynamique. En effet, utilisant le modèle de Bézier permettant la visualisation de trous et de déchirures, nous avons implémenté notre modèle sur la lecture d'un fichier « dcd » issus d'une trajectoire de dynamique moléculaire.

Un temps important a été consacré à la finalisation de ces développements afin de pouvoir les utiliser pleinement. Des milliers de fichiers (entre 1500 et 2000) ont été testés (issus de la PDB ou non), et l'ensemble des exceptions ou des défauts d'écriture constatés ont permis de rendre nos méthodes plus robustes.

5.2 Perspectives

Parmi les perspectives envisagées, celle qui apparaît la plus importante concerne certainement l'algorithme d'affectation de structures secondaires. L'algorithme utilisé est celui implémenté dans BALLView : DSSP. Il présente cependant plusieurs inconvénients. Tout d'abord il possède une certaine propension à déterminer ce que nous pouvons qualifier de micro-feuillets β. Ce sont des feuillets composé de deux brins de deux résidus chacun. Ces micro-feuillets parasitent la bonne observation des protéines, même s'il est très simple de ne pas les représenter en utilisant l'interface dédiée aux feuillets que nous avons développé sous BALLView. Ensuite, il amène à une « déstructuration » locale des feuillets dès que les critères géométriques et énergétiques d'établissement des liaisons hydrogène ne sont pas respectés, même légèrement. L'algorithme de substitution pourrait-être STRIDE [Heinig2004]. La différence majeure entre STRIDE et DSSP est que STRIDE prend non seulement en compte les liaisons hydrogène, mais également la géométrie du squelette carboné de la protéine. STRIDE semble être, à ce jour, l'algorithme le plus performant pour prédire la structure secondaire d'une protéine. Le code source est disponible librement sur le

serveur FTP du European Bioinformatics Institute[2]. De plus, l'utilisation de plusieurs algorithmes d'assignation pourrait s'avérer très intéressante, non pas de façon absolue pour chaque algorithme, mais comme élément de comparaison des feuillets β ainsi déterminés. En effet, les premiers essais effectués avec STRIDE et comparés avec les résultats de DSSP sont significatifs : les zones de « fragilité » des feuillets β sont mises en évidence, et les quelques exemples que nous avons pu tester ont permis d'observer la contribution environnementale des éléments structuraux concernés.

Une perspective intéressante serait également de pouvoir projeter des informations provenant des structures voisines aux feuillets β, telles des chaînes latérales des acides aminés proches, des potentiels hydrophobes, ou tout autre, directement sur nos surfaces. Deux niveaux de difficulté sont envisageables. La projection perpendiculaire des chaînes latérales des amino-acides du feuillet en utilisant notre représentation est tout à fait concevable. Nous pourrions ainsi obtenir une information supplémentaire en observant la projection de la chaîne latérale et « son amplitude de couverture ». Qui plus est, il serait possible de projeter ces informations sur les deux faces du feuillet. De façon moins évidente, la projection géométrique d'informations structurales environnantes nécessite que nous puissions extraire un plan de nos feuillets, ce qui peut être problématique si le feuillet se replie sur lui même. Dans ce cas nous pourrions utiliser des définitions locales pour nos plans, et ainsi pouvoir réaliser les projections. Une telle réalisation pourrait permettre de constater l'organisation des structures secondaires entre elles, et que la manière dont ces structures s'arrangent n'est pas due au hasard. Au delà de ces informations d'environnement, nous pourrions ajouter des données concernant les caractéristiques géométriques, comme des angles entre brins au sein du feuillet, des paramètres torsionnels ou des paramètres caractérisant des angles entre deux feuillets dans un ensemble de structures secondaires.

D'autres informations, telles que les positions des liaisons hydrogène, pourraient être représentées sur nos surfaces. En effet, nous connaissons l'importance de ces liaisons dans la formation des feuillets β. Pouvoir les représenter est très important, même si actuellement il est possible « d'appréhender » la force de ces liaisons hydrogène sur le modèle de Bézier étant donné qu'il est possible d'observer des trous ou des déchirures. La distance séparant deux Cα consécutifs étant statistiquement conservée, seules les modifications des distances entre les brins entraînent la formation de trous et donc qualifient les liaisons hydrogène. Nous pouvons imaginer utiliser des représentations sous forme de ressorts : plus les éléments formant la liaison hydrogène seraient

2 http://webclu.bio.wzw.tum.de/stride/

éloignés, plus le ressort serait distendu, ou pourraient être appliquées sur la surface avec un gradient de couleur.

