Tessellations de Voronoï appliquées aux structures protéiques

Franck Dupuis

UNIVERSITE PARIS 7 – DENIS DIDEROT

DOCTORAT

Biophysique moléculaire

Franck Dupuis

Tessellations de Voronoï appliquées aux structures protéiques

Thèse dirigée par Jean-Paul MORNON

Date de soutenance : 6 novembre 2003

JURY

Dr Nathalie COLLOC'H	Rapporteur
Dr Jean DOUCET	Rapporteur
Pr Serge HAZOUT	Examinateur
Pr Remi JULLIEN	Examinateur
Dr Etienne GHUYS	Examinateur
Dr Jean-Paul MORNON	Directeur de thèse
Pr Jean-François SADOC	Invité

Laboratoire d'accueil : Laboratoire de Minéralogie Cristallographie Paris

Remerciements

Je remercie Bernard Capelle de m'avoir accueilli au sein du laboratoire de Minéralogie et Cristallographie Paris.

Je remercie tout particulièrement Jean-Paul Mornon de m'avoir accueilli au sein de l'équipe Systèmes Moléculaires et Biologie Structurale et d'avoir dirigé mes travaux tout au long de ces trois années.

Je remercie également Isabelle Callebaut pour son aide, ses conseils et sa gentillesse.

Je remercie Nathalie Colloc'h et Jean Doucet d'avoir accepté d'être rapporteurs de cette thèse, ainsi que Remi Jullien, Serge Hazout et Etienne Ghuys d'avoir bien voulu en être les examinateurs.

Je remercie vivement Jean-François Sadoc pour sa disponibilité, ses conseils, ses relectures, ses programmes.

Je remercie Alain Soyer pour ses conseils et son aide informatique précieuse.

Je remercie une dernière fois Philippe Devaux.

Un grand merci à toutes les personnes du couloir 16-15 de Jussieu : Claire, Gaëlle, Sophie, Karine, Annette, Annick, Françoise, Nathalie, Tatiana, Nicky, Sabrina, Claudine, Denis V, Marc, Maxime, Nicolas, Eric, Paul, Quentin, Guillaume, Richard, Jean-Luc, Jean, Jérôme, Jacques, Claude.

Une pensée spéciale pour Emilie : ne t'inquiète pas, ça n'a pas été si difficile que ça.

Enfin, Feriel, Marina et Victoria, sans vous ces quatre années (et plus particulièrement la dernière) n'auraient pas été du tout les mêmes … un énorme et très sincère MERCI.

Ce travail est dédié à mon oncle et à ma grand-mère.

Abréviations

AA — **A**cide **A**miné

ADN — **A**cide **D**ésoxyribo**N**ucléique

ARN — **A**cide **R**ibo**N**ucléique

Cα — **C**arbone alpha

Cβ — **C**arbone beta

CG — **C**entre **G**éométrique des acides aminés

CGL — **C**entre **G**éométrique des chaînes **L**atérales des acides aminés

Cter — **C ter**minale

Liaisons H — **Liaisons H**ydrogène

Nter — **N ter**minale

PDB — **P**rotein **D**ata **B**ank

TdV — **T**essellation(s) **d**e **V**oronoï

Sommaire

Introduction générale

Les différents projets de séquençage mis en place depuis quelques années ont produit une quantité de données considérable. Pour pouvoir gérer, organiser, comparer, analyser et explorer les informations contenues dans ces données, la bio-informatique a progressivement émergé afin d'élaborer de nouveaux concepts ou produire de nouvelles connaissances. Cette nouvelle branche dite « in silico » qui vient compléter les approches plus classiques « in vitro » et « in vivo » de la biologie est initialement apparue dans les années soixante pour répondre aux besoins de la phylogénie moléculaire[1-3]. La bio-informatique ne s'intéresse pas exclusivement aux données issues des séquençages car elle intègre aussi l'étude des structures des macromolécules biologiques telles que les protéines[4].

Le travail présenté ici s'inscrit dans ce cadre, car nous nous sommes intéressés à l'étude des structures des protéines en utilisant un outil mathématique régulièrement utilisé en physique : les tessellations de Voronoï (TdV). A notre connaissance, la première application de ces tessellations aux structures protéiques date de 1974[5], Frederic M. Richards a calculé les volumes d'atomes ou de groupes d'atomes à partir des coordonnées atomiques de deux protéines : le lysozyme et la ribonucléase S. Depuis, les tessellations de Voronoï ont largement été utilisées pour le calcul des volumes des atomes (ou de groupes d'atomes), des résidus ou des protéines ; ou pour le calcul des densités ou de la compaction au sein de ces protéines[5-20]. Au fil du temps les applications des tessellations aux structures protéiques se sont multipliées et diversifiées soit sur des thématiques proches des précédentes, comme l'étude de la compressibilité des protéines[21], la mesure de la qualité des structures (en déterminant les déviations par rapport aux volumes standards)[22], la détection des cavités à l'intérieur des protéines[23, 24] ; soit sur des thématiques plus éloignées : identification de domaines structuraux[25], étude des interactions entre résidus aromatiques[26], mise au point de potentiels statistiques[27, 28], détermination des contacts entre atomes[29], caractérisation des sites d'interaction[30-32]. Les tessellations de Delaunay qui sont étroitement liées aux tessellations de Voronoï ont elles aussi été utilisées régulièrement dans l'étude des structures protéiques mais dans une moindre mesure[33-39].

Dans cette thèse je présente une nouvelle méthode d'application des tessellations de Voronoï aux structures protéiques. Le premier chapitre présente une brève description des structures protéiques et du vocabulaire qui leur est lié. Le deuxième chapitre, qui propose une approche théorique des tessellations de Voronoï (et de Delaunay), est complété par le troisième chapitre qui détaille plus précisément nos méthodes. Le quatrième chapitre est un tour d'horizon du paysage protéique vu au travers du filtre des tessellations de Voronoï. Enfin, les deux derniers chapitres sont consacrés à des applications concrètes.

C h a p i t r e 1

Structure des protéines : généralités

1 - Introduction

L'acide désoxyribonucléique (ADN) est le support physique des gènes et la structure de cette macro-molécule composée de deux molécules d'ADN (ou deux brins) est une double hélice. Chacune de ces deux hélices est constituée d'une succession de nucléotides, composés d'une part d'un sucre (le désoxyribose) et d'un groupement phosphate qui forment l'armature du brin et d'autre part, d'une base qui caractérise le nucléotide. Il en existe quatre types (différenciés par leur base) : l'adénosine (A), la cytidine (C), la guanosine (G) et la thymidine (T). Le long de la double hélice ces bases s'assemblent toujours par paires : la guanosine avec la cytidine et la thymidine avec l'adénosine. C'est la succession de ces nucléotides le long de l'hélice qui constitue le code génétique.

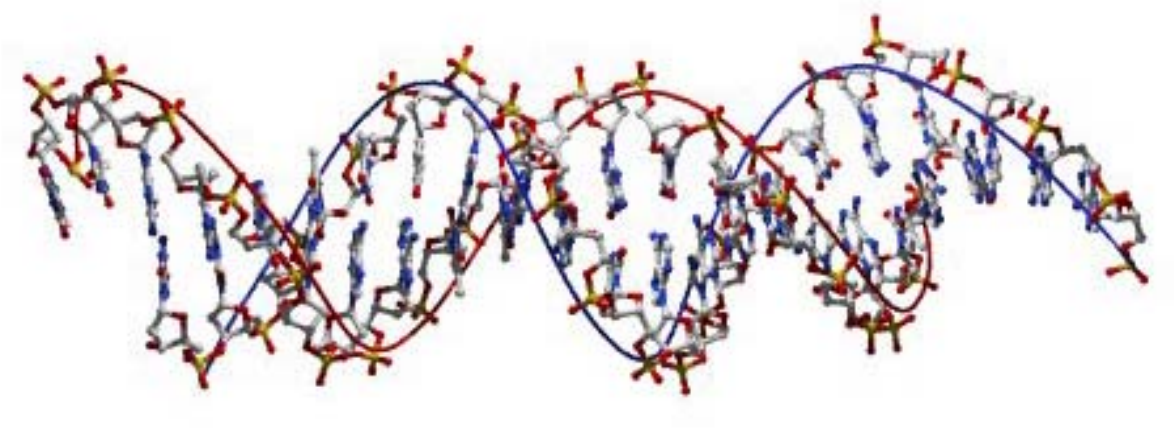


Figure 1 : Double hélice d'ADN. Les atomes de carbone sont en blanc, ceux d'oxygène en rouge, ceux d'azote en bleu et ceux de phosphate en jaune.

C'est à l'aide de ce code que la machinerie cellulaire peut traduire une partie de l'ADN en protéines via l'ARN messager (ARNm). Ce dernier est constitué d'une seule chaîne et est également caractérisé par une succession de nucléotides (A, C, G et uridine (U)). Chaque triplet de nucléotides successifs est appelé un codon auquel le code génétique associe un acide

aminé (AA) particulier. A la succession des codons sur le polymère qu'est l'ARNm correspond donc une succession d'AA qui forme un autre polymère que l'on appelle protéine.

2 - Le séquençage des génomes

Le séquençage consiste à déterminer l'ordre dans lequel se succèdent les nucléotides pour former l'ADN. Une fois le séquençage terminé, l'information dont on dispose alors est un ensemble de données brutes (on parle dans ce contexte de séquence anonyme) qu'il faut déchiffrer pour en extraire l'information utile que sont les gènes puis les séquences protéiques (c'est à dire la succession des AA). A l'heure actuelle l'imprécision de la prédiction des séquences codantes reste un problème important. Depuis le séquençage du premier génome bactérien (*Haemophilus influenzae* : bactérie responsable de nombreuses infections chez l'homme, notamment la méningite)[40], de nombreux génomes ont été séquencés. Pour des organismes tels que *Escherichia coli* (bactérie présente dans la flore intestinale de tous les animaux), *Saccharomyces cerevisiae* (levure du boulanger), *Caenorhabditis elegans* (ver nématode) ou *Drosophila melanogaster* (mouche du vinaigre) le séquençage complet est terminé. L'étude de ces organismes avait été privilégiée non seulement en raison de la taille de leur génome, qui est sensiblement plus petite que celle du génome de l'homme, mais aussi pour des raisons biologiques. Aujourd'hui le séquençage des génomes humain[41, 42] et murin[43] a été également effectué.

D'autres organismes bactériens ont été choisis pour des raisons économiques ou médicales. Le séquençage des génomes de bactéries pathogènes telles que *Haemophilus influenzae*, *Mycoplasma genitalium* (responsable de l'urétrite), *Staphylococcus aureus* (responsable de nombreuses infections nosocomiales), *Mycobacterium tuberculosis* (bacille de Koch responsable de la tuberculose) et *Helicobacter pylori* (présente dans tous les cas de cancers gastriques) ont aussi été terminés. En raison de son importante utilisation dans le secteur agroalimentaire, le séquençage du génome de *Bacillus subtilis* (bactérie du sol, non pathogène) a été entrepris. De même, celui d'organismes tels que *Pyrococcus furiosus* ou *Methanococcus jannaschii* (archébactéries vivant à très haute température) a été également réalisé dans l'espoir d'identifier des enzymes industriellement intéressantes.

Le nombre de séquences protéiques est donc en accroissement permanent : en 1996 la banque épurée Swiss-Prot[44] contenait moins de 10.000 entrées, en mars 2003 elle en contenait plus de 123.000. Ces chiffres sont à comparer au nombre de structures protéiques résolues. En

effet, la ou les fonctions des protéines (et donc des gènes qui les ont codées) sont étroitement associées à leur structure tridimensionnelle car le repliement du polymère peptidique en une structure compacte va avoir pour conséquence de positionner correctement dans l'espace les AA fonctionnels. Ils pourront alors interagir avec le ou les substrats spécifiques de la protéine. Or depuis la première détermination d'une structure tridimensionnelle (la myoglobine en 1960)[45], le nombre de structures tridimensionnelles déterminées, soit par résonance magnétique nucléaire (RMN)[46] soit par diffraction des rayons X n'a cessé de croître également. Les coordonnées atomiques de ces structures sont déposées dans la Protein Data Bank (PDB)[47] et en avril 2003 son effectif dépassait les 20.000 structures. Toutefois, cette augmentation du nombre de structures résolues ne doit pas masquer le fait que la proportion de nouveaux types de repliements est en baisse depuis plusieurs années alors que dans le même temps le nombre de structures déterminées s'accroissait très fortement (voir Figure 2).

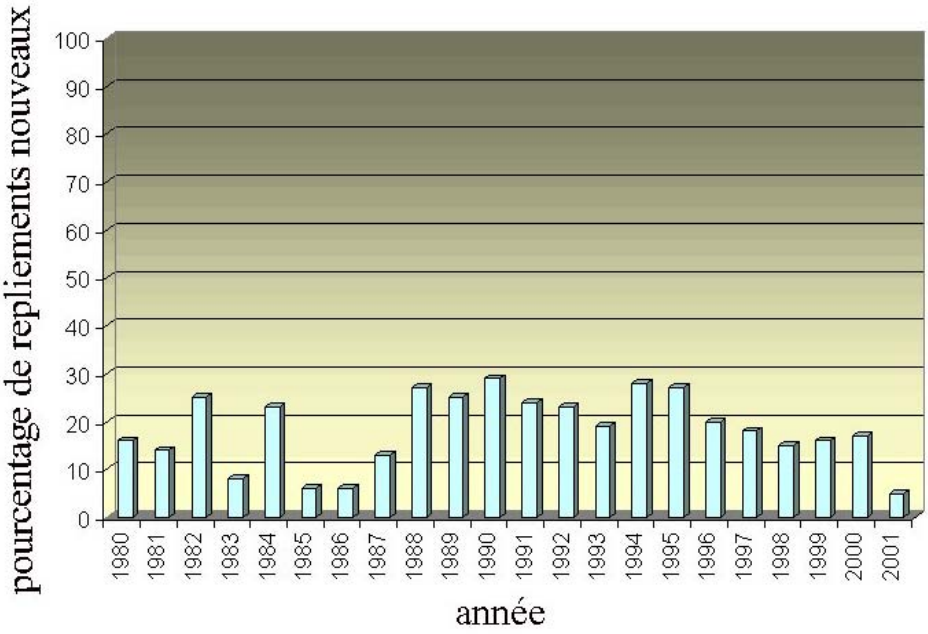


Figure 2 : Pourcentage de nouveaux types de repliement par année.

D'après de récentes estimations[48-51] le nombre total de types de repliement protéique différents serait compris entre 650 et 1.000. Ce nombre est plus petit (de plusieurs ordres de grandeur) que le nombre de séquences protéiques déterminées à l'heure actuelle. Ceci semble donc indiquer que l'information contenue dans les séquences protéiques est fortement

dégénérée puisque c'est à travers ce millier d'agencements que des protéines peuvent effectivement exprimer leurs spécificités fonctionnelles. Un même type de repliement peut donc être adopté par un très grand nombre de séquences très différentes les unes des autres.

3 - Structure des protéines

Nous venons de voir qu'une protéine est un polymère linéaire constitué de différentes unités de base disposées les unes à la suite des autres : les acides aminés, appelés également résidus. Un AA est constitué d'un carbone central (carbone alpha ou Cα) lié à un groupement carboxyle (COOH), à un groupement amine (NH_2), à un atome d'hydrogène (H) et à un radical R. Les protéines sont constituées de vingt AA naturels (voir Tableau 1). Ils se différencient les uns des autres par la nature même de ce radical qui leur confère différentes propriétés telles que, entre autres, la charge, la flexibilité, l'encombrement stérique ou bien encore l'hydrophobie qui est considérée comme le moteur du repliement protéique[52, 53].

NOM	code 3 lettres	code 1 lettre
alanine	ALA	A
cystéine	CYS	C
acide aspartique	ASP	D
acide glutamique	GLU	E
phénylalanine	PHE	F
glycine	GLY	G
histidine	HIS	H
isoleucine	ILE	I
lysine	LYS	K
leucine	LEU	L
méthionine	MET	M
asparagine	ASN	N
proline	PRO	P
glutamine	GLN	Q
arginine	ARG	R
sérine	SER	S
thréonine	THR	T
valine	VAL	V
tryptophane	TRP	W
tyrosine	TYR	Y

Tableau 1 : Noms et codes des 20 AA courants présents dans les structures protéiques.

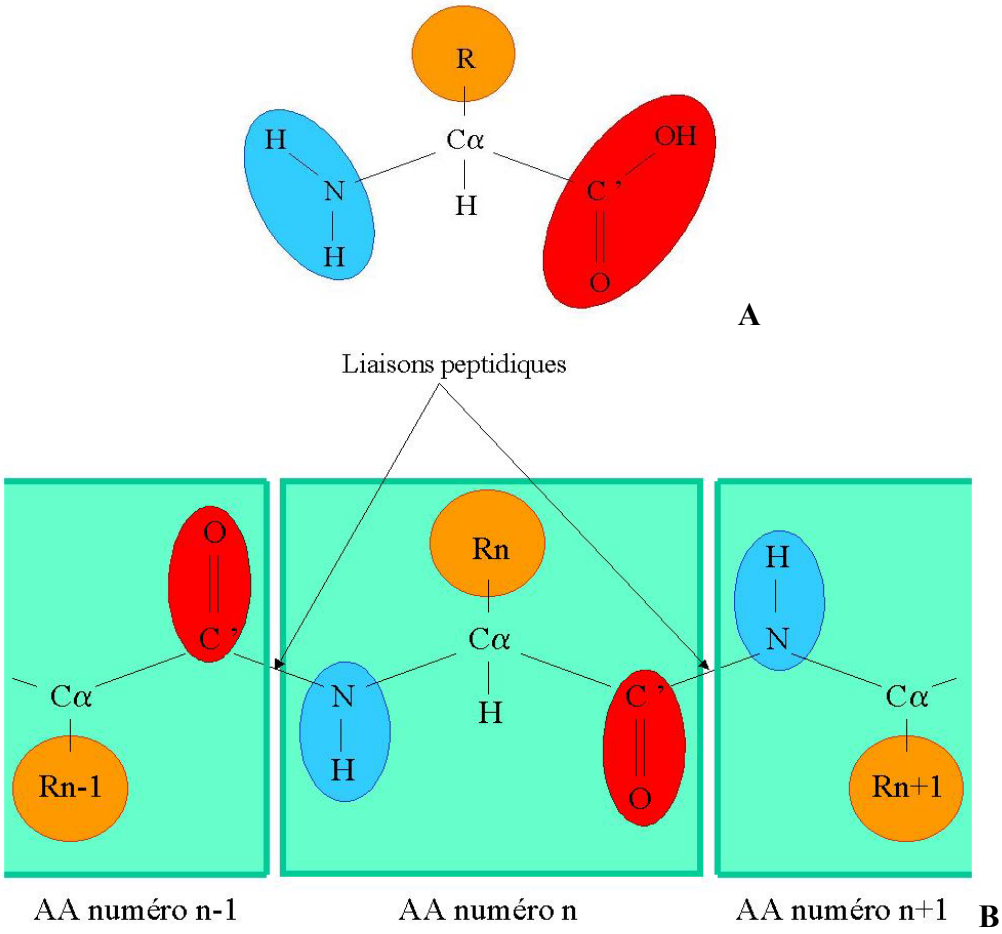


Figure 3
A : Diagramme représentant de manière schématique la structure d'un AA. Le Cα central est lié à un groupe amine (en bleu), à un groupe carboxyle (en rouge), à un atome d'hydrogène et à une chaîne latérale (en orange) qui caractérise chaque AA.
B : Dans une chaîne polypeptidique le groupe carboxyle de l'AA numéro n a formé une liaison peptidique plane avec le groupe amine de l'AA numéro n+1, une molécule d'eau est éliminée durant ce processus.

Les AA sont reliés entre eux par une liaison peptidique plane formée entre le groupement carboxyle COOH d'un résidu et le groupement amine NH_2 du résidu suivant. Cette organisation permet de regrouper les atomes des protéines en deux classes distinctes : d'une part les atomes d'azote des groupements amines, les carbones alpha, les atomes de carbone et d'oxygène des groupements carboxyles constituent la chaîne principale de la protéine (également appelée squelette ; je ne mentionne pas les atomes d'hydrogène à dessein puisqu'ils sont très rarement présents dans les fichiers structuraux), d'autre part, les différents radicaux sont appelés par opposition chaînes latérales (Figure 4). La synthèse des protéines débute par l'extrémité N-terminale (Nter) appelée ainsi car le premier atome de la chaîne est un atome d'azote, et elle se termine par l'extrémité C-terminale (Cter) car le dernier atome de la chaîne est le carbone du groupe carboxyle.

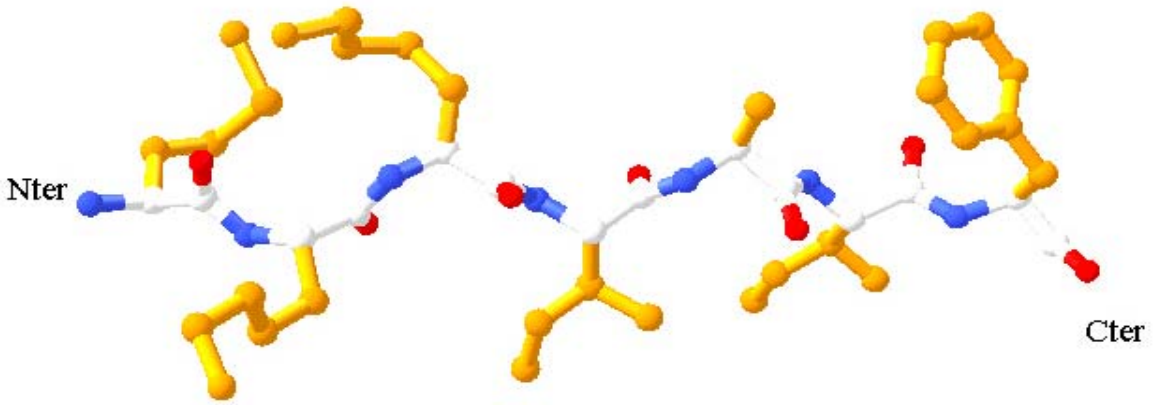


Figure 4 : Polypeptide de 7 AA (M.K.K.I.A.I.F). La chaîne principale est en bleu (N), blanc (C), rouge (O) alors que les chaînes latérales sont en orange. Le premier AA de la chaîne est à gauche à l'extrémité N-terminale (Nter), le dernier est à droite à l'extrémité C-terminale (Cter).

Parmi les différents AA, les cystéines ont une particularité remarquable puisque deux cystéines éloignées sur la chaîne principale mais proches dans l'espace tridimensionnel peuvent former un pont disulfure, ce qui signifie qu'une liaison covalente apparaît entre les deux atomes de soufre des deux chaînes latérales (Figure 5).

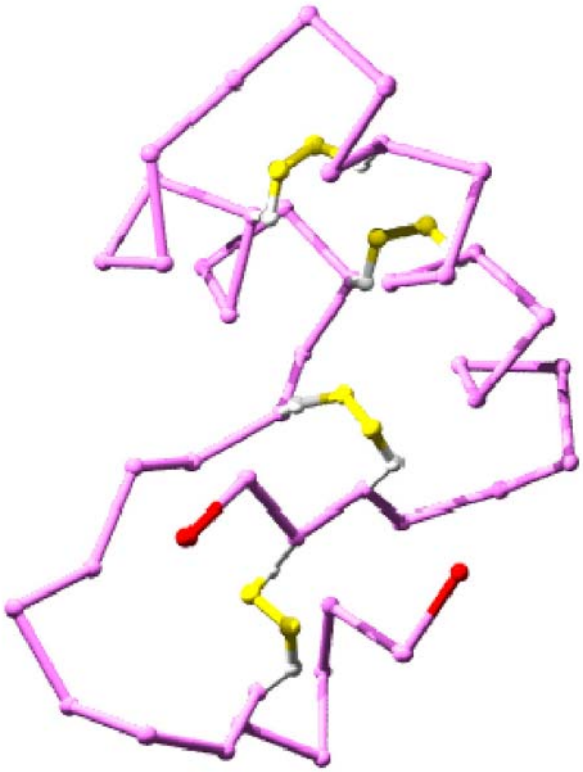


Figure 5 : Pour cette structure seuls les Cα sont représentés (en violet), les extrémités Nter et Cter sont en rouge. Les chaînes latérales des cystéines sont représentées (l'atome de carbone en blanc et celui de soufre en jaune). Cette structure contient 4 ponts disulfure représentés par une liaison covalente en jaune entre deux atomes de soufre.

La succession des AA le long de la chaîne principale entre l'extrémité N-terminale et l'extrémité C-terminale définit la séquence protéique dans laquelle toutes les informations nécessaires pour que la protéine puisse se replier correctement sont contenues[54]. Ce repliement peut être analysé à plusieurs échelles associées à divers degrés de structuration. Le premier de ces degrés est la structure primaire qui est en fait la simple succession des AA le long de la chaîne principale : séquence et structure primaire renferment une seule et même information.

4 - Structures secondaires

Comme je l'ai indiqué plus haut, la compaction hydrophobe est le moteur essentiel du repliement des protéines globulaires. Une des principales conséquences de ce phénomène, déjà observée dès la première résolution de la myoglobine par Kendrew[45], est que l'intérieur de ces protéines est dans une grande majorité constitué par des AA hydrophobes. Schématiquement, les structures protéiques peuvent donc se décomposer à l'image d'une micelle en un cœur hydrophobe et une surface hydrophile. Toutefois cette décomposition pose un problème majeur puisque pour compacter les AA hydrophobes au cœur de la protéine, il est indispensable que la chaîne principale passe aussi dans ce cœur hydrophobe, or cette chaîne est fortement hydrophile avec pour chaque unité peptidique un donneur de liaison hydrogène (NH) et un accepteur de liaison hydrogène (C'=O). Pour que cela soit possible il est donc nécessaire que le repliement protéique neutralise ces groupes polaires en formant des liaisons hydrogène (liaisons H). Ceci est réalisé, effectivement, par l'intermédiaire de la formation de structures secondaires régulières dont les principaux types sont les hélices alpha (α) et les brins beta (β). Ces structures, qui sont les parties les mieux définies par cristallographie des rayons X ou par résonnance magnétique nucléaire, fournissent pour les structures protéiques une base relativement stable et rigide. Les groupes fonctionnels des protéines peuvent être associés à ces structures secondaires par l'intermédiaire de leurs chaînes latérales ou, plus fréquemment, par les boucles qui relient ces différents éléments entre eux.

4.1 Hélices α

L'hélice α est caractérisée par des liaisons hydrogène entre le groupe C'=O du résidu n° i et le groupe NH du résidu n° i+4 (Figure 6). Chaque tour comporte 3.6 résidus ce qui

correspond à un pas de 5.4 Å (soit 1.5 Å par AA). Dans les protéines globulaires les longueurs des hélices α peuvent varier de quelques AA (4 ou 5) jusqu'à plus de 40 AA, avec une taille moyenne proche de 10 AA correspondant à trois tours. Il existe d'autres types d'hélices présentant un enroulement plus ou moins marqué par rapport à celui de l'hélice α. L'hélice π favorise les liaisons hydrogène entre l'AA n° i et l'AA n° i+5, cet enroulement est moins prononcé que celui de l'hélice α et il apparaît un vide le long de son axe (rayon de 2.8 Å au lieu de 2.3 Å pour l'hélice α). L'hélice 3_{10} favorise quant à elle les liaisons hydrogène entre l'AA n° i et l'AA n° i+3, elle comporte trois résidus par tour et il y a dix atomes entre le donneur et l'accepteur de liaison hydrogène. L'enroulement de cette hélice est plus marqué que celui de l'hélice α et la compaction des atomes du squelette qui lui est associée n'est pas si marquée, alors que dans le cas de l'hélice π celle-ci l'est plus. Ces hélices (3_{10}, π) ne sont donc pas favorables d'un point de vue énergétique et c'est pourquoi elles sont relativement rares (plus particulièrement l'hélice π). On les trouve aux extrémités des hélices α ou isolées, mais pour de très faibles longueurs (surtout l'hélice π, un tour). Il est important de remarquer ici que les hélices font intervenir des résidus proches les uns des autres le long de la structure primaire.

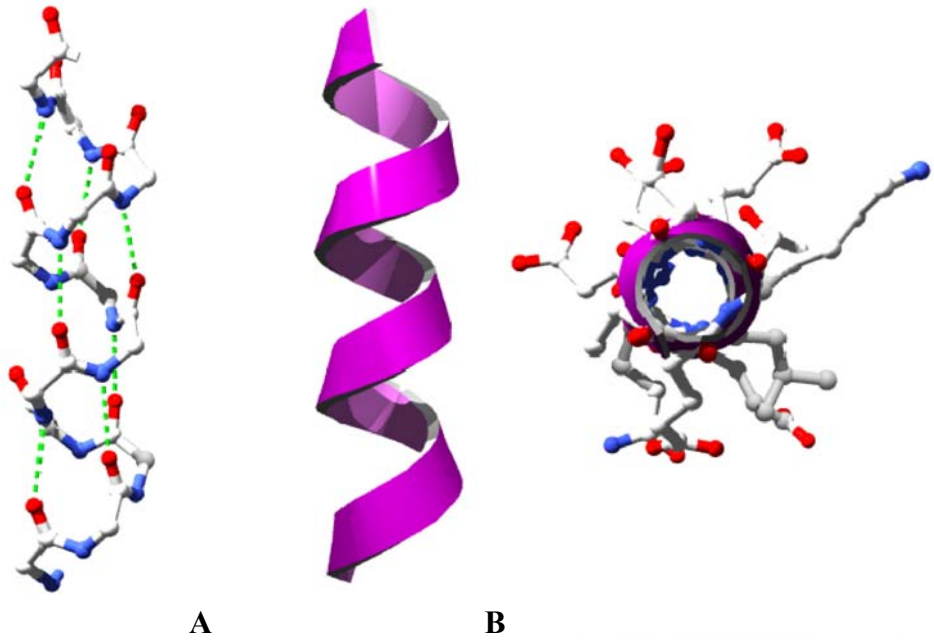


Figure 6 : Hélice α.
A : Seuls les atomes du squelette ont été représentés, les liaisons H sont symbolisées par des tirets verts.
B : La même hélice α représentée en ruban, ce style de schématisation est très utilisé pour représenter les protéines, il permet de mieux appréhender l'architecture globale des structures.
C : La même hélice mais vue selon son axe, les chaînes latérales sont présentes et on peut constater qu'elles se projettent vers l'extérieur.

4.2 Les feuillets β

Contrairement aux hélices, les feuillets β font intervenir des régions éloignées les unes des autres le long de la structure primaire. On les appelle les brins β. En réalité, ces derniers n'existent pas en tant que structures secondaires isolées car ils doivent s'associer en feuillets pour assurer leur stabilité. Les feuillets β sont composés de plusieurs brins, alignés les uns à côtés des autres entre lesquels s'établissent des liaisons hydrogène. Cet alignement peut s'organiser de deux façons différentes : si deux brins adjacents sont orientés dans le même sens biologique (de Nter vers Cter) on dit alors que le feuillet β est parallèle, si les deux brins sont dans des sens opposés on dit que le feuillet β est anti-parallèle.

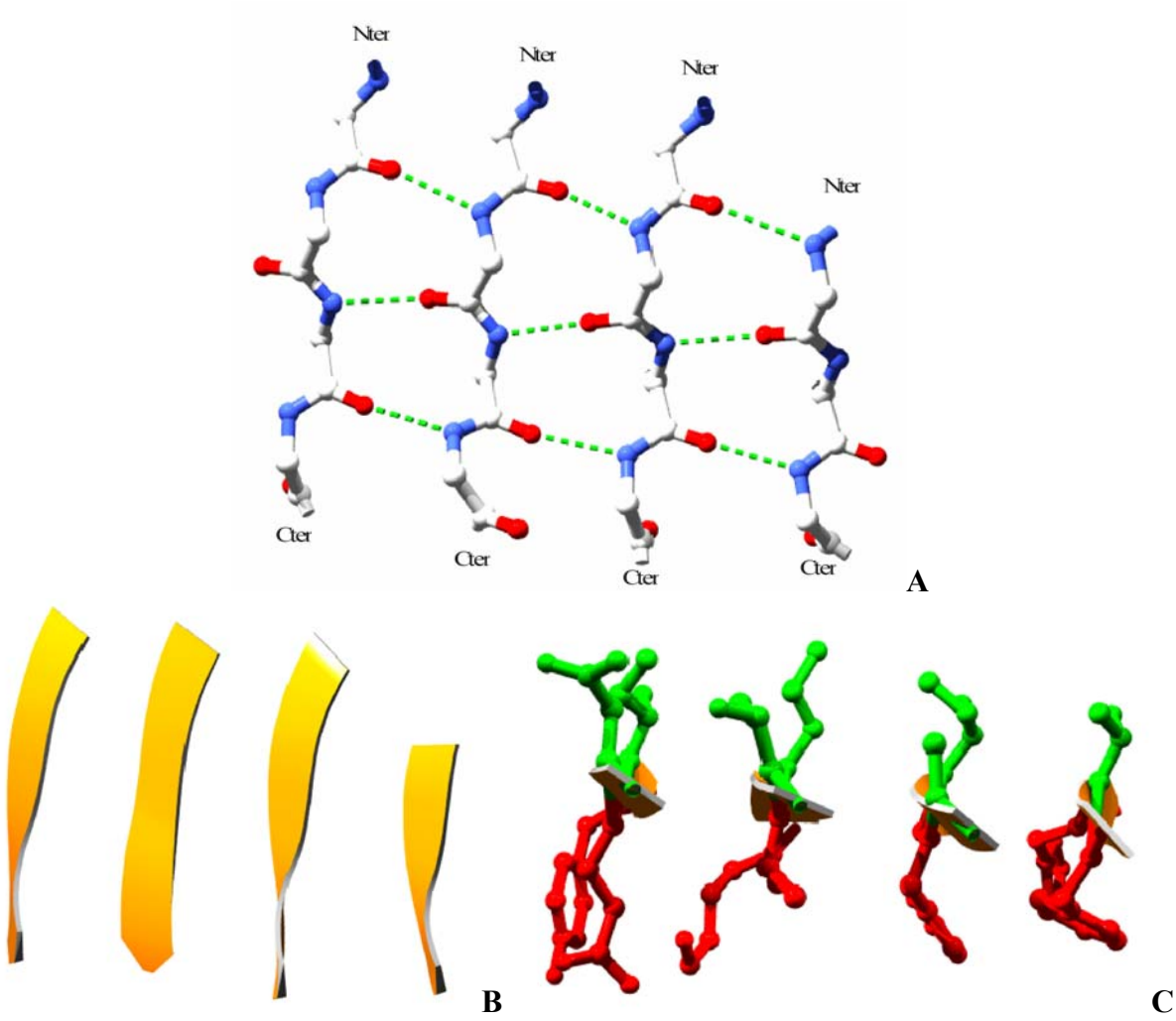
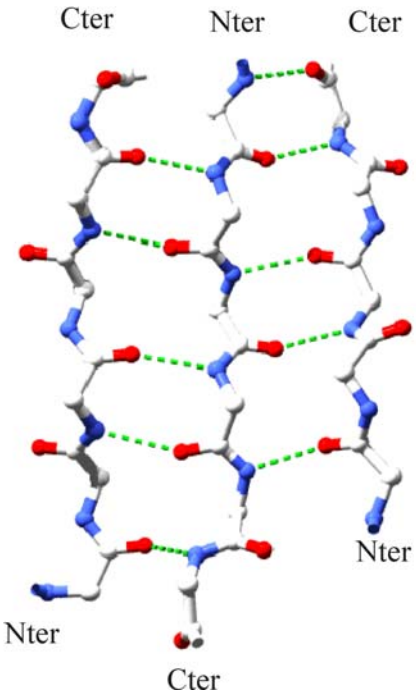


Figure 7 : Feuillet β parallèle.
A : Seuls les atomes du squelette ont été représentés, les liaisons H sont symbolisées par des tirets verts.
B : Le même feuillet β représenté en ruban.
C : Vue à 90° de B, les chaînes latérales sont représentées en rouge et en vert en fonction de leurs positions (dessus ou dessous) par rapport au plan défini par le squelette des brins.



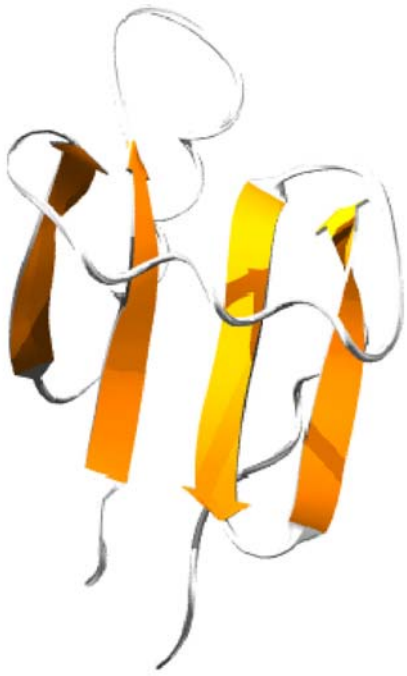


Figure 8 à gauche : Feuillet β anti-parallèle.
Seuls les atomes du squelette ont été représentés, les liaisons H sont symbolisées par des tirets verts.
Figure 9 à droite : Feuillet β mixte.
Ce feuillet est mixte, le sens Nter vers Cter est indiqué sur les brins par une flèche.

On peut constater en comparant la Figure 7 et la Figure 8 que les feuillets β parallèles et anti-parallèles présentent des motifs de liaisons hydrogène assez dissemblables. Pour les feuillets parallèles (Figure 7A) les liaisons hydrogène sont régulièrement espacées et forment des angles différents avec l'axe du squelette alors que pour les feuillets anti-parallèles (Figure 8) les liaisons, relativement parallèles entre elles, sont irrégulièrement espacées faisant succéder de petits intervalles à de plus grands. Il existe également des feuillets β mixtes qui contiennent des brins orientés de manière parallèle et anti-parallèle comme celui de la Figure 9.

La plupart des protéines sont construites à partir de combinaisons d'hélices et/ou de brins qui contiennent environ les deux tiers des résidus. Les hélices et les brins sont regroupés sous le terme générique de structures secondaires régulières par opposition aux boucles (« coils ») qui peuvent être également considérées comme des structures secondaires mais dont la régularité est nettement moins évidente, elles relient les hélices et les brins et contiennent le reste des AA. Ces boucles sont généralement situées à la surface des protéines et peuvent aussi être classées en sous-groupes.

5 - Structure tertiaire

La structure tertiaire d'une protéine est l'agencement de ses différentes structures secondaires entre elles. L'architecture adoptée permet à des AA éloignés en séquence d'être proches dans l'espace alors que la stabilité de l'ensemble est assurée par différents types de relations entre ces résidus : liaisons ioniques, liaisons hydrogène, compaction hydrophobe ou bien encore ponts disulfure. C'est lors du repliement de la chaîne polypeptidique que les groupes apolaires vont avoir tendance à se regrouper au cœur de la protéine alors que les groupes hydrophiles vont plutôt s'orienter vers l'extérieur. Un type particulier de repliement est caractérisé par la disposition relative des structures secondaires régulières entre elles, par leurs connexions et leur succession le long de la séquence.

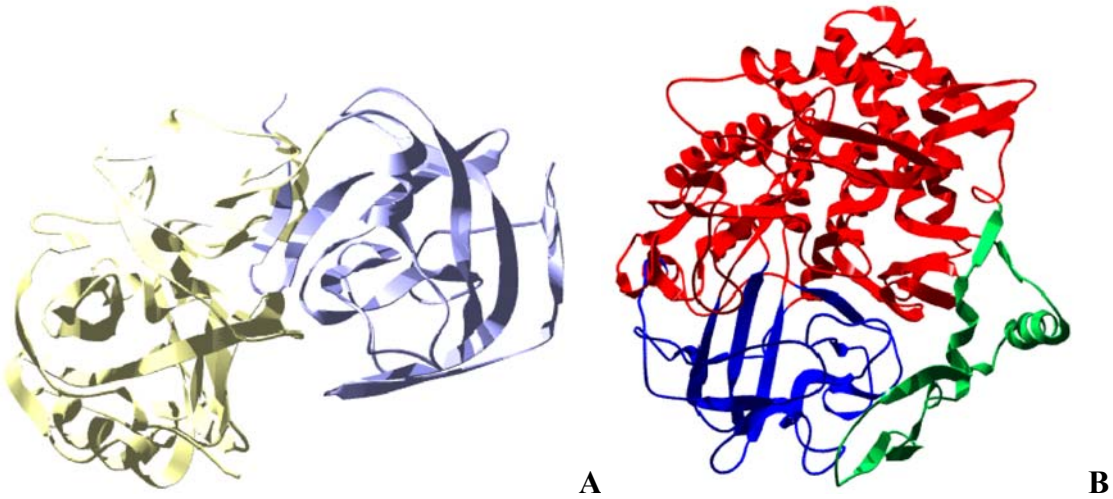


Figure 10 : Protéines multidomaines.
A : Structure de code PDB 1fkn[60], cette protéine est composée de deux domaines continus.
B : Structure de code PDB 1fbo[61], cette protéine est composée de 3 domaines discontinus, par exemple il est impossible de parcourir entièrement le domaine bleu sans passer par le domaine rouge, les morceaux de la séquence correspondant au domaine bleu ne sont pas tous connectés les uns aux autres.

Les structures protéiques peuvent être décomposées en unités de base, que l'on appelle domaines protéiques, qui sont des structures globulaires et compactes adoptant des repliements indépendants[55]. Ces domaines structuraux sont les briques de l'architecture des protéines et de leurs fonctions. Ils peuvent s'associer avec d'autres domaines de nature différente et selon des ordres très variés. Les domaines ont des longueurs comprises entre 20 et 400 AA et alors que les domaines comportant un nombre d'AA proche de 400 sont très rares, les plus petits sont plus fréquents et sont en général stabilisés par des ions métalliques ou par des ponts disulfure. Ces domaines peuvent être classés en deux grandes catégories : les

domaines continus sont formés par des régions peptidiques consécutives qui se replient pour former un domaine indépendant et unique alors que les domaines discontinus sont composés de régions peptidiques non consécutives (voir Figure 10). L'inspection visuelle des structures permet de détecter des régions protéiques compactes et globulaires qui constituent des domaines structuraux. Il existe plusieurs méthodes automatiques qui, à partir de calculs de compacité, de globularité ou de flexibilité des chaînes, permettent de déterminer les différents domaines d'une structure[56-59].

Une protéine peut être constituée d'une seule chaîne polypeptidique, elle est qualifiée dans ce cas de protéine monomérique, la protéine étant donc un monomère. L'ultime étape du repliement est constituée par l'assemblage éventuel de plusieurs monomères pour former un multimère. Les monomères peuvent être identiques ou différents et dans le cas de deux chaînes on parlera alors respectivement d'homodimérisation ou d'hétérodimérisation. L'assemblage des différentes chaînes correspond à la structure quaternaire (voir Figure 11).

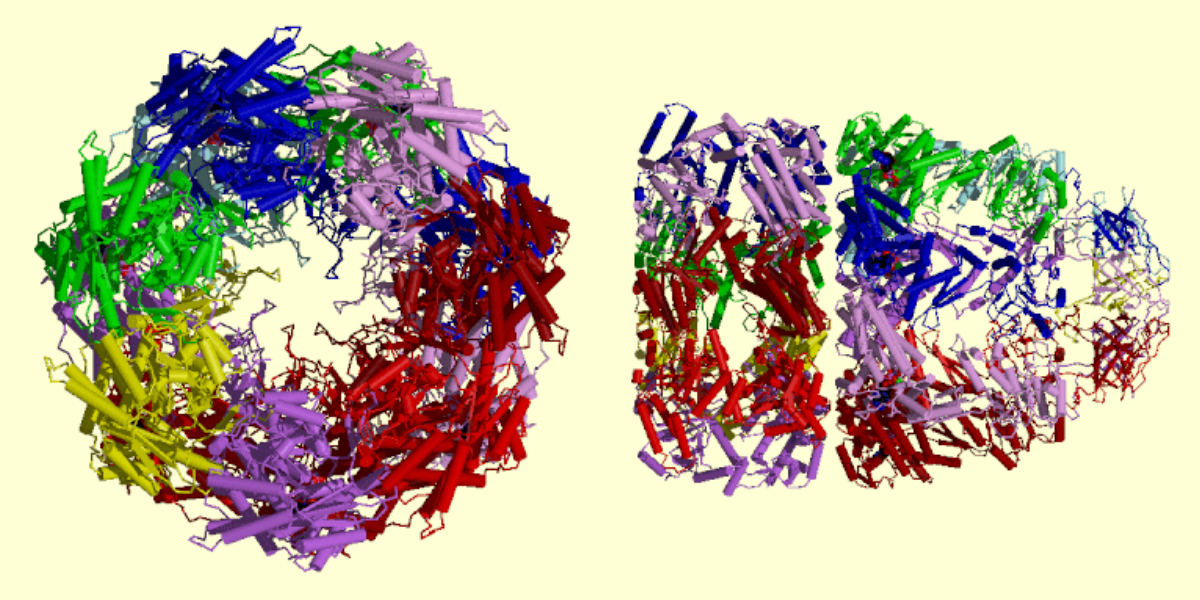


Figure 11 : Structure du complexe GROEL/GROES[62].
Cette structure est composée de 21 chaînes polypeptidiques.

6 - Répartition des protéines

Sur le plan structural, les protéines peuvent être réparties en cinq classes. La classe tout α regroupe les structures qui comme l'hémoglobine sont essentiellement constituées d'hélices (au moins 90% de leurs structures secondaires régulières). La classe tout β contient toutes les

protéines constituées essentiellement de brins β, comme les immunoglobulines. La classe α/β rassemble les protéines dont les structures sont construites par l'alternance d'hélices α et de brins β alors que la classe α+β correspond aux structures comportant des hélices et des brins répartis dans des régions plus ou moins distinctes (l'alternance n'est plus observée). La dernière catégorie rassemble les protéines de petites tailles (constituées souvent de moins de 70 AA) comportant généralement peu de structures secondaires régulières.

La plupart des protéines sont fonctionnelles en milieu aqueux mais certaines (dans leur globalité ou en partie) le sont au sein de membranes lipidiques (membranes cellulaires ou membranes de divers organites intracellulaires). Ces protéines membranaires remplissent des fonctions souvent essentielles pour les cellules telles que le transport de molécules à travers la bicouche lipidique, la catalyse enzymatique, la transmission de signaux chimiques, etc.

C h a p i t r e 2

Tessellation de Voronoï et triangulation de Delaunay

1 - Introduction

Si l'on considère un espace donné, une tessellation est un recouvrement de cet espace par des objets géométriques. Ce recouvrement présente deux caractéristiques importantes : tout l'espace est entièrement recouvert et il n'y a pas de superposition entre les différents objets. La tessellation la plus simple de l'espace Euclidien à trois dimensions est cet espace lui-même. Il existe un nombre infini de tessellations de cet espace, dans ce chapitre je m'intéresserai uniquement aux tessellations de Voronoï (TdV) et à la triangulation de Delaunay (appelée également parfois tessellation ou « tétraèdrisation » de Delaunay).