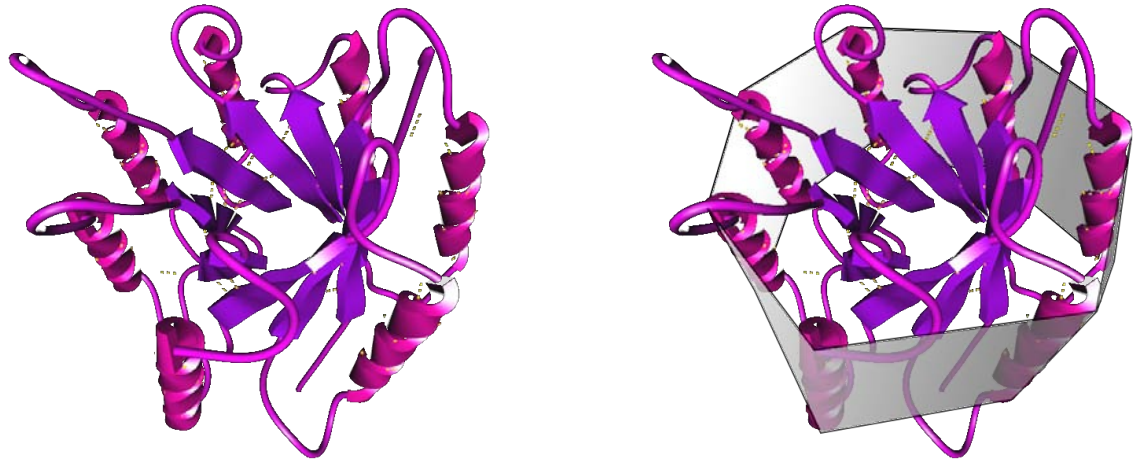
Nous pourrions également représenter la distribution de l'ensemble des normales d'un feuillet β sous la forme d'un cône. Nous aurions ainsi une information sur la topologie du feuillet. Pour un feuillet plan, le cône serait très restreint, alors que pour un feuillet formant un cylindre le cône serait ovoïde.

En ce qui concerne les simulations dynamiques, nous pourrions imaginer intégrer nos graphiques d'évolution de l'aire en fonction du temps de simulation à l'interface du logiciel. Les graphiques seraient construits de façon dynamique au cours de la simulation, nous pourrions définir des zones d'intérêt d'un clic et afficher directement les pas de simulation correspondants. Nous pouvons également imaginer développer des graphes de suivi des feuillets β au cours du temps. Cela donnerait une vision globale de l'évolution des feuillets tout au long d'une simulation. De plus, nous avons su représenter les fichiers issus de la RMN, c'est à dire contenant l'ensemble des structures ne violant pas les conditions expérimentales et qui peuvent amener des variations des feuillets β. De même, nous avons pu représenter les modes normaux de vibration des molécules. Une contribution couplée des simulations de dynamique moléculaire et des mouvements issus, par exemple, des modes normaux est envisageable quand les mouvements mettent en jeu les feuillets β.

Il est évident que l'ensemble des descripteurs supplémentaires peuvent être associés à la visualisation dynamique. En l'occurrence, une description en épaisseur du feuillet dans les zones de fluctuation serait peut-être une information pertinente. Cette représentation pourrait s'effectuer comme la visualisation des « boudins » dans MolMol [Koradi1996] pour représenter la fluctuation de la chaîne carbonée. Ce mode de représentation en boudins pourrait s'avérer efficace dans la visualisation d'un alignement structural comme nous avons pu l'envisager dans le paragraphe étudiant la base de données des immunoglobulines.

Dans cette étude nous nous sommes concentrés sur la problématique des structures secondaires de type feuillet β. Le choix a été dicté par une étude de faisabilité. Il serait impératif de pouvoir envisager les autres éléments de structure secondaire : hélices α et coudes, si tant est que les coudes puissent être considérés comme des éléments de structure secondaire. Les hélices α n'ont pas été le centre d'intérêt immédiat car, dans les logiciels actuels, leur représentation s'effectue par le

biais de cylindres ou de rubans s'enroulant sur eux-mêmes, et leur identification est assez simple et directe. Nous pourrions cependant les représenter en utilisant une « feuille » par groupe composé d'hélices α préalablement sélectionnées par l'utilisateur. Cette « feuille hélicoïdale » pourrait utiliser le centre des hélices α comme point de départ, et nous pouvons imaginer la faire se « dilater » ou se « contracter » pour permettre une représentation des structures non visibles jusqu'alors. Une représentation schématique d'un résultat possible est donnée sur la figure 5.2.1.



(a) (b)

Figure 5.2.1 – Représentation d'une construction protéique, en (a) nous utilisons le mode cartoon *et en (b) nous définissons une surface par l'intermédiaire des hélices α*

Afin d'appréhender l'apport de nos modèles à un niveau supérieur de visualisation, nous nous sommes appliqués à utiliser la plate-forme technologique « Centre Image » dotée d'une salle de réalité virtuelle située à l'IUT de Reims. Nous avons également, dans le cadre d'une collaboration du groupe « Signal Image et Connaissance » du laboratoire CReSTIC avec l'entreprise 3DTV Solutions[3], adapté nos modèles sur des écrans auto-stéréoscopiques permettant une visualisation en trois dimensions sans utilisation de lunettes spéciales. Il est indéniable que ces visualisations constituent une avancée majeure pour la compréhension des structures protéiques, et que nos modèles s'adaptent parfaitement à ces nouvelles technologies.

Enfin, nous devrions intégrer nos modèles au sein du logiciel BALLView, et les faire évoluer dans les futures « *releases* » du programme. Nous proposons de baptiser nos modèles, dont les aspects dynamiques peuvent faire penser à un tapis volant, SheHeRASADe pour « *Sheets Helper for RepresentAtion of SurfAce Descriptors* », pour devenir, lorsque le développement des hélices α aura été effectué, « *Sheets and Helices RepresentAtion of SurfAces Descriptors* ».

3 http://www.3dtvsolutions.com

Bibliographie

« List your three favorite books and how they have influenced your life.
– Is TV Guide a book?
– No.
– Son of sniglet?
– No.
– Katharine Hepburn's Me?
– No.
– Oh, I suck! »

Lisa & Homer J. Simpson

Dalton1808
Dalton, J. (1808). *A New System of Chemical Philosophy*. London.

Avogadro1811
Avogadro, A. (1811). Essai d'une manière de déterminer les masses relatives des molécules élémentaires des corps, et les proportions selon lesquelles elles entrent dans ces combinaisons. *Journal de physique, de chimie et d'histoire naturelle*, 58-76.

Perrin1911
Perrin, J. (1911). Les preuves de la réalité moléculaire (Étude spéciale des émulsions).

Einstein1905
Einstein, A. (1905). Ist die Trägheit eines Körpers von seinem Energieinhalt abhängig? *Annalen der Physik*, **18**, 639-641.

Mulder1838
Mulder, G. (1838). Annalen der Pharmacie. *Annalen der Pharmacie*, **28**, 73.

Fischer1891
Fischer, E. (1891). *Chem. Ber.*, **24**, 2683-2695.

Venn1880
Venn, J. (1880). On the Employment of Geometrical Diagrams for the Sensible Representation of Logical Propositions. *Proceedings of the Cambridge Philosophical Society*, **4**, 47-59.

Sanger1953a
Sanger, F. and Thompson, E. (1953). The amino-acid sequence in the glycyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. *The Biochemical journal*, **53(3)**, 353-366.

Sanger1953b
Sanger, F. and Thompson, E. (1953). The amino-acid sequence in the glycyl chain of insulin. II. The investigation of peptides from enzymic hydrolysates. *The Biochemical Journal*, **53(3)**, 366-374.

Dayhoff1965
Dayhoff, M., Eck, R., Chang, Y. and Sochard, S. (1965). *Atlas of Protein Sequence and Structure*.

Laue1912
Laue, M.V., Friedrich, W. and Knipping, P. (1912). *Interferenz-Erscheinungen bei Röntgenstrahlen*. Sitzungberichte der Königlich Bayerischen Akademie der Wissenschaften Mathematisch-physikalische Klasse.

Bragg1913a
Bragg, W. (1913). The diffraction of electromagnetic waves by a crystal. *Proc. Camb. Phil. Soc.*, **17**, 43-57.

Bragg1913b
Bragg, W. (1913). The structure of some crystals as indicated by their diffraction of X-rays. *Proc. Roy. Soc.*, **A 89**, 248-277.

Astbury1931a
Astbury, W. and Woods, H. (1931). The Molecular Weights of Proteins. *Nature*, **127**, 663-665.

Astbury1931b
Astbury, W. and Street, A. (1931). X-ray studies of the structures of hair, wool and related fibres. I. General. *Trans. R. Soc. Lond.*, **A230**, 75-101.

Astbury1933
Astbury, W. (1933). Some Problems in the X-ray Analysis of the Structure of Animal Hairs and Other Protein Fibers. *Trans. Faraday Soc.*, **29**, 193-211.

Astbury1934
Astbury, W. and Woods, H. (1934). X-ray studies of the structures of hair, wool and related fibres. II. The molecular structure and elastic properties of hair keratin. *Trans. R. Soc. Lond.*, **A232**, 333-394.

Astbury1935
Astbury, W. and Sisson, W. (1935). X-ray studies of the structures of hair, wool and related fibres. III. The configuration of the keratin molecule and its orientation in the biological cell. *Proc. R. Soc. Lond.*, **A150**, 533-551.

Pauling1951a
Pauling, L. and Corey, R. (1951). Atomic Coordinates and Structure Factors for Two Helical Configurations of Polypeptide Chains. *PNAS*, **37**, 235-240.

Pauling1951b
Pauling, L. and Corey, R. (1951). The Structure of Synthetic Polypeptides. *PNAS*, **37**, 241-250.

Pauling1951c
Pauling, L. and Corey, R. (1951). The Pleated Sheet, A New Layer Configuration of Polypeptide Chains. *PNAS*, **37**, 251-256.

Pauling1951d
Pauling, L. and Corey, R. (1951). The Structure of Feather Rachis Keratin. *PNAS*, **37**, 256-261.

Pauling1951e
Pauling, L. and Corey, R. (1951). The Structure of Hair, Muscle, and Related Proteins. *PNAS*, **37**, 261-271.

Pauling1951f
Pauling, L. and Corey, R. (1951). The Structure of Fibrous Proteins of the Collagen-Gelatin Group. *PNAS*, **37**, 272-281.

Pauling1951g
Pauling, L. and Corey, R. (1951). The Polypeptide-Chain Configuration in Hemoglobin and Other Globular Proteins. *PNAS*, **37**, 282-285.

Ramachandran1963
Ramachandran, G., Ramakrishnan, C. and Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, **7**, 95-99.

Perutz1960
Perutz, M., Rossman, M., Cullis, A., Muirhead, H., Will, G. and North, A. (1960). Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5A resolution, obtained by x-ray analysis. *Nature*, **185**, 416-422.

Kendrew1958
Kendrew, J., Bodo, G., Dintzis, H., Parrish, R. and Wyckoff, W. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, **181**, 662-666.

Wüthrich2001
Wüthrich, K. (2001). The way to NMR structures of proteins. *Nat. Struct. Biol.*, **8**, 923-925.

Bernstein1977
Bernstein, F., Koetzle, T., Williams, G., Meyer Jr, E., Brice, M., Rodgers, J., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535-542.

Berman2000
Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Shindyalov, I. and Bourne, P. (2000). The Protein Data Bank . *Nucleic Acids Research*, **28**, 235-242.

Berman2003
Berman, H., Henrick, K. and Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. *Nature Structural Biology*, **10 (12)**, 980.

Chou1974
Chou, P. and Fasman, G. (1974). Prediction of protein conformation. *Biochemistry*, **13**, 222-245.