2 - Un peu d'histoire

Les TdV sont des motifs que l'on retrouve régulièrement dans la nature et il est fort probable que ces structures aient intéressé les esprits les plus curieux dès l'antiquité. Il faut cependant attendre le XVIIème siècle pour que les premières traces écrites (actuellement connues) de ces tessellations fassent leur apparition. En effet Descartes dans *Le Monde de Mr Descartes, ou Le Traité de la Lumière* publié en 1644 et dans *Principia Philosophiae*, également publié en 1644, utilise de telles tessellations pour décrire la répartition de la matière dans le système solaire et ses environs. La Figure 12 est extraite de ces ouvrages et présente une tessellation proche d'une TdV pondérée (j'explique ce terme dans la suite). Les premières apparitions indiscutables du concept apparaissent vraisemblablement dans le travail de Dirichlet (1850) et Voronoï (1908).

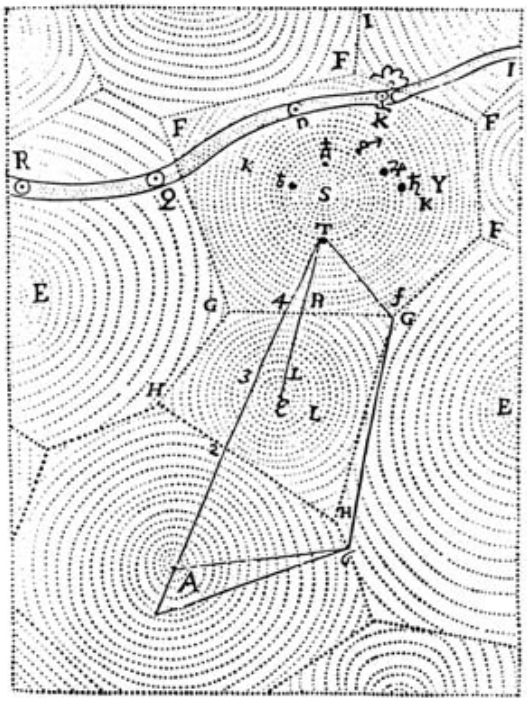


Figure 12 : « Répartition de la matière dans le système solaire et ses environs » (Descartes).
S en haut représente le soleil, ε au centre représente une étoile, les lignes continues courbes représentent la trajectoire d'une comète. On retrouve en pointillé les polygones caractéristiques des TdV, les cercles permettent de comprendre comment Descartes les a probablement tracés.

Figure 13
A gauche : Peter-Gustav Lejeune-Dirichlet (1805-1859).
A droite : Georgii Feodosevitch Voronoï (1868-1908).

Malgré l'antériorité des travaux de Dirichlet, c'est le nom de Voronoï qui restera plus fréquemment attaché aux tessellations, sans doute parce que Dirichlet s'intéressait aux tessellations à deux et trois dimensions alors que Voronoï traitait les cas beaucoup plus généraux des tessellations dans des espaces à N dimensions. Les premiers développements de ce concept concernaient des ensembles de points disposés de manière périodique dans

l'espace. Il était donc naturel à partir de la fin du XIXème siècle que les premières applications concernent le champ de la cristallographie. Parallèlement, le concept était à nouveau redécouvert dans deux autres domaines indépendants. Tout d'abord en météorologie, Thiessen en 1911 utilisait ces tessellations pour calculer des moyennes pluviométriques ; ceci explique pourquoi en météorologie et en géographie les applications à une et deux dimensions désignent les cellules de Voronoï par le terme de polygones de Thiessen. Le second domaine concerne la géologie, les tessellations ont été entre autres utilisées pour estimer les réserves de minerai à partir d'informations obtenues par forage. Depuis cette période de nombreuses redécouvertes ont eu lieu dans différents domaines aussi variés que la chimie, la physique des alliages, l'écologie, ou les sciences sociales. Les applications et les diverses versions des TdV sont devenues de plus en plus nombreuses à partir des années 1970 avec l'apparition de calculateurs de plus en plus puissants.

Le concept de tessellation de Delaunay, indissociable des TdV, trouve également son origine dans le travail de Voronoï, mais c'est Delaunay en 1934 qui le décrivit le premier en introduisant la notion de sphère vide largement utilisée ensuite et qui sera décrite plus loin.

Figure 14 : Boris Nikolaevitch Delone (1890-1980)
Delaunay est la version francisée de Delone.

3 - TdV à deux dimensions

3.1 Intuitivement

Dans ce court paragraphe j'introduis la notion de TdV et quelques termes dont les définitions plus théoriques seront données plus loin. Soit un ensemble de points du plan

Euclidien ; je les appellerai dans la suite les sites ou les germes de la tessellation. Je suppose qu'il y a plus de deux points dans cet ensemble, que leur nombre est fini, et qu'ils sont tous distincts. Etant donné l'ensemble de ces sites, j'associe chaque point du plan au site le plus proche. Si un point du plan est à égale distance de deux sites, je l'associe à ces deux sites. Pour chaque site, j'obtiens ainsi une région qui lui est propre, constituée des points du plan qui sont plus proches de ce site que de tous les autres. L'ensemble de ces régions couvre le plan de manière exhaustive car chaque point est attribué à au moins une région et cette décomposition est unique. Les points qui sont attribués à au moins deux sites forment les limites ou les frontières des régions qui sont donc les seuls points que peuvent avoir en commun deux régions distinctes. L'ensemble des régions ainsi définies recouvre totalement le plan et les régions ne se recouvrent pas sauf, encore une fois, au niveau de leurs frontières. Cet ensemble de régions forme une TdV que j'appellerai par la suite TdV à deux dimensions. Les différentes régions sont appelées polygones ou cellules de Voronoï.

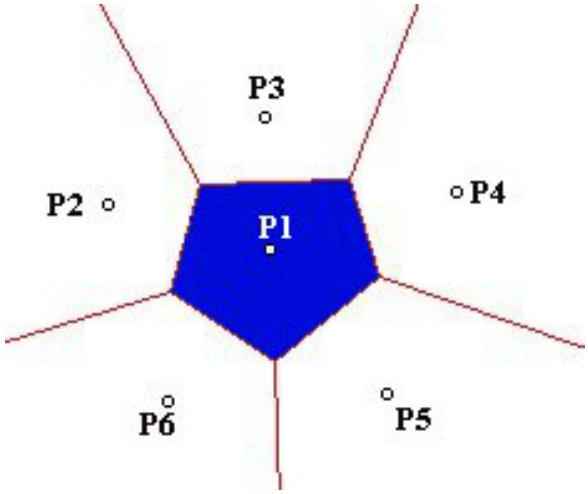


Figure 15 : TdV plane pour six sites.

Dans la figure ci-dessus, six germes sont représentés $\{P_1, P_2, P_3, P_4, P_5, P_6\}$ ainsi que leur cellule de Voronoï, la cellule générée par P_1 est colorée en bleu. Un point à l'intérieur de ce polygone est plus proche de P_1 que des cinq autres sites. Un point situé sur une arrête du polygone bleu est à égale distance de P_1 et du site associé au polygone partageant également cette arrête.

3.2 Mathématiquement

3.2.1 Première définition et notations

Je vais reprendre la définition intuitive du paragraphe précédent de manière plus rigoureuse en termes mathématiques, je vais également introduire les différentes notations que j'utiliserai dans cet exposé.

Soit un ensemble de n points (sites) notés $P_1,\ldots, P_n$ du plan Euclidien tel que $2 \leq n < \infty$. Le plan est muni d'un repère orthonormé et les points P_i ont pour coordonnées (x_{i1} , x_{i2}). Soit O l'origine du repère orthonormé, le vecteur $\overrightarrow{OP_j}$ sera noté $\vec{x_j}$. Les n points sont distincts, c'est à dire quels que soient i et j de $I_n=\{1,....,n\}$ avec $i\neq j$, on a $\vec{x}_i \neq \vec{x_j}$. Soit P (x_1 , x_2) un point du plan avec $\overrightarrow{OP}=\vec{x}$, la distance euclidienne entre P et P_i notée $d(P,P_i)$ est égale à $\left\|\vec{x}-\vec{x_i}\right\|=\sqrt{(x_1-x_{i1})^2+(x_2-x_{i2})^2}$. Si P est plus proche de P_i que P_j quel que soit $i\neq j$ j'obtiens naturellement la relation $\left\|\vec{x}-\vec{x_i}\right\|\leq\left\|\vec{x}-\vec{x_j}\right\| \quad \forall j\in I_n$ avec $i\neq j$. Dans ce cas P est attribué à P_i. J'obtiens donc la définition mathématique suivante :

Définition d'une TdV à deux dimensions :

- Soit $\mathcal{P}= \{P_1,\ldots, Pn\}\subset\mathbb{R}^2$ tel que $2 \leq n < \infty$
- $\forall(i,j)\in I_n^2 \quad \overrightarrow{OP_i}=\vec{x_i}$ et $\overrightarrow{OP_j}=\vec{x_j}$,si $i\neq j$ alors $\vec{x}_i \neq \vec{x_j}$
- Le polygone de Voronoï associé à P_i est défini par :

$$V(P_i)=\left\{P/\left\|\vec{x}-\vec{x_i}\right\|\leq\left\|\vec{x}-\vec{x_j}\right\|,\forall j\neq i, j\in I_n\right\} \qquad \textbf{Équation 1}$$

- $\mathcal{V}= \{V(P_1),\ldots, V(P_n)\}$ est donc la TdV à deux dimensions générée par l'ensemble $\mathcal{P}$.

Dans l'équation 1, le signe ≤ implique que les cellules de Voronoï sont des ensembles fermés. Les frontières de ces régions, notées $\delta V(P_i)$, sont constituées de segments, de demi-droites ou de droites, chacun de ces éléments sera appelé une arrête de Voronoï et sera noté e_i, ces arrêtes sont par définition à égale distance des deux sites qui les ont générées, elles sont donc incluses dans les médiatrices des segments joignant ces deux sites. En réalité, une arrête de Voronoï est un segment, une demi-droite ou une droite partagée par deux régions,

ceci implique que si $V(P_i) \cap V(P_j) \neq \varnothing$ alors $V(P_i) \cap V(P_j)$ est une arrête de Voronoï (qui peut être dégénérée en un point), pour plus de précision la notation e_i deviendra $e(P_i, P_j)$. Si $e(P_i, P_j)$ n'est ni un point, ni l'ensemble vide alors les cellules $V(P_i)$ et $V(P_j)$ sont dites adjacentes. Les intersections des arrêtes de Voronoï entre elles sont appelées les sommets des cellules de Voronoï et sont notées q_i. S'il existe un sommet de la tessellation $\mathcal{V}$ correspondant à l'intersection d'au moins quatre arrêtes, la tessellation est dite dégénérée. La figure suivante présente un exemple de TdV générée par un ensemble de huit sites. Cette tessellation est dégénérée puisque le sommet S1 est l'intersection de quatre arrêtes de Voronoï.

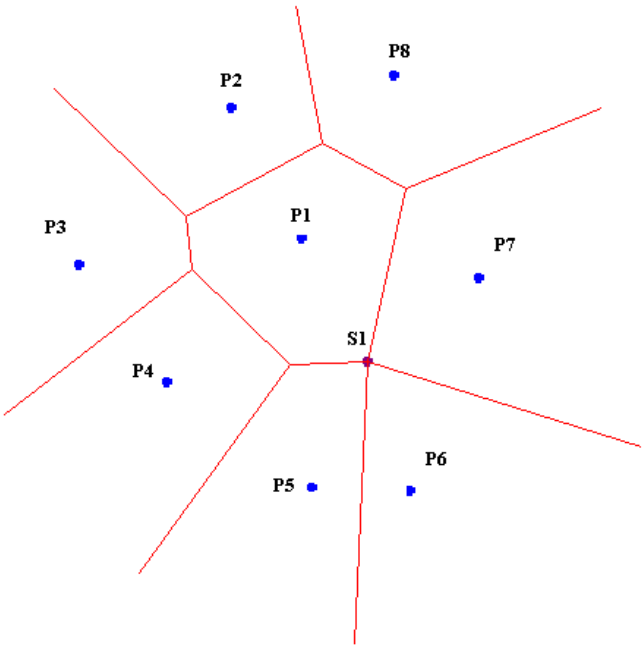


Figure 16 : TdV plane dégénérée, pour huit sites.

3.2.2 Seconde définition

Une autre façon de définir les TdV à deux dimensions consiste à considérer des demi-plans. Soient P_i et P_j deux sites de l'ensemble $\mathcal{P}$, on considère le segment $[P_iP_j]$ et sa médiatrice notée $b(P_i, P_j)$ définie par :

$$b(P_i,P_j)=\left\{P/\left\|\vec{x}-\vec{x_i}\right\|=\left\|\vec{x}-\vec{x_j}\right\|, j\neq i\right\} \qquad \textbf{Équation 2}$$

La médiatrice divise le plan en deux demi-plans, j'appelle le demi-plan dans lequel se trouve P_i la région de domination de P_i sur P_j notée $H(P_i, P_j)$ et définie par :

$$H(P_i,P_j)=\left\{P/\left\|\vec{x}-\vec{x_i}\right\|\leq\left\|\vec{x}-\vec{x_j}\right\|,j\neq i\right\}$$ **Équation 3**

Cette équation signifie que tout point P situé dans $H(P_i, P_j)$ sera plus proche de P_i que de P_j. Dans la figure suivante, les trois demi-plans de domination de P_1 sur P_2, P_3 et P_4 sont hachurés respectivement en vert, rouge et bleu. Le polygone de Voronoï généré par P_1 est le triangle d'intersection de ces trois demi-plans. Cet exemple permet de mieux comprendre la définition suivante qui indique qu'une cellule de Voronoï associée à un site P_i est l'intersection de toutes les régions de domination de P_i sur tous les autres sites :

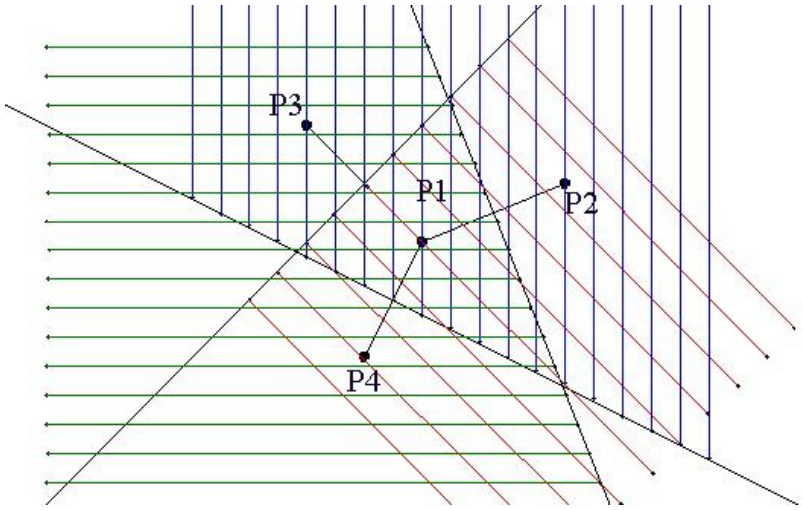


Figure 17 : Une cellule de Voronoï obtenue à partir de demi-plans.

Autre définition d'une TdV à deux dimensions :

- Soit $\mathcal{P}=\{P_1,\ldots,P_n\}\subset\mathbb{R}^2$ tel que $2\leq n<\infty$
- $\forall(i,j)\in I_n^2\ \ \overrightarrow{OP_i}=\vec{x_i}$ et $\overrightarrow{OP_j}=\vec{x_j}$,si $i\neq j$ alors $\vec{x}_i\neq\vec{x}_j$.
- Le polygone de Voronoï associé à P_i est défini par :

$$V(P_i)=\bigcap_{j\in I_n-\{i\}}H(P_i,P_j)$$ **Équation 4**

- $\mathcal{V}=\{V(P_1),\ldots,V(P_n)\}$ est donc de la même manière que précédemment la TdV à deux dimensions générée par l'ensemble $\mathcal{P}$.

L'équivalence entre les deux définitions réside dans le fait que $\left\|\vec{x}-\vec{x_i}\right\| \leq \left\|\vec{x}-\vec{x_j}\right\|$ si et seulement si $P(\vec{x}) \in H(P_i, P_j)$ avec $i \neq j$.

4 - Tessellations de Voronoï à trois dimensions

Les deux définitions vues plus haut sont très facilement généralisables à des dimensions supérieures et donc plus particulièrement à trois dimensions. Les différents points et vecteurs ont maintenant trois coordonnées, les médiatrices deviennent des plans médians et les demi-plans des demi-espaces. Les deux définitions précédentes peuvent être réunies en une seule :

- Soit $\mathcal{P} = \{P_1, \ldots, Pn\} \subset \mathbb{R}^3$ tel que $2 \leq n < \infty$
- $\forall (i,j) \in I_n^2 \quad \overrightarrow{OP_i} = \vec{x_i}$ et $\overrightarrow{OP_j} = \vec{x_j}$,si $i \neq j$ alors $\vec{x}_i \neq \vec{x}_j$
- Le polyèdre de Voronoï associé à P_i est défini par :

$$V(P_i) = \bigcap_{j \in I_n - \{i\}} H(P_i, P_j) = \left\{ P / \left\|\vec{x}-\vec{x_i}\right\| \leq \left\|\vec{x}-\vec{x_j}\right\|, \forall j \neq i, j \in I_n \right\} \qquad \textbf{Équation 5}$$

- $\mathcal{V} = \{V(P_1), \ldots, V(P_n)\}$ est la TdV à trois dimensions générée par l'ensemble $\mathcal{P}$.

Pour les tessellations à trois dimensions, les cellules de Voronoï deviennent des polyèdres, la frontière d'une cellule est donc constituée de faces appelées faces de Voronoï, chacune de ces faces est incluse dans le plan médian du segment joignant deux sites. Ces faces peuvent être finies ou infinies et leurs frontières sont des segments ou des demi-droites ou bien encore des droites appelées les arrêtes de Voronoï. Les limites de ces arrêtes sont des points appelés les sommets de Voronoï. L'ensemble de l'espace est ainsi pavé de manière exhaustive et unique par des polyèdres jointifs.

5 - Triangulation de Delaunay

Comme je vais le montrer par la suite, les TdV et les triangulations de Delaunay sont deux outils mathématiques fortement liés. De plus, les algorithmes informatiques permettant de calculer les différentes propriétés des cellules de Voronoï sont généralement fondés sur une

triangulation de Delaunay, il est donc important d'en présenter ici les définitions et quelques propriétés.

La figure ci-dessous représente une TdV (en pointillés) générée par un ensemble de quinze sites. Si les segments joignant deux sites dont les cellules sont adjacentes sont tracés, on obtient un ensemble de triangles que j'appellerai une triangulation de Delaunay.

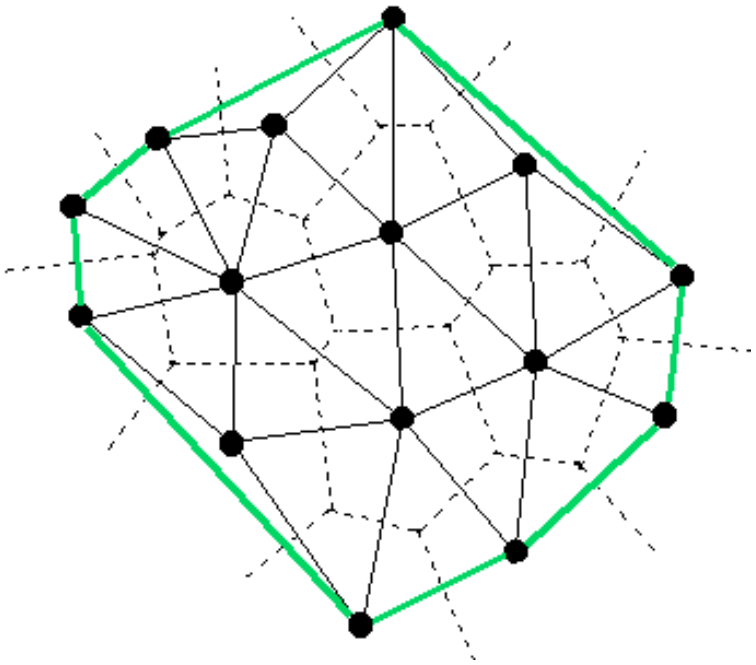


Figure 18 : Triangulation de Delaunay (ligne continue) et TdV (en pointillé). En vert est représentée l'enveloppe convexe de l'ensemble des sites.

Soit $\mathcal{V} = \{V(P_1),\ldots, V(P_n)\}$ une TdV dans le plan Euclidien, je supposerai dans la suite que les différents sites ne sont pas alignés ce qui implique notamment que le cardinal de $\mathcal{P} = \{P_1,\ldots, P_n\}$ est supérieur à deux, en effet une triangulation n'est possible par définition que pour au moins trois éléments non alignés. Soit Q_i un des sommets de la tessellation, je noterai $V(P_{i1}),\ldots, V(P_{ij}), V(P_{i(j+1)}),\ldots, V(P_{ik(i)})$ les polygones de Voronoï partageant ce sommet, ces polygones sont numérotés dans un ordre correspondant au sens trigonométrique autour de ce sommet. L'ensemble des segments $[P_{ij}P_{i(j+1)}]$ forme un polygone entourant Q_i, la construction de tous les polygones associés à tous les sommets de toutes les cellules donne, si la TdV n'est pas dégénérée, un ensemble de triangles, (Figure 18) nommé triangulation de Delaunay. Il est intéressant de noter que la triangulation de Delaunay est une tessellation de l'enveloppe convexe de $\mathcal{P}$ notée $\mathcal{EC}$, dont la frontière notée $\delta\mathcal{EC}$ est représentée en vert sur la Figure 18. Parmi l'infinité de polygones convexes contenant $\mathcal{P}$ on appelle enveloppe convexe le plus

petit de ces polygones, $\mathcal{EC}$ est donc également l'intersection de l'infinité de polygones convexes contenant $\mathcal{P}$.

Définition d'une triangulation de Delaunay à deux dimensions :

- Soit $\mathcal{P}= \{P_1,\ldots, P_n\}\subset\mathbb{R}^2$ tel que $3 \leq n < \infty$ et tels que $P_1,\ldots, P_n$ ne soient pas colinéaires
- Soit $Q= \{Q_1,\ldots, Q_m\}$l'ensemble des sommets de la tessellation $\mathcal{V}$générée par $\mathcal{P}$
- Soient $\overrightarrow{x_{i1}},\ldots,\overrightarrow{x_{ik(i)}}$ les vecteurs de positions des germes $P_{i1},\ldots, P_{ik(i)}$ dont les polygones $V(P_{i1}),\ldots, V(P_{ik(i)})$ partagent le même sommet Q_i

$$T_i = \left\{ P / \vec{x} = \sum_{j=1}^{k(i)} \lambda_i \overrightarrow{x_{ij}}, avec \sum_{j=1}^{k(i)} \lambda_j = 1, \lambda_j \geq 0, j \in I_{k(i)} \right\} \qquad \textbf{Équation 6}$$

- $\mathcal{D}= \{T_1,\ldots, T_m\}$

Si $\forall i\in I_m$ on a $k(i)$=3, $\mathcal{D}$ est appelée la triangulation de Delaunay de $\mathcal{P}$. Dans les équations précédentes $k(i)$ représente le nombre de sites entourant le sommet Q_i. Les frontières des triangles sont des segments, que j'appellerai dans la suite des arrêtes de Delaunay. Plus précisément si une arrête est commune à deux triangles, je l'appellerai arrête interne par opposition aux arrêtes externes qui n'appartiennent qu'à un seul triangle, l'ensemble de ces arrêtes externes constitue la frontière de l'enveloppe convexe $\delta\mathcal{EC}$ de $\mathcal{P}$. Dans le cas d'une tessellation non dégénérée, chaque arrête de Voronoï correspond à une arrête de Delaunay et inversement (voir Figure 18), il y a donc autant d'arrêtes de Delaunay que d'arrêtes de Voronoï pour un ensemble de germes $\mathcal{P}$. Il existe cependant une grande différence entre ces deux catégories d'arrêtes. En effet les arrêtes de Delaunay sont toujours des segments alors que celles de Voronoï peuvent être également des demi-droites ou des droites. Les extrémités des arrêtes de Delaunay que j'appellerai des sommets de Delaunay, sont les germes ou les sites constituant l'ensemble $\mathcal{P}$.

La définition que je viens de présenter implique la construction préalable d'une TdV, il est bien sûr possible de déterminer une triangulation de Delaunay directement à partir d'un ensemble $\mathcal{P}$de germes, j'expose ce point dans le paragraphe suivant. Il est également possible de définir une triangulation dans des espaces à m dimensions, et en particulier dans l'espace Euclidien à 3 dimensions.

- Soit $\mathcal{V}$ la TdV à trois dimensions générée par $\mathcal{P} = \{P_1,\ldots, P_n\} \subset \mathbb{R}^3$ tel que $4 \leq n < \infty$ et tels que $P_1,\ldots, P_n$ ne soient pas coplanaires.
- Soit $Q = \{Q_1,\ldots, Q_m\}$ l'ensemble des sommets de la tessellation $\mathcal{V}$ générée par $\mathcal{P}$.
- Soient les polyèdres $V(P_{i1}),\ldots, V(P_{ik(i)})$ partageant le même sommet Q_i
- Soit P_{ij} le germe générant le polyèdre $V(P_{ij})$ dont un des sommets est Q_i. L'ensemble des segments $[P_{ij}P_{i(j+1)}]$ forme un polyèdre T_i, entourant Q_i, la construction de tous les polyèdres associés à tous les sommets de toutes les cellules donne, si la TdV n'est pas dégénérée, un ensemble de tétraèdres noté $\mathcal{D}$ que j'appellerai triangulation de Delaunay.

6 - Propriétés des tessellations de Voronoï

- Soit $\mathcal{P} = \{P_1,\ldots, P_n\} \subset \mathbb{R}^3$ tel que $2 \leq n < \infty$ et tels que $P_1,\ldots, P_n$ soient distincts
- $\mathcal{V} = \{V(P_1),\ldots, V(P_n)\}$ satisfait les relations :
  - $\bigcup_{i=1}^{n} V(P_i) = \mathbb{R}^3$
  - $\forall (i,j) \in I_n^2, i \neq j$ on a $\left[V(P_i)/\delta V(P_i)\right] \cap \left[V(P_j)/\delta V(P_j)\right] = \phi$

Ces deux propriétés indiquent que les cellules de Voronoï pavent entièrement l'espace et qu'elles non pas de points en commun si l'on ne considère pas leurs frontières.

6.1 Cellules finies, cellules infinies

Puisque les différentes cellules de Voronoï pavent entièrement l'espace, certaines cellules sont infinies, si les différents sites de $\mathcal{P}$ ne sont pas colinéaires la propriété suivante est vérifiée :

Soit $\mathcal{P} = \{P_1,\ldots, P_n\} \subset \mathbb{R}^3$ tel que $3 \leq n < \infty$ et tels que $P_1,\ldots, P_n$ ne soient pas colinéaires, une cellule de $\mathcal{V}$, $V(P_i)$ sera infinie si et seulement si P_i appartient à $\delta \mathcal{E}C$.

Ceci est illustré à deux dimensions par la Figure 19 qui reprend les mêmes points que la Figure 18. Dans cette dernière, pour améliorer la clarté de la lecture, certaines cellules n'étaient pas fermées, elles apparaissent maintenant en bleu et on constate que les points qui

ont généré ces polygones n'appartiennent pas à $\delta\mathcal{E}C$ toujours représentée en vert. La cellule représentée en rouge est bien une cellule infinie et son germe associé fait partie de $\delta\mathcal{E}C$.

Cette dernière propriété met en évidence un problème que j'exposerai plus en détail par la suite et qui concerne les cellules en surface dans les structures protéiques. Parmi ces cellules, nous verrons que certaines sont infinies et que d'autres ont des formes très allongées.

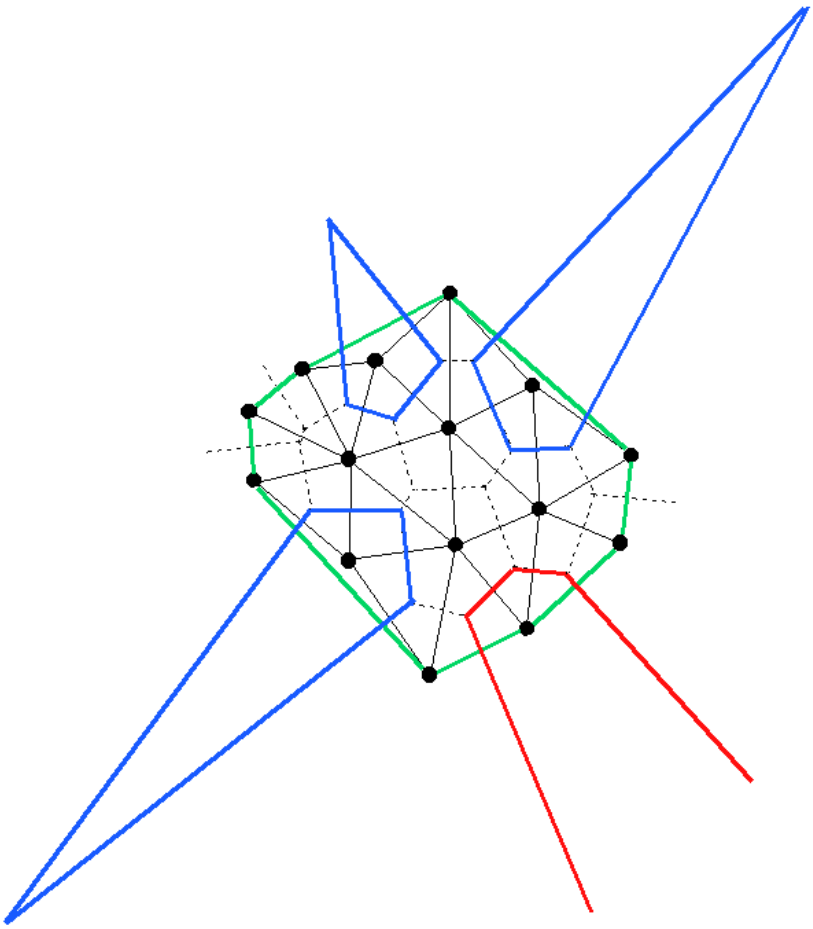


Figure 19 : Triangulation de Delaunay et TdV sur les mêmes sites que la Figure 18. En vert est représentée l'enveloppe convexe de l'ensemble des sites.

6.2 Sphères circonscrites

J'ai indiqué plus haut que les faces des cellules de Voronoï d'une tessellation à trois dimensions sont incluses dans les plans médians joignant les deux sites qui ont généré cette face, tous les points de cette face sont donc à égales distances de ces deux sites. Une arrête est l'intersection de deux faces, les points de cette arrête sont donc à égales distances des trois sites qui l'ont générée. Dans une tessellation non dégénérée, l'intersection de deux arrêtes est

un sommet de Voronoï, ce sommet est donc à égale distance des quatre sites qui l'ont généré. On en déduit donc la propriété suivante : un sommet d'une cellule de Voronoï est le centre de la sphère circonscrite au tétraèdre de Delaunay ayant généré le sommet de Voronoï considéré. La caractéristique fondamentale de cette sphère est qu'elle ne contient aucun des germes de la tessellation.

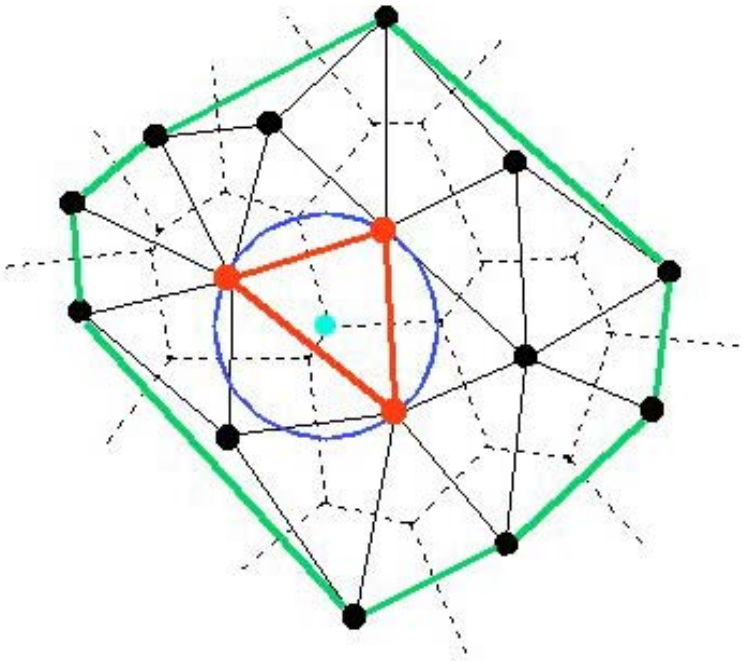


Figure 20 : Reprise de la Figure 18. Le cercle en bleu ne contient aucun site de l'ensemble, son centre est donc un des sommets de la TdV.

La Figure 20 illustre cette caractéristique, pour plus de clarté elle représente toujours une tessellation à deux dimensions. Le point bleu clair indique un sommet de la tessellation, les points oranges indiquent les sites qui l'ont généré. Le triangle de Delaunay correspondant est indiqué lui aussi en orange. Comme il est facile de le constater sur la figure, les arrêtes issues du sommet considéré sont incluses dans les médiatrices des trois segments oranges, le sommet est donc à égale distance des trois sites, il est par conséquent le centre du cercle circonscrit au triangle, ce cercle est représenté en bleu. A l'intérieur de ce cercle, il n'y a pas d'autres sites ; si tel était le cas, le triangle considéré ne serait pas un triangle de Delaunay et la tessellation ne serait pas une TdV.

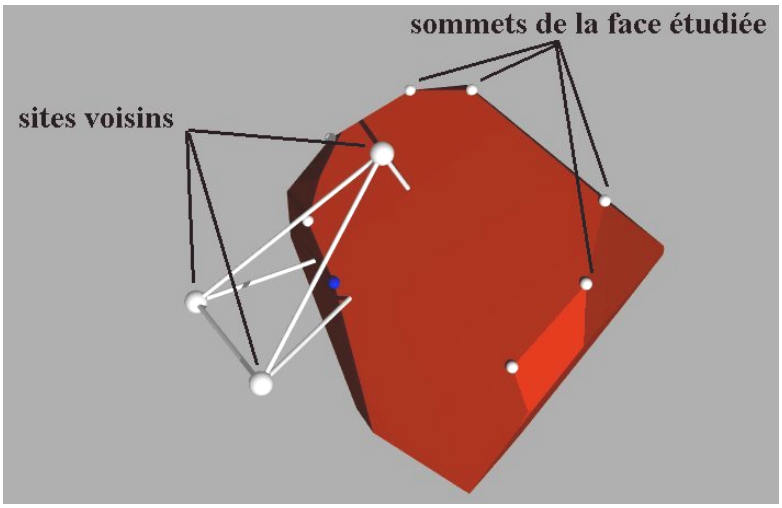


Figure 21 : Une cellule de Voronoï à trois dimensions. Le sommet en bleu est déterminé à l'aide du tétraèdre de Delaunay représenté en blanc.

La Figure 21 illustre cette propriété à trois dimensions, une cellule est représentée en rouge et les sommets d'une face sont mis en évidence par des petites sphères. Le sommet signalé par une sphère bleue a été généré par le tétraèdre représenté en blanc. Le sommet qui a généré la cellule n'est pas visible.

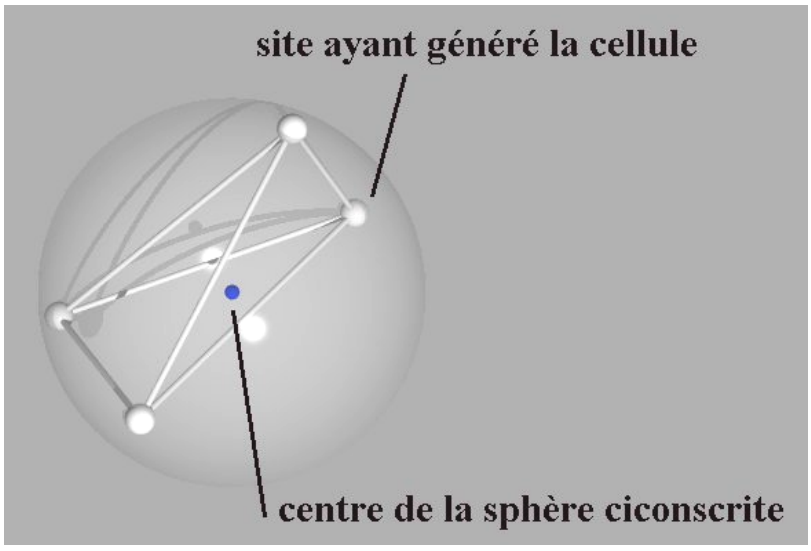


Figure 22 : Même tétraèdre que dans la Figure 21. Sa sphère circonscrite a pour centre le point bleu qui est donc un sommet de Voronoï.

La Figure 22 montre le tétraèdre précédent sans la cellule de Voronoï, le sommet représenté en bleu est le centre de la sphère circonscrite au tétraèdre. Je montrerai plus loin que cette propriété est à la base de l'algorithme qui m'a permis de construire les tessellations. Pour tous les sommets de la face considérée, on obtient la Figure 23 dans laquelle chaque

sommet est associé un tétraèdre. Il est d'ores et déjà intéressant de noter que le sommet associé à un tétraèdre n'est pas nécessairement contenu dans ce tétraèdre, par exemple le sommet représenté en jaune est bien le centre de la sphère circonscrite au tétraèdre lui donnant naissance, ce centre se situe pourtant hors de ce tétraèdre. On constate également que si les arrêtes rouge et verte traversent bien les faces auxquelles elles sont associées, en revanche l'arrête bleue ne traverse pas la face bleue.

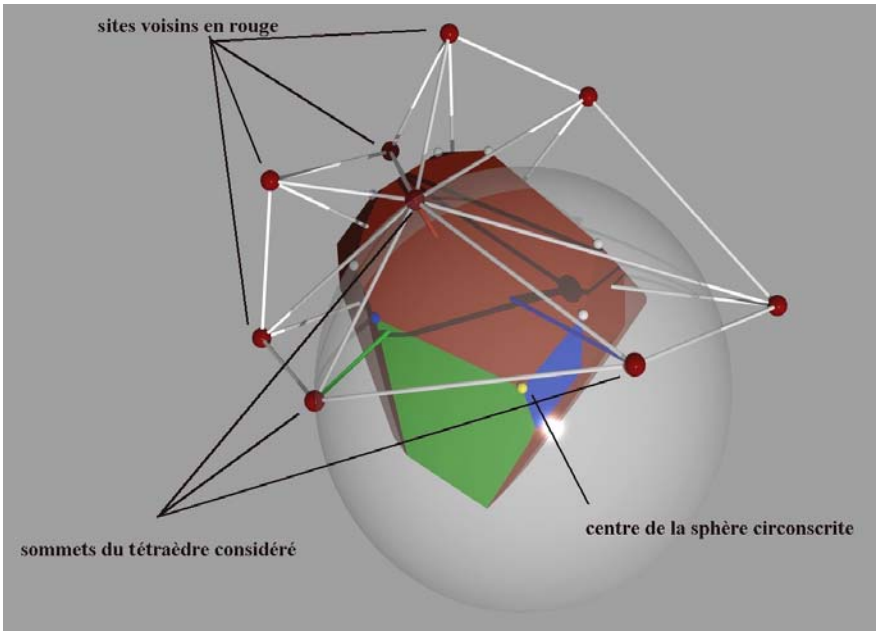


Figure 23 : Même cellule que dans la Figure 21. Tous les tétraèdres permettant de déterminer les sommets (petites sphères) de la face sont représentés. La sphère transparente est celle ayant permis de construire le sommet en jaune. Ce dernier n'est pas inclus dans son tétraèdre.

6.3 Dégénérescence

La dégénérescence de la TdV est associée également à cette propriété. En effet, dans le cas d'une tessellation dégénérée, il existe au moins un sommet qui est le centre (à deux dimensions) d'un cercle auquel appartiennent au moins quatre sites et non plus trois, et à trois dimensions d'une sphère à laquelle appartiennent au moins cinq sites et non plus quatre.

La Figure 24 reprend la Figure 16, le sommet S_1 appartient à quatre cellules, il est donc dégénéré et il est le centre du cercle représenté en vert passant par P_1, P_5, P_6 et P_7, la triangulation est ici ambiguë puisque l'on peut considérer à la fois d'une part les triangles (P_1, P_5, P_6) et (P_1, P_6, P_7) ou d'autre part les triangles (P_5, P_6, P_7) et (P_1, P_5, P_7). Ce cas est très peu probable dans les structures protéiques et il ne sera pas développé dans ce rapport.

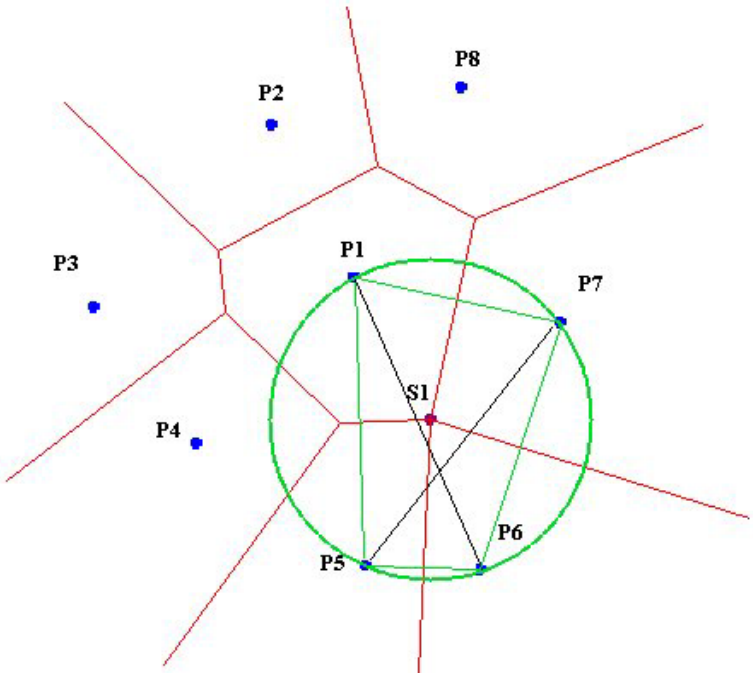


Figure 24 : Reprise de la Figure 16. Le sommet de Voronoï S1 est dégénéré car il est le centre d'un cercle circonscrit à quatre triangles différents.

6.4 Nombre de sites, nombre de faces etc.

Soit n le nombre de sites contenus dans l'espace Euclidien, on suppose que $3 \leq n < \infty$, n est également le nombre de cellules de la TdV générée par ces sites, car il y une cellule par germe. Soit n_f le nombre de faces de la tessellation, n_a le nombre d'arrêtes et n_s le nombre de sommets. Ces différentes valeurs ne sont pas indépendantes et, si la tessellation n'est pas dégénérée et si les différents sites ne sont pas colinéaires, elles sont reliées par la relation suivante : $n_s - n_a + n_f - n = -1$

Les cellules d'une tessellation dans l'espace Euclidien, sont donc des objets infinis ou finis, associés à des points. Si ces points sont repérables, par un numéro par exemple, il est possible également de repérer chaque cellule. Ces cellules peuvent être caractérisées par leur volume, leur surface, leur nombre de faces, leur forme (plus ou moins allongée par exemple) ou bien le simple fait qu'elles soient fermées ou ouvertes, c'est à dire finies ou infinies. Les faces de ces cellules peuvent être également caractérisées par leur nombre de côtés, leur aire, leur périmètre, la distance entre les sites qui les ont générées ou bien encore le fait qu'elles soient finies ou infinies. Ces propriétés sont en fait le reflet de l'environnement immédiat de chaque site et de l'agencement local.

7 - Propriétés des triangulations de Delaunay

- Soit $\mathcal{P} = \{P_1,\ldots, P_n\} \subset \mathbb{R}^3$ tel que $4 \leq n < \infty$ et tels que $P_1,\ldots, P_n$ soient distincts et qu'il n'existe pas de sphère contenant cinq sites (hypothèse de non dégénérescence)
- $\mathcal{D} = \{T_1,\ldots., T_i ,\ldots., T_m\}$ avec T_i définit par l' Équation 6, satisfait les relations :
  - $\bigcup_{i=1}^{n} T_i = \mathcal{EC}$, $\mathcal{EC}$ étant l'enveloppe convexe de $\mathcal{P}$ définie plus haut
  - $\forall (i,j) \in I_n^2, i \neq j$ on a $\lfloor T \ / \delta T_i \rfloor \cap \lfloor T_j \ / \delta T_j \rfloor = \phi$

Ces deux propriétés indiquent que les tétraèdres de Delaunay pavent entièrement l'enveloppe convexe de $\mathcal{P}$ et qu'ils non pas de points en commun si l'on ne considère pas leurs frontières. Les faces externes, c'est à dire celles qui ne sont pas communes à deux tétraèdres, constituent la frontière de $\mathcal{EC}$. Ceci signifie donc implicitement que contrairement aux cellules de Voronoï, toutes les cellules de Delaunay sont finies. Une tessellation de Delaunay dans l'espace Euclidien est uniquement constituée de tétraèdres et toutes les faces de contact sont des triangles.

7.1 Notion de plus proche voisinage

Dans les paragraphes précédents, la triangulation de Delaunay a été construite à partir de la TdV. Comme je l'ai déjà indiqué, il est possible de déterminer cette triangulation directement à partir des différents sites. Ceci se conçoit aisément dans l'espace Euclidien, si l'on considère que parmi tous les tétraèdres possibles, c'est à dire ceux dont les sommets sont des sites distincts, les tétraèdres de Delaunay sont ceux qui ne contiennent aucun autre site. Cette méthode de construction est plus ou moins fastidieuse en fonction du nombre de sites considérés. Il est intéressant de noter que cette triangulation définit de manière précise un ensemble de paires de sites voisins. Si l'on considère un site donné, et l'ensemble des tétraèdres de Delaunay dont un des sommets est le site considéré, comme ces tétraèdres ne contiennent par définition aucun autre site, l'ensemble des autres sommets de ces tétraèdres définit le plus proche voisinage du site considéré. Ce voisinage peut être également caractérisé à l'aide des cellules de Voronoï, en effet à chacune des arrêtes de ces tétraèdres correspond une face de la cellule générée par le site considéré, les propriétés de ces faces (aire, nombre de côtés etc.) caractérisent le voisinage. On peut donc en conclure que les TdV

définissent pour un site donné son plus proche voisinage à l'aide des faces des cellules, et qu'elles présentent l'avantage par rapport aux triangulations de Delaunay de pouvoir caractériser ce voisinage à l'aide des propriétés des faces.

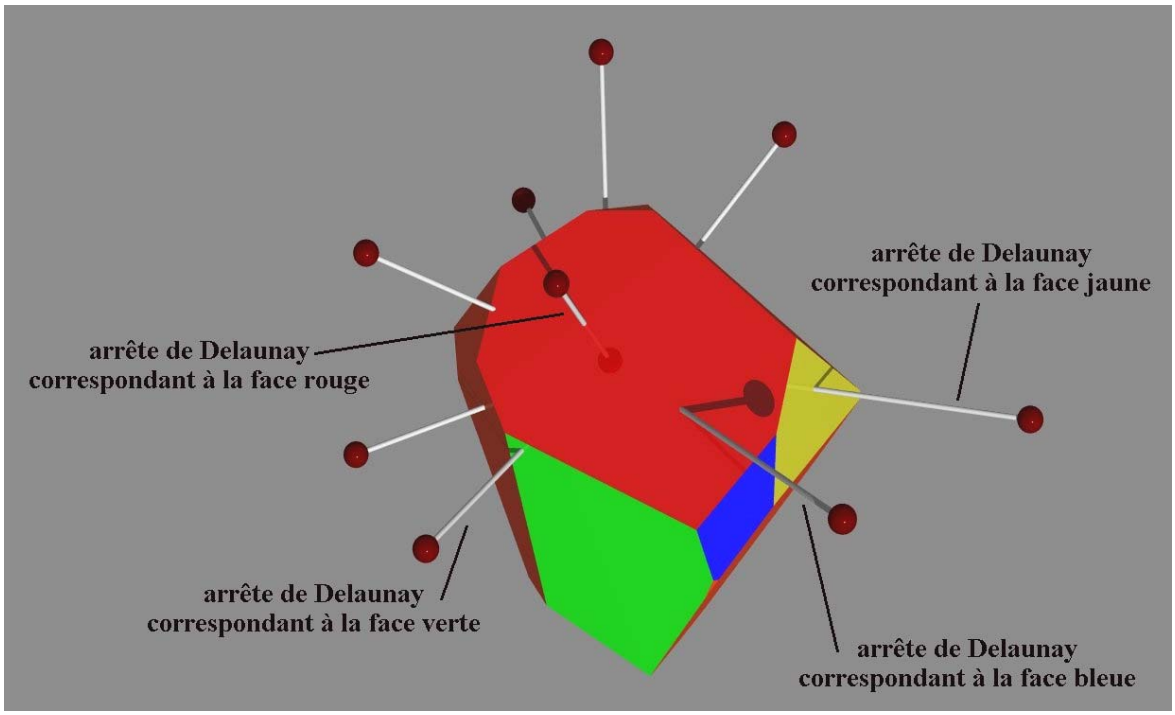


Figure 25 : Même cellule que celle de la Figure 21.

La Figure 25 illustre cette propriété que j'ai abondamment utilisée au cours de ma thèse et qui constitue un des atouts majeurs des TdV. Chaque sphère rouge représente un germe ayant permis de définir la face caractérisée par les petites sphères blanches et bleues de la Figure 21, chaque segment représenté définit un voisinage et à chacun de ces segments est associée une face de Voronoï. Ces faces présentent des différences entre elles, par exemple elles peuvent avoir une surface plus ou moins importante (la face verte est ainsi plus grande que les faces jaune et bleue), elles peuvent avoir un plus ou moins grand nombre de côtés et avoir des formes plus ou moins régulières. Ces propriétés permettent de caractériser le voisinage local.

7.2 Liens entre TdV et tessellation de Delaunay

La triangulation de Delaunay permet de déterminer certaines propriétés des cellules de Voronoï. Ainsi, le nombre de sommets de la cellule générée par un site est égal au nombre de tétraèdres ayant ce site pour sommet ; le nombre de sommets distincts entre eux et distincts du site considéré est le nombre de faces de la cellule ; le nombre de tétraèdres partageant une

arrête de Delaunay issue du site considéré est le nombre de côtés de la face associée à cette arrête. Ce dernier point est également illustré par la Figure 23, le nombre de tétraèdres ayant le segment rouge pour arrête est égal au nombre de côtés de la face dont les sommets sont représentés par les petites sphères. J'expliquerai dans le chapitre consacré à la conception des programmes informatiques, comment on peut calculer les volumes et les aires des cellules à partir de la triangulation de Delaunay.

8 - Pondération des tessellations de Voronoï

Jusqu'ici la définition des TdV que j'ai présentée considérait que tous les sites étudiés étaient égaux, ils étaient donc tous traités de la même façon ; autrement dit la construction des cellules ne donnait pas plus d'importance à un site plutôt qu'à un autre. Mathématiquement, ceci se traduit par le simple fait que les faces des cellules de Voronoï sont incluses dans les plans médians des segments de Delaunay qui relient deux sites voisins. Tous les points d'une face sont attribués aux deux cellules adjacentes et sont à égales distances des deux sites voisins. Dans cette partie, je vais présenter d'autres méthodes de construction des TdV qui prennent en compte l'importance que l'on décide d'accorder à un site par rapport aux autres à l'aide d'un coefficient que l'on appelle le poids. Chaque site sera associé à un poids, qui pourra refléter différentes propriétés le caractérisant. Il existe différentes manières de construire une TdV pondérée, je vais dans la suite de cet exposé décrire rapidement les méthodes les plus courantes et détailler un peu plus celle que nous avons utilisée.

Soit $\mathcal{P} = \{P_1, \ldots, P_n\} \subset \mathbb{R}^3$ tel que $2 \leq n < \infty$, on attribue à chaque site P_i un poids w_i, qui caractérisera la distance pondérée $d_w(P, P_i)$ du point P au point P_i. A l'aide de cette distance pondérée on peut ici aussi définir une zone de domination $H(P_i, P_j)$ définie par :

$$H(P_i,P_j)=\{P / d_w(P,P_i) \leq d_w(P,P_j), j \neq i\} \qquad \textbf{Équation 7}$$

Dans le cas des TdV non pondérées, cette région de domination est le demi-espace contenant le point P_i, délimité par le plan médian de $[P_i, P_j]$, la définition de la cellule de Voronoï associée à P_i s'écrit de la même façon que l'on soit dans le cas d'une tessellation pondérée ou non pondérée :

$$V(P_i)=\bigcap_{j \in I_n - \{i\}} H(P_i,P_j) \qquad \textbf{Équation 8}$$

L'ensemble $\{V(P_1),\ldots, V(P_n)\}$ est maintenant noté : $\mathcal{V}_w$. A travers cette définition on constate que le principe même des TdV est respecté. En fait, les TdV pondérées se distinguent les unes des autres par la définition de la distance pondérée.

8.1 Tessellation de Voronoï pondérée de manière multiplicative

Dans ce cas la distance pondérée est définie de la manière suivante :

$$d_w(P,P_i)=\frac{1}{w_i}\left\|\vec{x}-\vec{x_i}\right\|$$

La région de domination devient :

$$H(P_i,P_j)=\left\{P/\frac{1}{w_i}\left\|\vec{x}-\vec{x_i}\right\|\leq\frac{1}{w_j}\left\|\vec{x}-\vec{x_j}\right\|, j\neq i\right\}$$

Si les poids sont égaux, on retrouve la définition de la TdV non pondérée et si l'on note $b(P_i, P_j)$ le lieu des points pour lesquels $d_w(P, P_i) = d_w(P, P_j)$, on retrouve bien un plan médian de l'espace Euclidien. Si les poids sont différents, on a :

$$b(P_i,P_j)=\left\{P/\left\|\vec{x}-\frac{w_i^2}{w_i^2-w_j^2}\vec{x_j}+\frac{w_j^2}{w_i^2-w_j^2}\vec{x_i}\right\|=\frac{w_iw_j}{\left|w_i^2-w_j^2\right|}\left\|\vec{x_i}-\vec{x_j}\right\|, j\neq i\right\} \qquad \textbf{Équation 9}$$

Les sites P_i et P_j étant fixés, le second terme de l'égalité ci-dessus est une constante. De plus, le terme $\frac{w_i^2}{w_i^2-w_j^2}\vec{x_j}-\frac{w_j^2}{w_i^2-w_j^2}\vec{x_i}$ représente un point fixe, $b(P_i, P_j)$ est donc une sphère dont le centre est défini par $\frac{w_i^2}{w_i^2-w_j^2}\vec{x_j}-\frac{w_j^2}{w_i^2-w_j^2}\vec{x_i}$ et de rayon $\frac{w_iw_j}{\left|w_i^2-w_j^2\right|}\left\|\vec{x_i}-\vec{x_j}\right\|$. Les cellules de cette tessellation ne sont donc plus des polyèdres convexes mais des intersections de boules qui peuvent être concaves. La Figure 26 montre un exemple de cette tessellation dans le plan Euclidien. Les sites sont représentés par des croix rouges et les différents poids sont indiqués. Les cellules sont ici des intersections de disques et plusieurs observations sont possibles : tout d'abord, la cellule de poids cinq signalée par un poids en rouge n'est pas convexe, contrairement aux cellules de Voronoï dites classiques. On peut constater également qu'il y a une cellule de poids deux entièrement contenue dans une cellule de poids sept, il y a donc un trou dans la cellule de poids sept. La cellule de poids quinze, grisée sur la figure, est infinie. On remarque également que certaines frontières de cellule, notamment celles entre les cellules de poids sept et celle de poids quinze, en bleu sur la figure, ne sont pas continues. Enfin les cellules convexes sont des cellules dont les cellules adjacentes ont des poids supérieurs.

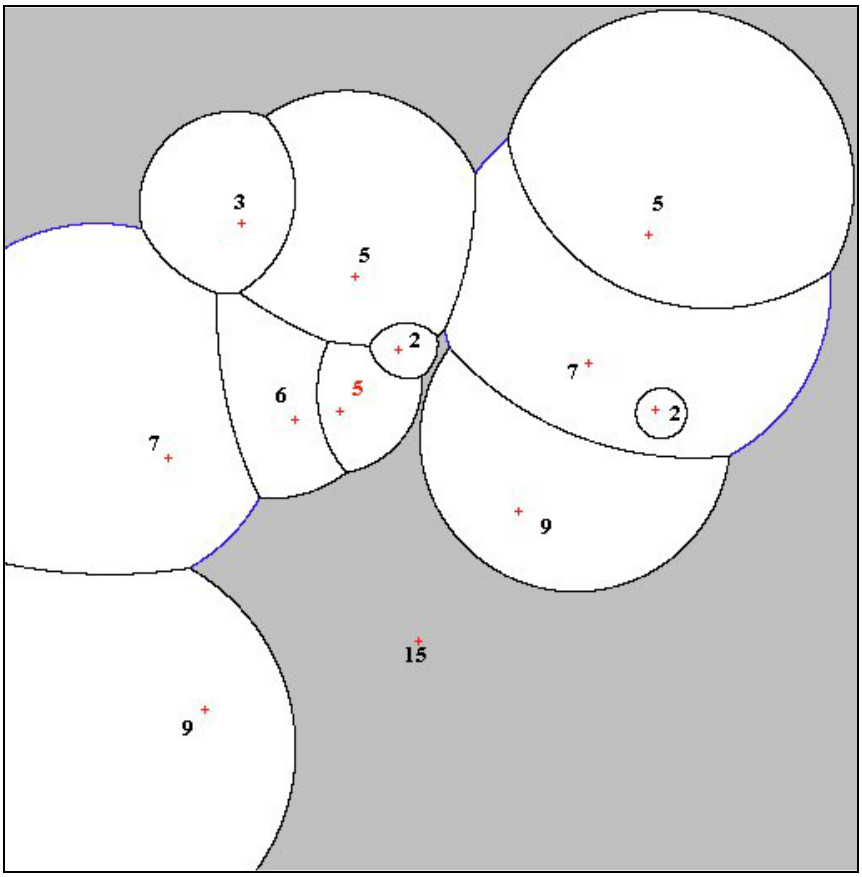


Figure 26 : Exemple de tessellation pondérée de manière multiplicative. Les poids sont indiqués à côté des sites.

8.2 Tessellation de Voronoï pondérée de manière additive

Dans ce cas, la distance pondérée est définie par :

$$d_w(P,P_i)=\left\|\vec{x}-\vec{x_i}\right\|-w_i$$

La région de domination devient :

$$H(P_i,P_j)=\left\{P/\left\|\vec{x}-\vec{x_i}\right\|-w_i\leq\left\|\vec{x}-\vec{x_j}\right\|-w_j, j\neq i\right\}$$

Si les poids sont égaux, on retrouve la définition de la TdV non pondérée et $b(P_i, P_j)$ reste bien un plan médian de l'espace Euclidien. Si les poids sont différents on a :

$$b(P_i,P_j)=\left\{P/\left\|\vec{x}-\vec{x_i}\right\|-\left\|\vec{x}-\vec{x_j}\right\|=w_i-w_j, j\neq i\right\}$$ **Équation 10**

La forme de $b(P_i, P_j)$ dépend de deux paramètres $\alpha=\left\|\vec{x_i}-\vec{x_j}\right\|$ et $\beta=w_i-w_j$; on suppose dans la suite que β est positif sans perte de généralité. On peut distinguer trois cas :

1. $0<\alpha<\beta$ dans ce cas $H(P_i,P_j)=\mathbb{R}^3$, la région de domination de P_j disparaît ce qui n'arrivait pas avec les autres méthodes de tessellation.
2. $\alpha=\beta$ dans ce cas $H(P_i,P_j)=\mathbb{R}^3$ - $\left\{P/\overrightarrow{x_j}+\chi(\overrightarrow{x_j}-\overrightarrow{x}),\chi\geq 0\right\}$, c'est à dire tout l'espace moins la demi-droite partant de P_j incluse dans la droite (P_i,P_j) mais ne contenant pas P_i.
3. $\alpha>\beta$ dans ce cas les frontières des cellules sont les lieux des points P tels que la différence entre la distance de P à P_i et la distance de P à P_j soit une constante. Ce lieu est la nappe de l'hyperboloïde de révolution contenue dans le demi-espace dans lequel se trouve P_j et ayant pour foyers P_i et P_j. La région de domination est donc la portion d'espace contenant P_i et limitée par cette nappe.

Les cellules de Voronoï sont donc des intersections de nappes d'hyperboloïde de révolution. Dans la Figure 27, les poids sont représentés par des cercles rouges, les cellules de Voronoï sont en bleu et leurs arrêtes sont des portions d'hyperbole.

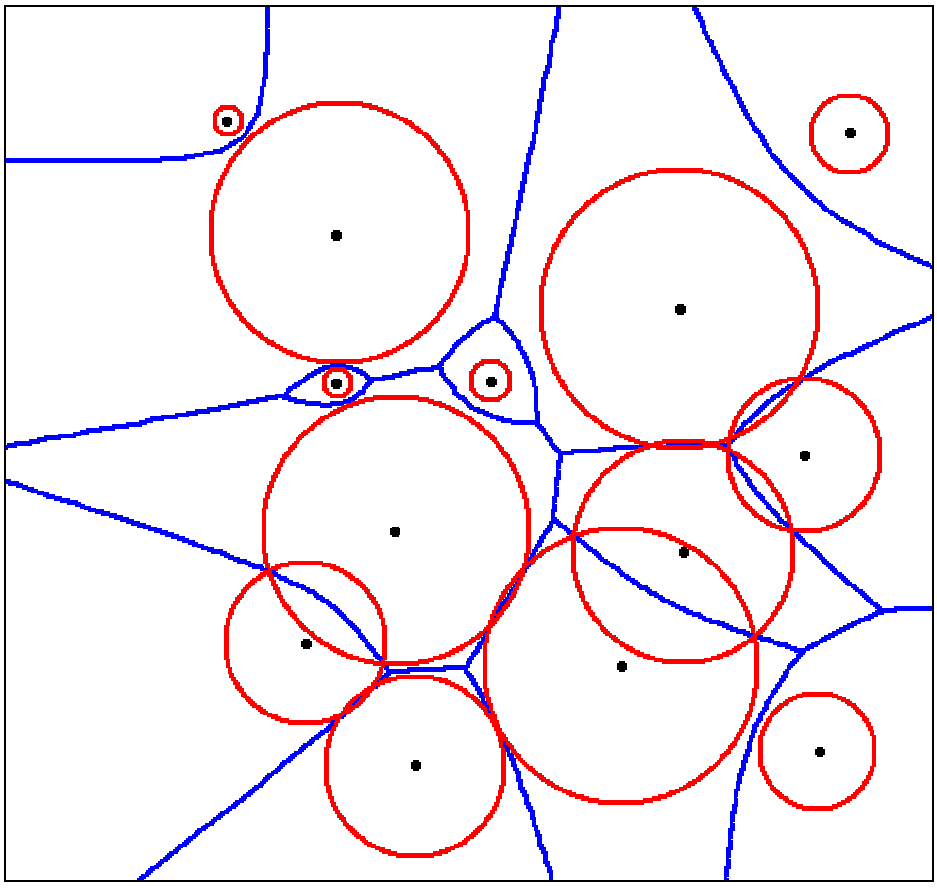


Figure 27 : Exemple de tessellation pondérée de manière additive. Les cellules sont en bleu, les poids sont représentés par des cercles rouges.

Une troisième méthode de pondération que je ne développerai pas consiste bien sûr à cumuler pondération additive et pondération multiplicative. Dans ce cas, la distance pondérée est définie de la manière suivante :

$$d_w(P,P_i)=\frac{1}{w_{i1}}\left\|\vec{x}-\vec{x_i}\right\|-w_{i2}$$

Les frontières des régions de domination sont alors définies par des équations polynomiales en $\vec{x}$ du quatrième ordre et deviennent relativement complexes.

8.3 Diagramme de puissance ou décomposition de Laguerre.

Je vais développer ce chapitre de manière un peu plus détaillée que les chapitres précédents ; en effet comme je l'exposerai par la suite, c'est cette méthode de pondération que j'ai utilisée dans mon travail de thèse. Dans ce cas, la distance pondérée est définie d'une manière assez proche de la manière précédente : $d_w(P,P_i)=\left\|\vec{x}-\vec{x_i}\right\|^2-w_i$, la seule différence réside dans le fait que la distance entre P et P_i est ici élevée au carré. La région de domination devient :

$$H(P_i,P_j)=\left\{P/\left\|\vec{x}-\vec{x_i}\right\|^2-w_i\leq\left\|\vec{x}-\vec{x_j}\right\|^2-w_j, j\neq i\right\} \qquad \textbf{Équation 11}$$

L'équation définissant la frontière s'écrit :

$$b(P_i,P_j)=\left\{P/\left\|\vec{x}-\vec{x_i}\right\|^2-\left\|\vec{x}-\vec{x_j}\right\|^2=w_i-w_j, j\neq i\right\} \qquad \textbf{Équation 12}$$

Quelques opérations de calcul conduisent à :

$$b(P_i,P_j)=\left\{P/\vec{x}.(\vec{x_j}-\vec{x_i})=\frac{\left\|\vec{x_j}\right\|^2-\left\|\vec{x_i}\right\|^2+w_i-w_j}{2}, j\neq i\right\} \qquad \textbf{Équation 13}$$

Ceci signifie que la projection orthogonale de tout point de $b(P_i, P_j)$ sur $\overrightarrow{P_iP_j}$ est une constante, $b(P_i, P_j)$ est donc un plan perpendiculaire à $\overrightarrow{P_iP_j}$. Ce plan se projette en un seul point M repéré par $\overrightarrow{x_m}$ tel que $\overrightarrow{x_m}=\lambda(\vec{x_j}-\vec{x_i})$, si on remplace $\vec{x}$ par $\overrightarrow{x_m}$ dans l'équation définissant $b(P_i, P_j)$, on obtient :

$$\vec{x_m}=\frac{\left\|\vec{x_j}\right\|^2-\left\|\vec{x_i}\right\|^2+w_i-w_j}{2\left\|\vec{x_j}-\vec{x_i}\right\|^2}(\vec{x_j}-\vec{x_i})$$ **Équation 14**

Comme les frontières des régions de domination sont des plans, on en déduit que comme dans le cas des TdV non pondérées les régions de domination sont des demi-espaces. Les cellules de Voronoï sont donc ici aussi des polyèdres.

Pour décrire quelques propriétés de cette technique de pondération, je reprends l'Équation 14 avec quelques simplifications. On supposera que P_i est à l'origine c'est à dire que $\vec{x_i}=\vec{0}$, on obtient donc :

$$\vec{x_m}=\frac{\left\|\vec{x_j}\right\|^2+w_i-w_j}{2\left\|\vec{x_j}\right\|^2}\vec{x_j}$$ **Équation 15**

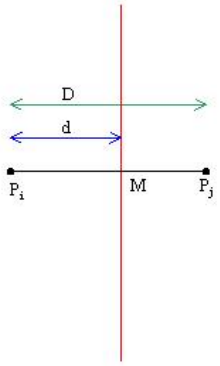


Figure 28

Soit D la distance entre P_i et P_j et d la distance entre P_i et M (voir la Figure 28 pour une illustration à deux dimensions), l'équation précédente fournit alors :

$$\frac{d}{D}=\frac{D^2+w_i-w_j}{2D^2}$$ **Équation 16**

Plus w_i est grand, plus d sera grand et donc plus le plan qui séparera les deux cellules sera éloigné de P_i. Si on suppose que $w_i-w_j>D^2$, on a alors $d>D$; ceci implique que les deux points sont dans la région de domination de P_i. On a alors la situation illustrée par la Figure 29, toujours pour un espace à deux dimensions. Pour un ensemble de sites donné, il sera donc possible d'avoir des cellules de Voronoï ne contenant pas les sites qui les ont

générées et bien sûr des cellules contenant plusieurs sites. Il faut noter que la tessellation ne dépend pas directement de la valeur des poids associés à chaque site mais des différences entre les poids des voisins, ces poids pourront donc être définis à un facteur additif près sans que cela change la tessellation.

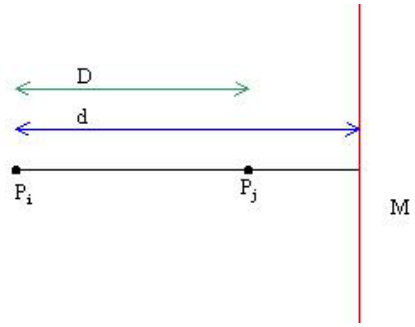


Figure 29

Une propriété intéressante de cette méthode de pondération est illustrée par la Figure 30. Dans cette figure, *A* et *B* représentent deux sites, leur poids (4 et 16) sont représentés par deux cercles, le rapport des rayons de ces cercles est 2 de telle manière que le rapport de leur aire soit 4 comme le rapport des poids. La droite qui sépare les deux cellules associées à chacun de ces sites est en rouge. Cette droite possède la propriété intéressante suivante : si l'on considère un point de cette droite, *M1* par exemple, et si l'on trace les tangentes aux deux cercles passant par ce point (en jaune sur la figure) et bien on a $\left\|\overrightarrow{A1M1}\right\|=\left\|\overrightarrow{B1M1}\right\|$, de même on a également $\left\|\overrightarrow{A2M2}\right\|=\left\|\overrightarrow{B2M2}\right\|$.

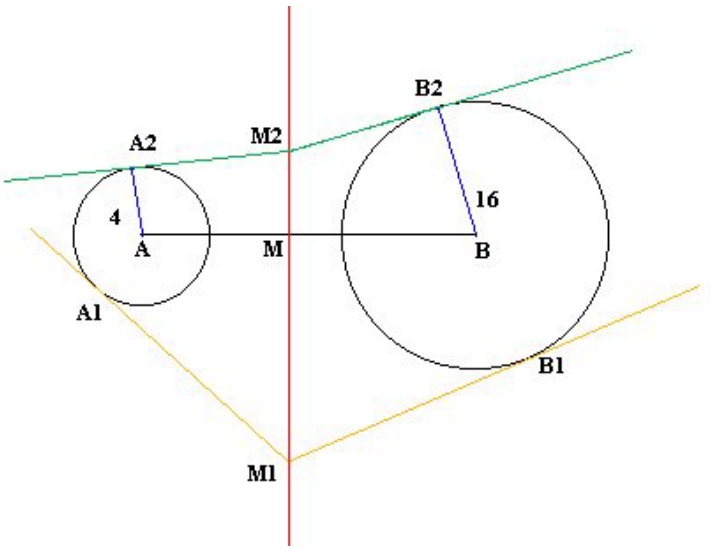


Figure 30

9 - Conclusion

Ces définitions montrent que les TdV constituent un concept relativement simple dont les multiples propriétés en font un outil puissant pour, entre autres, étudier les structures protéiques. Cette simplicité des TdV ne doit pas cacher les nombreuses difficultés que doivent cependant gérer et résoudre les programmes conçus afin de les déterminer. Le chapitre suivant décrit les différentes procédures que j'ai utilisées au cours de ma thèse afin de pouvoir construire ces cellules et en calculer les caractéristiques.

C h a p i t r e 3

Implémentation des TdV

1 - Introduction

Les programmes que j'ai utilisés sont inspirés principalement de ceux réalisés par Jean-François Sadoc, Rémi Jullien, Alain Soyer et Borislav Anguelov. Ces programmes ont été écrits en langage Fortran ou Mathematica, je les ai traduits en langage C et largement modifiés. Comme je l'ai déjà expliqué, les TdV et les triangulations de Delaunay sont étroitement liées. La conception de l'algorithme utilisé est fondée sur cette dualité et la réalisation de la TdV ainsi que les différents calculs associés sont réalisés à partir de la triangulation de Delaunay. Comme les tessellations ont été appliquées à des structures protéiques, l'élément de base à partir duquel s'effectue la construction est un fichier au format PDB, contenant les coordonnées atomiques des différents atomes.

2 - Première étape : Détermination des sites

Une TdV se construit à partir d'un ensemble de points. La plupart des travaux concernant l'application des TdV aux structures protéiques, utilisent directement les positions des atomes, l'originalité de notre approche réside essentiellement dans le fait que nous avons travaillé à l'échelle des AA. Ceci signifie simplement que chaque AA de la structure étudiée est représenté par un point auquel sera associée une cellule. La première étape du programme est donc de déterminer, dans le fichier PDB, l'ensemble des points à partir desquels seront effectués les calculs. La méthode la plus simple consiste à retenir les coordonnées des Cα, une autre méthode consiste à calculer le centre géométrique des atomes d'un même AA à partir des coordonnées de ses atomes lourds tels que le carbone, l'azote, l'oxygène ou le soufre. Dans ce cas, les atomes d'hydrogène ne sont pas pris en compte, car ils sont rarement présents dans les fichiers et sont plus légers. Au final, on dispose donc d'un ensemble de points dont le nombre est égal au nombre d'AA présents dans la structure, il est à ce stade important de

repérer les éventuels manques, c'est à dire les parties de la protéine pour lesquelles les positions des atomes n'ont pas été déterminées. En effet, ces « trous » vont sensiblement modifier les différentes valeurs associées aux cellules, puisqu'une partie des points sera absente, par exemple la surface et le volume d'une cellule proche d'un de ces trous seront augmentés du fait de l'absence de ses voisins. Tous les sites retenus sont finalement repérables par un nombre entier qui correspond à la position dans la séquence de l'AA associé.

3 - Deuxième étape, la recherche des voisins : Triangulation de Delaunay

L'étape suivante consiste pour chaque site à détecter ses voisins potentiels, ceci revient à éliminer les sites trop éloignés pour pouvoir faire une face de contact avec la cellule associée au germe considéré. Pour un site donné, on procède de la manière suivante : on calcule la distance entre ce site et tous les autres sites de l'ensemble, et on ne conserve que ceux dont la distance est inférieure à une certaine valeur. La distance choisie est de 16 Å, elle permet de n'omettre aucun voisin et de diminuer le nombre de calculs à effectuer par la suite, cette valeur limite est donc fixée pour réduire le temps de calcul et n'affecte pas les résultats dans le cas d'une structure protéique. Pour un germe donné que j'appellerai I, les sites ainsi sélectionnés sont stockés dans un tableau nommé NPV (Numéro des Proches Voisins). Pour un site donné de NPV noté IV1 on cherche tous les autres germes issus de NPV qui sont à une distance inférieure à 16 Å de IV1. Tous les sites ainsi sélectionnés sont stockés dans le tableau NPV2, le tableau NPV2R quant à lui est créé pour ne stocker que les sites dont le numéro (c'est à dire l'ordre dans la séquence) est supérieur à celui de IV1. Les tableaux NPV2 et NPV2R ne contiennent donc que des voisins (ou plutôt des sites pas trop éloignés) de I et de IV1. Je vais maintenant décrire l'opération consistant à déterminer les véritables voisins, c'est à dire les voisins définis par une face de contact, d'un site particulier I. Cette opération est bien sûr répétée pour tous les sites, tour à tour.

Déterminer les voisins au sens Voronoï d'un site particulier consiste à trouver ses plus proches voisins. Ceci est rendu possible en déterminant tous les tétraèdres de Delaunay ayant le site I considéré pour sommet. En effet comme je l'ai expliqué dans le chapitre précédent, les sphères circonscrites aux tétraèdres de Delaunay ne contiennent aucun autre site, les sommets de ces tétraèdres sont donc les plus proches voisins du site I (Figure 31).

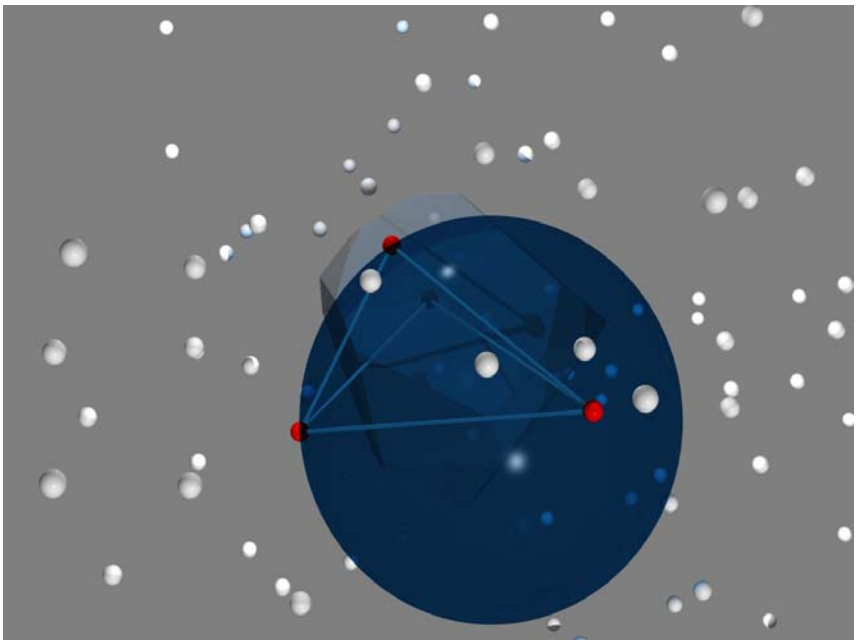


Figure 31 : Les petites sphères représentent différents sites, un tétraèdre est représenté ainsi que sa sphère circonscrite. Celle-ci ne contient pas d'autres sites, ceci signifie donc que les trois sites en rouge font partie des plus proches voisins du site noir au centre de la cellule.

On en déduit donc qu'une fois tous les tétraèdres de Delaunay déterminés on est certain qu'il n'existe pas de sites plus proches et on dispose par conséquent de tous les voisins qui donneront ensuite naissance à une face de Voronoï (Figure 32).

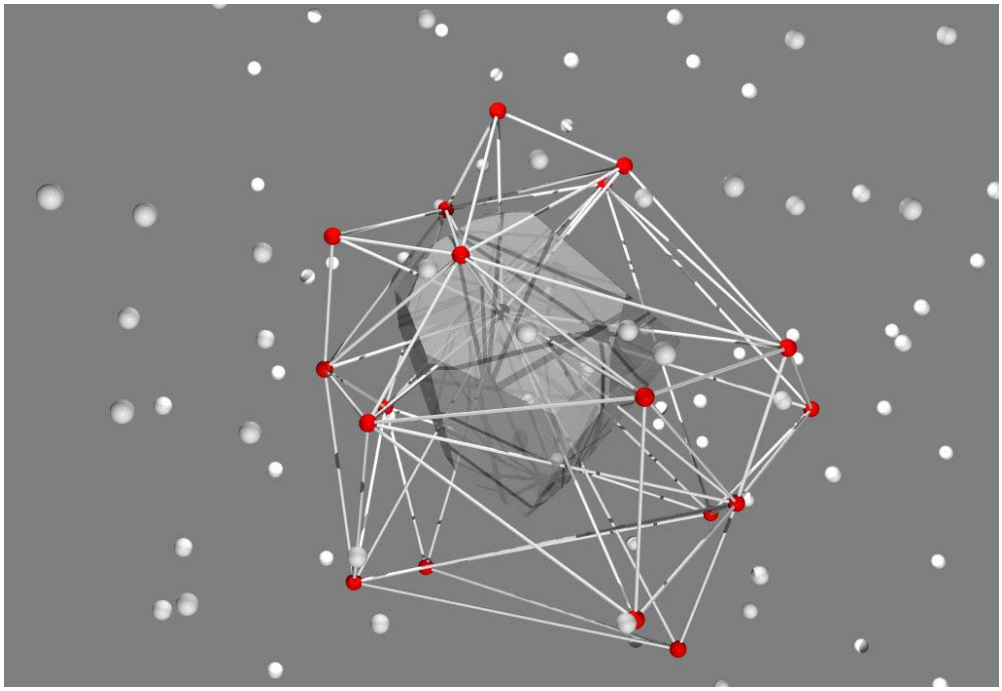


Figure 32 : Tous les voisins du point représenté par une sphère noire (au centre de la cellule) sont représentés par des sphères rouges. Tous les tétraèdres de Delaunay sont représentés et tout tétraèdre supplémentaire ayant pour sommet la sphère noire aurait une sphère circonscrite qui contiendrait nécessairement une des sphères rouges, ce ne serait donc pas un tétraèdre de Delaunay.

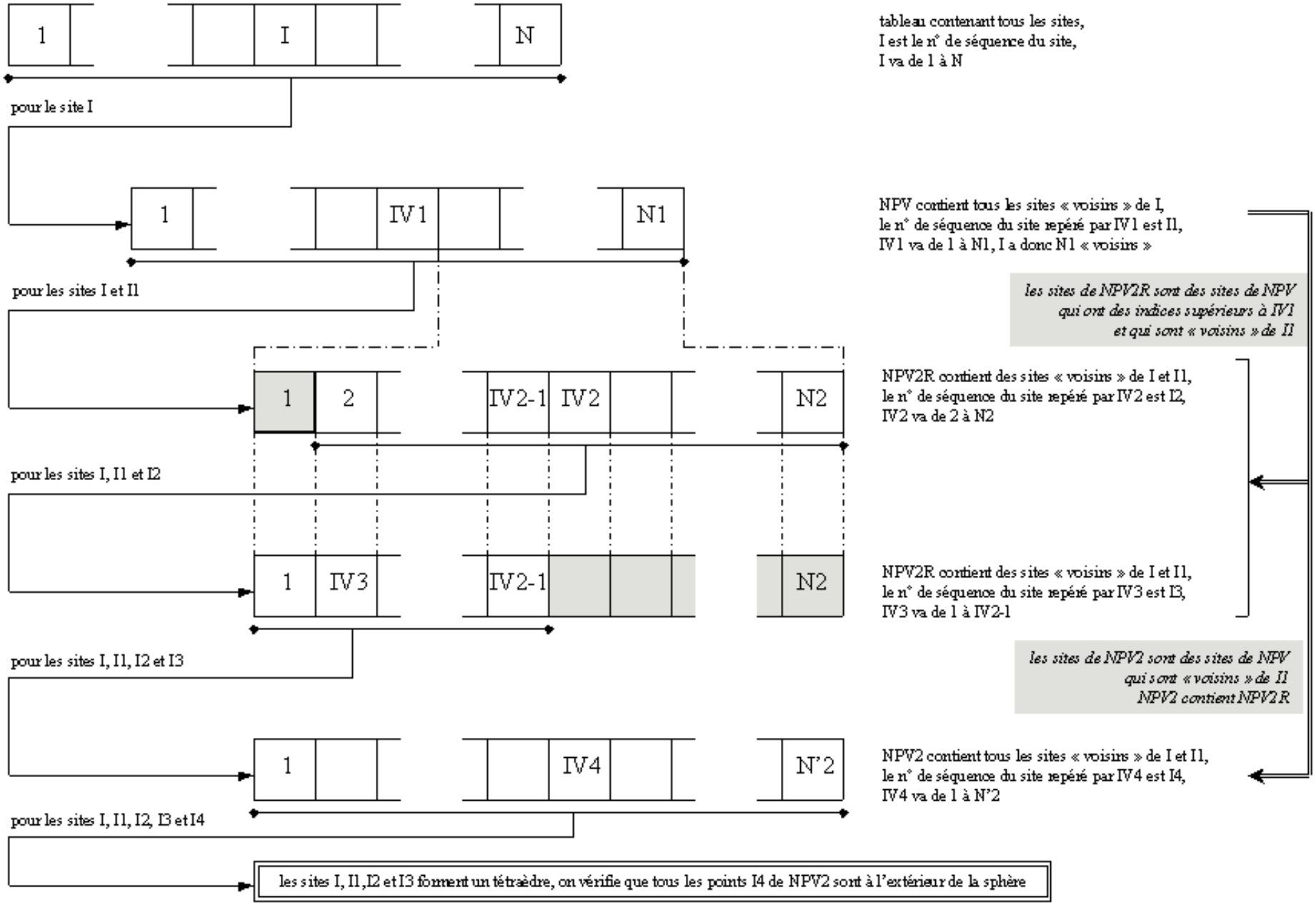

1
I
N
tableau contenant tous les sites,
I est le n° de séquence du site,
I va de 1 à N
pour le site I
1
IV1
N1
NPV contient tous les sites « voisins » de I,
le n° de séquence du site repéré par IV1 est I1,
IV1 va de 1 à N1, I a donc N1 « voisins »
pour les sites I et I1
les sites de NPV2R sont des sites de NPV
qui ont des indices supérieurs à IV1
et qui sont « voisins » de I1
1
2
IV2-1
IV2
N2
NPV2R contient des sites « voisins » de I et I1,
le n° de séquence du site repéré par IV2 est I2,
IV2 va de 2 à N2
pour les sites I, I1 et I2
1
IV3
IV2-1
N2
NPV2R contient des sites « voisins » de I et I1,
le n° de séquence du site repéré par IV3 est I3,
IV3 va de 1 à IV2-1
pour les sites I, I1, I2 et I3
les sites de NPV2 sont des sites de NPV
qui sont « voisins » de I1
NPV2 contient NPV2R
1
IV4
N'2
NPV2 contient tous les sites « voisins » de I et I1,
le n° de séquence du site repéré par IV4 est I4,
IV4 va de 1 à N'2
pour les sites I, I1, I2, I3 et I4
les sites I, I1, I2 et I3 forment un tétraèdre, on vérifie que tous les points I4 de NPV2 sont à l'extérieur de la sphère

Figure 33 (page précédente) : Description de la procédure utilisée pour déterminer les tétraèdres de Delaunay.

Le programme procède donc de la manière suivante toujours pour un site donné I : on considère tour à tour chacun des sites de NPV. Pour un site donné IV1 de NPV, on considère tour à tour les sites de NPV2R à partir du deuxième site, de même pour un site donné IV2 ainsi sélectionné on considère tour à tour les sites de NPV2R mais cette fois-ci à partir du premier jusqu'à celui précédent IV2. Cette manière de procéder évite de faire plusieurs fois les mêmes calculs. On dispose donc à ce stade de quatre sites (I, I1 repéré par l'indice IV1 de NPV, I2 repéré par IV2, I3 repéré par IV3 de NPV2R), dont on sait qu'ils ne sont pas trop éloignés les uns des autres. La prochaine étape du programme consiste à vérifier que I1, I2 et I3 font partie des plus proches voisins de I, ceci consiste à vérifier que le tétraèdre (I, I1, I2, I3) est un tétraèdre de Delaunay. On cherche donc à savoir si la sphère circonscrite à ce tétraèdre contient un autre site ; si c'est le cas, le tétraèdre en question n'est pas un tétraèdre de Delaunay, si au contraire cette sphère est vide, on a bien un des tétraèdres recherchés que le programme stocke en mémoire, ainsi que le centre de cette sphère qui est un des sommets de la cellule de Voronoï associée à I. Le programme détermine en fait si un site noté I4 du tableau NPV2 repéré par l'indice IV4 (c'est à dire tous les voisins de I et I1) différent de I2 et I3 est à l'intérieur ou non de la sphère considérée, et les points de NPV2 sont pris tour à tour. Le programme sélectionne alors le site IV3+1 de NPV2R et répète cette opération avec maintenant les sites I, I1, I2 et un autre site I3, et ainsi de suite jusqu'à épuisement de toutes les possibilités. Lorsque tous les sites de NPV sont épuisés, on est sûr d'avoir déterminé les plus proches voisins de I (Figure 32) ; le programme passe alors au site I+1 et recommence les mêmes opérations.

3.1 Comment déterminer le centre de la sphère circonscrite ?

Tous les sites *I* sont repérés par un vecteur $\overrightarrow{R_I}$ de coordonnées $\overrightarrow{R_I}(r_{Ix},r_{Iy},r_{Iz})$, pour chaque tétraèdre (*I, I1, I2, I3*) on effectue un changement de coordonnées tel que *I* devienne l'origine. On obtient ainsi les trois vecteurs : $\overrightarrow{D_1}=\overrightarrow{R_{I1}}-\overrightarrow{R_I}$, $\overrightarrow{D_2}=\overrightarrow{R_{I2}}-\overrightarrow{R_I}$ et $\overrightarrow{D_3}=\overrightarrow{R_{I3}}-\overrightarrow{R_I}$.

L'équation d'une sphère dont le centre *C* est repéré par le vecteur $\vec{R}$ de coordonnées (*a, b c*) et de rayon $R=\left\|\vec{R}\right\|$ s'écrit :

$$(x-a)^2+(y-b)^2+(z-c)^2=R^2$$

Dans le cas qui nous intéresse, l'origine est sur la sphère que l'on cherche à déterminer, on a donc :

$$a^2+b^2+c^2=R^2$$

Or :

$$(x-a)^2+(y-b)^2+(z-c)^2=x^2+y^2+z^2+a^2+b^2+c^2-2ax-2by-2cz=R^2$$

Après simplification on obtient donc :

$$x^2+y^2+z^2=2ax+2by+2cz$$

Soit $\overrightarrow{D}$ le vecteur de coordonnées (*x, y, z*) repérant un point situé sur la sphère, on a alors :

$$\left\|\overrightarrow{D}\right\|^2=2\vec{R}\cdot\vec{D}$$

On sait que les points *I1, I2 et I3* sont également sur la sphère on a donc :

$$\begin{cases} 2\vec{R}\cdot\overrightarrow{D_1}=\left\|\overrightarrow{D_1}\right\|^2=D_1{}^2 \\ 2\vec{R}\cdot\overrightarrow{D_2}=\left\|\overrightarrow{D_2}\right\|^2=D_2{}^2 \\ 2\vec{R}\cdot\overrightarrow{D_3}=\left\|\overrightarrow{D_3}\right\|^2=D_3{}^2 \end{cases}$$

Ce système est équivalent au système à trois équations et à trois inconnues suivant :

$$\begin{cases} ax_1+by_1+cz_1=D_1{}^2/2 \\ ax_2+by_2+cz_2=D_2{}^2/2 \\ ax_3+by_3+cz_3=D_3{}^2/2 \end{cases}$$

Soit Δ le déterminant principal de ce système :

$$\Delta=\begin{vmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{vmatrix}$$

D'après la règle de Cramer, si $\Delta\neq 0$ on a :

$$a=\frac{\begin{vmatrix} D_1{}^2/2 & y_1 & z_1 \\ D_2{}^2/2 & y_2 & z_2 \\ D_3{}^2/2 & y_3 & z_3 \end{vmatrix}}{\Delta}=\frac{\begin{vmatrix} D_1{}^2 & y_1 & z_1 \\ D_2{}^2 & y_2 & z_2 \\ D_3{}^2 & y_3 & z_3 \end{vmatrix}}{2\Delta} \quad b=\frac{\begin{vmatrix} x_1 & D_1{}^2/2 & z_1 \\ x_2 & D_2{}^2/2 & z_2 \\ x_3 & D_3{}^2/2 & z_3 \end{vmatrix}}{\Delta}=\frac{\begin{vmatrix} x_1 & D_1{}^2 & z_1 \\ x_2 & D_2{}^2 & z_2 \\ x_3 & D_3{}^2 & z_3 \end{vmatrix}}{2\Delta} \quad c=\frac{\begin{vmatrix} x_1 & y_1 & D_1{}^2/2 \\ x_2 & y_2 & D_2{}^2/2 \\ x_3 & y_3 & D_3{}^2/2 \end{vmatrix}}{\Delta}=\frac{\begin{vmatrix} x_1 & y_1 & D_1{}^2 \\ x_2 & y_2 & D_2{}^2 \\ x_3 & y_3 & D_3{}^2 \end{vmatrix}}{2\Delta}$$

3.2 Comment vérifier que la sphère circonscrite est vide ?

A ce stade on dispose donc des coordonnées du centre de la sphère, il reste à savoir si cette sphère contient le site *I4* repéré par $\overrightarrow{D_4}=\overrightarrow{R_4}-\overrightarrow{R_1}$.