Lewis1970
Lewis, P., Go, N., Go, M., Kotelchuck, D. and Scheraga, H. (1970). Helix Probability Profiles of Denatured Proteins and Their Correlation with Native Structures . *PNAS*, **65**, 810-815.

Robson1971
Robson, B. and Pain, R. (1971). Analysis of the Code Relating Sequence to Conformation in Globular Proteins: Possible Implications for the Mechanism of Formation of Helical Regions. *J. Mol. Biol.*, **58**, 237-256.

Richardson1977
Richardson, J. (1977). β-sheet topology and the relatedness of proteins. *Nature*, **268**, 495-500.

Chothia1977
Chothia, C., Levitt, M. and Richardson, D. (1977). Structure of proteins: packing of alpha-helices and beta-sheets. *PNAS*, **74**, 4130-4134.

Bond2003
Bond, C.S. (2003). TopDraw: a sketchpad for protein structure topology cartoons. *Bioinformatics*, **19**, 311-312.

Edelman1970
Edelman, G. (1970). The covalent structure of a human gamma G-immunoglobulin. XI. Functional implications. *Biochemistry*, **9**, 3197-3205.

PDB ID : 3A5A
Kuwada, T., Hasegawa, T., Takagi, T., Sato, I. and Shishikura, F. (2010). pH-dependent structural changes in haemoglobin component V from the midge larva Propsilocerus akamusi (Orthocladiinae, Diptera). *Acta Crystallogr D Biol Crystallogr*, **66**, 258-267.

PDB ID : 2WP3
Pernigo, S., Fukuzawa, A., Bertz, M., Holt, M., Rief, M., Steiner, R. and Gautel, M. (2010). Structural insight into M-band assembly and mechanics from the titin-obscurin-like-1 complex. *Proc Natl Acad Sci USA*, **107**, 2908-2913.

PDB ID : 8TIM
Artymiuk, P., Taylor, W. and Phillips, D. The structure of Triose Phosphate Isomerase.

PDB ID : 1AMK
Williams, J., Zeelen, J., Neubauer, G., Vriend, G., Backmann, J., Michels, P., Lambeir, A. and Wierenga, R. (1999). Structural and mutagenesis studies of leishmania triosephosphate isomerase: a point mutation can convert a mesophilic enzyme into a superstable enzyme without losing catalytic power. *Protein Eng.*, **12**, 243-250.

PDB ID : 3A67
Yokota, A., Tsumoto, K., Shiroishi, M., Nakanishi, T., Kondo, H. and Kumagai, I. (2010). Contribution of asparagine residues to the stabilization of a proteinaceous antigen-antibody complex, HyHEL-10-hen egg white lysozyme. *J Biol Chem.*, **285**, 7686-7696.

PDB ID : 1THB
Waller, D. and Liddington, R. (1990). Refinement of a partially oxygenated T state human haemoglobin at 1.5 A resolution. *Acta Crystallogr B.*, **46**, 409-418.

PDB ID : 1I1B
Finzel, B., Clancy, L., Holland, D., Muchmore, S., Watenpaugh, K. and Einspahr, H. (1989). Crystal structure of recombinant human interleukin-1 beta at 2.0 A resolution. *J.Mol.Biol.*, **209**, 779-791.

PDB ID : 256B
Lederer, F., Glatigny, A., Bethge, P., Bellamy, H. and Matthew, F. (1981). Improvement of the 2.5 A resolution model of cytochrome b562 by redetermining the primary structure and using molecular graphics. *J.Mol.Biol.*, **148**, 427-448.

PDB ID : 2RHE
Furey Jr, W., Wang, B., Yoo, C. and Sax, M. (1983). Structure of a novel Bence-Jones protein (Rhe) fragment at 1.6 A resolution. *J.Mol.Biol.*, **167**, 661-692.

PDB ID : 1UBQ
Vijay-Kumar, S., Bugg, C. and Cook, W. (1987). Structure of ubiquitin refined at 1.8 A resolution. *J.Mol.Biol.*, **194**, 531-544.

PDB ID : 2BUK
Jones, T. and Liljas, L. (1984). Structure of satellite tobacco necrosis virus after crystallographic refinement at 2.5 A resolution. *J.Mol.Biol.*, **177**, 735.

PDB ID : 2FOX
Ludwig, M., Pattridge, K., Metzger, A., Dixon, M., Eren, M., Feng, Y. and Swenson, R. (1997). Control of oxidation-reduction potentials in flavodoxin from Clostridium beijerinckii: the role of conformation changes. *Biochemistry*, **36**, 1259-1280.

PDB ID : 1APS
Pastore, A., Saudek, V., Ramponi, G. and Williams, R. (1992). Three-dimensional structure of acylphosphatase. Refinement and structure analysis. *J.Mol.Biol.*, **224**, 427-440.

PDB ID : 7TIM
Davenport, R., Bash, P., Seaton, B., Karplus, M., Petsko, G. and Ringe, D. (1991). Structure of the triosephosphate isomerase-phosphoglycolohydroxamate complex: an analogue of the intermediate on the reaction pathway. *Biochemistry*, **30**, 5821-5826.

Thornton1997
Thornton, J., Orengo, C., Michie, A., Jones, D. and Swindells, M. (1997). CATH: A Hierarchic Classification of Protein Domain Structures. *Structure*, **5**, 1093-1108.

Chothia1995
Chothia, C., Murzin, A., Brenner, S. and Hubbard, T. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536-540.