Pour que *I4* soit en dehors de la sphère, il faut avoir :

$$(x_4-a)^2+(y_4-b)^2+(z_4-c)^2>R^2$$

Or on a $a^2+b^2+c^2=R^2$ donc après simplification on a :

$$x_4^2+y_4^2+z_4^2>2ax_4+2by_4+2cz_4$$

On remplace a, b, et c par leur expression :

$$D_4^2-x_4\frac{\begin{vmatrix} D_1^2 & y_1 & z_1 \\ D_2^2 & y_2 & z_2 \\ D_3^2 & y_3 & z_3 \end{vmatrix}}{\Delta}-y_4\frac{\begin{vmatrix} x_1 & D_1^2 & z_1 \\ x_2 & D_2^2 & z_2 \\ x_3 & D_3^2 & z_3 \end{vmatrix}}{\Delta}-z_4\frac{\begin{vmatrix} x_1 & y_1 & D_1^2 \\ x_2 & y_2 & D_2^2 \\ x_3 & y_3 & D_3^2 \end{vmatrix}}{\Delta}>0$$

En multipliant de chaque côté de l'inégalité par Δ, on obtient :

$$\begin{cases} \Delta D_4^2-x_4\begin{vmatrix} D_1^2 & y_1 & z_1 \\ D_2^2 & y_2 & z_2 \\ D_3^2 & y_3 & z_3 \end{vmatrix}-y_4\begin{vmatrix} x_1 & D_1^2 & z_1 \\ x_2 & D_2^2 & z_2 \\ x_3 & D_3^2 & z_3 \end{vmatrix}-z_4\begin{vmatrix} x_1 & y_1 & D_1^2 \\ x_2 & y_2 & D_2^2 \\ x_3 & y_3 & D_3^2 \end{vmatrix}>0, \Delta>0 \\ \Delta D_4^2-x_4\begin{vmatrix} D_1^2 & y_1 & z_1 \\ D_2^2 & y_2 & z_2 \\ D_3^2 & y_3 & z_3 \end{vmatrix}-y_4\begin{vmatrix} x_1 & D_1^2 & z_1 \\ x_2 & D_2^2 & z_2 \\ x_3 & D_3^2 & z_3 \end{vmatrix}-z_4\begin{vmatrix} x_1 & y_1 & D_1^2 \\ x_2 & y_2 & D_2^2 \\ x_3 & y_3 & D_3^2 \end{vmatrix}<0, \Delta<0 \end{cases}$$

Or :

$$\Theta=\Delta D_4^2-x_4\begin{vmatrix} D_1^2 & y_1 & z_1 \\ D_2^2 & y_2 & z_2 \\ D_3^2 & y_3 & z_3 \end{vmatrix}-y_4\begin{vmatrix} x_1 & D_1^2 & z_1 \\ x_2 & D_2^2 & z_2 \\ x_3 & D_3^2 & z_3 \end{vmatrix}-z_4\begin{vmatrix} x_1 & y_1 & D_1^2 \\ x_2 & y_2 & D_2^2 \\ x_3 & y_3 & D_3^2 \end{vmatrix}=\begin{vmatrix} x_1 & y_1 & z_1 & D_1^2 \\ x_2 & y_2 & z_2 & D_2^2 \\ x_3 & y_3 & z_3 & D_3^2 \\ x_4 & y_4 & z_4 & D_4^2 \end{vmatrix}$$

On a donc finalement :

$$\begin{cases} \Theta>0, \Delta>0 \\ \Theta<0, \Delta<0 \end{cases}$$

En conclusion :

Si $\begin{vmatrix} x_1 & y_1 & z_1 & D_1^2 \\ x_2 & y_2 & z_2 & D_2^2 \\ x_3 & y_3 & z_3 & D_3^2 \\ x_4 & y_4 & z_4 & D_4^2 \end{vmatrix}$ et $\begin{vmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{vmatrix}$ sont de même signe, *I4* est en dehors de la sphère circonscrite à *I, I1, I2, I3*, si ceci est vrai pour tous les points *I4* (c'est à dire tous les points de NPV2) alors (*I, I1, I2, I3*) est un tétraèdre de Delaunay et *C* est un sommet de la cellule de Voronoï générée par *I*.

Toutes ces opérations sont répétées pour tous les sites présents, on dispose donc ainsi au final de tous les sommets de toutes les cellules de Voronoï.

3.3 Détermination des propriétés des cellules.

Certaines valeurs associées aux cellules sont directement calculables à partir de la triangulation de Delaunay, par exemple le nombre de faces d'une cellule correspond au nombre de voisins, le nombre d'arrêtes par face correspond au nombre de tétraèdres ayant le segment perpendiculaire à cette face comme arrête commune, le nombre de sommets d'une cellule est bien évidemment le nombre de tétraèdres ayant le site générateur de la cellule comme sommet commun.

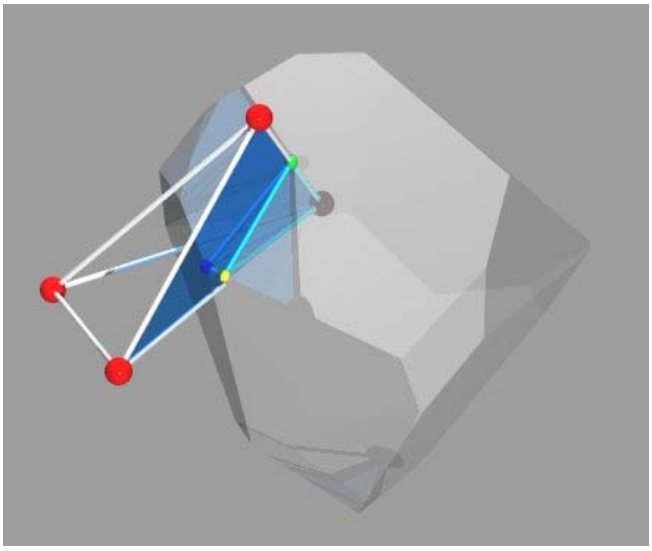


Figure 34 : Un tétraèdre de Delaunay est représenté (arrêtes blanches), un tétraèdre élémentaire est représenté également (arrêtes bleu clair), le point vert est l'intersection de l'arrête du tétraèdre de Delaunay et de la face perpendiculaire, le point bleu est le centre de la sphère circonscrite au tétraèdre de Delaunay, le point jaune est sa projection orthogonale sur la face bleue.

Les volumes et les surfaces quant à eux demandent un peu plus de calculs. Pour les réaliser on procède en décomposant les tétraèdres de Delaunay à l'aide d'autres tétraèdres. Pour un tétraèdre particulier, celui de la Figure 34 par exemple, on utilisera des tétraèdres du même type que celui représenté en bleu clair. Les sommets de ce tétraèdre sont les points suivants : la sphère noire est le site générateur de la cellule ; la sphère verte est le milieu du segment joignant les deux sites voisins, ce point appartient donc à la face de la cellule orthogonale à ce segment ; la sphère bleue est le centre de la sphère circonscrite au tétraèdre de Delaunay, c'est donc également un des sommets de la cellule et enfin la sphère jaune est la projection orthogonale de ce sommet sur la face du tétraèdre représentée en bleu, de ce fait ce point est également le centre du cercle circonscrit à ce triangle. La Figure 35 représente d'un point de vue légèrement différent la même cellule et les mêmes sites. Sur cette figure sont représentés également tous les tétraèdres permettant de calculer les différentes valeurs associées à cette cellule et émanant du sommet toujours représenté par une sphère bleue. Tous ces tétraèdres ont le site générateur de la cellule (sphère noire) comme sommet commun, et le segment reliant ce site au sommet considéré (sphère bleue) comme arrête commune. Les autres sommets de ces tétraèdres sont les milieux des arrêtes du tétraèdre de Delaunay dont une extrémité est le site générateur, et les projections orthogonales sur les faces de ce même tétraèdre partageant également ce sommet. Il y a donc en tout six tétraèdres par cellule et par tétraèdre de Delaunay. Les volumes de tous les tétraèdres associés à un site particulier permettent de calculer le volume total de la cellule. Il est important de noter que le centre de la sphère circonscrite au tétraèdre de Delaunay peut se situer à l'extérieur du tétraèdre, c'est donc une somme algébrique des différents volumes qu'il faut effectuer.

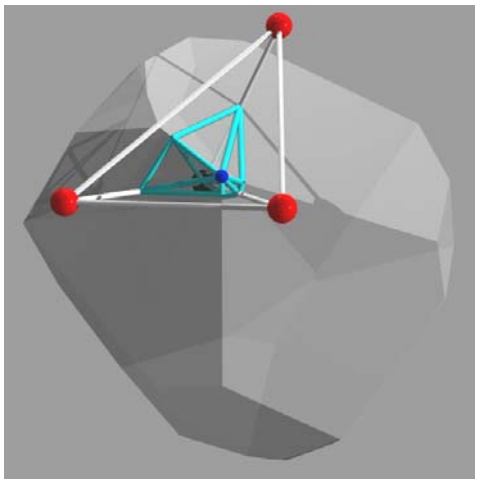
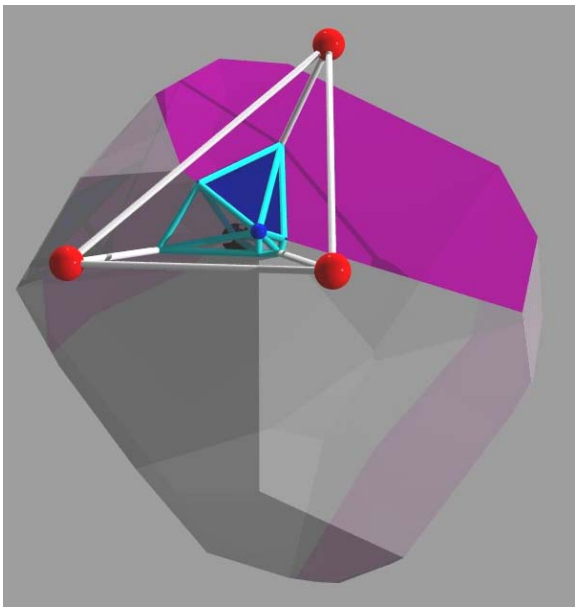


Figure 35 : Angle de vue légèrement différent du précédent, tous les tétraèdres élémentaires ayant pour sommet le site générateur de la cellule et issus du tétraèdre de Delaunay représenté sont indiqués en bleu.

Le calcul des surfaces s'effectue également à l'aide de ces tétraèdres. La Figure 36 reprend la même cellule, les faces des tétraèdres représentées en bleu sont incluses dans la face de la cellule représentée en violet, la somme algébrique des aires de tous les triangles inclus dans la face considérée, associés à tous les tétraèdres de Delaunay ayant le segment reliant les deux sites voisins pour arrête commune donnera l'aire de cette face. Sur la Figure 37, les différentes contributions de chaque tétraèdre de Delaunay impliqué sont représentées par des polygones de différentes couleurs, chaque polygone est l'union de deux triangles. Le calcul des périmètres des faces s'effectue de la même façon, la somme algébrique des arrêtes des triangles coïncidant avec les arrêtes des faces des cellules donne le périmètre de la face considérée.



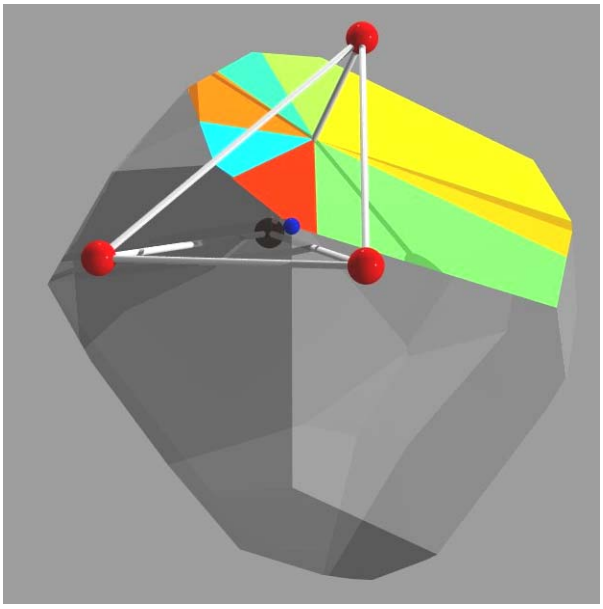


Figure 36 (à gauche) :
Les triangles permettant de calculer l'aire de la face représentée en violet sont indiqués en bleu foncé, ces triangles sont les faces des tétraèdres élémentaires incluses dans la face considérée.
Figure 37 (à droite) :
Les triangles bleus sont maintenant unis pour former un polygone orange foncé (dont un des sommets est la sphère bleu). La face est constituée de huit polygones correspondant aux contributions des huit tétraèdres partageant l'arrête orthogonale à cette face, et ayant généré les huit sommets de la face.

3.4 Comment calculer les volumes et les surfaces ?

Pour calculer ces valeurs à partir des tétraèdres élémentaires, il faut au préalable déterminer les coordonnées des sommets de ces tétraèdres. Un tétraèdre élémentaire particulier est défini par le site générateur de le cellule (*I* dans la Figure 38), du milieu d'une arrête (*I3'*), du centre de la sphère circonscrite (*C*) au tétraèdre de Delaunay dans lequel on se

situe (*I,I1,I2,I3*) et de la projection orthogonale de ce centre sur une des deux faces du tétraèdre partageant l'arrête dont on a retenu le milieu ((*I, I2, I3*) dans le cas de la Figure 38, une autre possibilité est la face (*I, I1, I3*), plus rigoureusement on considère en fait la projection orthogonale de *C* sur le plan contenant la face du tétraèdre car cette projection ne se situe pas nécessairement à l'intérieur du triangle). On connaît déjà les coordonnées du site générateur, du sommet de la cellule (le centre de la sphère), et du milieu de l'arrête concernée. Il reste donc à déterminer les coordonnées de la projection orthogonale du sommet de la cellule sur la face retenue. Nous avons vu que cette projection est le centre du cercle circonscrit au triangle constituant la face sur laquelle on projette, c'est cette propriété que l'on va utiliser pour déterminer ses coordonnées.

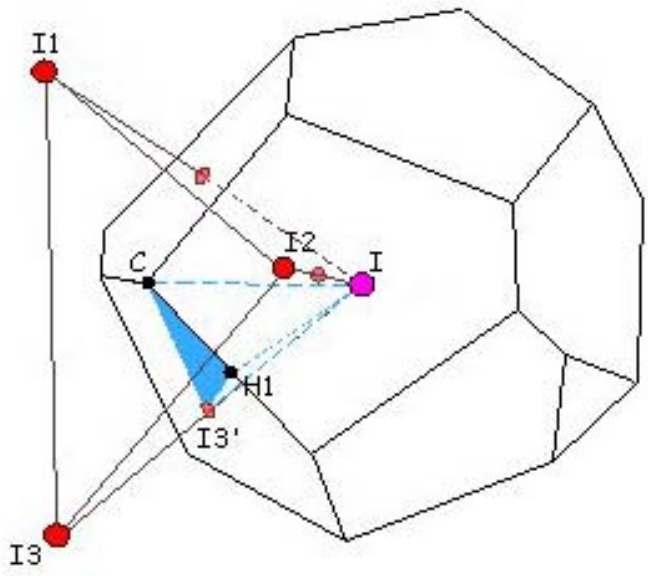


Figure 38 : Le site I a généré la cellule, C est le centre de la sphère circonscrite au tétraèdre (I,I1,I2,I3), H1 est la projection orthogonale de C sur la face (I,I2,I3), I3' est le milieu de [I,I3], le tétraèdre élémentaire (I,C,H1,I3') est représenté en bleu, le triangle coloré en bleu (C,H1,I3') est une des composantes de la face de contact entre I et I3, enfin la longueur de [C,H1] intervient dans le calcul du périmètre de cette face.

Supposons donc que l'on cherche le centre $H1(H1_x,H1_y,H1_z)$ du cercle circonscrit au triangle (*I, I2, I3*). Ce cercle de rayon *R* est inclus dans la sphère de centre *H*1 et de rayon *R* qui a pour équation :

$(x-H1_x)^2+(y-H1_y)^2+(z-H1_z)^2=R^2$

Puisque *I* est toujours l'origine du repère orthonormé, on peut simplifier cette équation comme nous l'avons fait plus haut et on obtient :

$$\left\|\vec{D}\right\|^2=2\overrightarrow{H1}\cdot\vec{D} \text{ avec } \vec{D} \text{ de coordonnées } (x, y, z)$$

Or on sait que *I2* et *I3* sont sur cette sphère, on a donc :

$$\begin{cases} 2\overrightarrow{H1}\cdot\overrightarrow{D_2}=\left\|\overrightarrow{D_2}\right\|^2=D_2{}^2 \\ 2\overrightarrow{H1}\cdot\overrightarrow{D_3}=\left\|\overrightarrow{D_3}\right\|^2=D_3{}^2 \end{cases}$$

On sait par définition de $H1$ que I, $H1$, $I2$ et $I3$ sont coplanaires, on peut donc écrire :

$$2\overrightarrow{H1}=(\overrightarrow{H1}\cdot\overrightarrow{D_2})\overrightarrow{D_2}+(\overrightarrow{H1}\cdot\overrightarrow{D_3})\overrightarrow{D_3}$$

On obtient alors :

$$\begin{cases} [(\overrightarrow{H1}\cdot\overrightarrow{D_2})\overrightarrow{D_2}+(\overrightarrow{H1}\cdot\overrightarrow{D_3})\overrightarrow{D_3}]\cdot\overrightarrow{D_2}=D_2{}^2 \\ [(\overrightarrow{H1}\cdot\overrightarrow{D_2})\overrightarrow{D_2}+(\overrightarrow{H1}\cdot\overrightarrow{D_3})\overrightarrow{D_3}]\cdot\overrightarrow{D_3}=D_3{}^2 \end{cases}$$

On développe et avec $D2=D_2{}^2$, $D3=D_3{}^2$ et $S1=\overrightarrow{D_2}\cdot\overrightarrow{D_3}$ on a :

$$\begin{cases} (\overrightarrow{H1}\cdot\overrightarrow{D_2})D_2{}^2+(\overrightarrow{H1}\cdot\overrightarrow{D_3})(\overrightarrow{D_3}\cdot\overrightarrow{D_2})=D_2{}^2=(\overrightarrow{H1}\cdot\overrightarrow{D_2})D2+(\overrightarrow{H1}\cdot\overrightarrow{D_3})S1=D2 \\ (\overrightarrow{H1}\cdot\overrightarrow{D_2})(\overrightarrow{D_2}\cdot\overrightarrow{D_3})+(\overrightarrow{H1}\cdot\overrightarrow{D_3})D_3{}^2=D_3{}^2=(\overrightarrow{H1}\cdot\overrightarrow{D_2})S1+(\overrightarrow{H1}\cdot\overrightarrow{D_3})D3=D3 \end{cases}$$

C'est un système à deux équations et dont les deux inconnues sont $\overrightarrow{H1}\cdot\overrightarrow{D_2}$ et $\overrightarrow{H1}\cdot\overrightarrow{D_3}$. Le déterminant principal de ce système s'écrit :

$$\Delta=\begin{vmatrix} D2 & S1 \\ S1 & D3 \end{vmatrix}=D2D3-S1S1$$

D'après la règle de Cramer on a également :

$$\begin{cases} \overrightarrow{H1}\cdot\overrightarrow{D_2}=\dfrac{\begin{vmatrix} D2 & S1 \\ D3 & D3 \end{vmatrix}}{\Delta}=\dfrac{D3(D2-S1)}{\Delta} \\ \overrightarrow{H1}\cdot\overrightarrow{D_3}=\dfrac{\begin{vmatrix} D2 & D2 \\ S1 & D3 \end{vmatrix}}{\Delta}=\dfrac{D2(D3-S1)}{\Delta} \end{cases}$$

On a toujours :

$$2\overrightarrow{H1}=(\overrightarrow{H1}\cdot\overrightarrow{D_2})\overrightarrow{D_2}+(\overrightarrow{H1}\cdot\overrightarrow{D_3})\overrightarrow{D_3}$$

Soit :

$$\overrightarrow{H1}=\frac{(\overrightarrow{H1}\cdot\overrightarrow{D_2})}{2}\overrightarrow{D_2}+\frac{(\overrightarrow{H1}\cdot\overrightarrow{D_3})}{2}\overrightarrow{D_3}$$

Par conséquent :

$$\begin{cases} H1_x = \dfrac{(\overrightarrow{H1}\cdot\overrightarrow{D_2})}{2} D_{2x} + \dfrac{(\overrightarrow{H1}\cdot\overrightarrow{D_3})}{2} D_{3x} \\ H1_y = \dfrac{(\overrightarrow{H1}\cdot\overrightarrow{D_2})}{2} D_{2y} + \dfrac{(\overrightarrow{H1}\cdot\overrightarrow{D_3})}{2} D_{3y} \\ H1_z = \dfrac{(\overrightarrow{H1}\cdot\overrightarrow{D_2})}{2} D_{2z} + \dfrac{(\overrightarrow{H1}\cdot\overrightarrow{D_3})}{2} D_{3z} \end{cases}$$

Si on remplace dans ce système $\overrightarrow{H1}\cdot\overrightarrow{D_2}$ et $\overrightarrow{H1}\cdot\overrightarrow{D_3}$ par leur valeur on obtient :

$$\begin{cases} H1_x = \dfrac{D3(D2-S1)}{2\Delta} D_{2x} + \dfrac{D2(D3-S1)}{2\Delta} D_{3x} \\ H1_y = \dfrac{D3(D2-S1)}{2\Delta} D_{2y} + \dfrac{D2(D3-S1)}{2\Delta} D_{3y} \\ H1_z = \dfrac{D3(D2-S1)}{2\Delta} D_{2z} + \dfrac{D2(D3-S1)}{2\Delta} D_{3z} \end{cases}$$

Toutes les valeurs sont connues, on obtient donc ainsi les coordonnées du point $H1$, les coordonnées de toutes les projections orthogonales du centre de la sphère circonscrite au tétraèdre de Delaunay peuvent être calculées de la même façon. Sur la Figure 39 est représenté un autre de ces tétraèdres ; pour chaque cellule de Voronoï et chaque tétraèdre de Delaunay associé à un de ses sommets, il y a six tétraèdres qui contribuent au volume de cette cellule, et deux qui contribuent à l'aire de chacune des trois faces ayant le sommet considéré en commun.

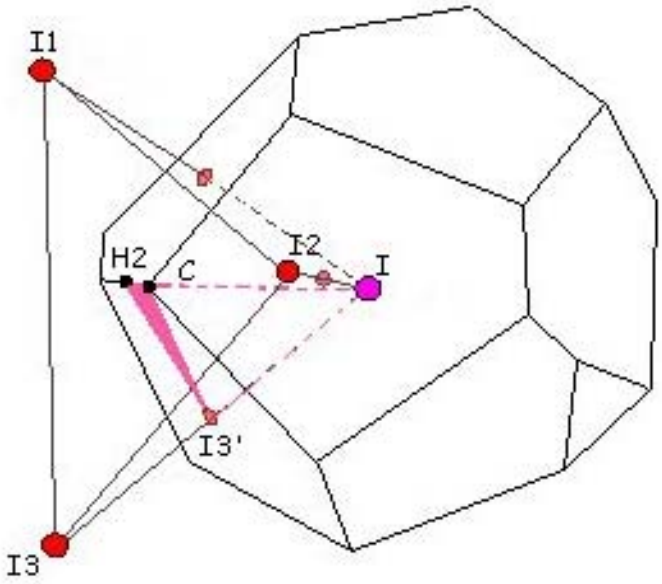


Figure 39 : Le tétraèdre rose (I, C, H2, I3') est un autre tétraèdre élémentaire obtenu en projetant le point C sur la face (I, I1, I3) du tétraèdre de Delaunay représenté.

Le volume du tétraèdre élémentaire dont les sommets sont *I*, *C* le centre de la sphère repéré par le vecteur $\vec{R}$, le milieu d'une arrête [*I*,*I*3] par exemple repéré par $\frac{\overrightarrow{D3}}{2}$ et *H*1 repéré par $\overrightarrow{H1}$ est déterminé par la formule :

$$\left|V_{tétraèdre}\right| = \left|\frac{1}{6}\vec{R}.\frac{\overrightarrow{D3}}{2} \Lambda \overrightarrow{H1}\right| = \left|\frac{1}{12}\begin{vmatrix} Rx & D3x & H1x \\ Ry & D3y & H1y \\ Rz & D3z & H1z \end{vmatrix}\right|$$

Pour les raisons mentionnées plus haut, le programme n'additionne par les valeurs absolues des volumes mais effectue une somme algébrique des contributions de chaque tétraèdre élémentaire.

Les surfaces des triangles se calculent simplement à partir du volume d'une pyramide. Dans le cas présent c'est l'aire du triangle dont les sommets sont : *H*1, le milieu de l'arrête [*I*,*I*3] et *C* que l'on cherche à calculer.

$$V_{pyramide} = \frac{base \times hauteur}{2}$$

Un tétraèdre est une pyramide à base triangulaire et c'est l'aire de cette base que l'on cherche ; on a donc :

$$base = \frac{V_{pyramide} \times 2}{hauteur}$$

La face de la cellule dans laquelle est inclus le triangle dont on cherche l'aire est par définition orthogonale à l'arrête du tétraèdre de Delaunay dont le milieu est un des sommets du tétraèdre élémentaire dont on vient de calculer le volume. La hauteur du tétraèdre élémentaire est donc cette demie arrête de longueur $\left\|\frac{\overrightarrow{D3}}{2}\right\|$.

$$base = \frac{V_{tétraèdre}}{\left\|\frac{\overrightarrow{D3}}{2}\right\|}$$

La somme algébrique de ces aires donne finalement l'aire de la face de contact considérée.

Le périmètre de cette face de contact se calcule en additionnant également les longueurs des arrêtes des triangles précédents. La longueur de cette arrête est la distance entre le sommet de la cellule et la face du tétraèdre de Delaunay concernée, dans le cas présent c'est la

longueur de l'arrête [*CH*1]. Cette longueur se calcule trivialement à partir des coordonnées calculées précédemment.

En résumé, nous venons de voir comment il était possible de calculer toutes les caractéristiques concernant les cellules de Voronoï dans le cas non pondéré. Dans le paragraphe suivant, je vais décrire les changements à apporter pour pouvoir déterminer celles concernant les cellules dans le cas pondéré.

4 - Tessellation de Voronoï pondérée

Le principe de construction des cellules de Voronoï pondérées est le même que dans le cas non pondéré, il est fondé sur la construction d'une triangulation proche de celle de Delaunay mais légèrement différente. Cette triangulation permet toujours de définir les germes qui auront des faces communes avec un site particulier. La recherche des voisins potentiels schématisée par la Figure 33 reste inchangée, les modifications apparaissent lors des premiers calculs.

Dans le cas d'une tessellation non pondérée, le centre de la sphère circonscrite au tétraèdre considéré peut être défini comme étant l'intersection de quatre sphères de même rayon (le rayon de la sphère circonscrite) et dont les centres sont les quatre sommets du tétraèdre. On obtient ainsi les quatre équations de sphère suivantes (on considère toujours que le point *I* est à l'origine) :

$$\begin{cases} x^2+y^2+z^2=R^2 \\ (x-D_{1x})^2+(y-D_{1y})^2+(z-D_{1z})^2=R^2 \\ (x-D_{2x})^2+(y-D_{2y})^2+(z-D_{2z})^2=R^2 \\ (x-D_{3x})^2+(y-D_{3y})^2+(z-D_{3z})^2=R^2 \end{cases}$$

Après simplification on obtient :

$$\begin{cases} x^2+y^2+z^2=R^2 \\ 2\vec{R}\cdot\vec{D_1}=\left\|\vec{D_1}\right\|^2 \\ 2\vec{R}\cdot\vec{D_2}=\left\|\vec{D_2}\right\|^2 \\ 2\vec{R}\cdot\vec{D_3}=\left\|\vec{D_3}\right\|^2 \end{cases}$$

On retrouve bien évidemment les mêmes équations que celles présentées dans le cas d'une tessellation non pondérée. Dans le cas pondéré, on utilise cette même approche. En effet, si on considère que w_i^2 est le poids associé au site I_i (w^2 est le poids associé à l'origine I), le sommet de la cellule *C* (*a, b, c*) défini par le tétraèdre considéré est l'intersection de

quatre sphères dont les centres sont les sommets du tétraèdre et dont les rayons sont dépendants à la fois des poids associés à chaque sommet et de la position des sommets du tétraèdre de telle manière que l'intersection des quatre sphères existe toujours. Les quatre équations deviennent alors :

$$\begin{cases} a^2+b^2+c^2=L^2+w^2 \\ (a-D_{1x})^2+(b-D_{1y})^2+(c-D_{1z})^2=L^2+w_1^2 \\ (a-D_{2x})^2+(b-D_{2y})^2+(c-D_{2z})^2=L^2+w_2^2 \\ (a-D_{3x})^2+(b-D_{3y})^2+(c-D_{3z})^2=L^2+w_3^2 \end{cases}$$

Ceci est illustré à deux dimensions par la Figure 40 ; le triangle (*I*, *I1*, *I2*) est un triangle de Delaunay, les poids associés à chacun de ces points sont respectivement 4, 0.5, 2 et sont représentés par des cercles noirs centrés sur ces points et de rayons égaux au poids. Le point C est l'intersection des trois cercles en bleu centrés eux aussi sur *I*, *I1*, et *I2* dont les rayons varient avec les poids. Les tangentes aux cercles noirs sont représentées en rouge, par définition (*CM*) et (*IM*) sont orthogonales, le triangle (*CIM*) est donc un triangle rectangle. Si on note *L* la distance entre *C* et *M*, *w* la distance entre *I* et *M* (le rayon du cercle représentant le poids, donc 4 ici) et *R* la distance entre *C* et *I* c'est à dire le rayon du cercle bleu passant par *C*, on a $R^2=w^2+L^2$ qui correspond bien à l'équation $a^2+b^2+c^2=L^2+w^2$ du système vu plus haut. Comme nous l'avons déjà vu dans le chapitre précédent en ce qui concerne les tangentes aux cercles de rayons égaux aux poids issues de *C*, on constate à nouveau que la distance entre *C* et le point de tangence est la même quel que soit le cercle considéré, tous ces points de tangence appartiennent donc à un cercle de centre *C* et de rayon *L* (en vert sur la figure).

Après simplification, on a :

$$\begin{cases} L^2=a^2+b^2+c^2-w^2 \\ 2\vec{R}\cdot\overrightarrow{D_1}=\left\|\overrightarrow{D_1}\right\|^2-w_1^2+w^2 \\ 2\vec{R}\cdot\overrightarrow{D_2}=\left\|\overrightarrow{D_2}\right\|^2-w_2^2+w^2 \\ 2\vec{R}\cdot\overrightarrow{D_3}=\left\|\overrightarrow{D_3}\right\|^2-w_3^2+w^2 \end{cases}$$

Les coordonnées *(a, b, c)* du point d'intersection *C* des quatre sphères se calculent alors de la même façon que dans le cas non pondéré à partir d'un système de trois équations à trois inconnues ; l'inconnue *L* se calcule aisément à partir des coordonnées de *C* et du poids associé à *I* :

$$\begin{cases} ax_1+by_1+cz_1=(D_1^2-w_1^2+w^2)/2 \\ ax_2+by_2+cz_2=(D_2^2-w_2^2+w^2)/2 \\ ax_3+by_3+cz_3=(D_3^2-w_3^2+w^2)/2 \end{cases}$$

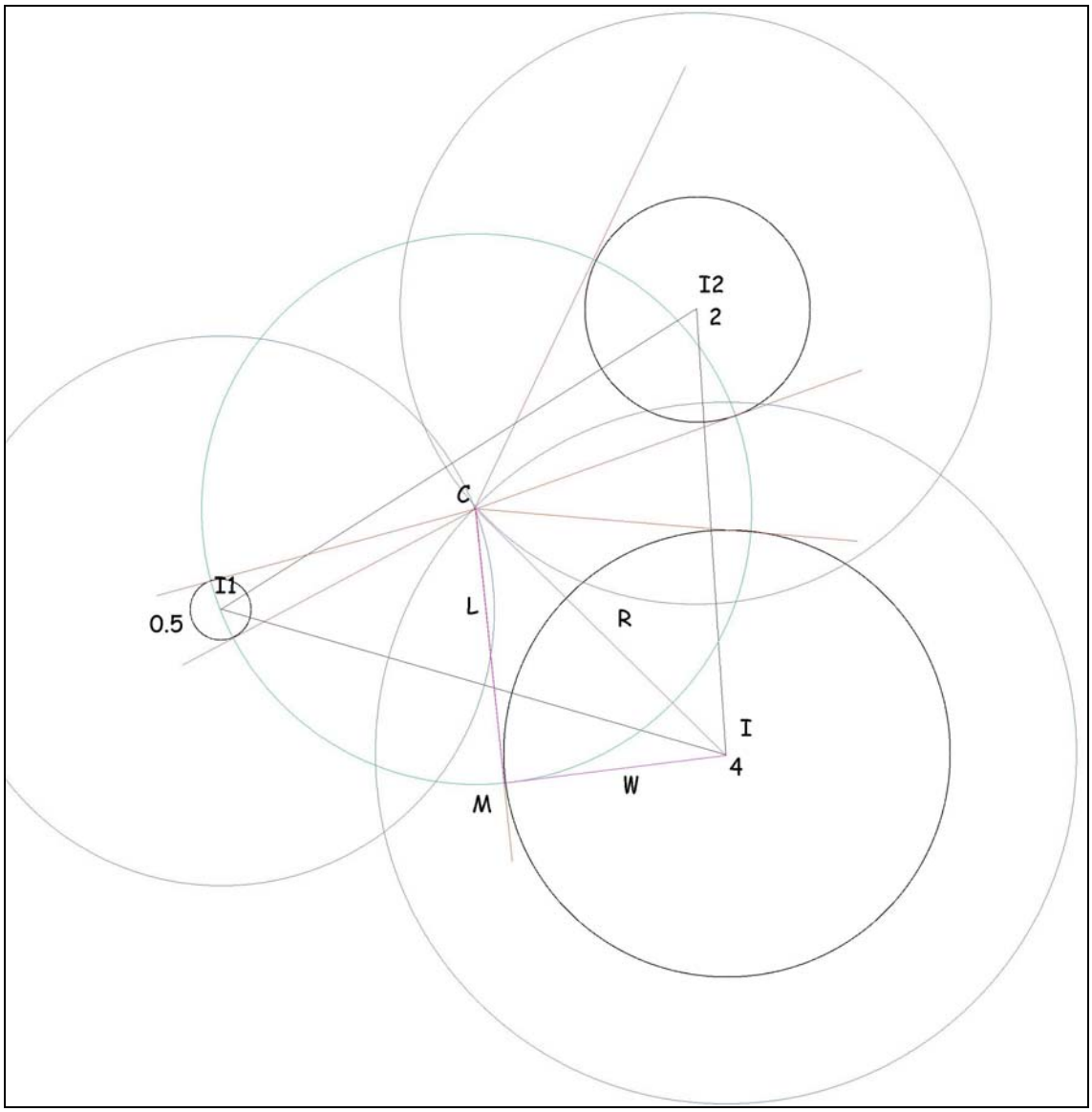


Figure 40 : Détermination de *C* à deux dimensions.

Δ, le déterminant principal de ce système, reste inchangé :

$$\Delta=\begin{vmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{vmatrix}$$

D'après la règle de Cramer, si $\Delta \neq 0$ on a maintenant les nouvelles coordonnées du sommet :

$$a=\frac{\begin{vmatrix} D_1^2-w_1^2+w^2 & y_1 & z_1 \\ D_2^2-w_2^2+w^2 & y_2 & z_2 \\ D_3^2-w_3^2+w^2 & y_3 & z_3 \end{vmatrix}}{2\Delta} \quad b=\frac{\begin{vmatrix} x_1 & D_1^2-w_1^2+w^2 & z_1 \\ x_2 & D_2^2-w_2^2+w^2 & z_2 \\ x_3 & D_3^2-w_3^2+w^2 & z_3 \end{vmatrix}}{2\Delta} \quad c=\frac{\begin{vmatrix} x_1 & y_1 & D_1^2-w_1^2+w^2 \\ x_2 & y_2 & D_2^2-w_2^2+w^2 \\ x_3 & y_3 & D_3^2-w_3^2+w^2 \end{vmatrix}}{2\Delta}$$

Il est d'ores et déjà intéressant de noter que si tous les poids sont égaux (et pas nécessairement nuls), on retrouve les coordonnées précédemment déterminées pour une tessellation non pondérée ; on constate également que les poids peuvent être définis à une constante additive près, puisque c'est leur différence qui intervient lors du calcul.

A ce stade, on dispose donc des coordonnées du point d'intersection des quatre sphères, pour déterminer si ce point est un des sommets de cellule de la tessellation pondérée, il reste à vérifier que le tétraèdre étudié fait bien partie de la triangulation finale, pour cela on procède de la même façon que pour la triangulation de Delaunay. Dans ce dernier cas on cherchait à savoir si un cinquième point *I4* était bien à l'extérieur de la sphère de rayon *R* circonscrite au tétraèdre de Delaunay, ceci revenait en fait à savoir si le centre de cette sphère était contenu dans la sphère de centre *I4* et également de rayon *R*. Dans le cas d'une tessellation pondérée, le principe est le même, on cherche à savoir si le point *C* est à l'extérieur de la sphère de centre *I4* et de rayon *R'* tel que $R'^2=L^2+w_4^2$.

Pour cela, il faut avoir :

$$(x_4-a)^2+(y_4-b)^2+(z_4-c)^2>L^2+w_4^2$$

Or on a $a^2+b^2+c^2=L^2+w^2$ donc, après simplification, on obtient :

$$x_4^2+y_4^2+z_4^2-w_4^2+w^2>2ax_4+2by_4+2cz_4$$

On remplace a, b, et c par leur expression :

$$D_4^2-w_4^2+w^2-x_4\frac{\begin{vmatrix} D_1^2-w_1^2+w^2 & y_1 & z_1 \\ D_2^2-w_2^2+w^2 & y_2 & z_2 \\ D_3^2-w_3^2+w^2 & y_3 & z_3 \end{vmatrix}}{\Delta}-y_4\frac{\begin{vmatrix} x_1 & D_1^2-w_1^2+w^2 & z_1 \\ x_2 & D_2^2-w_2^2+w^2 & z_2 \\ x_3 & D_3^2-w_3^2+w^2 & z_3 \end{vmatrix}}{\Delta}-z_4\frac{\begin{vmatrix} x_1 & y_1 & D_1^2-w_1^2+w^2 \\ x_2 & y_2 & D_2^2-w_2^2+w^2 \\ x_3 & y_3 & D_3^2-w_3^2+w^2 \end{vmatrix}}{\Delta}>0$$

En multipliant de chaque côté de l'inégalité par Δ, on obtient :

$$\begin{cases} \Delta(D_4^2-w_4^2+w^2)-x_4\begin{vmatrix} D_1^2-w_1^2+w^2 & y_1 & z_1 \\ D_2^2-w_2^2+w^2 & y_2 & z_2 \\ D_3^2-w_3^2+w^2 & y_3 & z_3 \end{vmatrix}-y_4\begin{vmatrix} x_1 & D_1^2-w_1^2+w^2 & z_1 \\ x_2 & D_2^2-w_2^2+w^2 & z_2 \\ x_3 & D_3^2-w_3^2+w^2 & z_3 \end{vmatrix}-z_4\begin{vmatrix} x_1 & y_1 & D_1^2-w_1^2+w^2 \\ x_2 & y_2 & D_2^2-w_2^2+w^2 \\ x_3 & y_3 & D_3^2-w_3^2+w^2 \end{vmatrix}>0, \Delta>0 \\ \Delta(D_4^2-w_4^2+w^2)-x_4\begin{vmatrix} D_1^2-w_1^2+w^2 & y_1 & z_1 \\ D_2^2-w_2^2+w^2 & y_2 & z_2 \\ D_3^2-w_3^2+w^2 & y_3 & z_3 \end{vmatrix}-y_4\begin{vmatrix} x_1 & D_1^2-w_1^2+w^2 & z_1 \\ x_2 & D_2^2-w_2^2+w^2 & z_2 \\ x_3 & D_3^2-w_3^2+w^2 & z_3 \end{vmatrix}-z_4\begin{vmatrix} x_1 & y_1 & D_1^2-w_1^2+w^2 \\ x_2 & y_2 & D_2^2-w_2^2+w^2 \\ x_3 & y_3 & D_3^2-w_3^2+w^2 \end{vmatrix}<0, \Delta<0 \end{cases}$$

Or :

$$\Theta=\Delta(D_4^2-\omega_4^2+\omega^2)-x_4\begin{vmatrix} D_1^2-w_1^2+w^2 & y_1 & z_1 \\ D_2^2-w_2^2+w^2 & y_2 & z_2 \\ D_3^2-w_3^2+w^2 & y_3 & z_3 \end{vmatrix}-y_4\begin{vmatrix} x_1 & D_1^2-w_1^2+w^2 & z_1 \\ x_2 & D_2^2-w_2^2+w^2 & z_2 \\ x_3 & D_3^2-w_3^2+w^2 & z_3 \end{vmatrix}-z_4\begin{vmatrix} x_1 & y_1 & D_1^2-w_1^2+w^2 \\ x_2 & y_2 & D_2^2-w_2^2+w^2 \\ x_3 & y_3 & D_3^2-w_3^2+w^2 \end{vmatrix}$$

$$= \begin{vmatrix} x_1 & y_1 & z_1 & D_1^2-w_1^2+w^2 \\ x_2 & y_2 & z_2 & D_2^2-w_2^2+w^2 \\ x_3 & y_3 & z_3 & D_3^2-w_3^2+w^2 \\ x_4 & y_4 & z_4 & D_4^2-w_4^2+w^2 \end{vmatrix}$$

On a donc finalement :

$$\begin{cases} \Theta>0, \Delta>0 \\ \Theta<0, \Delta<0 \end{cases}$$

En conclusion :

Si $\begin{vmatrix} x_1 & y_1 & z_1 & D_1^2-w_1^2+w^2 \\ x_2 & y_2 & z_2 & D_2^2-w_2^2+w^2 \\ x_3 & y_3 & z_3 & D_3^2-w_3^2+w^2 \\ x_4 & y_4 & z_4 & D_4^2-w_4^2+w^2 \end{vmatrix}$ et $\begin{vmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{vmatrix}$ sont de même signe, la sphère de centre *I4* et de rayon *R'* tel que $R'^2=L^2+w_4^2$ ne contient pas le point *C*, on considère alors que *C* est donc bien un sommet de la cellule de Voronoï associée au site *I*. La suite des opérations concernant la

tessellation reste inchangée, seuls les calculs concernant les volumes et les surfaces sont à modifier.

4.1 Calculs des volumes et des surfaces.

Nous avons vu plus haut que le calcul de ces valeurs impliquait le calcul des coordonnées des projections orthogonales du sommet *C* de la cellule sur les faces du tétraèdre. Ce calcul était effectué en considérant que ces projections étaient les centres des cercles circonscrits aux triangles constituant les faces sur lesquelles on projetait. Dans le cas d'une tessellation pondérée, le sommet *C* n'est plus le centre de la sphère circonscrite au tétraèdre et ses projections orthogonales sur les faces du tétraèdre ne correspondent plus aux centres des cercles circonscrits de ces faces. Il faut donc procéder à un calcul différent. Ce nouveau calcul est toujours fondé sur une décomposition en tétraèdres élémentaires mais ces derniers sont légèrement différents de ceux que l'on utilisait précédemment. Cette décomposition ne s'effectue plus à partir des tétraèdres de Delaunay que l'on prenait tour à tour mais à partir de chacune des faces des cellules. Le principe consiste à ordonner les sommets d'une face dans le sens trigonométrique, de prendre deux sommets successifs, le site associé à la cellule que l'on considère et le point d'intersection du plan de la face de la cellule et de l'arrête du tétraèdre liant les deux sites que la face de Voronoï met en voisinage (la petite sphère bleue sur la Figure 41).

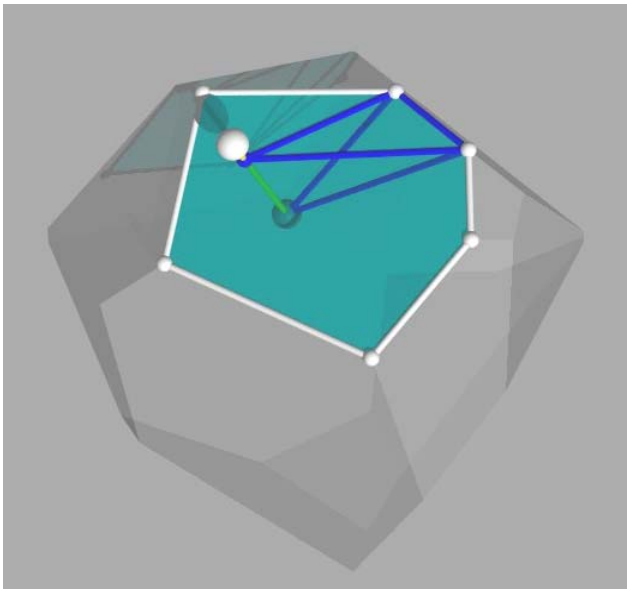


Figure 41 : Cellule d'une tessellation pondérée : la face de contact entre les deux sites est en vert, l'intersection entre la face et le segment liant les deux sites est représentée par une sphère bleue, le tétraèdre élémentaire est en bleu.

L'ensemble de ces tétraèdre élémentaires ayant comme sommet commun le site générateur de la cellule donnera le volume de celle-ci ainsi que l'aire totale et l'aire de chacune des faces de cette cellule.

Le site *I* est toujours l'origine, le site voisin *I1* est repéré par $\overrightarrow{D_1}$ de coordonnées (x_1,y_1,z_1), les sommets consécutifs de la face sont C_i et C_{i+1} de coordonnées respectives (cx_i,cy_i,cz_i) et $(cx_{i+1},cy_{i+1},cz_{i+1})$, enfin l'intersection de la face de la cellule et du segment [*I*,*I1*] est *M* de coordonnées (Mx,My,Mz). Seules les coordonnées de *M* restent à déterminer. Comme nous l'avons vu au chapitre précédent, on sait que le plan dans lequel est incluse la face de Voronoï séparant *I* et *I1* est à la distance *d* de *I* telle que : $\frac{d}{\left\|\overrightarrow{D_1}\right\|}=\frac{1}{2}\left(\left\|\overrightarrow{D_1}\right\|+\frac{w^2-w_1^2}{\left\|\overrightarrow{D_1}\right\|}\right)$. Le point *M* est donc défini par : $\overrightarrow{IM}=\frac{1}{2}\left(\left\|\overrightarrow{D_1}\right\|+\frac{w^2-w_1^2}{\left\|\overrightarrow{D_1}\right\|}\right)\overrightarrow{D_1}$ et ses coordonnées se calculent donc à partir de celles de $\overrightarrow{D_1}$.

$$M\begin{vmatrix} \frac{1}{2}\left(\left\|\overrightarrow{D_1}\right\|+\frac{w^2-w_1^2}{\left\|\overrightarrow{D_1}\right\|}\right)D_{1x} \\ \frac{1}{2}\left(\left\|\overrightarrow{D_1}\right\|+\frac{w^2-w_1^2}{\left\|\overrightarrow{D_1}\right\|}\right)D_{1y} \\ \frac{1}{2}\left(\left\|\overrightarrow{D_1}\right\|+\frac{w^2-w_1^2}{\left\|\overrightarrow{D_1}\right\|}\right)D_{1z} \end{vmatrix}$$

On dispose donc maintenant des coordonnées des quatre points du tétraèdre élémentaire. Le volume de ce tétraèdre se calcule à l'aide de la formule :

$$\left|V_{tétraèdre}\right|=\left|\frac{1}{6}\overrightarrow{IM}\cdot\overrightarrow{IC_i}\Lambda\overrightarrow{IC_{i+1}}\right|=\frac{1}{6}\begin{vmatrix} M_x & cx_i & cx_{i+1} \\ M_y & cy_i & cy_{i+1} \\ M_z & cz_i & cz_{i+1} \end{vmatrix}$$

L'aire de la face de ce tétraèdre contribuant à l'aire totale de la face de la cellule de Voronoï, se calcule à partir de ce volume. En effet, puisque cette face est incluse dans un plan orthogonal à [*I*,*I1*], ce dernier segment constitue une hauteur pour le tétraèdre. Or comme nous l'avons vu plus haut, un tétraèdre est une pyramide à base triangulaire et c'est l'aire de cette base que l'on cherche, on a donc :

$$base = \frac{V_{tétraèdre} \times 2}{hauteur} = \frac{V_{tétraèdre} \times 2}{\left\| \overrightarrow{II1} \right\|}$$

La longueur de l'arrête contribuant au périmètre total de la face de Voronoï est la distance entre les deux points C_i et C_{i+1}.

5 - Cellule ouverte ou cellule fermée ?

Nous avons vu dans le chapitre précédent que certaines cellules de Voronoï sont complètement fermées et que d'autres sont ouvertes, on peut dire aussi qu'elles sont infinies. La procédure de construction que je viens de détailler (avec ou sans pondération) détermine en fait les différents sommets des cellules et les diverses grandeurs associées à celles-ci. Telle quelle, cette procédure n'est pas capable de différencier les polyèdres clos des cellules ouvertes, il est d'ailleurs intéressant de noter que cette procédure est faite de telle façon que des cellules ouvertes et donc infinies peuvent avoir des volumes et des surfaces déterminés et finis. Pour ne pas prendre en considération ces chiffres il est donc nécessaire de distinguer les deux classes de cellules. Pour cela on procède à des calculs d'angle dièdre.

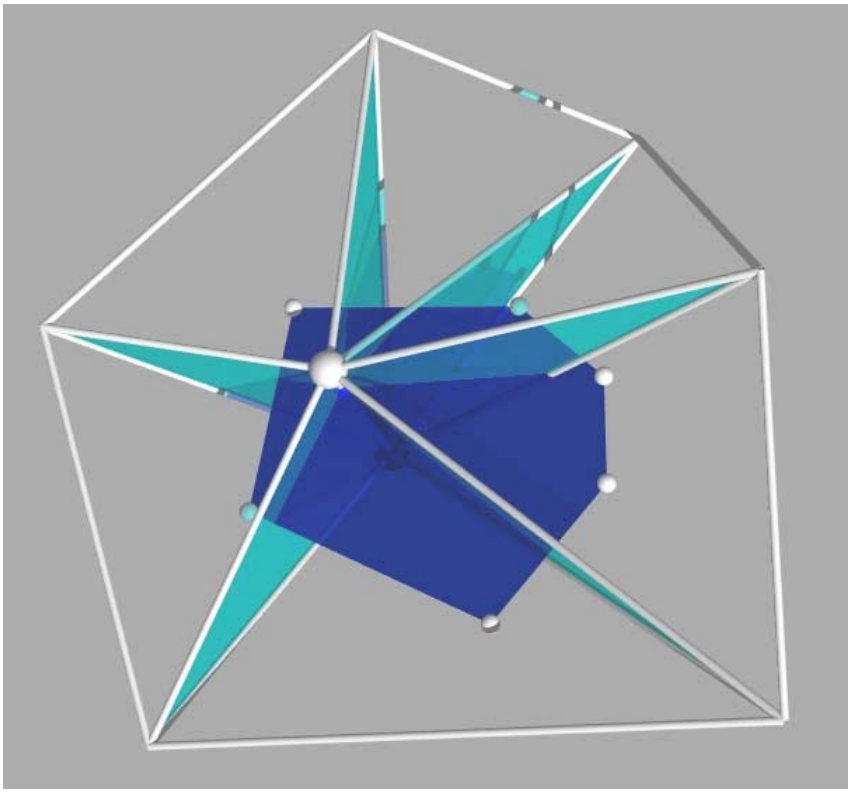


Figure 42 : Une face d'une cellule est représentée en bleu foncé, les tétraèdres de Delaunay associés à cette face sont en blanc ainsi que les sommets de la face. Pour cette dernière on calcule les angles dièdres entre les faces communes aux différents tétraèdres, celles-ci sont représentées en bleu clair.

Pour chaque cellule de Voronoï on procède face par face, et on vérifie qu'elles sont fermées. Sur la Figure 42 est représentée une face de Voronoï fermée, on constate que les tétraèdres associés à ce polygone « tournent » autour de leur arrête commune et forment un tour complet autour de celle-ci. Dans le cas d'une cellule ouverte il est facile d'imaginer que ce tour ne serait pas complet. Mathématiquement cela se traduit par le fait que la somme des angles dièdres entre les faces consécutives des tétraèdres de Delaunay dont au moins une des arrêtes est l'arrête commune à tous les tétraèdres (les faces en bleu clair de la Figure 42) est égale à 2π pour une cellule fermée.

Considérons une de ces faces, nous disposons des coordonnées des sommets de cette face, *I*, *I1* et *I2* par exemple, *I* est toujours à l'origine et *I1* et *I2* sont toujours repérables par $\overrightarrow{D_1}$ et $\overrightarrow{D_2}$, le vecteur $\overrightarrow{D_1} \wedge \overrightarrow{D_2}$ est colinéaire à la normale de cette face. Considérons maintenant la face consécutive, si [*I*,*I*1] est l'arrête commune à tous les tétraèdres de Delaunay associés à la face de Voronoï considérée, cette face consécutive est nécessairement définie par *I*, *I1* et *I3*, le vecteur $\overrightarrow{D_1} \wedge \overrightarrow{D_3}$ est colinéaire à la normale de cette face. L'angle α entre ces deux normales est l'angle dièdre entre les deux faces du tétraèdre étudié

On a :

$$(\overrightarrow{D_1} \wedge \overrightarrow{D_2}) \cdot (\overrightarrow{D_1} \wedge \overrightarrow{D_3}) = \left\| (\overrightarrow{D_1} \wedge \overrightarrow{D_2}) \right\| \times \left\| (\overrightarrow{D_1} \wedge \overrightarrow{D_3}) \right\| \cos\alpha$$

Donc :

$$\alpha = \arccos\left(\frac{(\overrightarrow{D_1} \wedge \overrightarrow{D_2}) \cdot (\overrightarrow{D_1} \wedge \overrightarrow{D_3})}{\left\| (\overrightarrow{D_1} \wedge \overrightarrow{D_2}) \right\| \times \left\| (\overrightarrow{D_1} \wedge \overrightarrow{D_3}) \right\|} \right)$$

L'angle dièdre entre les deux faces est ainsi défini, si la somme de tous les angles dièdres est égale à 2π, la face de Voronoï étudiée est un polygone fermé. Si toutes les faces de la cellule sont fermées cette cellule est un polyèdre clos. Dans le cas contraire, si une face n'est pas fermée, la cellule est ouverte.

6 - Utilisation d'un environnement

Dans l'étude des structures protéiques, le fait que certaines cellules puissent être ouvertes pose un problème car il est alors impossible d'associer un polyèdre à chaque AA. A cause de l'absence de voisins à l'extérieur de la protéine, les résidus concernés sont ceux situés en surface. Pour remédier à cet inconvénient, un environnement artificiel a été modélisé

afin de pouvoir clore toutes les cellules[63]. Cet environnement consiste en un empilement aléatoire relaxé de sphères de 6.5 Å de diamètre. Le volume d'une telle sphère est de 143.8 $Å^3$, ce qui est proche du volume moyen des 20 AA présents dans les protéines qui est à peu près de 143 $Å^3$. Ce diamètre est légèrement différent de celui indiqué dans la publication d'Angelov et coll[63] qui était de 7 Å. Cette modification fait suite à diverses observations : d'une part, le volume d'une sphère de 7 Å de diamètre est de 179.6 $Å^3$ ce qui est plus éloigné de la moyenne des volumes des AA ; de plus, un diamètre de 6.5 Å permet de minimiser les désaccords entre certaines valeurs associées aux cellules en volume dans le cas d'une tessellation sans environnement par rapport aux mêmes cellules dans le cas d'une tessellation avec environnement (ceci est résumé dans le Tableau 2).

	nb de cotés par face	distance entre voisins	nb de faces par cellule
6.5 Å	1.05%	1.07%	0.62%
7.0 Å	0.99%	6.17%	10.55%

Tableau 2 : Comparaison des désaccords entre une tessellation sans environnement et une tessellation avec environnement dans le cas de sphères de 6.5 Å ou de 7.0 Å de diamètre (effectif de 4106 AA).

Cet environnement a été conçu de telle sorte que ses propriétés de compaction soient proches de celles observées pour les protéines et quatre critères différents ont été considérés afin de le sélectionner parmi les environnements testés.

1 Le rapport du nombre de cellules en surface (en contact avec l'environnement) et du nombre total de cellules doit être proche des 2/3 ce qui est approximativement la valeur du rapport entre le nombre d'AA exposés au solvant et le nombre d'AA total pour les protéines globulaires. Ce critère influence notamment la densité de l'environnement retenu.

2 Les distributions du nombre de faces par cellule pour les cellules en volume d'une part et pour celles en surface d'autre part doivent avoir des profils similaires (les valeurs moyennes peuvent être différentes).

3 La distribution du nombre de faces de contact avec l'environnement par cellule en surface ne doit pas présenter de singularités fortes afin d'éliminer les cellules très allongées que l'on peut rencontrer par exemple lors d'une tessellation sans environnement.

4 Enfin, pour des AA voisins en séquence, le nombre de cellules correspondantes n'ayant pas de face en commun doit être proche de zéro.

L'installation de l'environnement est conçue de manière à ce que les cavités internes de taille suffisante soient remplies par des sphères et qu'il y ait au moins trois couches de sphères autour de la structure protéique. C'est la raison pour laquelle, pour permettre aux plus grosses

protéines d'être entièrement immergées dans l'environnement, la boîte initiale dans laquelle sont contenus les 1024 points correspondant aux centres des 1024 sphères est multipliée et translatée afin d'obtenir une plus grande boîte de 27×1024 points (ce processus est schématisé dans la Figure 43).

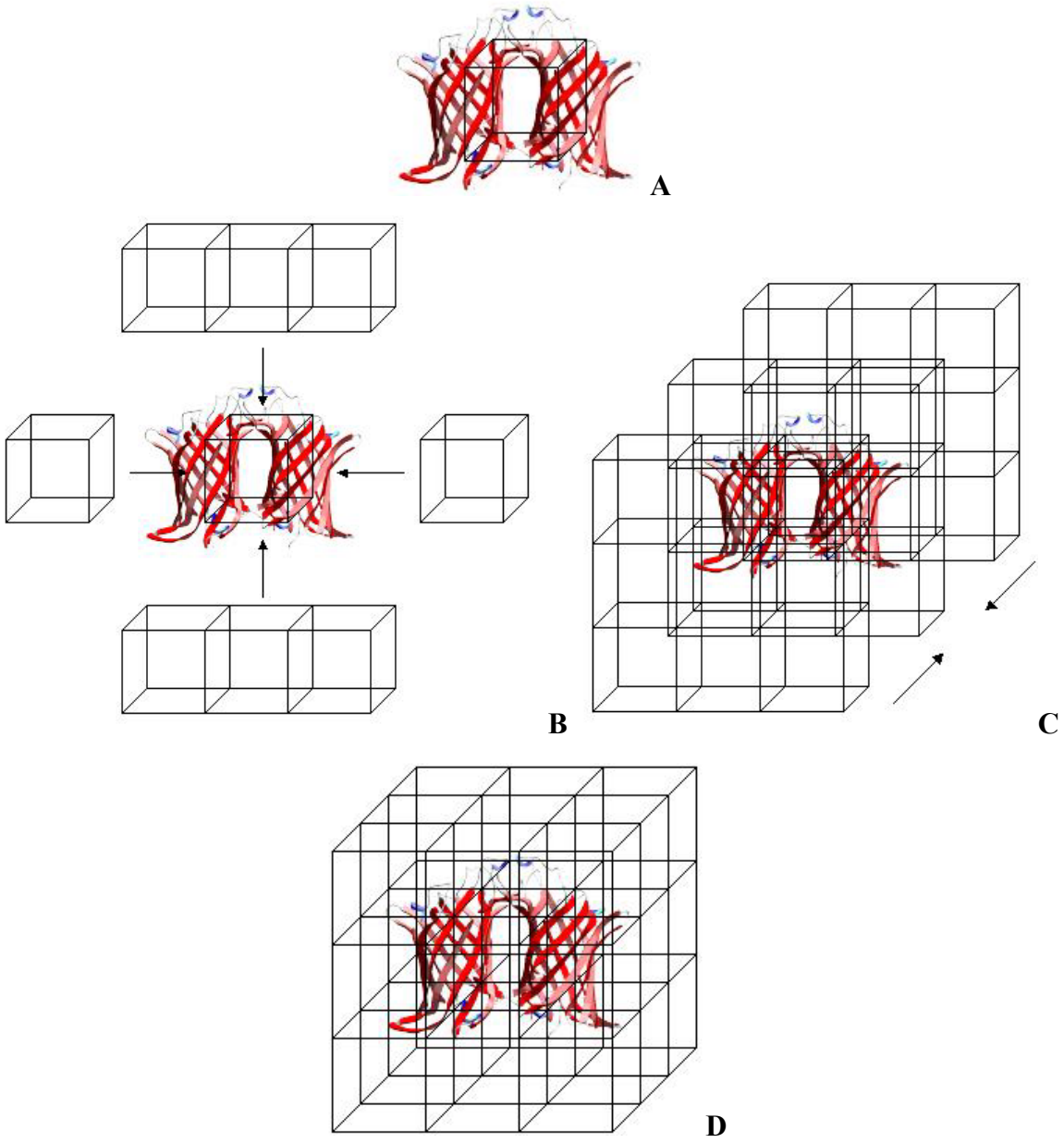
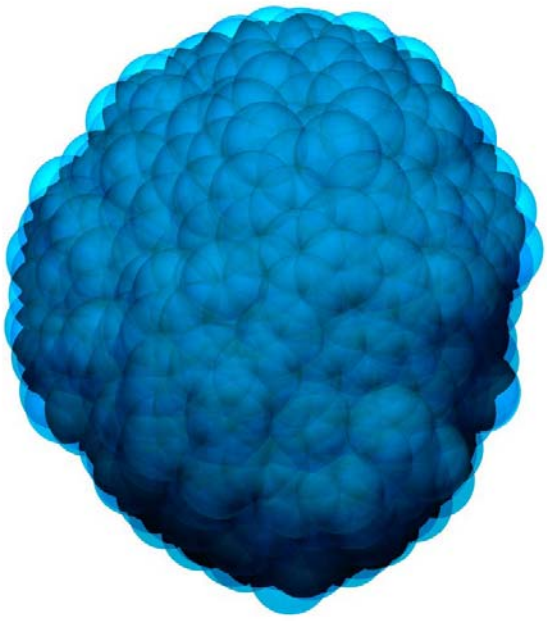


Figure 43
A : La protéine est trop volumineuse pour être contenue dans la boîte initiale.
B et C : On ajoute des boîtes identiques dans toutes les directions.
D : La protéine est maintenant entourée par 27648 (27×1024) sphères.

A l'interface entre la protéine et l'environnement, les sphères chevauchant les AA sont supprimées, il apparaît donc des disparités qui ne permettraient pas de remplir les conditions

citées plus haut. Pour résoudre ce problème, on procède à une relaxation des sphères de l'environnement de la manière suivante. Une première tessellation est réalisée sur les points de la protéine et sur les centres des sphères de l'environnement. Les centres géométriques des cellules de l'environnement sont alors déterminés afin de servir de nouveaux points pour l'environnement. On procède alors à une nouvelle tessellation en considérant uniquement les nouvelles positions des points de l'environnement et en gardant toujours les points correspondant aux AA. Cette opération est ensuite répétée jusqu'à ce que les nouvelles positions de l'environnement ne fluctuent plus d'une tessellation à la suivante. Il faut en général six itérations pour parvenir à ce résultat. Pour pouvoir automatiser cette opération et être sûr de la qualité de l'environnement utilisé, tous les résultats présentés dans la suite de ce rapport ont été obtenus avec un environnement relaxé neuf fois. La Figure 44 présente l'aspect de l'environnement autour de la protéine 2Fe-2S ferredoxine issue de *Haloarcula marismortui*[64], les sphères sont transparentes mais cependant on constate que l'on ne peut pas voir la structure protéique, indiquant que celle-ci est suffisamment entourée de sphères. La Figure 45 reprend la même figure, mais cette fois les sphères sont représentées avec un rayon beaucoup plus faible.



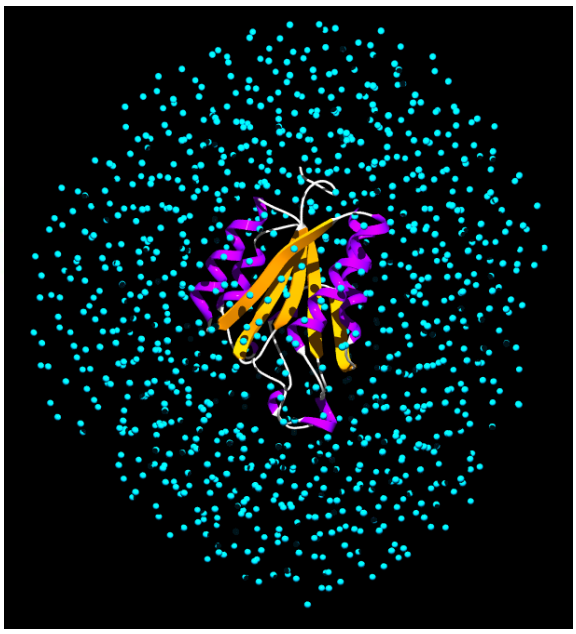


Figure 44 (à gauche) : Environnement autour de la protéine 2Fe-2S ferredoxine issue de *Haloarcula marismortui* (code PDB 1doi, 1 chaîne, 128 AA).
Figure 45 (à droite) : Même protéine, les hélices sont en violet et les brins en orange, les points de l'environnement sont représentés pas des petites sphères bleues.

La Figure 46 A représente la toxine beta-purothionine de *Triticum aestivum*[65] avec une tessellation effectuée sans environnement. Seules les cellules closes sont représentées et on peut constater qu'il en manque un grand nombre. Parmi celles qui sont représentées certaines d'entre elles présentent des formes très asymétriques. La Figure 46 B présente la même protéine avec une tessellation effectuée avec environnement (les cellules de ce dernier ne sont pas représentées). On constate maintenant que toutes les cellules sont bien présentes et qu'elles ont des formes beaucoup plus régulières. Toutefois, certaines d'entre elles ne contiennent pas totalement l'AA auquel elles sont associées.

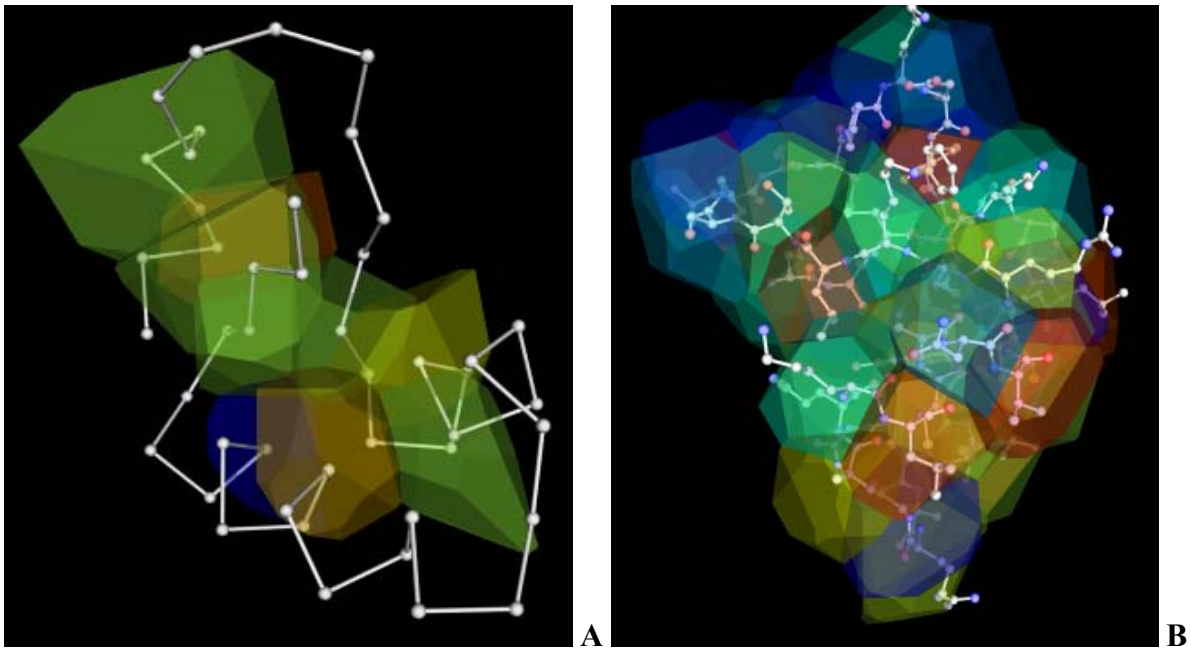
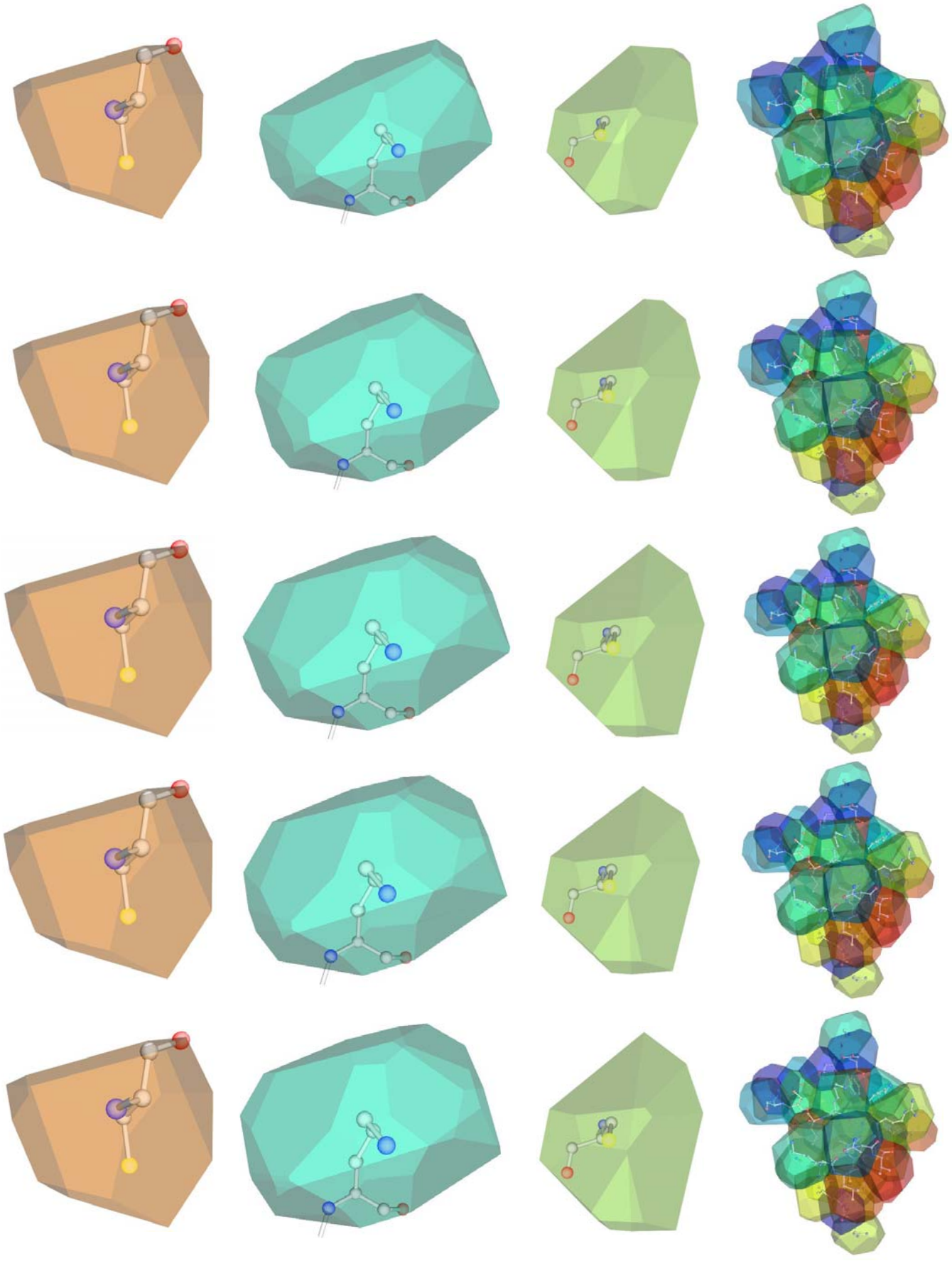


Figure 46 : Toxine beta-purothionine de *Triticum aestivum* (code PDB 1bhp, 45 AA).
A Tessellation sur les C*α*, sans environnement. Seules les cellules fermées sont représentées.
B Tessellation sur les C*α*, avec environnement, les cellules de ce dernier ne sont pas représentées.

La Figure 47 représente plusieurs cellules à différents stades de la relaxation (entre 0 et 9). Le volume et le nombre de faces de ces cellules sont donnés dans le Tableau 3. On peut tout d'abord constater que l'influence de la relaxation ne se fait pas ressentir pour la cystéine n°25 (C25) qui possède une cellule en volume puisqu'elle ne fait aucun contact avec des cellules liées à des points de l'environnement. Pour la lysine n°41 (K41) qui au contraire est très exposée à l'environnement, les choses sont légèrement différentes ; en effet le nombre de faces ne change pas, ni les proportions de faces avec des AA ou avec des points de l'environnement. En revanche, le volume évolue sensiblement puisqu'il passe de 316.5 $Å^3$ à 276.6 $Å^3$ soit une diminution de 12.6% dont 11.6% dès les cinq premières relaxations.



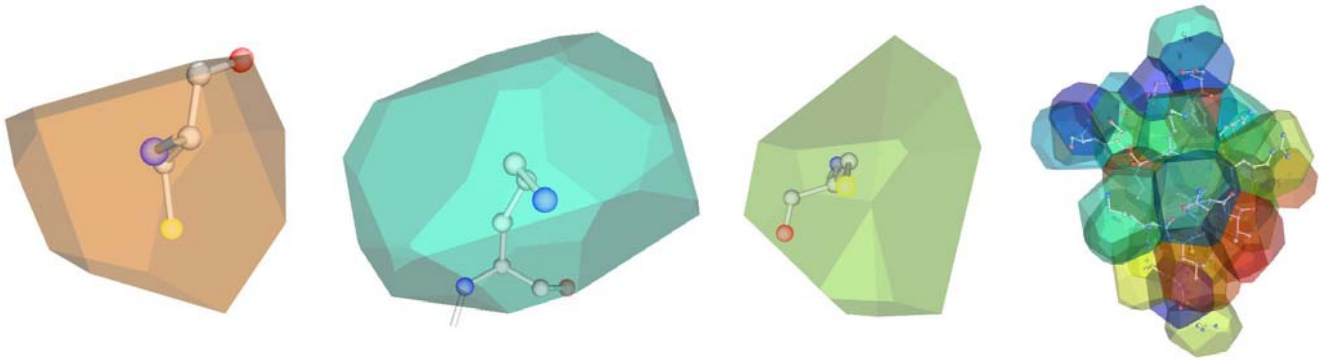


Figure 47 : TdV sur les centres géométriques des chaînes latérales en non pondéré sur la structure 1bhp. De gauche à droite : C25, K41, C12 et la structure entière. La 1ère ligne correspond aux tessellations sans relaxation, la 2ème ligne aux tessellations avec 1 relaxation, la 3ème à 3 relaxations, puis 5, 7 et 9 relaxations pour la dernière ligne.

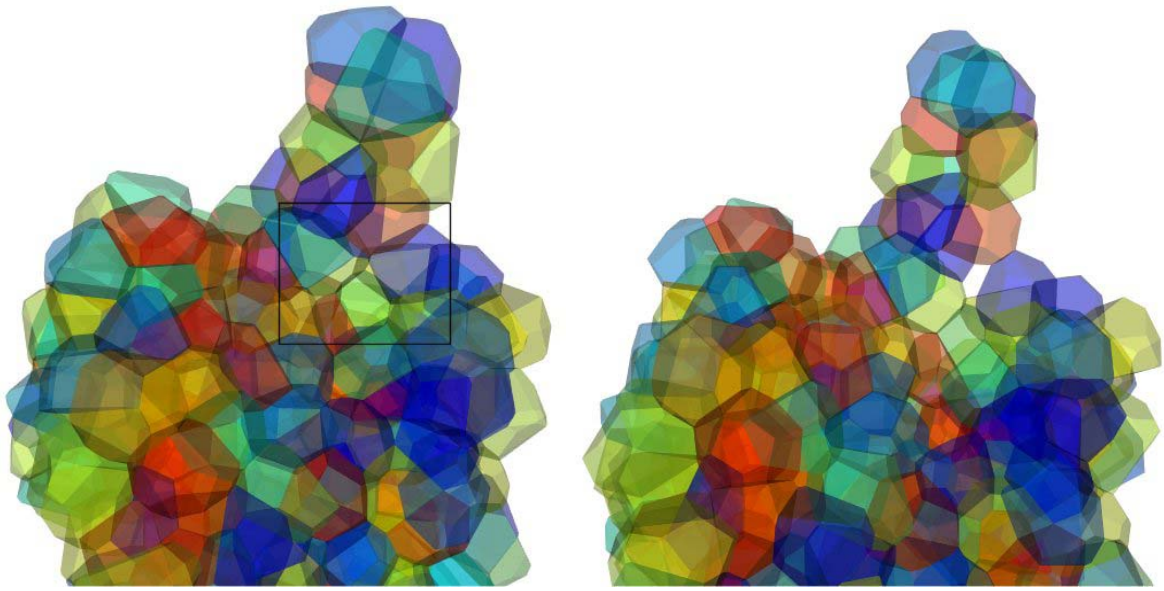
	C 25	K 41	C 12
0	113.0 11 (11 - 0)	316.5 20 (3 -17)	181.0 14 (9 - 5)
1	113.0 11 (11 - 0)	310.0 20 (3 -17)	166.6 12 (9 - 3)
3	113.0 11 (11 - 0)	290.1 20 (3 -17)	153.9 11 (8 - 3)
5	113.0 11 (11 - 0)	281.0 20 (3 -17)	151.2 11 (8 - 3)
7	113.0 11 (11 - 0)	277.7 20 (3 -17)	149.4 11 (8 - 3)
9	113.0 11 (11 - 0)	276.6 20 (3 -17)	148.3 11 (8 - 3)

Tableau 3 : Pour chaque stade de la relaxation et pour chaque AA sont indiqués le volume en Å^3 (ligne supérieure) et le nombre de faces (ligne inférieure) avec entre parenthèses le nombre de faces de contact avec des cellules liées à des AA (1er nombre) et à des points de l'environnement (2nd nombre).

Pour la cystéine n°12 (C12) dont la cellule est exposée à l'environnement mais dans une moindre mesure que celle de la lysine n°41 (K41), on observe une évolution concernant à la fois le volume et le nombre de faces ainsi que leur proportion. La variation de volume est de 18% et cette cellule perd trois faces entre la première et la troisième relaxation. L'aspect général des cellules ne varie pas de manière drastique et les valeurs associées semblent évoluer rapidement pendant les premières relaxations. Le nombre de faces se stabilise très vite et les variations de volume, si elles existent toujours au bout de neuf relaxations, restent toutefois minimes (0.4% et 0.7% respectivement pour la lysine n°41 (K41) et la cystéine n°12 (C12) entre la septième et la neuvième relaxation). Les effets de la relaxation se font beaucoup plus sentir lorsque l'on considère la structure dans son ensemble (colonne de droite

de la Figure 47). Il est facile de constater, en comparant les « bords » de la structure sans relaxation et de la structure au bout de neuf relaxations, que des cavités ou des fentes se sont progressivement formées ou approfondies ; le profil des cellules semble se dessiner de manière de plus en plus précise au fur et à mesure que les relaxations progressent (plus particulièrement à gauche).

Ceci peut également se constater sur la Figure 48. Cette figure présente les tessellations sur une structure (code PDB 1a05) dans le cas où il n'y a pas de relaxation (figure A) et au bout de neuf relaxations (figure B). On constate ici également que le profil de la structure s'affine et que des trous peuvent apparaître, par exemple dans la zone délimitée par un rectangle noir sur la figure A. Deux sphères de l'environnement qui sont relativement éloignées l'une de l'autre de part et d'autre de la structure au moment de l'installation de l'environnement, tendent à se rapprocher au fur et à mesure des relaxations, les cellules associées à ces deux points finissent par avoir une face en commun, c'est cette face qui apparaît comme un trou.



A **B**

Figure 48 : TdV sur les centres géométriques des chaînes latérales des résidus en non pondéré sur la structure de la 3-isopropylmalate déhydrogénase de *Thiobacillus ferroxidans* (code PDB : 1a05)[66]. A : Tessellation sans relaxation. B : Tessellation après 9 relaxations de l'environnement.

7 - Conception d'une application informatique de calcul et de visualisation : Voro3D

A mesure que mon travail de thèse avançait, il me semblait de plus en plus indispensable d'une part d'unifier et de rassembler en un tout cohérent les différents éléments qui permettaient de construire les TdV (installation de l'environnement, relaxation, construction des cellules) qui étaient jusque là des programmes différents écrits en Fortran et Mathematica et d'autre part de pouvoir « voir » ces tessellations afin de mieux comprendre ce qui se passait concrètement au sein des structures protéiques, pour pouvoir avancer sur les bases les plus solides possibles. Un de mes travaux de thèse a donc été de concevoir un logiciel permettant à partir de n'importe quelle structure protéique (c'est à dire à partir de n'importe quel fichier au format PDB) d'installer l'environnement, de le relaxer, de construire les cellules et de pouvoir déterminer leurs propriétés les plus utiles telles que le volume, la surface, le nombre de faces, etc… Ce logiciel conçu pour fonctionner sur PC se présente sous la forme d'une interface présentée Figure 49. Cette interface se décompose en quatre parties principales : en haut au centre, un tableau présente ligne par ligne les différentes propriétés propres aux cellules telles que le nom de l'AA, son numéro, son volume, sa surface totale, son nombre de faces. Les colonnes suivantes permettent de faire varier divers paramètres utiles pour la visualisation. La colonne violette permet, lorsque l'on clique sur une de ses cases, de visualiser les propriétés des faces composant la cellule. Ces données apparaissent dans le tableau situé en bas à gauche. Chaque ligne de ce tableau correspond à une face et donne : le nom de l'AA correspondant (la lettre « U » pour les sphères de l'environnement), la chaîne à laquelle appartient cet AA (l'esperluette « & » pour l'environnement), le nombre de côtés de la face, son aire, son périmètre et la distance entre les deux points représentatifs. La zone de visualisation se situe en bas à droite et permet d'opérer différentes manœuvres avec les structures. Il est possible de faire tourner la structure, de la déplacer ou encore d'effectuer un zoom. Avec les contrôles du tableau supérieur central, il est possible de choisir les cellules que l'on veut faire apparaître, de déterminer leur couleur, leur transparence, il est également possible de représenter la protéine en style « ball-&-sticks » pour tous les atomes ou seulement pour la trace (c'est à dire les carbones alpha) et également les noms (code à une lettre) et le numéro de chaque AA. Ce logiciel permet également d'effectuer les attributions

automatiques des structures secondaires et d'obtenir les matrices de contacts (voir le chapitre VI).

Ce programme présente à l'heure actuelle deux inconvénients majeurs : il ne fonctionne que sur PC et le système d'exploitation Windows, et il est assez lent. Cependant j'ai conçu ce logiciel de façon à ce qu'il soit possible de le faire évoluer facilement, ainsi les programmes permettant les différents calculs sont des programmes réalisés en C qui peuvent être aisément remplacés par des programmes beaucoup plus rapides (encore faut-il les écrire) sans se préoccuper du fonctionnement de l'interface qui a été conçue en Visual Basic. Le système de visualisation tridimensionnel quant à lui est celui mis à disposition par Cortona Parallel Graphics et fonctionne en VRML (Virtual Reality Modelling Language). Ce logiciel sera disponible sur la page web du laboratoire et fera l'objet d'une publication.

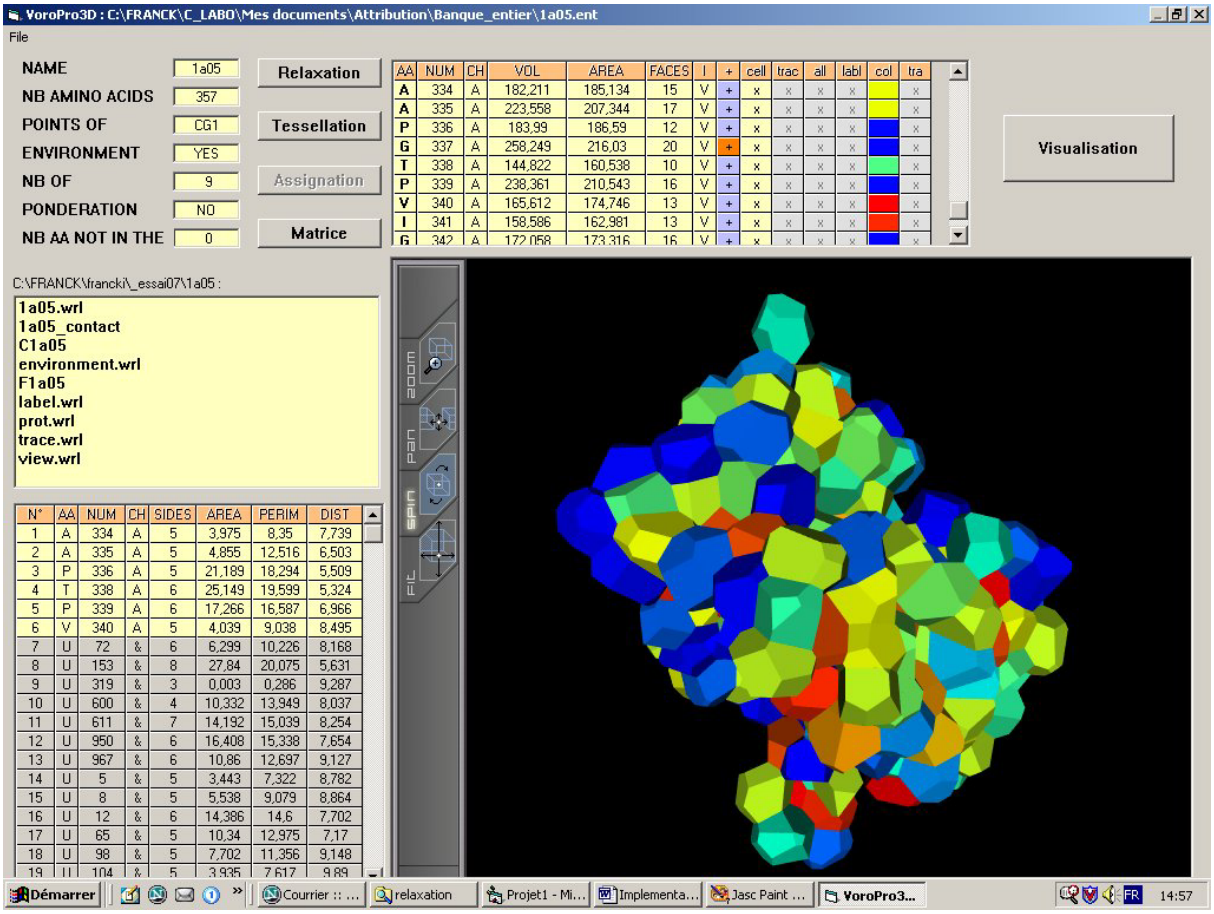


Figure 49 : Interface du logiciel de visualisation et de calculs Voro3D.

8 - Conclusion

L'utilisation d'un environnement présente l'avantage de pouvoir déterminer toutes les cellules, mais cet avantage a un coût en terme de temps de calcul. En effet, sans environnement, le nombre de points sur lesquels doivent s'effectuer les tessellations correspond au nombre d'AA des protéines. Avec l'environnement, ce nombre de points est considérablement augmenté (décuplé en moyenne) ce qui a pour fâcheux effet d'augmenter sensiblement le temps de calcul par tessellation, qui lui-même est à multiplier par le nombre de relaxations, conduisant au total à une procédure assez longue. A ceci s'ajoute bien entendu le temps de calcul nécessaire pour installer la protéine dans l'environnement. Sur un PC équipé d'un processeur cadencé à 1GHz, il faut compter pour une protéine de 150 AA de une à deux minutes !!

Chapitre 4

Les cellules

1 - Les différents points de tessellation

Lorsque l'on cherche à représenter un AA par un point, plusieurs possibilités sont envisageables. On peut utiliser un véritable atome du résidu ou un point virtuel comme un centre géométrique par exemple. Pour chaque unité peptidique, parmi toutes les possibilités offertes, j'ai retenu trois points à partir desquels une TdV était envisageable. La plus naturelle est bien sûr le carbone alpha (Cα) puisqu'il est toujours présent dans les fichiers de coordonnées atomiques même pour les plus basses résolutions et qu'il se trouve à une position essentielle des AA, à la fois sur le squelette et à la naissance de la chaîne latérale. Le deuxième point retenu est le centre géométrique des atomes de la chaîne latérale (sans le Cα) autres que l'hydrogène (ces derniers sont très rarement présents dans les fichiers). Ces points notés dans la suite CGL (Centre Géométrique de la chaîne Latérale) sont représentés par des sphères oranges dans la Figure 50 ; ce sont des points virtuels puisqu'ils ne correspondent à aucune entité physique, toutefois pour certains AA, il y concordance avec un véritable atome de la chaîne latérale, par exemple pour l'alanine (A). Pour la glycine (G) qui n'a pas de chaîne latérale, c'est le Cα qui est retenu. Cette représentation ne prend donc pas en compte les éléments du squelette, c'est pourquoi le troisième point retenu est le centre géométrique de tous les atomes de l'unité peptidique (encore une fois sans les atomes d'hydrogène) noté CG (Centre Géométrique). Ce point est représenté par des sphères en bleu clair sur la Figure 50.

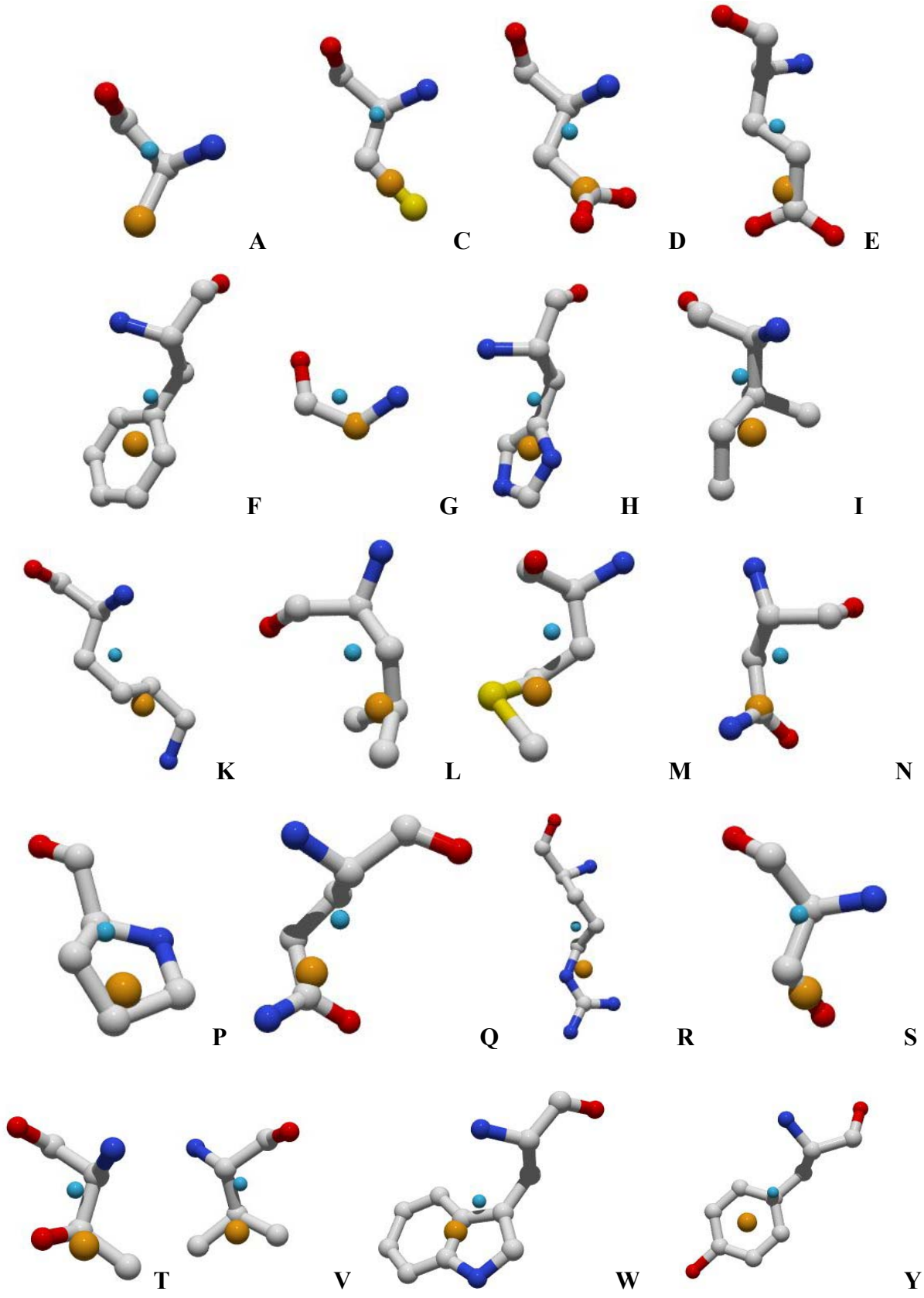


Figure 50 : Structures des 20 AA présents dans les protéines. Les atomes sont représentés avec le code de couleur suivant : blanc (C), rouge (O), bleu (N) et jaune (S). Les sphères en bleu clair représentent les CG et les sphères en orange les CGL.

2 - Banque et tessellation de Voronoï

Les différents résultats que je présente dans la suite de ce chapitre ont été obtenus à partir d'une banque de structures protéiques de bonnes résolutions (inférieures à 2.5 Å), sans AA absents par rapport à la séquence protéique correspondante issue de Swiss-Prot[67]. Pour éviter toute redondance structurale, chaque chaîne de cette banque contient des domaines issus de différentes super-familles selon la définition proposée dans SCOP[68]. Cette banque contient finalement 356 chaînes protéiques pour un total de 62 330 résidus répartis selon les effectifs présentés sur la Figure 52. Le nombre d'AA par protéine est en moyenne de 175.1, les protéines de taille moyenne (entre 100 et 200 AA) sont les plus représentées. La répartition des différents AA est présentée dans la Figure 51.