Redfern2007
Redfern, O., Harrison, A., Dallman, T., Pearl, F. and Orengo, C. (2007). CATHEDRAL: A Fast and Effective Algorithm to Predict Folds and Domain Boundaries from Multidomain Protein Structures. *PLoS Comput. Biol.*, **3**.

Orengo1996
Orengo, C. and Taylor, W. (1996). SSAP: Sequential Structure Alignment Program for Protein Structure Comparisons. *Methods in Enzymol.*, **266**, 617-634.

Swindells1995a
Swindells, M. (1995). A procedure for the automatic-determination of hydrophobic cores in protein structures. *Prot. Sci.*, **4**, 93-102.

Swindells1995b
Swindells, M.. (1995). A procedure for detecting structural domains in proteins. *Prot. Sci.*, **4**, 103-112.

Holm1994
Holm, L. and Sander, C. (1994). The FSSP database of structurally aligned protein fold families. *Nucleic Acids Res.*, **22**, 3600-3609.

Siddiqui1995
Siddiqui, A. and Barton, G. (1995). Continuous and Discontinuous Domains: An Algorithm for the Automatic Generation of Reliable Protein Domain Definitions. *Protein Science*, **4**, 872-884.

Krogh1994
Krogh, A., Brown, M., Mian, I., Sjolander, K. and Haussler, D. (1994). Hidden Markov models in computational biology: Applications to protein modeling. *J. Mol. Biol.*, **235**, 1501-1531.

Levinthal1968a
Levinthal, C.. (1968). Are there pathways for protein folding ?. *Journal of Chemical Physics*, **65**, 44-45.

Zwanzig1992
Zwanzig, R., Szabo, A. and Bagchi, B. (1992). Levinthal's paradox. *PNAS*, **89**, 20-22.

Anfinsen1954
Anfinsen, C., Redfield, R., Choate, W., Page, J. and Carroll, W. (1954). Studies on the Gross Structure, Cross-Linkages, and Terminal Sequences in Ribonuclease. *J. Biol. Chem.*, **207**, 201-210.

Anfinsen1961
Anfinsen, C. and Haber, E. (1961). Studies on the reduction and re-formation of protein disulfide bonds. *J. Biol. Chem.*, **236**.

Wyckoff1967a
Wyckoff, H., Hardman, K., Allewell, N., Inagami, T., Tsernoglou, D., Johnson, L. and Richards, F. (1967). The structure of ribonuclease-S at 6 A resolution. *J. Biol. Chem.*, **242**, 3749-3753.

Wyckoff1967b
Wyckoff, H., Hardman, K., Allewell, N., Inagami, T., Johnson, L. and Richards, F. (1967). The structure of ribonuclease-S at 3.5 A resolution. *J. Biol. Chem.*, **242**, 3984-3988.

Richards1968
Richards, F. (1968). The matching of physical models to three-dimensional electron-density maps: A simple optical device. *J. Mol. Biol.*, **37**, 225-230.

Rubin1972
Rubin, B. and Richardson, J. (1972). The simple construction of protein alpha-carbon models. *Biopolymers*, **11**, 2381-2385.

Levinthal1966
Levinthal, C. (1966). Molecular Model-Building by Computer. *Scientific American*, **214**, 42-52.

Levinthal1968b
Levinthal, C., Barry, C., Ward, S. and Zwick, M. (1968). *Computer Graphics in Macromolecular Chemistry. Emerging Concepts in Computer Graphics*.

Beem1977
Beem, K., Richardson, D. and Rajagopalan, K. (1977). Metal sites of copper-zinc superoxide dismutase. *Biochemistry*, **16**, 1930-1936.

Tainer1982
Tainer, J., Getzoff, E., Beem, K., Richardson, J. and Richardson, D. (1982). Determination and analysis of the 2 A-structure of copper, zinc superoxide dismutase. *J. Mol. Biol.*, **160**, 181-217.

Feldmann1980
TAMS: Teaching aids for macromolecular structure. Teachers manual. 1980.

Jones1978
Jones, T. (1978). A graphics model building and refinement system for macromolecules. *J. Appl. Crystallogr.*, **11**, 268-272.

Richardson1989
Richardson, J. and Richardson, D. (1989). Principles and patterns of protein conformation. In G. Fasman (ed.), *Prediction of protein structure and the principles of protein conformation*. Plenum Press, pp. 1-98.

Thomas1990
Thomas, A., Vaney, M.C., Le Bars, M., Mornon, J.P. and Morize, I. (1990). Symmetry and crystallography: new facilities in the graphic software MANOSK. *J Mol Graph.*, **8**, 108-110.

Richardson1992
Richardson, D. and Richardson, J. (1992). The kinemage: a tool for scientific communication. *Protein Sci.*, **1**, 3-9.

Sayle1995
Sayle, R. and Milner-White, E. (1995). RasMol: Biomolecular graphics for all. *Trends in Biochemical Sciences*, **20**, 374-376.

Humphrey1996
Humphrey, W., Dalke, A. and Schulten, K. (1996). VMD: visual molecular dynamics. *J. Mol. Graph.*, **14**, 33-38.

Phillips2005
Phillips, J., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R., Kale, L. and Schulten, K. (2005). Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, **26**, 1781-1802.

DeLano2008
DeLano, W. (2008). *The PyMOL Molecular Graphics System*. DeLano Scientific LLC.

Guex1997
Guex, N. and Peitsch, M. (1997). SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis*, **18**, 2714-2723.