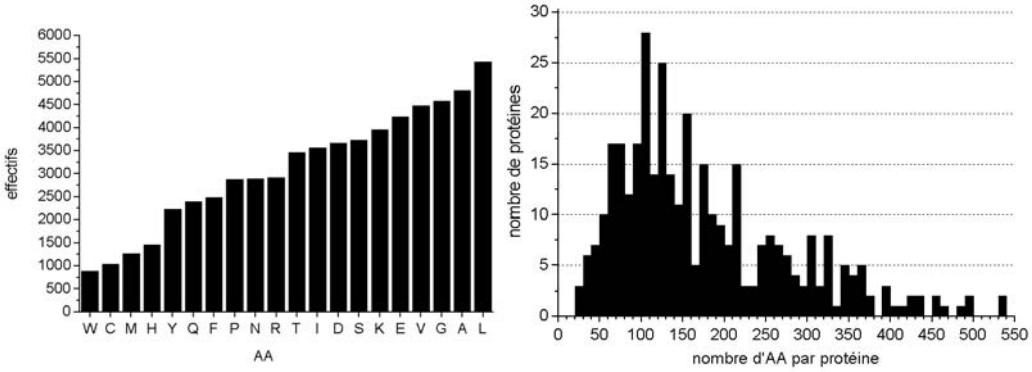


Figure 51 à gauche : répartition des différents AA dans la banque.
Figure 52 à droite : distribution du nombre d'AA par protéine.

codes PDB des 356 structures protéiques de la banque															
1a12	1aho	1b16	1bs9	1cku	1dd6	1e19	1eqo	1fup	1jpc	1otf	1qjb	1tyf	2cro	4mon
1a17	1ahs	1b33	1btn	1cl8	1dea	1e20	1ert	1fxd	1knb	1oun	1qje	1ugi	2ctc	4pah
1a1x	1ail	1b3a	1buo	1cmb	1dfu	1e2a	1esc	1fxo	1kp6	1oyc	1qk2	1uro	2e2c	4ubp
1a28	1air	1b3t	1bv1	1co6	1dif	1e6i	1eur	1fzd	1kpf	1p35	1qkj	1ute	2end	6ins
1a34	1aj8	1b66	1bvy	1coz	1dio	1e6t	1eyq	1g24	1kpt	1pbv	1qq8	1utg	2erl	6prc
1a44	1ako	1b67	1bx4	1cq3	1dj8	1e79	1eyv	1g31	1kuh	1pbw	1qqj	1vcc	2kau	7cei
1a4e	1amw	1b8o	1bx7	1cqd	1dk0	1ea5	1ez3	1g6g	1lau	1pdo	1qsd	1vhh	2mhr	
1a4m	1amx	1b93	1bxs	1cqy	1dk8	1eai	1f05	1g71	1lbe	1pfk	1qtn	1vhr	2nsy	
1a4y	1ann	1bb9	1bxy	1cru	1dlm	1eay	1f2k	1gak	1lki	1php	1qtw	1vid	2pii	
1a58	1aoe	1bbp	1byr	1ctf	1dlw	1ecm	1f3u	1gci	1luc	1phr	1rav	1vmo	2poo	
1a6j	1aoh	1bd8	1bzy	1cxy	1dm9	1ecy	1f3v	1gdo	1mat	1pml	1reg	1wap	2pth	
1a6o	1apa	1beo	1c08	1d0b	1doz	1ed1	1f3z	1gen	1mjc	1poc	1ris	1wba	2rn2	
1a99	1apx	1bhd	1c1k	1d0i	1dpj	1edm	1f41	1got	1mka	1poh	1rop	1wgj	2sic	
1aa7	1ars	1bhp	1c26	1d0q	1dqe	1edy	1f5m	1gym	1mml	1ppf	1rpx	1whi	2stb	
1aac	1aug	1bj1	1c2t	1d2z	1dqi	1efv	1f60	1h2r	1mnm	1puc	1rvv	1who	2tnf	
1aap	1aun	1bk5	1c5e	1d3v	1dqo	1egw	1f8y	1hcb	1moq	1pud	1sei	1xxa	2wrp	
1abe	1avb	1bkp	1c76	1d4o	1dsz	1ei7	1fas	1hfe	1mro	1pyt	1sfp	1ycq	3cla	
1ad1	1aw8	1bm8	1cc8	1d4t	1dtd	1ej1	1fd3	1hle	1msk	1qau	1srv	256b	3daa	
1ad6	1ay7	1bm9	1cfy	1d6r	1dun	1ejf	1fkj	1hoe	1mwp	1qb0	1tbg	2a0b	3eip	
1aew	1ayf	1bou	1cfz	1d7d	1dv8	1ekg	1fle	1htp	1nba	1qb2	1tcd	2abk	3fap	
1ag9	1ayx	1bov	1chd	1d8d	1dwv	1ekj	1flm	1hyp	1ndd	1qc7	1tfe	2acy	3pyp	
1agi	1azp	1bpl	1ciq	1dan	1dxe	1el6	1fof	1icf	1nec	1qd9	1tif	2ahj	3tdt	
1agj	1b00	1br9	1cjd	1dce	1dxj	1em9	1fqt	1imb	1noa	1qex	1tig	2arc	4aah	
1ah4	1b0n	1brf	1cjw	1dcp	1dy7	1emv	1fs1	1jac	1npk	1qf9	1toa	2asr	4fgf	
1ah7	1b0x	1bs4	1ck4	1dd3	1dzr	1eq6	1fua	1jdw	1opc	1qip	1tup	2bbk	4icb	

Tableau 4 : Liste des codes PDB de la banque.

3 - Volume des cellules : influence des points et de la pondération

3.1 Tessellation de Voronoï non pondérée

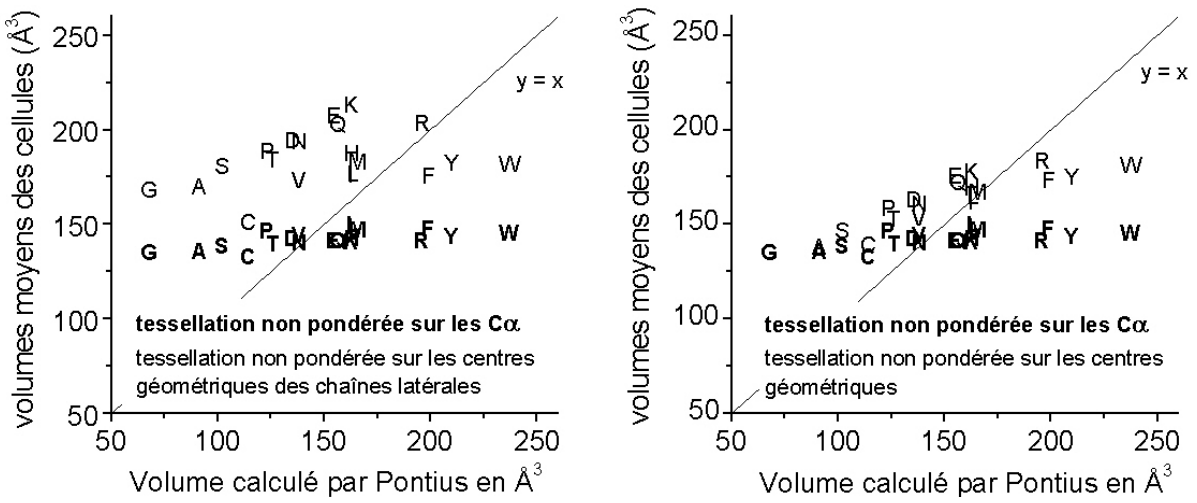


Figure 53 : Influence des points de TdV sur le volume des cellules
A gauche : Cα et CGL
A droite : Cα et CG

La Figure 53 représente les volumes moyens des cellules pour chaque type d'AA en fonction des volumes calculés atome par atome par Pontius[22], pour une TdV non pondérée. Le cas des tessellations sur les Cα est représenté sur les deux graphes et on constate que le volume des cellules ne varie pas beaucoup en fonction des AA. Pour ces cellules, la moyenne est de 142.6 Å^3 avec un minimum de 132.9 Å^3 pour la cystéine (C) et un maximum de 150.4 Å^3 pour la leucine (L). Si l'on compare l'écart entre ces deux extrêmes (7.8 Å^3) et l'écart entre les deux extrêmes des volumes de Pontius (169.7 Å^3 entre le tryptophane (W) et la glycine (G)) on constate bien que ces cellules ne sont pas très représentatives des volumes réels occupés par les résidus. Ceci s'explique simplement par la méthode de construction des cellules, en effet les faces qui définissent les polyèdres sont incluses dans les plans médians entre les points sur lesquels s'effectue la TdV. Dans le cas présent, les faces sont donc à mi-chemin entre Cα voisins quelle que soit la taille des AA en présence. Avec cette méthode, un résidu de taille importante peut donc être représenté par une cellule de volume moindre que son volume réel. C'est l'inverse qui se produit pour les petits AA dont les cellules seront artificiellement dilatées. Ceci se retrouve, sur les graphes sur lesquels j'ai représenté la droite

d'équation y = x afin de mieux apprécier cette propriété. Entre 140 et 150 $Å^3$ existe une limite sous laquelle les moyennes se situent au-dessus de cette droite ; pour ces résidus, les volumes moyens des cellules sont supérieurs aux volumes réels des AA. Au-dessus de cette limite, les moyennes se situent sous la droite, les volumes moyens des cellules sont donc inférieurs aux volumes des AA. Il y a ainsi neuf AA dont le volume moyen est plus grand que celui de leur cellules (G, A, S, C, P, T, D, N et V) et onze AA dont le volume moyen est plus petit (E, Q, K, H, M, L, I, F, R, Y et W). En fait, avec une TdV non pondérée sur les Cα, du point de vue du volume, tout se passe à peu près comme si tous les AA étaient de même nature. Sur le graphe de gauche de la Figure 53 sont également représentées les moyennes des volumes pour les TdV non pondérées sur les CGL. Ces moyennes sont supérieures à celles obtenues avec les Cα (moyenne générale de 185.4 $Å^3$) ce qui implique nécessairement qu'avec ce type de représentation la protéine a un volume plus important ! Ceci s'explique par les positions des centres des sphères de l'environnement. En effet, ces positions varient avec les points sur lesquels on décide d'effectuer la TdV. Dans le cas présent, pour éviter une superposition avec les CGL, les sphères de l'environnement sont éloignées des protéines par rapport aux positions qu'elles auraient occupées si l'on avait conservé les Cα. Les protéines se retrouvent ainsi dilatées ; plus exactement, c'est lors de la TdV sur les Cα que l'on rapetisse artificiellement les molécules. Le volume moyen des protéines avec une TdV sur les Cα est de 24 961 $Å^3$ alors qu'il est de 32 465 $Å^3$ avec les CGL, soit une différence proche de 23% ! Ceci se manifeste donc pour le volume moyen des cellules par une augmentation qui va bien sûr dépendre de la position des résidus dans les protéines. Plus un AA sera exposé à l'environnement plus le volume de sa cellule aura tendance à augmenter (toujours si on le compare à ce qu'il serait avec les Cα). Ceci explique pourquoi cette augmentation n'est pas uniforme. Avec les AA hydrophobes[69, 70], elle est à peu près constante (V, L, I, F, M, Y et W) mais avec les non hydrophobes, elle est plus dépendante du volume (S, P, T, D, N, E, Q, K, H et R). La cystéine (C) reste un cas particulier puisque l'augmentation qui lui est associée est plus faible que pour les autres AA. Il est intéressant de remarquer que les résidus dont le volume est supérieur au volume moyen de leurs cellules sont les trois AA hydrophobes contenant des cycles W, F, et Y.

Sur le graphe de droite de la Figure 53 sont aussi représentés les volumes moyens des cellules obtenues avec une TdV non pondérée sur les CG. La moyenne générale est de 160.8 $Å^3$ avec un maximum pour l'arginine (R) à 183.9 $Å^3$ et un minimum pour la glycine (G) de 135.4 $Å^3$ soit un écart de 48.5 $Å^3$, légèrement supérieur à celui obtenu avec les CGL (45 $Å^3$

entre la lysine (K) et la glycine (G)). Comme l'on pouvait s'y attendre, les CG sont un intermédiaire entre les Cα et les CGL. Par exemple, le volume moyen pour la glycine (G) est d'à peu près 135 $Å^3$ pour les CG. Pour les petits AA, la TdV sur les CG se comporte comme pour les Cα, ce qui est logique puisque pour ces résidus, le CG est très proche du Cα (Figure 50). Pour le tryptophane (W), le volume moyen est d'à peu près 181 $Å^3$ pour les CG et les CGL, ce qui s'explique par le fait que pour les plus gros résidus, ces deux points se rapprochent (Figure 50). L'augmentation par rapport aux volumes obtenus avec les Cα est proportionnelle au volume réel des AA, et la distinction entre hydrophiles et hydrophobes est moins marquée, même si elle existe toujours puisque l'arginine (R), la lysine (K) ou l'acide glutamique (E) qui sont hydrophiles sont au-dessus des hydrophobes. Ce plus faible écart entre les deux catégories d'AA s'explique encore par le fait que pour ces résidus aussi, le CG est moins éloigné du squelette que ne l'est le CGL, surtout pour l'arginine (R) et la lysine (K) qui ont des chaînes latérales relativement longues.

On voit donc déjà que le choix des points sur lesquels on effectue la TdV non pondérée a une importance sur les résultats que l'on obtient. Avec les Cα, l'influence de la taille des AA est très limitée, ceci n'est plus vrai avec les CG avec lesquels cette influence se fait nettement sentir, même si les volumes moyens des cellules ne sont pas toujours représentatifs des volumes réels des résidus. Enfin, les CGL permettent de rendre compte à la fois du volume mais également de certaines propriétés physico-chimiques.

3.2 Tessellation de Voronoï pondérée

3.2.1 Les poids de la pondération

Les poids que j'ai utilisés ont été déterminés par Sadoc et coll[71] par une méthode itérative permettant d'obtenir des cellules dont les volumes moyens sont proportionnels à ceux calculés par Pontius. Cette méthode appliquée à plusieurs protéines permet d'obtenir la relation suivante : $w = -14.325 + 0.5282 \times v_P^{2/3}$. Cette expression, dans laquelle w représente le poids et v_P les volumes des AA d'après Pontius, permet de déterminer les poids de chaque AA. Le Tableau 5 donne le poids de chacun des AA, celui de l'environnement est déterminé de la même façon à partir du volume indiqué qui est celui d'une sphère de 6.5 Å de diamètre ; il est proche du volume moyen des AA dans la banque (142.3 Å). Il faut rappeler ici que dans l'absolu, ces valeurs n'ont pas de réelle importance puisque les poids peuvent être définis à

une constante additive près. Pour les TdV pondérées, seules comptent les différences entre ces poids.

	A	C	D	E	F	G	H	I	K	L	M
Pontius	91.5	114.4	135.2	154.6	198.8	67.5	163.2	162.6	162.5	163.4	165.9
Poids	-3.600	-1.877	-0.411	0.890	3.667	-5.568	1.449	1.410	1.404	1.462	1.623
	N	**P**	**Q**	**R**	**S**	**T**	**V**	**W**	**Y**	**Environnement**	
Pontius	138.3	123.4	156.4	196.1	102	126	138.4	237.2	209.8	143.8	
Poids	-0.199	-1.233	1.008	3.504	-2.794	-1.050	-0.192	5.915	4.325	0.173	

Tableau 5 : Poids utilisés et calculés à partir des volumes de Pontius.

3.2.2 Influence des points de tessellation

La méthode employée par Sadoc et coll pour déterminer les poids utilisait les TdV sur les CGL. Le graphe B de la Figure 54 montre bien que les volumes calculés sont bien en moyenne proportionnels à ceux de Pontius (pente de 1.1 et coefficient de corrélation de 0.96). On remarque encore une fois qu'une sélection est faite entre les résidus hydrophobes (V, I, L, F, M, Y et W) et les autres notamment K et R.

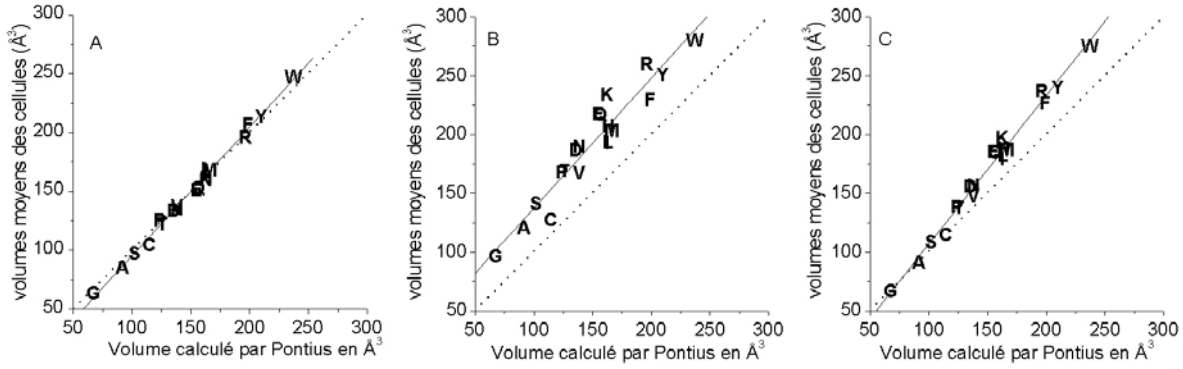


Figure 54 : Influence des points de tessellation. A : Cα, B : CGL, C : CG.
En pointillé est représentée la droite d'équation y = x.

Si les volumes calculés sont linéairement proportionnels à ceux de Pontius, ils ne leur sont pas égaux puisqu'en moyenne il existe un écart de plus de 40 $Å^3$. Cet écart n'existe plus lorsque l'on effectue la TdV sur les Cα (Figure 54, graphe A) puisque la pente est de 1.09 avec un coefficient de corrélation de 0.99. Comme dans le cas des TdV non pondérées on constate que les CG constituent un intermédiaire entre les Cα et les CGL, puisque là encore les petits résidus se comportent comme avec les Cα alors que les AA de volumes importants se comportent comme avec les CGL.

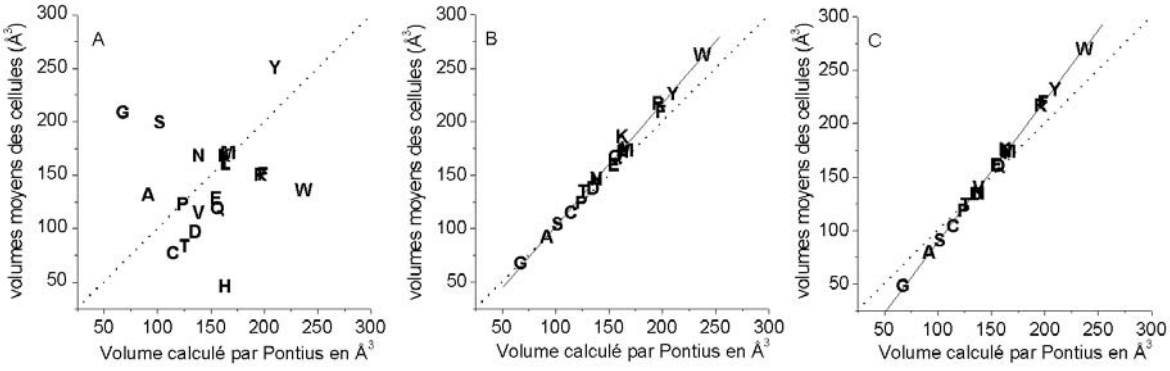


Figure 55 : Influence des points de tessellation. A : Cα, B : CGL, C : CG.
En pointillé est représentée la droite d'équation y = x. Les volumes indiqués sont ceux des cellules ne faisant pas de contacts avec les cellules de l'environnement.

La Figure 55 reprend les mêmes graphes que la Figure 54 mais uniquement pour les cellules ne faisant aucun contact avec les cellules associées à des points de l'environnement, ce que j'appellerai dans la suite les cellules en volume en opposition aux cellules en surface qui, elles, ont des faces communes avec les cellules associées aux points de l'environnement. On constate immédiatement que pour les graphes B et C, les volumes sont plus petits et qu'ils correspondent aux volumes de Pontius avec des pentes très proches de celles de la Figure 54. Ceci montre donc que l'augmentation de volume que l'on constatait précédemment est bien due aux cellules des résidus situés à la surface des protéines. Pour les Cα (graphe A), il n'y a plus de corrélation avec les volumes de Pontius, ce qui peut s'expliquer par la faible proportion d'AA en volume dans ce cas (proche de 10 % de tous les AA). La Figure 56 illustre pour une même structure l'influence du mode de tessellation et des points de représentation sur la forme des cellules. Le Tableau 6 donne le volume et le nombre de faces de chacune de ces cellules.

	G 9	W 78	R 86	V 161
CA np	114.5 15	142.8 15	148.3 17	141.4 12
CA p	65.8 14	231.0 15	194.8 16	120.6 14
CGL np	114.1 15	168.8 16	268.8 20	268.6 18
CGL p	73.7 12	286.9 18	338.1 21	282.1 19
CG np	121.3 16	194.8 15	170.9 14	179.4 15
CG p	70.8 13	308.7 15	226.7 17	170.7 18

Tableau 6 : Volume en Å³ et nombre de faces des cellules de la Figure 56.

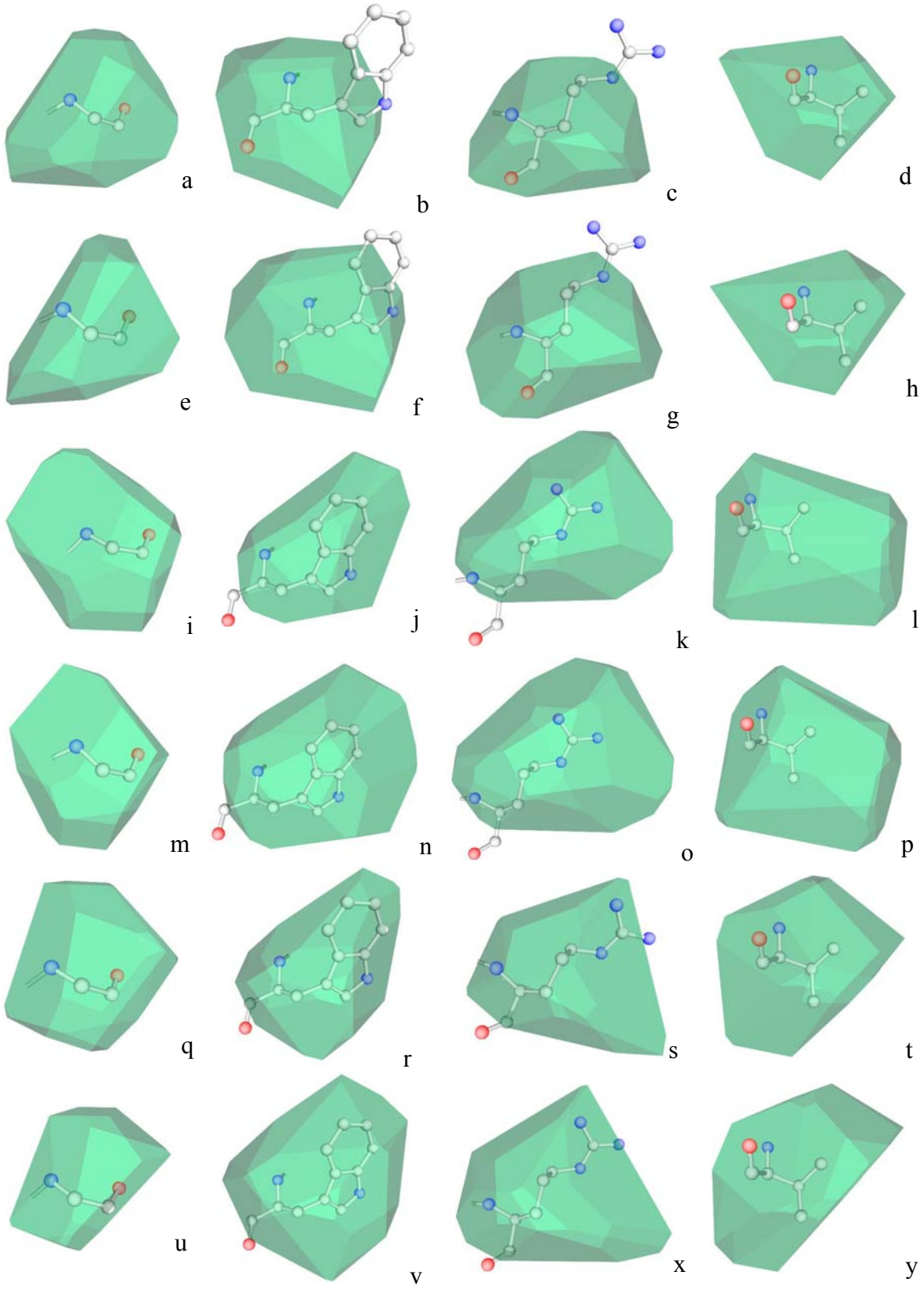


Figure 56 : De gauche à droite, cellules de la glycine n°9 (G9), du tryptophane n°78 (W78), de l'arginine n°86 (R86) et de la valine n°161 (V161) de la structure de code PDB 1a05 pour des TdV sur (de haut en bas) Cα non p, Cα p, CGL non p, CGL p, CG non p, CG p (p pour pondéré).

Une des premières observations que l'on peut faire est que les cellules des plus gros résidus ne contiennent pas tous les atomes, même pour des cellules dont la taille est très importante comme la cellule o (338.1 $Å^3$). Ceci est d'ailleurs vérifié également pour des résidus de tailles moyennes comme la valine avec les cellules h, p, y qui correspondent à des TdV pondérées. En fonction des points de représentation choisis et de la forme des résidus, les atomes situés à l'extérieur des cellules seront différents ; par exemple avec le Cα ce sont les atomes d'azote (en bleu) de l'arginine qui sortent des cellules c et g, avec les CG ce sont plutôt les atomes du squelette qui sortent et plus particulièrement l'atome d'oxygène (en rouge) mais également parfois l'atome de carbone (en blanc) comme avec les cellules h, j, k, n et o. Concernant la forme des cellules, il semble que le point de représentation soit plus important que le mode de tessellation (pondération ou non pondération). Par exemple, pour la valine ou l'arginine, on constate que les formes des cellules sont voisines lorsque l'on passe d'une tessellation non pondérée à une tessellation pondérée (de d à h ou de l à p ou encore de t à y) mais qu'elles se différencient nettement lorsque l'on change de point de tessellation (de d à l ou de h à p). La glycine est un cas intéressant puisque pour ce résidu, le CGL et Cα sont confondus ; il est curieux de voir qu'en non pondéré pour ces deux points de représentation, les nombres de faces sont les mêmes et les volumes quasiment identiques alors que les cellules sont de formes assez dissemblables ; en pondéré les valeurs et les forment changent. Pour le tryptophane, les formes des cellules j, r et n, v sont assez proches malgré des points de représentation différents, ceci s'explique par la forme du résidu (le plus gros de tous ceux présents dans les protéines). En effet, pour cet AA la chaîne latérale est très volumineuse et les atomes la composant nombreux, ceci implique donc que le CG et le CGL sont proches (Figure 50), les différences observées sont principalement dues au voisinage.

4 - Nombre de faces par cellule

Nombre de côtés par face

4.1 Nombre de faces par cellule

4.1.1 Notion de voisin ou de contact

Les cellules de Voronoï sont des polyèdres définis par un certain nombre de faces. Par construction chacune de ces faces est commune à deux cellules jointives, les points associés à

ces deux cellules sont ainsi unis par une relation géométrique forte, qui est celle de plus proche voisinage. Pour décrire le fait que deux cellules ont une face en commun j'écrirai dans la suite que les deux points associés sont voisins ou bien qu'ils sont en contact. Je l'écrirai également pour les AA associés à ces points (ou pour un AA et un point de l'environnement).

4.1.2 Moyennes et distributions

Le nombre de faces d'une cellule correspond au nombre de ses plus proches voisins, le nombre moyen de faces par cellule est donc une quantité particulièrement intéressante pour connaître les relations de plus proche voisinage qui existent entre les AA. La Figure 57 représente les distributions du nombre de faces par cellule en fonction des différentes TdV utilisées, les moyennes correspondantes sont indiquées dans le Tableau 7. On constate tout d'abord que les maximums des histogrammes se situent à 14 et 15 faces par cellule, avec des moyennes situées entre ces deux valeurs. Ces résultats sont différents de ceux trouvés dans les études précédentes[63, 72] et qui portent sur des TdV non pondérées effectuées sur les CGL. Dans le premier cas (moyenne de 13.97) cela s'explique par le fait que cette étude n'intégrait pas d'environnement, certaines cellules pouvaient alors présenter des formes très allongées comportant relativement peu de faces. Dans le second cas (moyenne de 14.27), l'explication ne peut provenir que de la composition des banques utilisées puisque mes programmes ont été testés afin de fournir les mêmes résultats que les programmes utilisés pour ces études. La banque d'Angelov et coll est composée de 39 protéines dont certaines (1bdm, 1fds, 1phm) contiennent des trous (tous les AA de la séquence ne sont pas présents dans la structure), d'autres protéines sont composées de plusieurs chaînes (1gse, 1plf …) et il est possible que des cavités trop étroites pour accueillir des sphères de l'environnement soient ainsi créées et modifient les résultats. Ces diverses particularités peuvent expliquer les différences observées.

	Calpha	CGL	CG
non pondéré	14.46	14.58	14.34
pondéré	14.34	14.59	14.38

Tableau 7 : Moyenne du nombre de faces par cellule.

Les variations entre les TdV pondérées et non pondérées sont assez faibles, toutefois pour les Cα, le nombre moyen de faces diminue légèrement avec la pondération. On constate également que les distributions sont légèrement plus larges dans le cas pondéré. Les différences entre les valeurs observées sont difficiles à expliquer, il a été montré par des

travaux théoriques que pour des empilements de sphères de différentes tailles, des variations importantes des tailles des cellules entraînaient une baisse du nombre moyen de faces ; à l'opposé, une irrégularité croissante dans la forme des cellules augmentai cette moyenne. Nous avons vu également que le choix du point représentant l'AA influençait la proportion d'AA en volume, or comme je le montrerai dans la suite le nombre de faces moyen varie avec la position de la cellule au sein de la structure protéique (opposition entre les cellules en surface et celles en volume). Les influences respectives de ces divers paramètres sur les valeurs observées restent difficiles à déterminer.

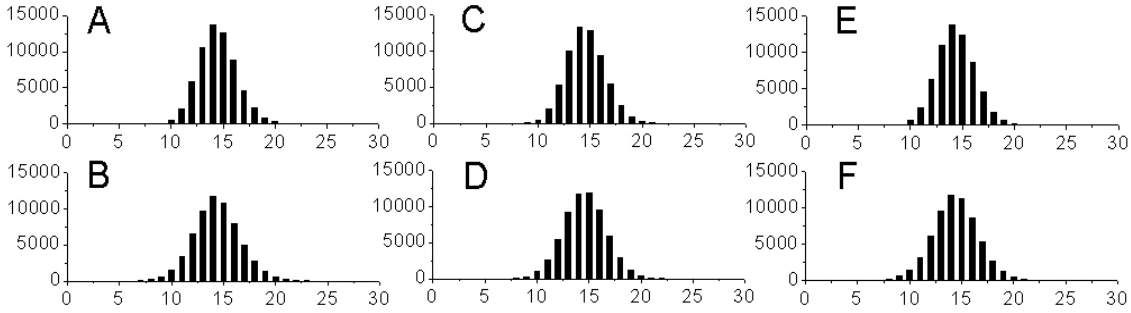


Figure 57 : Distributions du nombre de faces par cellule : A - Cα non pondéré, B - Cα pondéré, C - CGL non pondéré, D – CGL pondéré, E - CG non pondéré, F – CG pondéré, en abscisse est représenté le nombre de faces par cellule, en ordonnée le nombre de cellules correspondant.

4.1.3 Nombre de faces et volumes

On constate que dans le cas pondéré (seconde ligne de la Figure 57) les distributions sont légèrement plus étendues que dans le cas non pondéré, ceci semble logique puisque nous savons que dans le cas pondéré, le volume des cellules est en moyenne plus représentatif du volume réel occupé par les AA que dans le cas non pondéré. Il est naturel de penser que des cellules plus grosses/petites auront plus/moins de faces. Ceci est confirmé par la Figure 58 qui représente les relations entre le nombre de faces moyen et le volume moyen des cellules pour chaque type d'AA.

On constate que le nombre moyen de faces par cellule et le volume des cellules sont en étroite relation sauf pour la TdV non pondérée sur les Cα pour laquelle on ne constate pas de corrélation. Dans les autres cas, on a bien confirmation que les petites cellules ont un nombre de faces moyen plus faible, les AA associés à ces cellules ont donc un voisinage moins peuplé alors que les plus grosses cellules ont plus de faces et donc plus de voisins.

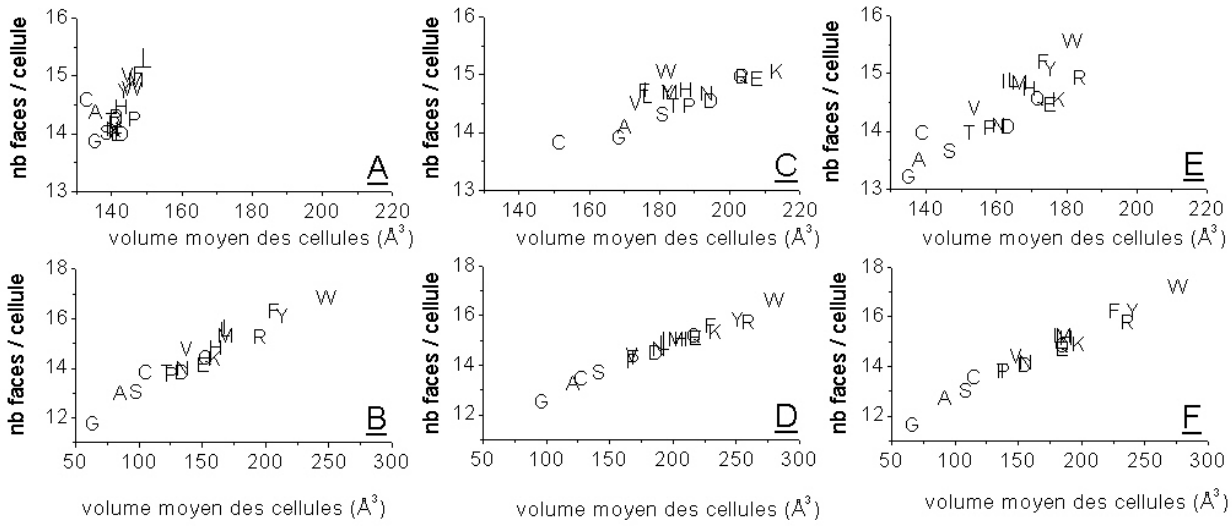


Figure 58 : Relation entre le nombre moyen de faces et le volume moyen des cellules. A - Cα non pondéré, B - Cα pondéré, C - CGL non pondéré, D – CGL pondéré, E - CG non pondéré, F – CG pondéré. Les échelles sont différentes entre les non pondérés (première ligne) et les pondérés (seconde ligne).

On observe également pour les TdV non pondérées que les résidus hydrophobes (V, I, L, F, M, Y, W voire C) ont à volume moyen égal des cellules qui comportent plus de faces que les cellules des autres résidus. Considérons par exemple le graphe C de cette figure qui est représentatif des TdV non pondérées sur les CGL. Si l'on différencie les cellules en volume et celles en surface, on obtient alors les résultats présentés sur la Figure 59. On constate immédiatement que les cellules en volume sont plus petites que celles en surface (c'est un point sur lequel je reviendrai plus en détail par la suite), on constate également que le nombre moyen de faces par cellule varie entre le volume et la surface en fonction des résidus. Le nombre de faces par cellule des plus gros résidus diminue lorsque l'on passe du volume à la surface, alors qu'il augmente pour les petits AA. Enfin, on constate que les résidus hydrophobes se démarquent pour les AA en surface mais pas pour ceux en volume. L'explication vient du fait que cette dichotomie volume/surface est relativement simpliste et qu'elle masque le fait qu'un résidu en surface est plus ou moins exposé au solvant. Parmi les résidus en surface, ceux qui sont chargés ou polaires auront en moyenne tendance à être plus proches de la surface alors que les hydrophobes auront tendance à être plus enfouis. Or les résidus les plus gros ont en moyenne plus de faces lorsqu'ils sont complètement enfouis (c'est à dire en volume), les résidus hydrophobes les plus gros auront donc même en surface des cellules ayant un nombre de faces plus élevé. Cet exemple montre que la compréhension des résultats observés nécessite la prise en compte de divers paramètres souvent corrélés mais

dont l'interdépendance est difficile à appréhender. Par exemple dans ce cas, il est curieux de constater que tous les volumes moyens diminuent lorsque les AA passent de la surface au volume, pourtant le nombre de faces par cellule, lui, ne diminue pas obligatoirement. De plus pour les CGL, en non pondéré, le nombre moyen de faces par cellule est supérieur en surface à celui observé en volume. Pour les Cα le phénomène inverse se produit.

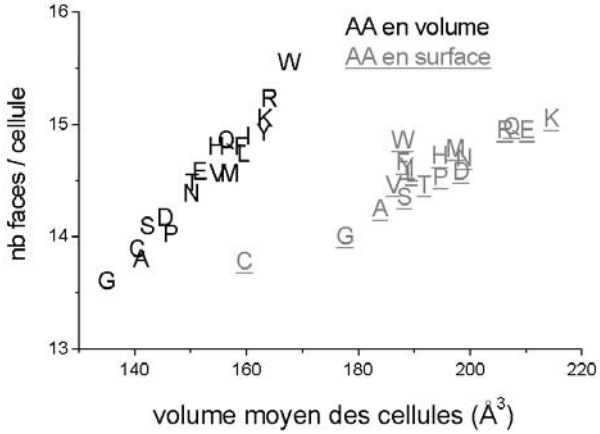


Figure 59 : Relation entre le nombre moyen de faces et le volume moyen des cellules.
TdV non pondérée sur les CGL.

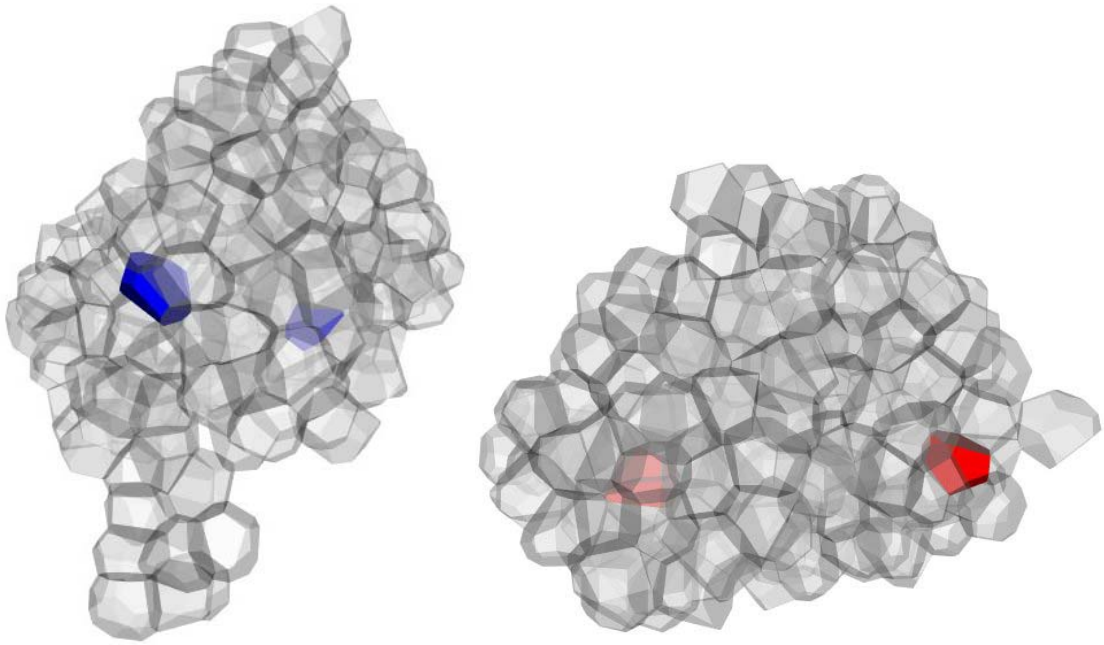


Figure 60 : TdV non pondérée sur les CGL de la structure de code PDB 1a05.
A gauche en bleu, G44 en volume et G97 en surface.
A droite en rouge, W200 en volume et W312 en surface.

La Figure 60 illustre une des propriétés précédemment observées avec des cellules associées au plus petit des résidus (la glycine en bleu) et le plus gros (le tryptophane en rouge). Tout d'abord à gauche, deux cellules associées à des glycines sont mises en valeur, la cellule bleue la plus à gauche est celle de la glycine n°97 (G97), elle se situe en surface car certaines de ces faces sont des faces de contact avec d'autres cellules liées à l'environnement. Son volume est de 238.9 $Å^3$ ce qui est énorme pour une cellule associée à une glycine et son nombre de faces est de 18. La cellule bleue la plus à droite quant à elle est une cellule en volume (il n'y a pas de faces de contact avec des cellules de l'environnement) et son volume est de 72.8 $Å^3$ pour 8 faces. Pour la figure de droite, deux cellules liées à des tryptophanes sont représentées en rouge. La cellule la plus à droite est une cellule en surface, son volume est de 162.8 $Å^3$ pour 14 faces dont 3 avec des cellules de l'environnement. La cellule rouge la plus à gauche a un volume de 170.4 $Å^3$ pour 19 faces. Pour ces deux exemples, on voit bien que le graphe de la Figure 59 est confirmé. Avec ce mode de tessellation, les plus gros résidus ont plus de voisins en volume qu'en surface, alors que c'est l'inverse pour les plus petits.

4.2 Nombre de côtés par face

4.2.1 Moyennes et distributions

Le nombre de côtés par face est également une donnée importante puisqu'il est relié à la symétrie locale autour des liens mettant en relation les plus proches voisins. Le Tableau 8 donne les moyennes pour chaque type de TdV. Des valeurs proches de 5 (typiquement entre 5.1 et 5.2) sont généralement associées à des structures denses et compactes que l'on rencontre régulièrement en physique de la matière condensée[63, 72]. Un nombre de côtés par face proche de 5 est également caractéristique de structures compactes établies avec des règles de construction locales, mais des frustrations géométriques interdisent à cette symétrie locale de s'étendre dans l'ensemble de la structure 3D[73]. On constate que ces valeurs ne sont pas très sensibles au type de TdV utilisé. La Figure 61 confirme cela puisque les histogrammes ne varient quasiment pas, on peut constater toutefois que le nombre de faces triangulaires est sensiblement plus important lorsque la TdV s'effectue sur les Cα. Pour tous les types de TdV on retrouve un maximum pour les faces à 5 côtés avec une distribution non symétrique puisque l'on ne peut pas trouver de faces avec moins de 3 côtés.

	Calpha	CGL	CG
non pondéré	5.17	5.177	5.163
pondéré	5.163	5.177	5.165

Tableau 8 : Nombre moyen de côtés par face.

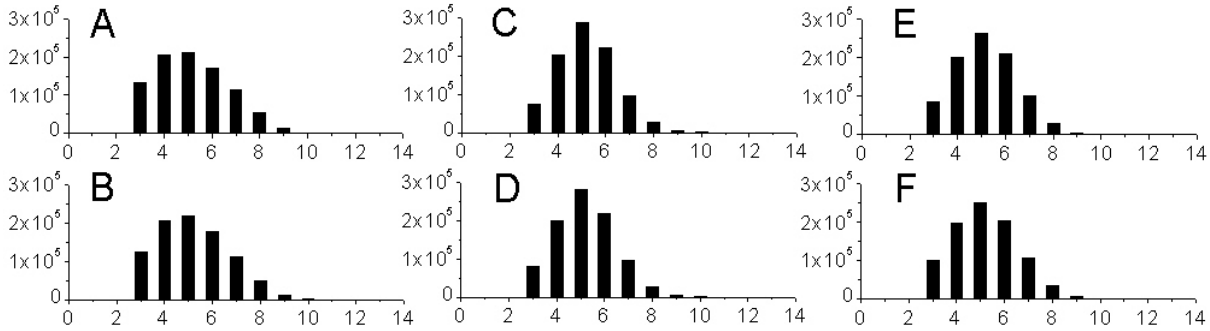


Figure 61 : Distributions du nombre de côtés par face : A - Cα non pondéré, B - Cα pondéré, C - CGL non pondéré, D – CGL pondéré, E - CG non pondéré, F – CG pondéré, en abscisse est représenté le nombre de côtés par face, en ordonnée le nombre de faces correspondant.

Le nombre moyen de faces par cellule (F) et le nombre moyen de côtés par face (C) obéissent à une relation déduite de la relation d'Euler :

$$F = 12 / (6 - C)$$

Le Tableau 9 présente les valeurs du nombre de faces par cellule déduites du nombre moyen de côtés par face. On retrouve les valeurs du Tableau 7, ce qui montre que la relation est bien vérifiée. On constate également que de faibles variations du nombre moyen de côtés par face provoquent des changements plus importants du nombre moyen de faces.

	Calpha	CGL	CG
non pondéré	14.46	14.58	14.34
pondéré	14.34	14.58	14.37

Tableau 9 : Valeurs de F déduites à partir des valeurs du Tableau 8.

4.2.2 Nombre de côtés par face, aire des faces et distance entre voisins

La Figure 62 présente les relations entre le nombre de côtés par face et l'aire des faces de contact entre les AA. Dans le cas des TdV pondérées (seconde ligne), la relation entre le nombre moyen de côtés par face et l'aire moyenne des faces est plus ou moins linéaire et logique : plus les faces sont grandes plus elles ont de côtés. Pour les TdV non pondérées (première ligne), il n'y a plus de relation nette, pour les Cα la relation décrite plus haut semble

inversée mais l'échelle étant beaucoup plus petite, les différences observées peuvent ne pas être significatives. On constate toutefois un comportement constant pour le groupe des AA hydrophobes (V, I, L, F, M, Y et W) pour lesquels à aires égales, on a un nombre moyen de côtés par face plus important, surtout pour les TdV pondérées.

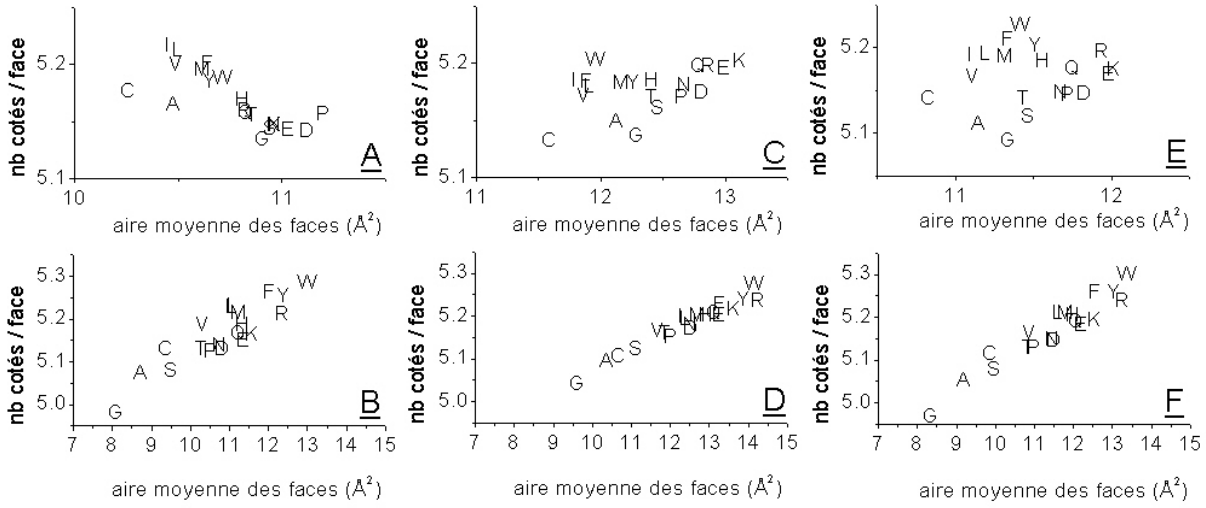


Figure 62 : Relation entre le nombre moyen de côtés par face et l'aire moyenne des faces. A - Cα non pondéré, B - Cα pondéré, C - CGL non pondéré, D – CGL pondéré, E - CG non pondéré, F – CG pondéré.

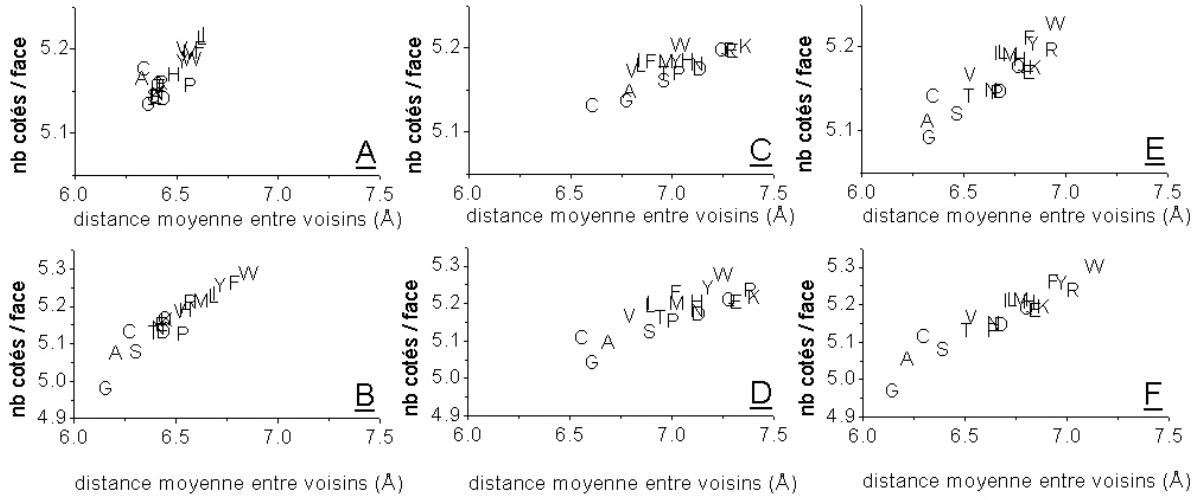


Figure 63 : Relation entre le nombre moyen de côtés par face et les distances moyennes entre voisins. A - Cα non pondéré, B - Cα pondéré, C - CGL non pondéré, D – CGL pondéré, E - CG non pondéré, F – CG pondéré. Les échelles des figures B, D et F d'une part et A, C et E d'autre part sont les mêmes.

La Figure 63 présente les relations entre nombre moyen de côtés par face et distance moyenne entre points voisins. Comme l'on pouvait s'y attendre, on constate bien que plus deux voisins sont proches l'un de l'autre, plus la face associée à ce voisinage aura de côtés.

On remarque encore une fois que les AA hydrophobes ont un comportement un peu particulier puisque pour des distances équivalentes, ils semblent faire des faces comportant plus de côtés. Contrairement au nombre de côtés qui est un paramètre relativement peu sensible au mode de TdV, le critère de distance semble beaucoup plus variable.

4.2.3 Interprétation du nombre de côtés par face

Comme nous venons de le voir le nombre de côtés par face dépend de la distance entre voisins et de l'aire des faces de contact ; si on utilise la valeur la plus représentée (5 côtés par face) comme référence, on peut supposer que des contacts avec des faces triangulaires ou quadrangulaires sont représentatives de contacts faibles alors que les faces ayant plus de 5 côtés traduisent des contacts plus marqués.

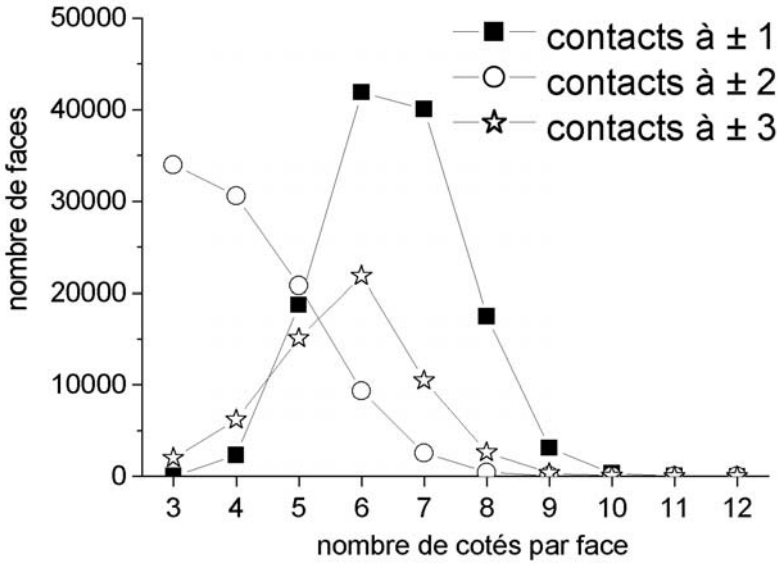


Figure 64 : Distributions du nombre de côtés par face selon l'écart en séquence, pour une TdV pondérée effectuée sur les CG.

La Figure 64 présente les distributions du nombre de côtés par face pour une TdV pondérée sur les CG. Sur ce graphe, chaque courbe représente une distribution pour un écart en séquence particulier (de ±1 AA à ±3 AA). Les courbes correspondant à un écart de ±1 AA et ±2 AA sont très différentes l'une de l'autre, ce qui montre bien l'influence des relations des résidus entre eux. En effet, les contacts à ±1 AA sont des contacts entre AA liés de manière covalente et le maximum de la courbe se situe à 6 côtés par face suivi de près par les faces à 7 côtés (moyenne de 6.5), alors que le maximum pour toutes les faces se situent à 5 côtés (moyenne de 5.2). A l'opposé, les contacts à ±2 AA sont des contacts entre AA non liés de manière covalente et aucune structure secondaire régulière ne favorise des liaisons

hydrogènes entre ces résidus. Le maximum pour cette courbe se situe à 3 côtés par face, c'est à dire les faces minimales, puis les effectifs diminuent (moyenne de 4.2). Enfin les contacts à ±3 AA qui sont nettement moins nombreux peuvent être favorisés dans les boucles et les hélices, ceci se traduit sur le graphe par un maximum à 6 côtés par faces (moyenne 5.7).

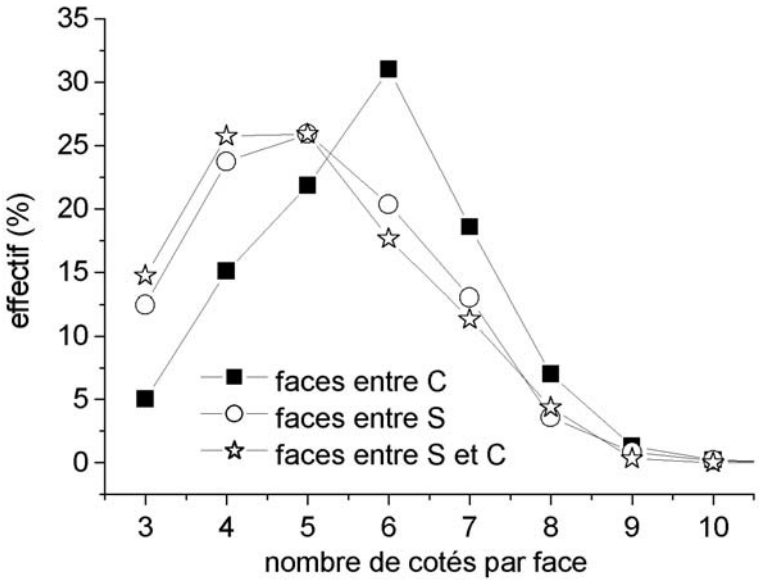


Figure 65 : Distributions du nombre de côtés par face pour une TdV pondérée sur les CG, pour les contacts entre sérines (S) et entre cystéines (C).

La cystéine (C) et la sérine (S) sont deux AA très proches d'un point de vue structural puisqu'ils ne différent que par un seul atome lourd (l'atome de soufre au bout de la chaîne latérale de la cystéine est remplacé par un atome d'oxygène dans la sérine). Leur volume de Pontius et le volume moyen de leurs cellules dans le cas d'une TdV pondérée sur les CG sont très proches. La Figure 65 montre les distributions du nombre de côtés par face pour les contacts entre cystéines (C), entre sérines (S) et entre sérines et cystéines. On constate que les distributions pour les contacts entre sérines et entre sérines et cystéines sont équivalentes avec un maximum à 5 côtés par face (puis 4). A l'inverse, la distribution pour les contacts entre cystéines est complètement différente puisque le maximum est à 6 et que près de 20% des contacts se font avec une face de 7 côtés (moins de 15% dans les deux autres cas). De plus entre cystéines, le nombre de faces triangulaires est très limité, ceci illustre comment le nombre de côtés par face peut refléter les liaisons existant entre résidus, dans le cas présent, le comportement singulier des contacts entre cystéines est bien sûr dû à la présence des liaisons covalentes introduites par les ponts disulfure.

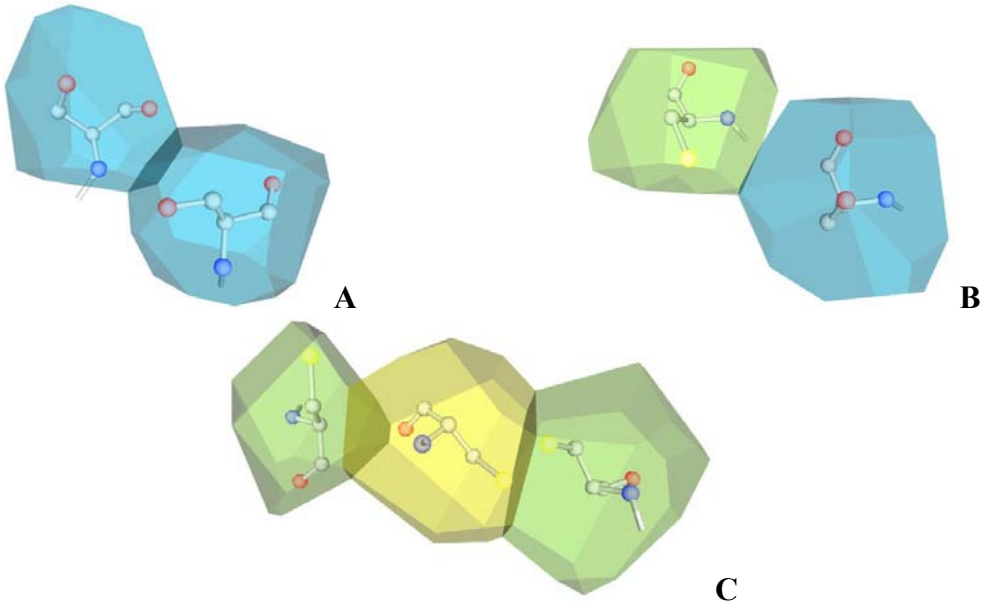


Figure 66 : TdV sur les CG de la structure de code PDB 1a05.
A : Cellules de S264 et S298. B : Cellules de S2 (en bleu) et C4 (en vert).
C : Cellules de C31 (en vert à gauche), C3 (en jaune), C39 (en vert à droite).

La Figure 66 illustre cette propriété avec trois exemples concrets issus de la structure 1a05. La figure A montre les cellules de la sérine n°264 (S264) et de la sérine n°298 (S298), la face de contact entre ces deux cellules a une aire de 6.9 $Å^2$ pour 5 côtés et la distance entre les deux CG (non représentés) est de 5.6 Å. La figure B montre la cellule de la cystéine n°4 (C4) en vert et la cellule de la sérine n°2 (S2) en bleu. La face de contact entre ces deux cellules a 4 côtés et une aire de 0.6 $Å^2$, les deux points représentatifs sont à 6.4 $Å^2$. Ces deux premiers exemples sont typiques des contacts que l'on peut trouver entre deux sérines ou entre une sérine et une cystéine ; il est intéressant de noter que dans ces cas précis, le nombre de faces est un critère à manier avec quelques précautions puisque les deux faces ont presque le même nombre de côtés alors que leur surface varie beaucoup (il existe un rapport de dix entre les deux aires) et ce malgré des distances entre les points représentatifs relativement proches. La figure C montre le cas très intéressant de la cystéine n°3 (C3) en jaune sur la figure, en contact avec deux autres cystéines : la n°31 (C31) à gauche et la n°39 (C39) à droite. Le contact entre les cystéines n°3 et n°31 n'a rien de particulier et la face de contact qui lui est associée a 5 côtés pour une aire de 6.6 $Å^2$ et une distance entre les points représentatifs de 5.5 Å. Ces valeurs sont comparables à celles que l'on observe entre sérines ou entre sérines et cystéines. A l'inverse, la face de contact entre la cystéine n°3 et la n°39 est bien différente puisque son nombre de côtés est de 6 pour une surface de 13.2 $Å^2$ et une distance entre les points représentatifs de 5.6 Å, soit une distance comparable à la face de

contact entre C3 et C31, mais pour une aire presque double. On voit sur la figure que les atomes de soufre (en jaune) de C3 et C39 sont très proches, ce qui s'explique par le fait qu'il existe un pont disulfure entre ces deux résidus. Ceci confirme donc bien que, comme pour les AA se suivant le long de la séquence, les faces de contact entre AA liés de manière covalente ont des caractéristiques marquées.

4.2.4 Influence des structures secondaires sur le nombre de côtés par face

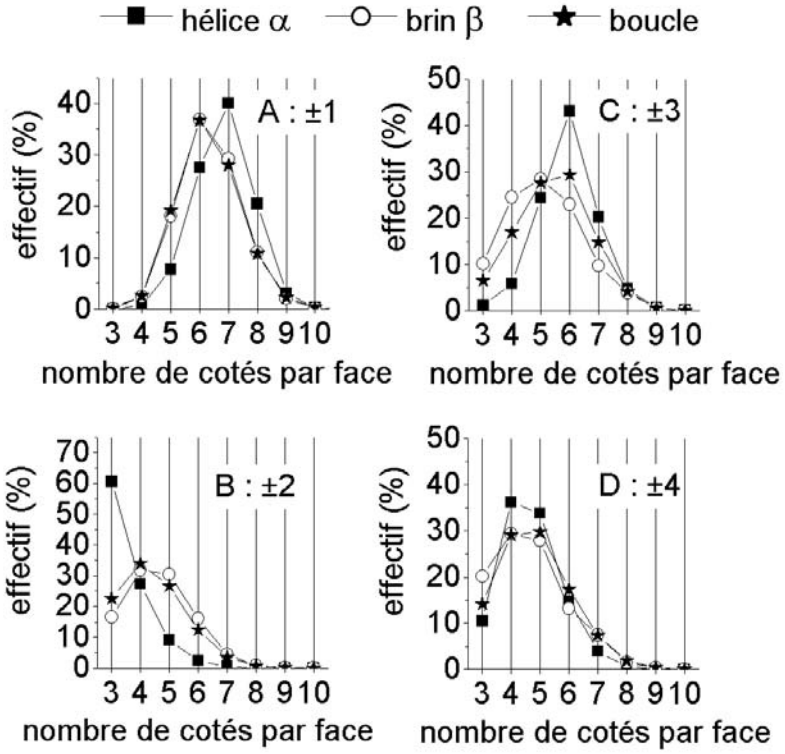


Figure 67 : Distributions du nombre de côtés par face pour une TdV pondérée sur les CG selon les différentes attributions des structures secondaires et en fonction de l'écart en séquence : A ±1, B ±2, C ±3 et D ±4. Chaque courbe est normalisée à 100%.

La Figure 67 présente les distributions du nombre de côtés par face, toujours pour une TdV pondérée sur les CG en fonction des écarts en séquence entre les AA et des structures secondaires. Les faces considérées sont celles mettant en relation des AA dont les attributions secondaires sont identiques ; ces attributions sont réalisées par DSSP[74], qui reste la méthode d'attribution automatique la plus employée à l'heure actuelle. Chaque graphe correspond à un écart en séquence déterminé. Pour l'ensemble des graphes, une remarque générale s'impose : les distributions correspondant aux brins β et au boucles sont à peu près les mêmes. Ceci est particulièrement vérifié en ce qui concerne les écart à ±1 AA pour lesquels les deux distributions sont quasiment identiques. Seul le graphe C présente une légère différence entre

ces deux courbes puisque pour les boucles le maximum se situe à 6 côtés par face alors que pour les brins on le trouve à 5. Il est toutefois important de noter que ces courbes sont normalisées, et que les effectifs des faces sont très différents entre les structures secondaires en fonction de l'écart considéré. Par exemple, pour les contacts avec un écart à ±3 AA, 58.8 % sont réalisés par des AA en conformation hélice, 36.0 % par des AA en conformation boucle et seulement 5.2 % en conformation brin. Les hélices α se singularisent pour les écarts inférieurs à ±4 AA mais, à partir de cette valeur (graphe D), les profils de distribution sont à peu près équivalents. Le trait caractéristique des hélices α se situe principalement pour les écarts à ±2 AA (graphe B) puisque dans ce cas, plus de 60 % des faces de contact sont triangulaires, ceci montre que les contacts à cet écart dans les hélices α ne sont pas des contacts favorisés par la structure secondaire et que leur existence est plutôt due à la proximité des AA le long de la structure primaire. De manière étonnante, les contacts à ±4 AA ne semblent pas avoir de comportement particulier pour les hélices α, à part une distribution légèrement plus étroite alors que ces structures comportent de nombreuses liaisons hydrogène à cet écart.

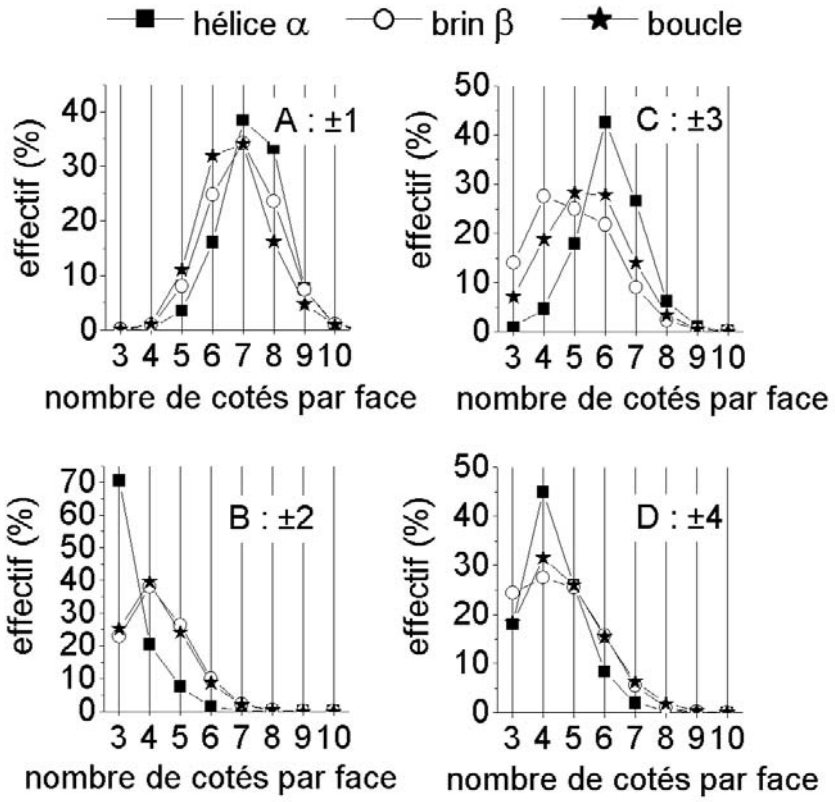


Figure 68 : Distributions du nombre de côtés par face pour une TdV pondérée sur les Cα selon les différentes attributions des structures secondaires et en fonction de l'écart en séquence : A ±1, B ±2, C ±3 et D ±4. Chaque courbe est normalisée à 100%.

J'ai effectué la même étude pour une TdV pondérée sur les Cα, les courbes sont présentées Figure 68. Les distributions sont à peu près équivalentes mais on peut observer quelques différences intéressantes, par exemple à ±1 AA (graphe A), les trois distributions ont maintenant le même maximum de 7 côtés par face, pour ±3 (graphe C), les trois distributions ont des comportements distincts avec un maximum à 4 pour les brins β, à 5 pour les boucles et à 6 pour les hélices α. Ici aussi les distributions pour les écarts à ±4 AA adoptent le même profil (maximum à 4 côtés par face). On peut donc en conclure que les TdV ne sont peut-être pas l'outil idéal pour rendre compte des liaisons hydrogène entre les atomes du squelette. Je montrerai dans un chapitre ultérieur que les TdV sont pourtant efficaces pour détecter les structures secondaires.

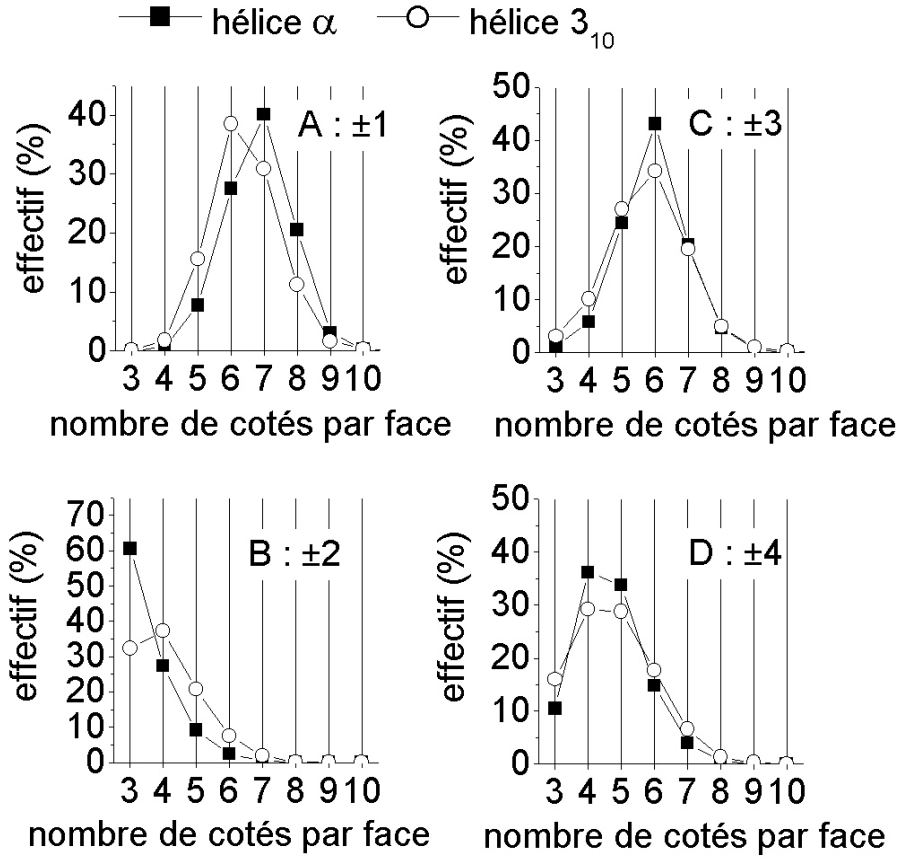


Figure 69 : Distributions du nombre de côtés par face pour une TdV pondérée sur les CG selon les différentes attributions en hélice (α et 3_{10}) et en fonction de l'écart en séquence : A ±1, B ±2, C ±3 et D ±4. Chaque courbe est normalisée à 100%.

La Figure 69 présente les distributions du nombre de côtés par face pour une TdV pondérée sur les CG mais en distinguant les hélices α et les hélices 3_{10} (les distributions des hélices π ne sont pas représentées car les effectifs sont trop faibles). De manière assez étonnante, les distributions ont des formes équivalentes à ±3 AA et ±4 AA, et les différences

apparaissent pour ±1 AA et ±2 AA. Ceci confirme donc la conclusion précédente puisque les hélices α sont caractérisées par des liaisons hydrogène à ±4 AA alors que les hélices 3_{10} sont caractérisées par des liaisons à ±3 AA. Si l'on compare la Figure 67 et la Figure 69, on s'aperçoit que les distributions des hélices 3_{10} sont relativement proches de celles que l'on trouve pour les brins β et les boucles, sauf à ±3 AA où la distribution assez étroite a son maximum à 6 côtés par face.

5 - Contacts entre acides aminés et distances

Un des principaux avantages des TdV réside dans le fait que la définition des contacts est indépendante de tout seuil (cut-off) plus ou moins arbitraire. En effet, la majorité des définitions de contact que l'on trouve dans la littérature utilisent le critère de la distance entre différents points pour déterminer si deux AA sont en contact ou pas. Les points choisis peuvent être des atomes tels que le Cα ou le carbone beta (Cβ), quand il est présent, ou n'importe quel atome (autre que l'hydrogène), ou bien ces points peuvent être des positions fictives comme le centre géométrique (CG) des chaînes latérales. Le contact est effectif si la distance retenue est inférieure à une valeur déterminée (notée R_c) qui elle aussi varie selon les auteurs. Le Tableau 10 propose un aperçu de différentes définitions.

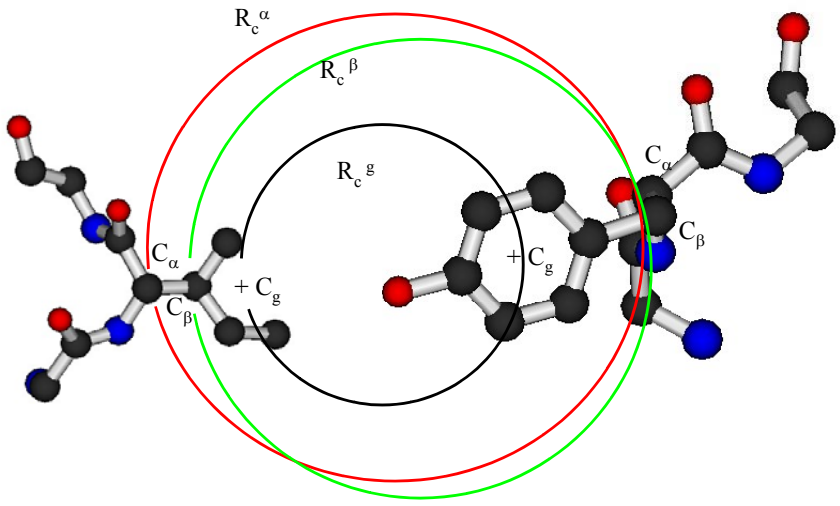


Figure 70 : Illustration des différentes définitions pour valider un contact éventuel entre l'isoleucine (à gauche) et la tyrosine (à droite). On constate que la valeur Rc^{α} choisie pour les contacts entre C_α n'est pas suffisante pour valider le contact (en rouge), alors que les valeurs de $Rc^{\beta,g}$ pour les contacts entre C_β et C_g permettent, elles, de le valider (en vert et en noir).

Auteurs	Définition du contact	Rc
Vendruscolo et coll.[75]	Tous atomes	4,5 Å
Vendruscolo et coll.[75]	Cα	8,5 Å
Hinds et Levitt. [76]	Tous atomes	4,5 Å
Mirny et Domany.[77]	Tous atomes	4,5 Å
Mirny et Shakhnovich.[78]	Tous atomes	4,5 Å
Miyazawa et Jernigan.[79]	CGL	6,5 Å
Skolnick et coll.[80]	Tous atomes	4,5 Å
Thomas et Dill.[81]	Cα ou Cβ	Dépendant de l'AA

Tableau 10 : Différentes définitions de contact entre AA.

5.1 Distribution des distances entre acides aminés en contact

La définition des contacts introduite par les TdV permet d'étudier les distributions des distances entre AA en contact de manière plus large qu'avec les définitions classiques.

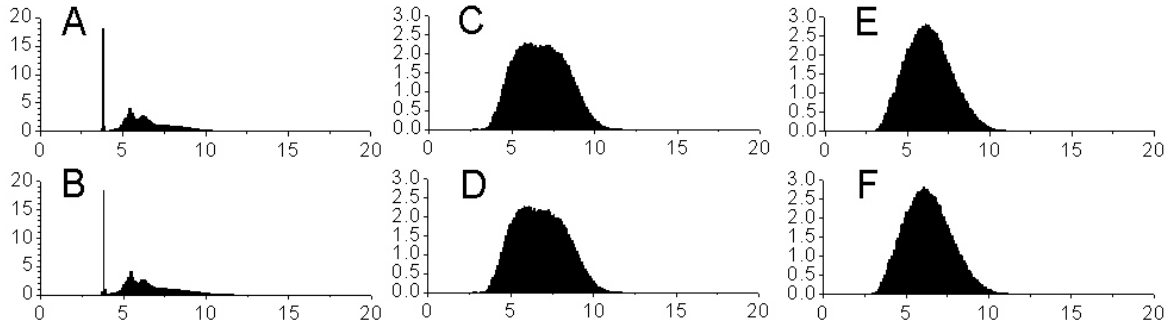


Figure 71 : Distributions des distances entre AA en contact. A - Cα non pondéré, B - Cα pondéré, C - CGL non pondéré, D – CGL pondéré, E - CG non pondéré, F – CG pondéré. En abscisse sont représentées les distances (Å), en ordonnée les effectifs (%).

Sur la Figure 71 qui représente les distributions des distances entre les AA en contact, on constate que les distributions ne varient quasiment pas lorsque l'on passe d'une TdV non pondérée à une TdV pondérée. Ceci signifie donc, comme l'on pouvait s'y attendre, que les relations de voisinage ne sont pas modifiées de manière drastique lorsque l'on change le mode de TdV. En revanche, lorsque l'on modifie les points de TdV, on observe de grandes variations, principalement entre les Cα qui semblent plus ordonnés et les centres géométriques CG et CGL pour lesquels la disparition de pics précis semble indiquer une répartition tridimensionnelle moins régulière.

5.1.1 Tessellation sur les Cα

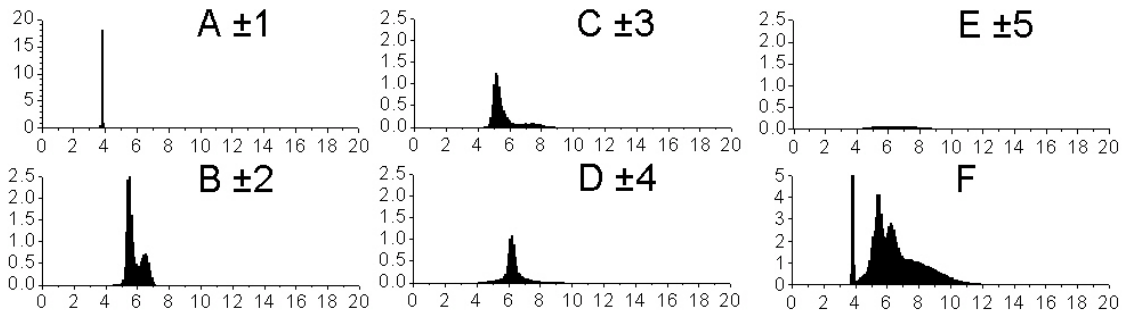


Figure 72 : Distributions des distances entre AA en contact pour une TdV non pondérée sur les Cα en fonction des écarts en séquence. Dans le graphe F qui reprend le graphe A de la Figure 71, le premier pic n'est pas représenté dans toute sa hauteur. En abscisse sont indiquées les distances (Å), en ordonnée les effectifs (%).

Afin de mieux comprendre les phénomènes observés, j'ai d'abord décomposé les distributions des distances pour les Cα pour une TdV non pondérée en fonction de l'écart en séquence des AA en contact. Ces distributions sont présentées Figure 72, jusqu'à l'écart de ±5 AA pour lequel les effectifs diminuent fortement.

Le graphe A montre bien sûr la distance entre Cα voisins le long de la séquence avec un premier situé pic à 3.8 Å. Le deuxième pic du graphe F est une superposition des contacts à ±2 AA et ±3 AA. Pour ce dernier écart (graphe C), le pic observé est dû principalement aux hélices α et aux boucles car les contacts à cet écart dans les brins sont extrêmement rares. Le détail à ±2 AA est donné dans la Figure 73, dans laquelle on peut constater que le premier pic est surtout dû aux hélices α et aux boucles alors que le second pic est dû aux brins β et également aux boucles. Le troisième pic du graphe F de la Figure 72 est la somme de celui observé à ±4 AA (graphe D), dû lui aussi principalement aux hélices α et à certaines boucles et d'un de ceux observés à ±2 AA pour les brins et les boucles (Figure 73 graphes C et D).

On peut donc en conclure que les distributions observées pour les TdV sur les Cα sont principalement modelées à la fois par les écarts en séquence et par les structures secondaires. Ceci est logique puisque les Cα sont situés sur le squelette de la chaîne polypeptidique : ce mode de TdV rend donc compte de la structure de la chaîne dans l'espace. Ceci n'est plus vrai lorsque l'on effectue les TdV sur les CGL car ces points ne sont plus situés sur le squelette (sauf pour la glycine (G)) et les distributions vont dépendre principalement de la nature des résidus.

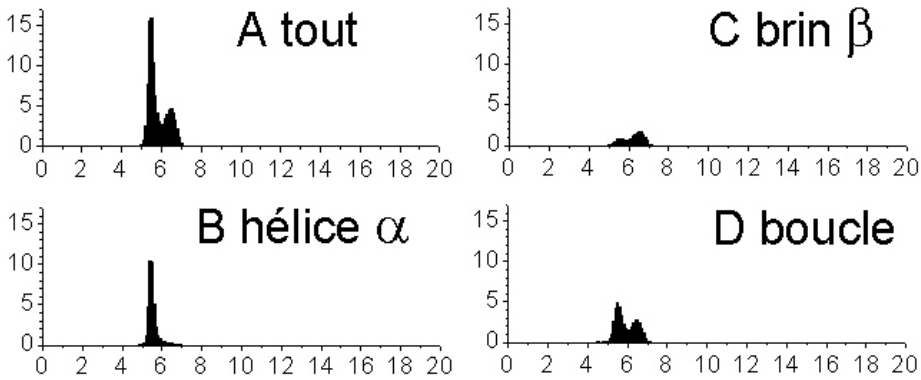


Figure 73 : Distributions des distances entre AA en contact (écart en séquence de ±2) pour une TdV non pondérée sur les Cα en fonction des attributions des structures secondaires (DSSP). Le graphe A reprend le graphe B de la Figure 72. En abscisse sont indiquées les distances (Å), en ordonnée les effectifs (%).

5.1.2 Tessellation sur les CGL

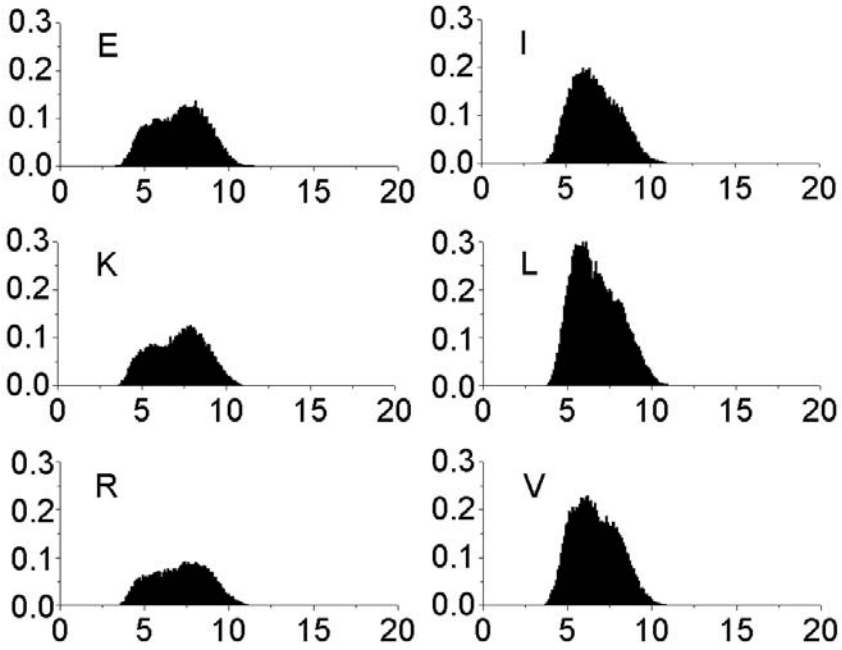


Figure 74 : Distributions des distances entre AA en contact pour une TdV non pondérée sur les CGL pour 6 types d'AA. En abscisse sont indiquées les distances (Å), en ordonnée les effectifs (%).

Pour illustrer ceci, la Figure 74 montre les distributions des distances pour des AA chargés (à gauche) et pour les petits AA hydrophobes (à droite). On peut constater que pour ces deux catégories les profils de distribution sont différents et spécifiques. Les AA chargés favorisent les contacts à plus longues distances, alors que les petits AA hydrophobes favorisent ceux à distances plus courtes. Pour essayer de mieux comprendre l'origine de ces

différences, le détail des distributions est présenté pour l'arginine (R) sur la Figure 75. On peut constater ici aussi que l'arginine établit des contacts à plus ou moins grandes distances selon la nature des AA. Pour les résidus chargés on observe deux comportements distincts, la distribution des distances entre l'arginine (R) et l'acide aspartique (D) ou l'acide glutamique (E) a tendance à privilégier les plus courtes distances alors que c'est l'inverse pour la lysine (K) et l'arginine (R), ceci s'explique bien sûr par la nature des charges mises en jeu, la lysine et l'arginine ont des chaînes latérales basiques (chargées positivement) alors que les deux acides portent une charge négative. La distribution des distances entre l'arginine et les gros hydrophobes (phénylalanine (F), tryptophane (W), et tyrosine (Y)) ne présente pas de maximum marqué, par contre avec les petits AA hydrophobes tels que l'isoleucine (I), la leucine (L) ou la valine (V), on constate que la distribution présente un maximum autour de 8 Å. Ces contacts sont très nombreux et expliquent en partie la forme de la distribution du graphe R de la Figure 74, mais également celle des graphes E et K de cette même figure. Ceci semble donc indiquer que la bi-modalité de la distribution des distances pour les TdV sur les CGL s'explique, en partie, moins par la forme et le volume des résidus que par leur nature physico-chimique et l'opposition entre hydrophobes et hydrophiles.

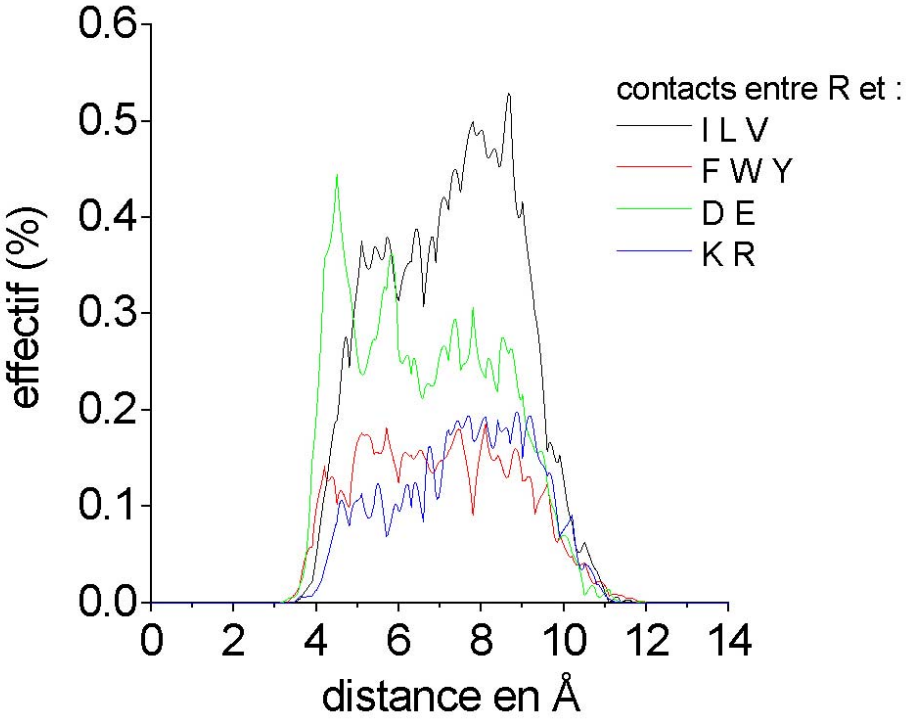


Figure 75 : Distributions des distances entre l'arginine (R) et divers AA en contact pour une TdV non pondérée sur les CGL.

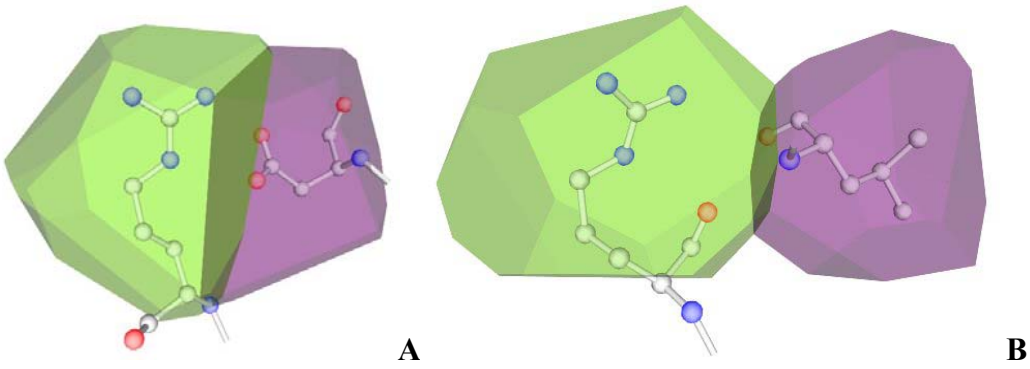


Figure 76 : TdV non pondérée sur les CGL de la structure de code PDB 1a05.
A : R201 (en vert) et D189 (en violet). B : R215 (en vert) et L185 (en violet).

La Figure 76 illustre les courbes précédentes avec un exemple concret. La figure A montre la cellule de l'arginine n°201 (R201 en vert) en contact avec l'acide aspartique n°189 (D189 en violet), les deux CGL (non représentés) sont à 4.1 Å l'un de l'autre, l'aire de la face de contact est de 24.7 $Å^2$ avec 8 côtés. Ces chiffres sont cohérents avec les courbes de la Figure 75 et illustrent le fait qu'il existe un pont salin entre les deux résidus (Figure 77). Les extrémités des deux chaînes latérales sont très proches l'une de l'autre, on retrouve ici l'idée que les paramètres liés aux faces de contact sont représentatifs du type de relation qu'entretiennent les AA entre eux. La figure B montre la cellule de l'arginine n°215 (R215 en vert) en contact avec la cellule de la leucine n°185 (L185 en violet). Les deux CGL sont à 8.1 Å l'un de l'autre et la face de contact a une aire de 8.3 $Å^2$ pour 6 côtés. Comme l'indiquait la Figure 75 les distances entre les arginines et les leucines sont en moyenne assez grandes, on retrouve donc ici cette propriété. Il existe cependant des liaisons hydrogène entre les deux résidus représentés mais celles-ci n'interviennent qu'avec les atomes d'oxygène et d'azote du squelette de la leucine ce qui rend cette liaison indépendante de la nature de la chaîne latérale en présence (voir figure ci-dessous).

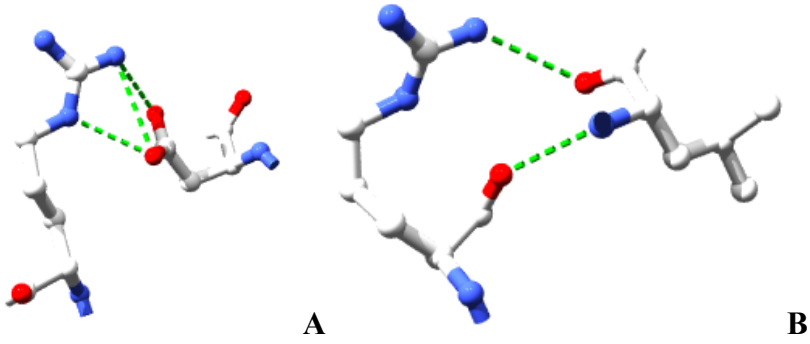


Figure 77 : A : Liaisons hydrogène entre R201 et D189.
B : Liaisons hydrogène entre R215 et L185.

On retrouve également ce type de propriétés pour les distributions de distances impliquant les contacts avec la leucine (L) présentées Figure 78. Ici encore les distributions dépendent plus de la dichotomie hydrophiles/hydrophobes que du volume des AA. Les distributions pour les hydrophobes présentent un maximum autour de 6 Å alors que pour les AA chargés, ce maximum se situe autour de 8 Å.

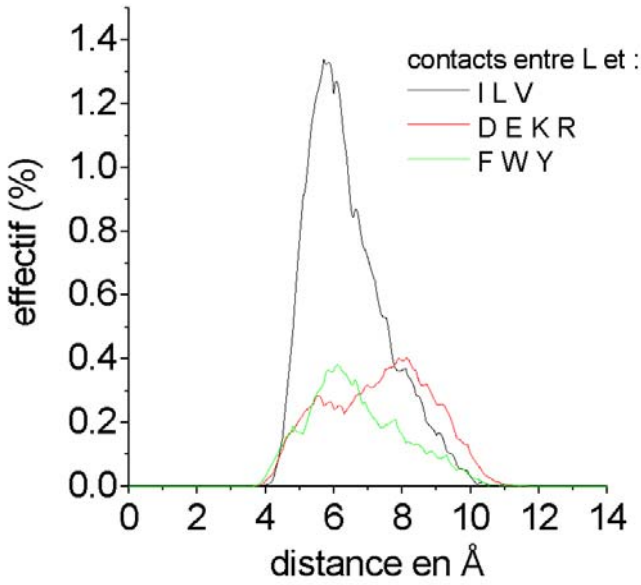


Figure 78 : Distribution des distances entre la leucine (L) et divers AA en contact pour une TdV non pondérée sur les CGL.