Gezelter
http://www.jmol.org

Rzepa1994
Rzepa, H., Whitaker, B. and Winter, M. (1994). Chemical Applications of the World-Wide-Web System. *J. Chem. Soc*.

Casher1995
Casher, O., Chandramohan, G., Hargreaves, M., Leach, C., Murray-Rust, P., Rzepa, H., Sayle, R. and Whitaker, B. (1995). Hyperactive Molecules and the World-Wide-Web Information System. *J. Chem. Soc.*

Kohlbacher2000
Kohlbacher, O. and Lenhof, H. (2000). BALL - Rapid Software Prototyping in Computational Molecular Biology. *Bioinformatics*, **16**, 815-824.

Moll2006
Moll, A., Hildebrandt, A., Lenhof, H. and Kohlbacher, O. (2006). BALLView: A tool for research and education in molecular modeling. *Bioinformatics*, **22**, 365-366.

PDB ID : 1914
Birse, D., Kapp, U., Strub, K., Cusack, S. and Aberg, A. (1997). The crystal structure of the signal recognition particle Alu RNA binding heterodimer, SRP9/14. *EMBO J.*, **16**, 3757-3766.

Lee1971
Lee, B. and Richards, F. (1971). The interpretation of protein structures: estimation of static accesibility. *J. Mol. Biol.*, **55**, 379-400.

Kraulis1991
Kraulis, P.J.. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.*, **24**, 946-950.

O'Donoghue2010
O'Donoghue, S.I., Goodsell, D.S., Frangakis, A.S., Jossinet, F., Laskowski, R.A., Nilges, M., Saibil, H.R., Schafferhans, A., Wade, R.C., Westhof, E. and Olson, A.J. (2010). Visualization of macromolecular structures. *Nat Methods.*, **7**, S42-55.

PDB ID : 2I5B
Newman, J., Das, S., Sedelnikova, S. and Rice, D. (2006). The crystal structure of an ADP complex of Bacillus subtilis pyridoxal kinase provides evidence for the parallel emergence of enzyme activity during evolution. *J.Mol.Biol.*, **363**, 520-530.

Heinig2004
Heinig, M. and Frishman, D. (2004). STRIDE: a Web server for secondary structure assignment from known atomic coordinates of proteins. *Nucl. Acids Res.*, **32**.

Richards1988
Richards, F.M. and Kundrot, C.E. (1988). Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. *Proteins*, **3**, 71-84.

Kabsch1983
Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577-2637.

PDB ID : 1PRN
Kreusch, A. and Schulz, G. (1994). Refined structure of the porin from Rhodopseudomonas blastica. Comparison with the porin from Rhodobacter capsulatus. *J.Mol.Biol.*, **243**, 891-905.

PDB ID : 1YTB
Kim, Y., Geiger, J., Hahn, S. and Sigler, P. (1993). Crystal structure of a yeast TBP/TATA-box complex. *Nature*, **365**, 512-520.

PDB ID : 1EZG
Liou, Y., Tocilj, A., Davies, P. and Jia, Z. (2000). Mimicry of ice structure by surface hydroxyls and water of a beta-helix antifreeze protein. *Nature*, **406**, 322-324.

PDB ID : 1E20
Albert, A., Martinez-Ripoll, M., Espinosa-Ruiz, A., Yenush, L., Culianez-Macia, F. and Serrano, R. (2000). The X-ray structure of the FMN-binding protein AtHal3 provides the structural basis for the activity of a regulatory subunit involved in signal transduction. *Structure*, **8**, 961.

Gaboriaud1987
Gaboriaud, C., Bissery, V., Benchetrit, T. and Mornon, J. **Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences**. 1987.

Callebaut1997
Callebaut, I., Labesse, G., Durand, P., Poupon, A., Canard, L., Chomilier, J., Henrissat, B. and Mornon, J. (1997). Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. *Cell. Mol. Life Sci.*, **53**, 621-645.

Efremov1993
Efremov, R. and Alix, A. (1993). Environmental characteristics of residues in proteins: three-dimensional molecular hydrophobicity potential approach. *J. Biomol. Struct. Dyn.*, **11**, 483-507.

Efremov2007
Efremov, R., Chugunov, A., Pyrkov, T., Priestle, J., Arseniev, A. and Jacoby, E. (2007). Molecular lipophilicity in protein modeling and drug design. *Curr. Med. Chem.*, **14**, 393-415.

Prudhomme2009
Prudhomme, N.. **Prédiction des résidus clés du repliement et classification structurale de fragments protéiques en interaction**. 2009.

Pawlicki2008
Pawlicki, S., Le Béchec, A. and Delamarche, C. (2008). AMYPdb: A database dedicated to amyloid precursor proteins. *BMC Bioinformatics*, **9**, 273.

PDB ID : 1ERJ
Sprague, E.R., Redd, M.J., Johnson, A.D. and Wolberger, C. (2000). Structure of the C-terminal domain of Tup1, a corepressor of transcription in yeast.

PDB ID : 1LGN
Hohenester, E., Hutchinson, W.L., Pepys, M.B. and Wood, S.P. (1997). Crystal structure of a decameric complex of human serum amyloid P component with bound dAMP.

PDB ID : 1EMA
Ormö, M., Cubitt, A.B., Kallio, K., Gross, L.A., Tsien, R.Y. and Remington, S.J. (1996). Crystal structure of the Aequorea victoria green fluorescent protein.

PDB ID : 1KSA
Li, Y.F., Zhou, W., Blankenship, R.E. and Allen, J.P. (1997). Crystal structure of the bacteriochlorophyll a protein from Chlorobium tepidum. *J Mol Biol.*, **271**, 456-471.

PDB ID : 3ENI
Tronrud, D.E., Wen, J., Gay, L. and Blankenship, R.E. (2009). The structural basis for the difference in absorbance spectra for the FMO antenna protein from various green sulfur bacteria. *Photosynth.Res.*, **100**, 79-87.

PDB ID : 2WUH
Carafoli, F., Bihan, D., Stathopoulos, S., Konitsiotis, A.D., Kvansakul, M., Farndale, R.W., Leitinger, B. and Hohenester, E. (2009). Crystallographic insight into collagen recognition by discoidin domain receptor 2.

PDB ID : 1TGH
Juo, Z.S., Chiu, T.K., Leiberman, P.M., Baikalov, I., Berk, A.J. and Dickerson, R.E. (1996). How proteins recognize the TATA box. *J.Mol.Biol.*, **261**, 239-254.

PDB ID : 1Q10
Byeon, I.J., Louis, J.M. and Gronenborn, A.M. (2003). A protein contortionist: core mutations of GB1 that induce dimerization and domain swapping. *J.Mol.Biol.*, **333**, 141-152.

Halaby1999
Halaby, D.M., Poupon, A. and Mornon, J. (1999). The immunoglobulin fold family: sequence analysis and 3D structure comparisons. *Protein Eng.*, **12**, 563-571.

Gerstein1995
Gerstein, M. and Altman, R.B. (1995). Average core structures and variability measures for protein families: application to the immunoglobulins. *J Mol Biol.*, **251**, 161-175.

PDB ID : 1BEC
Bentley, G.A., Boulot, G., Karjalainen, K. and Mariuzza, R.A. (1995). Crystal structure of the beta chain of a T cell antigen receptor. *Science*, **267**, 1984-1987.

PDB ID : 1CID
Brady, R.L., Dodson, E.J., Dodson, G.G., Lange, G., Davis, S.J., Williams, A.F. and Barclay, A.N. (1993). Crystal structure of domains 3 and 4 of rat CD4: relation to the NH2-terminal domains. *Science*, **260**, 979-983.

PDB ID : 1CTN
Perrakis, A., Tews, I., Dauter, Z., Oppenheim, A.B., Chet, I., Wilson, K.S. and Vorgias, C.E. (1994). Crystal structure of a bacterial chitinase at 2.3 A resolution. *Structure*, **2**, 1169-1180.

PDB ID : 1FC1
Deisenhofer, J. (1981). Crystallographic refinement and atomic models of a human Fc fragment and its complex with fragment B of protein A from Staphylococcus aureus at 2.9- and 2.8-A resolution. *Biochemistry*, **20**, 2361-2370.

PDB ID : 1HNG
Jones, E.Y., Davis, S.J., Williams, A.F., Harlos, K. and Stuart, D.I. (1992). Crystal structure at 2.8 A resolution of a soluble form of the cell adhesion molecule CD2. *Nature*, **360**, 232-239.

PDB ID : 2MCM
Van Roey, P. and Beerman, T.A. (1989). Crystal structure analysis of auromomycin apoprotein (macromomycin) shows importance of protein side chains to chromophore binding selectivity. *PNAS*, **86**, 6587-6591.

PDB ID : 2SOD
Tainer, J.A., Getzoff, E.D., Beem, K.M., Richardson, J.S. and Richardson, D.C. (1982). Determination and analysis of the 2 A-structure of copper, zinc superoxide dismutase. *J.Mol.Biol.*, **160**, 181-217.

PDB ID : 2BEG
Luhrs, T., Ritter, C., Adrian, M., Riek-Loher, D., Bohrmann, B., Dobeli, H., Schubert, D. and Riek, R. (2005). 3D structure of Alzheimer's amyloid-beta(1-42) fibrils. *PNAS*, **102**, 17342-17347.

Kajava2006
Kajava, A.V. and Steven, A.C. (2006). Beta-rolls, beta-helices, and other beta-solenoid proteins. *Advances in protein chemistry*, **73**, 55-96.

PDB ID : 1J2Z
Lee, B.I. and Suh, S.W. (2003). Crystal structure of UDP-N-acetylglucosamine acyltransferase from Helicobacter pylori. *Proteins*, **53**, 772-774.

PDB ID : 1P9H
Nummelin, H., Merckel, M.C., Leo, J.C., Lankinen, H., Skurnik, M. and Goldman, A. (2004). The Yersinia adhesin YadA collagen-binding domain structure is a novel left-handed parallel beta-roll. *Embo J.*, **23**, 701-711.

PDB ID : 1AIR
Lietzke, S.E., Scavetta, R.D., Yoder, M.D. and Jurnak, F. (1996). The Refined Three-Dimensional Structure of Pectate Lyase E from Erwinia chrysanthemi at 2.2 A Resolution. *Plant Physiol.*, **111**, 73-92.

PDB ID : 1VH4
Badger, J., Sauder, J.M., Adams, J.M., Antonysamy, S., Bain, K., Bergseid, M.G., Buchanan, S.G., Buchanan, M.D., Batiyenko, Y., Christopher, J.A., Emtage, S., Eroshkina, A., Feil, I., Furlong, E.B., Gajiwala, K.S., Gao, X., He, D., Hendle, J., Huber, A., Hoda, K., Kearins, P., Kissinger, C., Laubert, B., Lewis, H.A., Lin, J., Loomis, K., Lorimer, D., Louie, G., Maletic, M., Marsh, C.D., Miller, I., Molinari, J., Muller-Dieckmann, H.J., Newman, J.M., Noland, B.W., Pagarigan, B., Park, F., Peat, T.S., Post, K.W., Radojicic, S., Ramos, A., Romero, R., Rutter, M.E., Sanderson, W.E., Schwinn, K.D., Tresser, J., Winhoven, J., Wright, T.A., Wu, L., Xu, J. and Harris, T.J. (2005). Structural analysis of a set of proteins resulting from a bacterial genomics project. *Proteins*, **60**, 787-796.

PDB ID : 1T34
Ogawa, H., Qiu, Y., Ogata, C.M. and Misono, K.S. (2004). Crystal structure of hormone-bound atrial natriuretic peptide receptor extracellular domain: rotation mechanism for transmembrane signal transduction. *J.Biol.Chem.*, **279**, 28625-28631.

PDB ID : 1P8X
Narayan, K., Chumnarnsilpa, S., Choe, H., Irobi, E., Urosev, D., Lindberg, U., Schutt, C.E., Burtnick, L.D. and Robinson, R.C. (2003). Activation in isolation: exposure of the actin-binding site in the C-terminal half of gelsolin does not require actin. *FEBS LETT.*, **552**, 82-85.

PDB ID : 1ATH
Schreuder, H.A., de Boer, B., Dijkema, R., Mulders, J., Theunissen, H.J., Grootenhuis, P.D. and Hol, W.G. (1994). The intact and cleaved human antithrombin III complex as a model for serpin-proteinase interactions. *Nat.Struct.Biol.*, **1**, 48-54.

PDB ID : 1TFP
Sunde, M., Richardson, S.J., Chang, L., Pettersson, T.M., Schreiber, G. and Blake, C.C. (1996). The crystal structure of transthyretin from chicken. *Eur.J.Biochem.*, **236**, 491-499.

Koradi1996
Koradi, R., Billeter, M. and Wüthrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. *J Mol Graph.*, **14**, 51-55.

Résumé

L'enjeu de ces travaux consiste en la représentation de motifs structuraux réguliers des protéines : les feuillets β. Les représentations classiques de la modélisation moléculaire n'étant pas satisfaisantes, étant donné qu'elles ne représentent pas les feuillets β dans leur ensemble, nous proposons nos modèles représentant ces structures sous forme de surfaces.

Nous utilisons le logiciel *open source* « BALLView » pour créer nos propres modèles de feuillets β. La première approche utilise la description des feuillets β présente dans les fichiers issus de la « *Protein Data Bank* », la banque de données mondiale de structures protéiques, pour calculer une interpolation bidimensionnelle basée sur les splines de Catmull-Rom. La seconde approche utilise des carreaux de Bézier, construits à partir des résultats issus d'un algorithme d'attribution des structures secondaires des protéines, dont les feuillets β font partie. Ces approches sont les premières à représenter les feuillets β dans leur ensemble.

Les modèles classiques ne représentent que les brins β. Pour visualiser leur orientation nous plaquons cette information par le biais de textures. Cela nous amène à considérer nos surfaces comme de nouveaux médias sur lesquels nous pouvons dépeindre des données supplémentaires par l'intermédiaire de méthodes de coloration (« *Hydrophobic Cluster Analysis* », « *Molecular Hydrophobicity Potential* »…).

Nos modèles sont utilisables sur l'ensemble des fichiers au format PDB, en statique, mais également sur des fichiers de simulation de dynamique moléculaire. Nous pouvons alors constater l'évolution des feuillets β, leurs déformations, l'apparition de trous, d'invaginations ou de déchirures. Ces constatations nous amènent à baptiser nos modèles SheHeRASADe pour « *Sheets Helper for RepresentAtion of SurfAce Descriptors* ».

Nous nous intéressons, entre autres, à l'application de ces modèles sur les divers repliements protéiques des feuillets β répertoriés dans la classification CATH, ainsi qu'aux fibres amyloïdes, impliquées dans de nombreuses pathologies.

Mots-clés : modélisation moléculaire, feuillet β, format PDB, surface, visualisation, splines de Catmull-Rom, carreaux de Bézier, repliements, amyloïdes.

Abstract

The aim of this work consists in the representation of common structural motifs of proteins: the β sheets. The classical visualization modes are not satisfying, considering that they don't represent the whole β sheets. We propose innovative models materializing those structures using surfaces.

We use the open source software "BALLView" to create our own β sheet models. The first one uses the β sheets description stored in files from the Protein Data Bank, the worldwide data bank of proteic structures, to compute a bidimensionnal interpolated surface based on Catmull-Rom splines. The second one uses Bézier patches defined from β sheets produced by a secondary structure prediction algorithm. Those models are the first ones to fully represent β sheets.

Previous methods only represent β strands. In order to visualize their orientation, we map these important data to our surfaces by using textures. It leads us to consider our surfaces as a new medium on which we can depict additional information using coloring methods (Hydrophobic Cluster Analysis, Molecular Hydrophobicity Potential...).

Our models are available for any PDB formatted file, in both static and dynamic ways, using molecular dynamics simulations. We can observe the evolution of β sheets, deformations, holes appearances, invaginations or splits. Those observations lead us to call our models SheHeRASADe for "Sheets Helper for RepresentAtion of SurfAce Descriptors".

We apply those models to the different proteic folds of β sheets listed in the CATH classification, and on amyloid fibrils involved in many diseases.

Keywords: molecular modeling, β sheet, PDB format, surface, visualization, Catmull-Rom splines, Bézier patches, folds, amyloids.