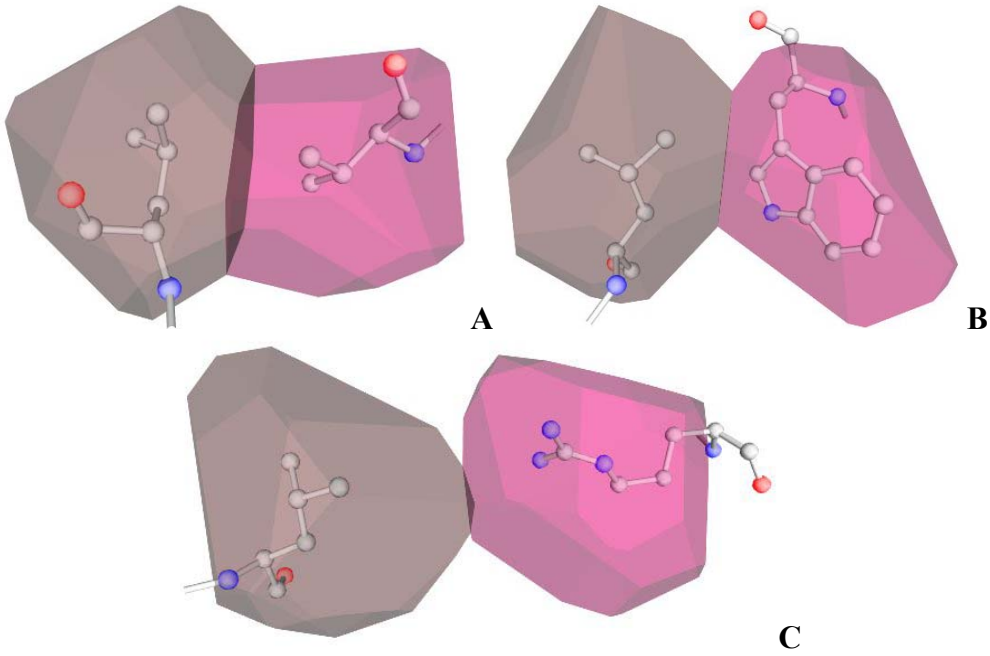


Figure 79 : TdV non pondérée sur les CGL de la structure de code PDB 1a05.
A : L24 (marron) et V299 (rose). B : L47 (marron) et W78 (rose). C : L92 (marron) et R145 (rose).

La Figure 79 montre trois exemples de contacts entre des leucines et d'autres résidus. Figure A, les cellules de la leucine n°24 (L24 en marron) et de la valine n°299 (V299 en rose) partagent une face possédant 6 côtés et dont l'aire est de 23.2 $Å^2$ pour une distance entre les CGL (absents de la figure) de 4.7 Å. Cet exemple est représentatif puisqu'il montre deux résidus possédant des chaînes aliphatiques dont la proximité traduit le phénomène de compaction hydrophobe. La figure B, montre les cellules de la leucine n°47 (L47) et du tryptophane n°78 (W78) dont la face commune comporte 4 côtés pour une aire de 7.1 $Å^2$ et une distance de 5.652 Å. Ici aussi on retrouve un témoignage de la compaction hydrophobe qui se traduit par une distance relativement courte entre les deux CGL. Enfin, la figure C montre la cellule de la leucine n°92 (L92) et celle de l'arginine n°145 (R145). Leur face commune est un triangle de 0.9 $Å^2$ et est associée à une distance de 9.8 Å. Cette dernière valeur qui est un bon exemple illustrant la courbe de la Figure 78, traduit bien la « répulsion » entre les chaînes hydrophobes (leucine) et hydrophiles (arginine).

La forme de la distribution pour un AA donné dépendra donc de ses propriétés physico-chimiques et de la nature et du nombre des voisins avec lesquels l'AA considéré a des affinités pour établir des contacts. Ceci est particulièrement bien illustré par la cystéine (C). La Figure 80 montre la distribution des distances pour ce résidu (seconde ligne), le pic correspondant aux ponts disulfure bien visible sur le graphe du bas est indiqué par une flèche sur la distribution totale.

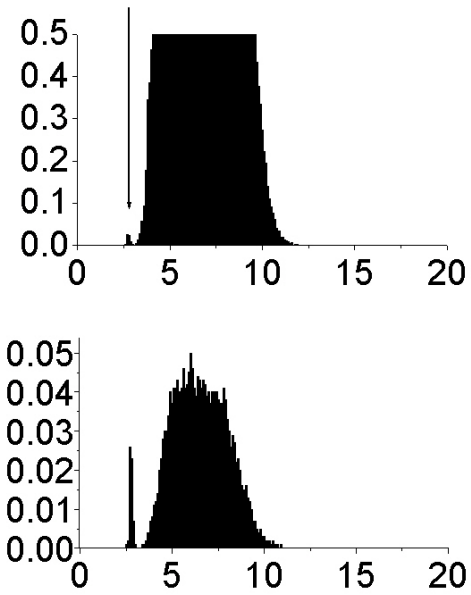


Figure 80 : Distribution des distances pour une TdV non pondérée sur les CGL pour tous les contacts entre AA (première ligne qui reprend le graphe C de la Figure 71), et la cystéine (seconde ligne). En abscisse sont indiquées les distances (Å), en ordonnée les effectifs (%).

5.1.3 Tessellation sur les CG

Lorsque la TdV est effectuée sur les CG, les effets dus à la nature des AA se font moins sentir qu'avec les CGL puisque les positions se rapprochent alors du squelette. On obtient une distribution (Figure 71 graphes E et F) beaucoup plus classique sans particularités remarquables. Les distributions des distances pour les différents AA sont de formes comparables et seules les valeurs moyennes changent d'un AA à l'autre, ces valeurs sont principalement dépendantes du volume des AA. Cette dernière propriété est illustrée par la Figure 81 où sont représentées les moyennes des distances en fonction de la nature des AA pour les TdV sur les différents points représentatifs.

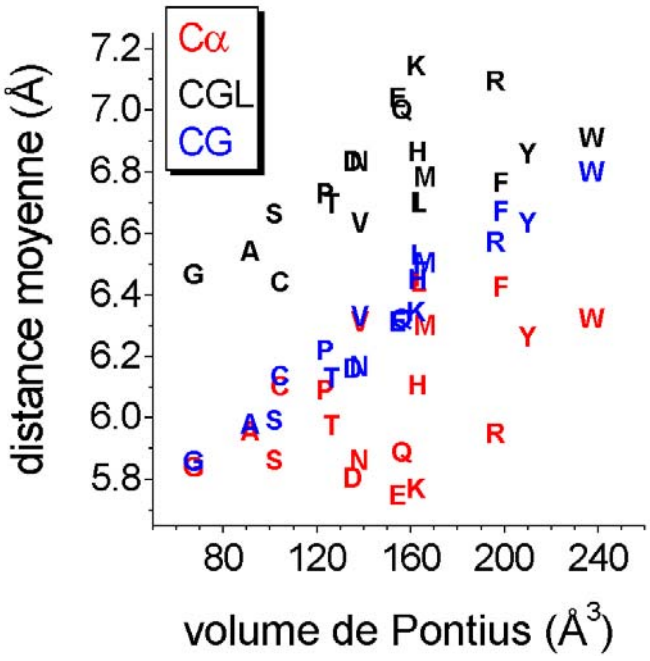


Figure 81 : Moyennes des distances en fonction du volume de Pontius pour les différents AA.

On peut constater que dans tous les cas les distances croissent avec le volume, cependant dans le cas des CGL, on retrouve une séparation entre les AA hydrophiles et les hydrophobes qui s'accentue avec l'augmentation du volume. Les hydrophobes font des contacts à plus petites distances alors que les hydrophiles les font, à volume égal, à plus grandes distances. Pour les CG, cette distinction existe toujours mais elle est très peu prononcée et dans ce cas, les hydrophobes font des contacts avec des distances très légèrement supérieures en moyenne à celles observées pour les hydrophiles. Les écarts de distance entre les deux modes de TdV sont dus à la forme des AA, par exemple la distance moyenne du tryptophane (W) dont la chaîne latérale est composée d'un grand nombre

d'atomes, ne subit pas une grande variation car le fait de rajouter les atomes du squelette ne change pas énormément la position du point. C'est l'inverse qui se produit pour les AA dont les chaînes latérales sont étendues comme la lysine (K) ou l'arginine (R). De plus, les chaînes latérales des résidus hydrophiles sont souvent exposées au solvant ; le fait de considérer les CG au lieu des CGL a donc tendance à réduire artificiellement le volume des protéines et donc de diminuer les distances avec les AA les plus exposés. On observe donc bien une diminution des distances surtout pour les hydrophiles. Pour les Cα, la relation avec le volume est légèrement moins marquée puisque comme je l'ai indiqué c'est la géométrie du squelette qui est le caractère dominant, mais on constate logiquement que les écarts avec les CG sont importants pour les résidus de volumes importants et également pour les hydrophiles du fait de la diminution du volume des protéines. De plus, la compaction hydrophobe qui est le moteur du repliement protéique est physiquement réalisée par les chaînes latérales des résidus hydrophobes. Le fait d'effectuer la TdV sur les Cα augmente donc les distances par rapport aux hydrophiles. Les deux phénomènes, compaction des chaînes latérales hydrophobes et diminution artificielle du volume des molécules, expliquent le renversement de situation observé. Avec les CGL, les distances des hydrophiles sont supérieures à celles des hydrophobes, c'est l'inverse avec les Cα. Comme on peut s'y attendre les CG adoptent un comportement intermédiaire. Enfin, pour les petits résidus, il n'y pas d'écarts avec les CG (glycine (G), alanine (A), cystéine (C)).

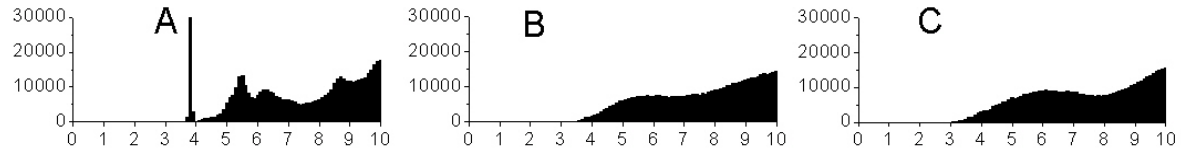


Figure 82 : Distribution de toutes les distances entre Cα (A), CGL (B), CG (C). En abscisse sont indiquées les distances (Å), en ordonnée les effectifs.

La Figure 82 présente les distributions de toutes les distances entre les différents points de représentation. Si l'on compare cette figure à la Figure 71, on constate à quel point les deux approches sont différentes. Par exemple pour les Cα (graphe A) le seuil régulièrement utilisé pour déterminer s'il y a contact entre deux AA est de 8.5 Å. Or sur la Figure 71 (graphes A et B), on constate que certains contacts se font avec des distances supérieures à cette valeur, cependant la Figure 82 montre qu'entre 8.5 Å et 10 Å, le nombre de contacts potentiels est énorme, posant ainsi le problème de l'utilisation de seuils et du choix de leur valeur.

6 - L'enfouissement ou l'exposition à l'environnement

6.1 Définition

Pour pouvoir associer une cellule à chaque AA d'une structure protéique, nous avons vu qu'il fallait installer cette structure dans un environnement. Ceci implique que les faces des cellules définissent des contacts de différentes natures : contacts entre AA, entre AA et points de l'environnement (centres des sphères) et entre points de l'environnement. Pour une cellule correspondant à un AA donné, on peut donc classer les faces en deux catégories : les faces définissant un contact avec un autre AA et celles définissant un contact avec l'environnement. Si l'on considère les aires de ces faces, il est alors possible de calculer la proportion de surface en contact avec des points de l'environnement par rapport à la surface totale de la cellule considérée. On peut ainsi définir l'exposition à l'environnement η d'un AA comme étant le rapport entre, d'une part, la somme des aires des faces en contact avec des points de l'environnement et d'autre part, la surface totale de la cellule considérée. Pour une cellule ne faisant aucun contact avec l'environnement (cellule en volume) on aura $\eta = 0$, pour une cellule totalement entourée de points de l'environnement (cas d'un AA isolé qui n'existe pas dans ma banque) on aura $\eta = 1$ (j'écrirai aussi indifféremment $\eta = 100$ %). L'enfouissement peut être défini par $1 - \eta$ (ou $100 - \eta$ en %).

6.2 Comparaison avec DSSP

Une des méthodes les plus utilisées pour calculer la surface accessible au solvant d'un résidu est celle fournie par le programme DSSP[74]. L'accessibilité au solvant peut être déterminée en divisant la valeur donnée par la surface du résidu afin d'obtenir sa proportion de surface accessible au solvant. Pour calculer cette accessibilité, j'ai utilisé les aires calculées par Shrake et al[82], qui sont les aires accessibles des résidus lorsqu'ils sont en configuration G - X - G (avec G pour glycine et X pour l'AA considéré). Ces surfaces sont présentées dans le Tableau 11.

A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
124	94	154	187	221	89	201	194	214	198	215	161	150	190	244	126	152	169	265	236

Tableau 11 : Surfaces des résidus en configuration G-X-G en Å².

Pour calculer les accessibilités à l'environnement, il est primordial de tenir compte des différents volumes des AA, c'est la raison pour laquelle je n'ai utilisé dans cette partie que des TdV pondérées.

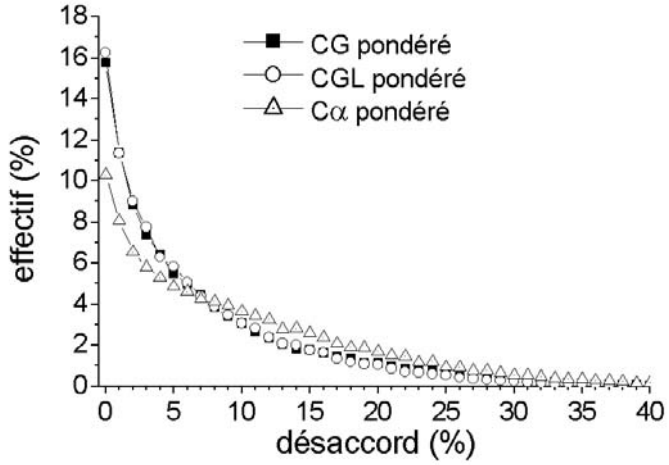


Figure 83 : Comparaison des deux méthodes de détermination de l'accessibilité au solvant/environnement. En abscisse est représenté le désaccord entre les deux méthodes (%).

La Figure 83 présente une comparaison des deux méthodes pour les différents points de représentation. Comme l'on pouvait s'y attendre, les TdV sur les Cα sont celles dont les accessibilités sont le plus en désaccord avec DSSP. Ceci est logique puisque comme nous l'avons vu les Cα donnent des résultats qui sont surtout représentatifs des structures secondaires et de la connectivité le long de la séquence. Les centres géométriques sont plus aptes à représenter la nature physico-chimique des AA et donnent des résultats plus proches de ceux calculés avec DSSP.

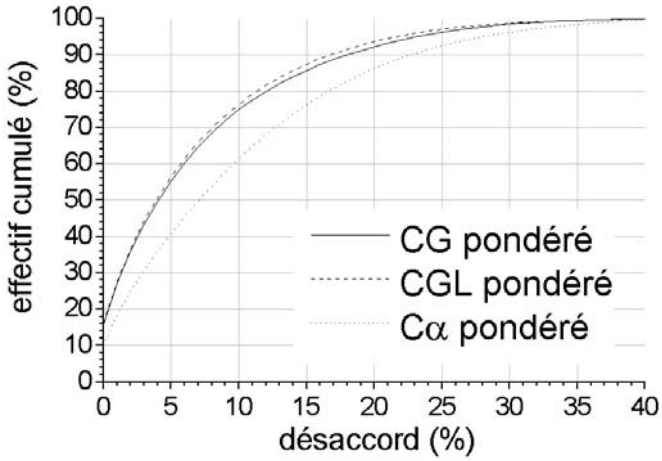


Figure 84 : Comparaison des deux méthodes de détermination de l'accessibilité au solvant/environnement. En abscisse est représenté le désaccord entre les deux méthodes (%).

La Figure 84 présente les mêmes résultats sous une forme légèrement différente. Dans le cas d'une TdV sur les CG ou les CGL, on constate que plus de 50% des accessibilités se font avec au plus 5% de désaccord entre notre méthode et DSSP. Sur ce graphe, on constate également que les TdV sur les CGL donnent des résultats très légèrement supérieurs à ceux obtenus avec des TdV sur les CG. La moyenne des désaccords entre notre méthode et DSSP est de 6.81 % pour les CGL, 7.17 % pour les CG et 9.95 % pour les Cα. Il semble donc naturel d'utiliser dans la suite les TdV pondérées sur les CGL pour calculer les accessibilités à l'environnement.

6.3 Influence de l'environnement sur les propriétés des cellules

6.3.1 Volume des cellules

Il est intéressant de suivre l'évolution des volumes des cellules en fonction de l'enfouissement des AA au sein de la protéine. Pour que cette étude ne soit pas biaisée par les différences de volume entre les vingt types d'AA, il est indispensable de s'intéresser au rapport entre le volume d'une cellule et le volume moyen des cellules pour chaque type de résidu, puis de faire les moyennes de ces rapports. La Figure 85 présente les résultats obtenus. On constate tout d'abord qu'à partir de 70 % d'accessibilité, on a une dispersion des points due au manque d'effectif pour ces valeurs. Mais, l'observation principale est bien sûr le fait qu'en moyenne les volumes diminuent avec l'enfouissement.

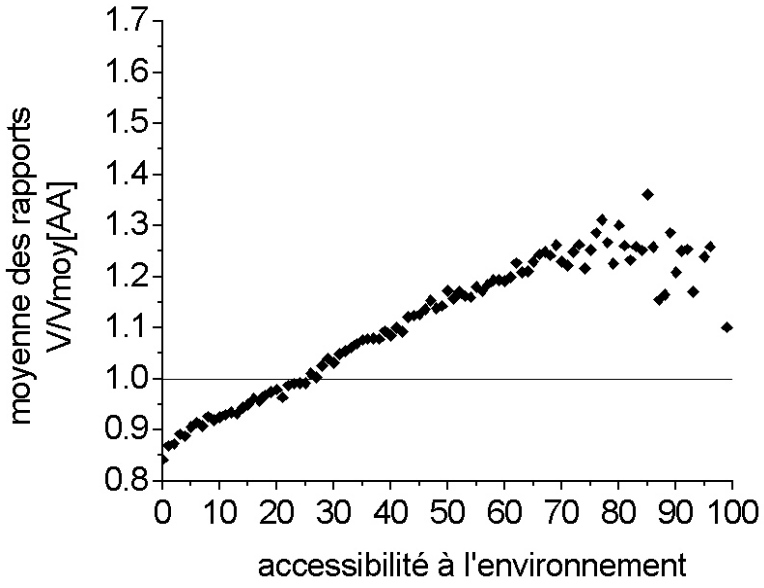


Figure 85 : Volume moyen des cellules en fonction de l'accessibilité à l'environnement.

Ceci peut être expliqué encore une fois par le moteur du repliement qu'est la compaction hydrophobe. Lors du repliement les résidus hydrophobes tendent à se regrouper de manière compacte en excluant les molécules d'eau alors que les AA hydrophiles sont plutôt orientés vers la surface. Ce regroupement de nombreux résidus au cœur de la protéine fait que ces résidus ont plus de voisins proches que ceux situés en surface. Les cellules qui leur sont associées ont donc moins d'espace à occuper et sont en moyenne plus petites. Plus on se rapproche de la surface plus cette compaction diminue[5], ainsi il a été montré que les atomes à la surface des protéines occupaient en moyenne un volume supérieur de 7 % aux atomes situés au cœur des structures[83]. Dans notre cas, la différence moyenne entre volume et surface est de 23 % mais bien sûr cette valeur est dépendante de l'environnent choisi et de la pondération de l'environnement utilisée. Avec les Cα, cette différence moyenne est de 7.5 %. Les cellules disposent donc de plus d'espace en étant plus proches de la surface. L'évolution est quasiment linéaire et reflète la progression du cœur hydrophobe vers la surface hydrophile. Il est évident que pour une cellule donnée, le calcul de l'enfouissement et le volume sont liés de manière forte, mais on obtient le même type de courbes avec l'accessibilité au solvant déterminée par DSSP (résultats non présentés).

6.3.2 Compaction des cellules

Pour faire suite et en complément du paragraphe précédent, il est naturel de donner un bref aperçu de la compaction des cellules en fonction de l'accessibilité à l'environnement. Je définirai la compaction d'une cellule, notée Q, comme le rapport entre le volume de cette cellule et le volume de Pontius correspondant. Une valeur supérieure à 1 signifiera donc que le volume de la cellule est supérieur au volume de l'AA, je parlerai donc plutôt d'expansion. Une valeur inférieure à 1 signifiera au contraire que le volume de la cellule est inférieur au volume occupé par l'AA et je parlerai donc de compaction.

La Figure 86 donne les différentes moyennes de compaction pour les différents AA. Il est facile de constater que la progression linéaire observée pour les Cα autour de la valeur un, n'est pas reproduite avec les CGL, pour lesquels on distingue clairement deux comportements. En effet, à volumes comparables, les cellules des résidus hydrophobes sont en moyenne plus compactes que celles des autres résidus.

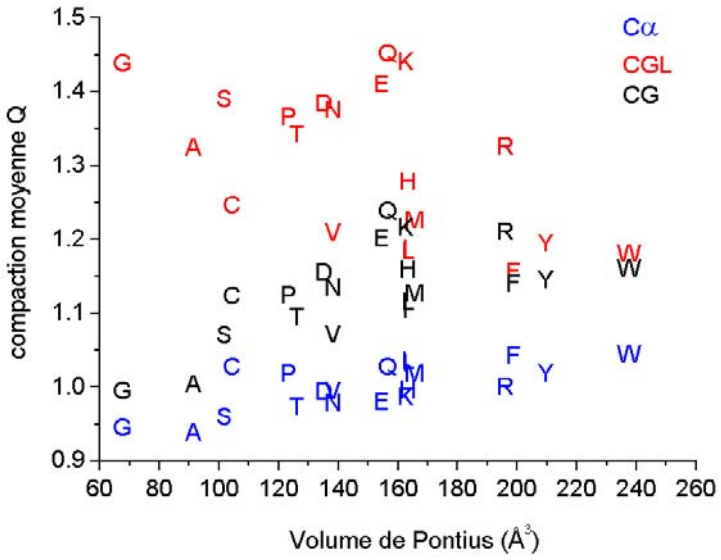


Figure 86 : Moyenne des compactions en fonction du volume de Pontius.

Lorsque l'on étudie l'évolution de Q en fonction de l'accessibilité à l'environnement (Figure 87), on retrouve le même profil que pour l'évolution des volumes (Figure 85). On constate bien que la compaction au cœur est plus forte et qu'elle diminue avec l'augmentation de l'accessibilité à l'environnement. Cette diminution est à peu près linéaire avec une compaction moyenne pour les hélices de 1.31, pour les brins de 1.21 et pour les boucles de 1.38. Ces valeurs s'expliquent par les valeurs moyennes d'accessibilité à l'environnement de ces structures : 21.8 % pour les hélices α, 14.5 % pour les brins et 31.3 % pour les boucles. Les fonctions des protéines sont souvent associées à des résidus situés sur les boucles, la mobilité de ces dernières (et donc l'espace dont elles disposent pour bouger) est primordiale.

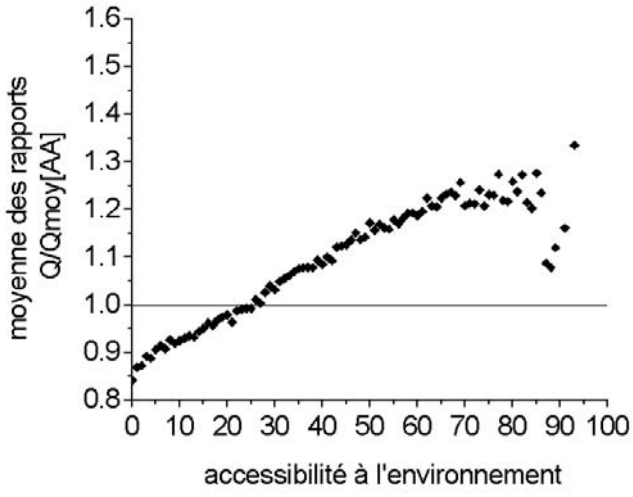


Figure 87 : Compaction moyenne des cellules en fonction de l'accessibilité à l'environnement déterminée pour les CGL.

6.3.3 Nombre de faces des cellules

On constate que les variations entre les AA les plus exposés et ceux qui sont le plus enfouis sont faibles et que cette évolution commence tout d'abord par une diminution du nombre de faces qui se poursuit pas une croissance à peu près linéaire. Cette croissance du nombre de faces s'explique en partie par le fait que, comme le montre la Figure 85, le volume des cellules croît avec l'augmentation de l'accessibilité au solvant, or le nombre de faces et le volume des cellules sont étroitement liés (voir Figure 58 graphe D, plus une cellule est grosse plus son nombre de faces est important). La légère diminution pour les faibles accessibilités résulte du conflit entre l'évolution du volume et la compaction hydrophobe au cœur. En effet, cette compaction implique que les cellules sont en moyenne plus petites mais elles font aussi plus de faces car le nombre de plus proches voisins est plus important. La Figure 89 présente la même courbe, mais en fonction de l'accessibilité au solvant déterminée par DSSP. On constate que les deux courbes présentent les mêmes variations. L'accessibilité au solvant déduite des valeurs de DSSP donne des résultats très proches de 100% voire supérieurs à cette valeur (non montrés). Ceci s'explique par le fait que je normalise les valeurs données par DSSP par des surfaces calculées en conformation G-X-G, or les AA aux extrémités de la chaîne polypeptidique ne sont pas dans cette conformation. Il est donc normal de trouver des valeurs d'accessibilité très proches et supérieures à 100%.

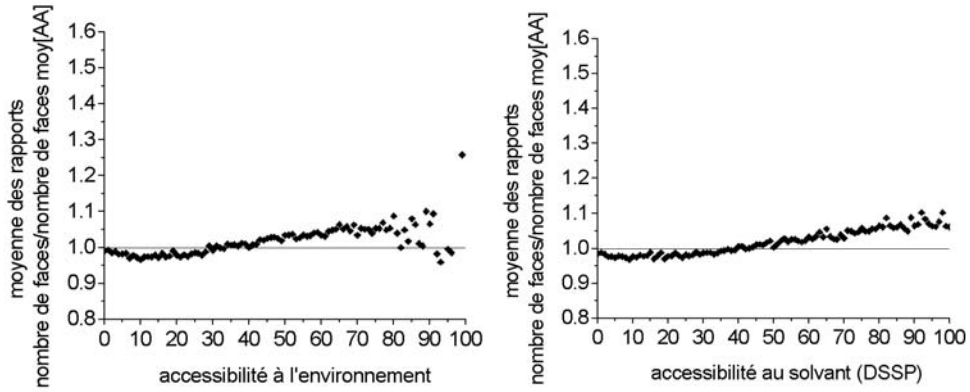


Figure 88 à gauche : Nombre de faces moyen en fonction de l'accessibilité à l'environnement.
Figure 89 à droite : Nombre de faces moyen en fonction de l'accessibilité au solvant (DSSP).
Pour les 2 graphes les TdV on été déterminées pour les CGL.

6.3.4 Rapport surface/volume : R

Une façon de « mesurer » la capacité d'une cellule à s'inscrire dans une sphère, est de calculer le rapport sans dimension R défini de la façon suivante : $R = 0.434 \times S^{1/2} / V^{1/3}$. Ce

rapport est défini pour être égal à 1 pour un dodécaèdre régulier, qui est la cellule idéale pour un empilement local compact de sphères. Ce rapport est plus grand pour les cellules les moins « sphériques » ou les moins régulières, c'est à dire celles qui optimisent le moins leur surface par rapport à leur volume.

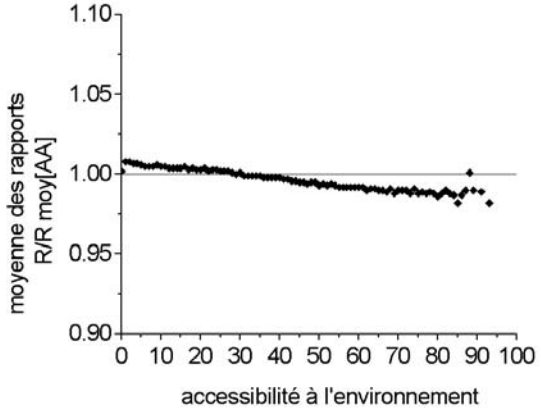
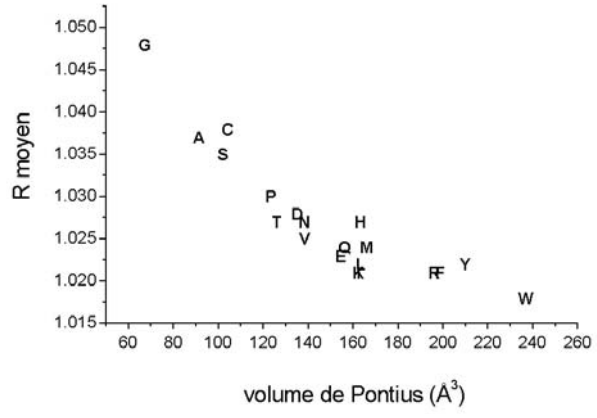


Figure 90 : R moyen en fonction de l'accessibilité à l'environnement déterminée pour les CGL.

L'évolution de cette valeur est présentée dans la Figure 90, on constate que là aussi les variations sont légères et que les cellules sont de plus en plus isotropes lorsque l'accessibilité à l'environnement augmente. Ceci s'explique à l'aide de la Figure 88 ou de la Figure 85. En effet, on peut constater que le volume (ou le nombre de faces par cellule) augmente avec l'accessibilité, or comme le montrent la Figure 91 et la Figure 92, R diminue avec l'augmentation du volume (ou du nombre de faces par cellule). Plus une cellule est exposée à l'environnement plus elle est de taille importante et plus elle a de faces, elle est donc plus isotrope.



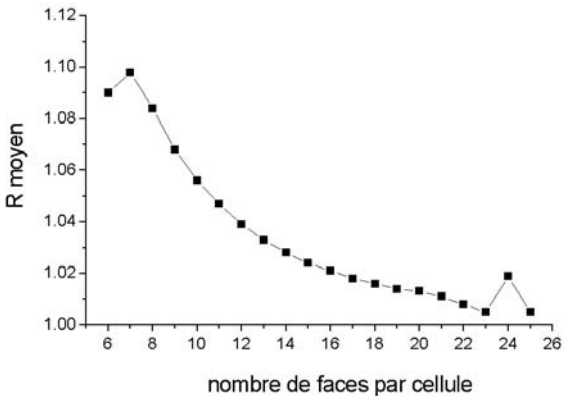


Figure 91 à gauche : R moyen pour chaque type d'AA, en fonction du volume de Pontius, pour les CGL.
Figure 92 à droite : R moyen en fonction du nombre de faces par cellule pour les CGL.

Nous venons donc de voir que la position de la cellule au sein de la structure protéique traduite en terme d'accessibilité à l'environnement ou d'enfouissement a une grande influence sur les propriétés des cellules. Les courbes du type de la Figure 90 permettent de suivre l'évolution d'une propriété en fonction de l'enfouissement, mais en fait cette évolution ne peut être réellement suivie que pour les cellules ayant une face de contact commune avec une cellule de l'environnement, c'est à dire pour les cellules de surface. Les cellules en volume, elles, ne sont représentées que par un point, or ces cellules représentent à elles seules 22% de la banque.

7 - Conclusion

La description rapide du paysage des TdV appliquées aux protéines qui est présentée dans ce chapitre montre qu'il n'y a pas de façon idéale de réaliser les tessellations. En fait, savoir quelle est la meilleure méthode n'a de sens que si l'on cherche à résoudre un problème précis, il sera alors plus astucieux d'utiliser les Cα ou les CGL, ou bien encore une tessellation pondérée ou non pondérée. Les TdV peuvent être utilisées comme un outil dont certains paramètres sont modulables, il est donc important de connaître leur influence sur les diverses propriétés observées pour employer cet outil de manière optimale.

C h a p i t r e 5

Proximité des extrémités N et C-terminales

1 - Introduction

Les régions terminales des structures protéiques ont des propriétés différentes en bien des points des régions centrales de ces mêmes structures. Par exemple, elles sont très souvent exposées au solvant et flexibles, c'est une des raisons pour lesquelles leurs structures sont souvent déterminées avec une plus faible résolution. De plus, elles sont fréquemment chargées et au sein d'une même famille structurale, leurs séquences peuvent être très variables. Peut-on imaginer que ces extrémités jouent un rôle important dans le processus de repliement de la chaîne polypeptidique ? Lors de la traduction de l'ARNm en protéine, l'extrémité N-terminale est la première à émerger du ribosome. Il a été par ailleurs montré[84, 85] que le processus de repliement et la synthèse du polymère se déroulent de manière simultanée. Il est donc possible que le polypeptide naissant adopte des conformations favorisant le repliement de la structure complète. Il a été également montré[86] que dans une forme dénaturée de l'apomyoglobine, des interactions comparables à celles de la forme repliée apparaissent entre les extrémités N et C. Des interactions entre les régions terminales du cytochrome c du cœur du cheval lors des premières étapes de son repliement ont été aussi détectées[87]. Ces observations tendent à confirmer l'hypothèse de Ptitsyn[88] formulée en 1981 selon laquelle les interactions entre extrémités N et C terminales pourraient être d'une grande importance lors du repliement des protéines. La proximité des extrémités N- et C-terminales conduirait à une diminution tangible du nombre de conformations accessibles à la protéine et expliquerait également la quasi absence de nœuds dans les structures protéiques[89-91].

A notre connaissance deux articles seulement ont été consacrés à l'étude de la proximité des extrémités N- et C-terminales. J. Thornton et B. L. Sibanda[92] ont montré en 1983 sur 52 structures que les extrémités des protéines natives étaient plus proches que celles de structures générées aléatoirement. La seconde étude menée par J. A. Christopher et T. O. Baldwin[93] en 1996 sur une banque de 72 structures montrèrent qu'en moyenne les distances entre

extrémités n'étaient pas particulièrement plus faibles que celles obtenues par des simulations aléatoires. Le but de ce chapitre est de montrer à partir de banques plus importantes et à l'aide des TdV que les extrémités N- et C-terminales sont statistiquement proches.

2 - Banque de structures

Les banques utilisées pour effectuer les différentes statistiques présentées dans ce chapitre ont été constituées en suivant le même schéma. Puisque nous nous sommes intéressés à la position des résidus et donc des atomes les uns par rapport aux autres, il était important que les structures retenues aient une résolution correcte, c'est la raison pour laquelle je n'ai conservé que des structures dont la résolution est inférieure à 2.5 Å. Afin d'éviter tout problème lors de la construction des cellules de Voronoï, il fallait s'assurer qu'il n'y avait pas de manques dans les structures choisies, cela signifie que tous les AA des séquences protéiques devaient être bien présents dans la structure. Pour cette vérification, les séquences, déduites des fichiers PDB, ont été comparées aux séquences disponibles dans les banques de séquences protéiques telles que SWISSPROT[67] maintenue par le SIB (Swiss Institute of Bioinformatics), TrEMBL qui contient la traduction de toutes les parties codantes figurant dans EMBL[94] maintenue à l'EBI (European Bioinformatics Institue), et GenPept qui correspond à celles de GenBank[95] maintenue au NCBI (National Center for Biotechnology Information) aux Etats-Unis. Les structures sélectionnées sont celles auxquelles manquaient tout au plus quatre AA à chaque extrémité et aucun ailleurs, par rapport aux séquences équivalentes retrouvées dans les banques par le programme de comparaison de séquences Blast[96]. Cette précaution a aussi pour but de s'assurer que les extrémités des chaînes polypeptidiques considérées sont bien les extrémités réelles de ces chaînes et non issues d'artefacts de production ou de gestion expérimentale de celles-ci.

Afin de supprimer la redondance structurale dans les banques utilisées, seules les protéines appartenant à des superfamilles différentes (selon la classification SCOP[68]) ont été finalement retenues. Ces protéines peuvent bien sûr être des monomères ou des multimères, les éventuelles distinctions faites dans la suite sont fondées sur les informations présentes dans les fichiers PDB ou dans la publication originale quand l'information y était disponible. J'obtiens donc finalement une banque de 177 protéines et une banque de 176 domaines, ce qui peut sembler peu, mais reste largement supérieur aux effectifs des banques utilisées lors des études similaires précédentes (52 et 72 structures). Dans la banque multidomaine, les

domaines extraits de protéines multidomaines sont ceux définis par SCOP. Après vérification visuelle (et comme nous le verrons dans la suite), pour quelques rares domaines dont la définition me paraissait discutable, j'ai utilisé la définition proposée par CATH[97] (structure de code PDB 1duv par exemple). J'ai privilégié SCOP au détriment de CATH car les domaines définis par cette dernière sont souvent constitués de morceaux de protéines disjoints répartis tout au long de la séquence. Ceci ne pouvait pas nous convenir puisque l'on ne peut alors définir de véritables extrémités N- et C-terminales.

PDB	SCOP	PDB	SCOP	PDB	SCOP	PDB	SCOP
1a2o	c.23.1.1 c.40.1.1	1d0d	g.8.1.2	1gpe	c.3.1.2 c.3.1.2 d.16.1.4	1ugi	d.17.5.1
1a2z	c.56.4.1	1dea	c.35.1.1	1gsa	c.30.1.3 d.142.1.1	1unk	a.28.2.1
1a44	b.17.1.1	1dfx	b.1.13.1 g.41.5.2	1gvp	b.40.4.7	1uok	b.71.1.1 c.1.8.1
1a53	c.1.2.4	1dhn	d.96.1.3	1hcv	b.1.1.1	1utg	a.101.1.1
1a73	d.4.1.3	1dli	a.100.1.4 c.2.1.6 c.26.3.1	1hnj	c.95.1.1 c.95.1.1	1vfr	d.90.1.1
1a7v	a.24.3.2	1dlm	b.3.6.1	1hoe	b.5.1.1	1vpi	a.133.1.2
1a8b	a.65.1.1	1dmu	c.52.1.4	1ifc	b.60.1.2	1whi	b.39.1.1
1a8l	c.47.1.2 c.47.1.2	1dnp	a.99.1.1 c.28.1.1	1iov	c.30.1.2 d.142.1.1	1xgs	a.4.5.25 d.127.1.1 d.127.1.1
1a8p	b.43.1.1 c.25.1.1	1doz	c.92.1.1	1isu	g.35.1.1	1xva	c.66.1.5
1aj2	c.1.21.1	1dpt	d.80.1.3	1jac	b.77.3.1	1ypr	d.110.1.1
1ako	d.151.1.1	1dqi	b.1.13.1	1jcv	b.1.8.1	1zin	c.37.1.1 c.37.1.1 g.41.2.1
1amm	b.11.1.1 b.11.1.1	1dqn	c.61.1.1	1kuh	d.92.1.1	2abk	a.96.1.1
1aoe	c.71.1.1	1drw	c.2.1.3 c.2.1.3 d.81.1.3	1lcl	b.29.1.3	2acy	d.58.10.1
1aor	a.110.1.1 d.152.1.1	1duv	c.78.1.1 c.78.1.1	1lop	b.62.1.1	2bnh	c.10.1.1
1aoz	b.6.1.3 b.6.1.3 b.6.1.3	1dxe	c.1.12.5	1mat	d.127.1.1	2cev	c.42.1.1
1apx	a.93.1.1	1dzr	b.82.1.1	1mka	d.38.1.2	2cga	b.47.1.2
1azp	d.9.2.1	1e0s	c.37.1.8	1mla	c.19.1.1 c.19.1.1 d.58.23.1	2end	a.18.1.1
1b16	c.2.1.2	1e15	b.72.2.1 c.1.8.5 c.1.8.5 d.26.3.1	1mng	a.2.11.1 d.44.1.1	2fdn	d.58.1.1
1b6t	e.26.1.3	1e19	c.73.1.1	1mpg	a.96.1.3 d.129.1.2	2hgs	c.30.1.4 d.142.1.6 d.142.1.6
1bf6	c.1.9.3	1e2a	a.7.2.1	1mug	c.18.1.2	2hvm	c.1.8.5
1bgv	c.2.1.7 c.58.1.1	1e7f	a.126.1.1 a.126.1.1 a.126.1.1	1nhk	d.58.6.1	2mhr	a.24.4.1
1bm9	a.4.5.7	1ebf	c.2.1.3 c.2.1.3 d.81.1.2	1one	c.1.11.1 d.54.1.1	2mnr	c.1.11.2 d.54.1.1
1br6	d.165.1.1	1ecp	c.56.2.1	1oyc	c.1.4.1	2nsy	c.26.2.1
1bs0	c.67.1.4	1ee2	b.35.1.2 b.35.1.2 c.2.1.1	1pbe	c.3.1.2 c.3.1.2 d.16.1.2	2ovo	g.15.1.1
1bs4	d.167.1.1	1emy	a.1.1.2	1pfk	c.89.1.1	2pia	b.43.1.2 c.25.1.2 d.15.4.2
1bv1	d.129.3.1	1eqo	d.58.30.1	1php	c.86.1.1	2pii	d.58.5.1
1bxo	b.50.1.2	1ert	c.47.1.1	1pii	c.1.2.4 c.1.2.4	2pth	c.56.3.1
1bxy	d.59.1.1	1ew4	d.82.2.1	1pjc	c.2.1.4 c.23.12.2 c.23.12.2	2rn2	c.55.3.1
1byf	d.169.1.1	1exm	b.43.3.1 b.44.1.1 c.37.1.8	1plq	d.131.1.2 d.131.1.2	2sn3	g.3.7.1
1c1k	a.120.1.1	1f4t	a.104.1.1	1poh	d.94.1.1	3chy	c.23.1.1
1c3q	c.72.1.2	1f8y	c.23.14.1	1qd9	d.79.1.1	3cla	c.43.1.1
1c7q	c.80.1.2	1feh	c.96.1.1 d.15.4.2 d.58.1.5	1qex	b.32.1.1	3lzm	d.2.1.3
1c8u	d.38.1.3 d.38.1.3	1flj	b.74.1.1	1qgx	e.7.1.1	3pmg	c.84.1.1 c.84.1.1 c.84.1.1 d.129.2.1
1c9h	d.26.1.1	1flm	b.45.1.1	1qh4	a.83.1.1 d.128.1.2	3pyp	d.110.3.1
1cfz	c.56.1.1	1fp3	a.102.1.3	1qh5	d.157.1.2	3tdt	b.81.1.2
1cmb	a.43.1.2	1fq0	c.1.10.1	1qkj	c.87.1.1	4icb	a.39.1.1
1cnu	d.109.1.2	1fqt	b.33.1.1	1qpo	c.1.17.1 d.41.2.1	4tms	d.117.1.1
1cnz	c.77.1.1	1frb	c.1.7.1	1qto	d.32.1.2	5mdh	c.2.1.5 d.162.1.1
1coz	c.26.1.2	1ftr	d.58.33.1 d.58.33.1	1qtw	c.1.15.1	5pnt	c.44.1.1
1cqx	a.1.1.2 b.43.1.4 c.25.1.5	1g3k	d.153.1.4	1rb9	g.41.5.1	6ins	g.1.1.1
1ctj	a.3.1.1	1g71	e.16.1.1	1rcf	c.23.5.1	7rsa	d.5.1.1
1ctt	c.97.1.1 c.97.1.1	1gcb	d.3.1.1	1tgx	g.7.1.1	7tim	c.1.1.1
1cuo	b.6.1.1	1gdh	c.2.1.4 c.23.12.1 c.23.12.1	1tib	c.69.1.17	8tln	a.67.1.1 d.92.1.2
1czj	a.138.1.1	1glq	a.45.1.1 c.47.1.5	1trk	c.36.1.2 c.36.1.2 c.48.1.1	9rnt	d.1.1.1
1czp	d.15.4.1						

Tableau 12 : Liste des 177 structures de la banque de protéines. Chaque colonne contient le code PDB et le code SCOP.

PDB	SCOP	PDB	SCOP	PDB	SCOP	PDB	SCOP
1a2o	c.40.1.1	1dea	c.35.1.1	1gcb	d.3.1.1	1tr2	c.48.1.1
1a2z	c.56.4.1	1dhn	d.96.1.3	1glq	a.45.1.1	1ugi	d.17.5.1
1a44	b.17.1.1	1djn	c.3.1.1	1gsa	d.142.1.1	1unk	a.28.2.1
1a53	c.1.2.4	1dl1	a.100.1.4	1gvp	b.40.4.7	1uok	b.71.1.1
1a73	d.4.1.3	1dl2	c.26.3.1	1hcv	b.1.1.1	1utg	a.101.1.1
1a7v	a.24.3.2	1dlm	b.3.6.1	1hnj	c.95.1.1	1vfr	d.90.1.1
1a8b	a.65.1.1	1dmu	c.52.1.4	1hoe	b.5.1.1	1vpi	a.133.1.2
1a8p	b.43.1.1	1dnp	c.28.1.1	1ifc	b.60.1.2	1whi	b.39.1.1
1aj2	c.1.21.1	1doz	c.92.1.1	1isu	g.35.1.1	1xva	c.66.1.5
1ako	d.151.1.1	1dpt	d.80.1.3	1jac	b.77.3.1	1ypr	d.110.1.1
1amm	b.11.1.1	1dqi	b.1.13.1	1jcv	b.1.8.1	1zin	g.41.2.1
1ao1	a.110.1.1	1dqn	c.61.1.1	1kuh	d.92.1.1	2abk	a.96.1.1
1ao2	d.152.1.1	1duv	c.78.1.1	1lcl	b.29.1.3	2acy	d.58.10.1
1aoe	c.71.1.1	1dxe	c.1.12.5	1lop	b.62.1.1	2bnh	c.10.1.1
1apx	a.93.1.1	1dzr	b.82.1.1	1mat	d.127.1.1	2cev	c.42.1.1
1az9	c.55.2.1	1e0s	c.37.1.8	1mka	d.38.1.2	2cga	b.47.1.2
1azp	d.9.2.1	1e11	b.72.2.1	1mla	d.58.23.1	2end	a.18.1.1
1b16	c.2.1.2	1e12	d.26.3.1	1mn1	a.2.11.1	2fdn	d.58.1.1
1b6t	e.26.1.3	1e19	c.73.1.1	1mn2	d.44.1.1	2hvm	c.1.8.5
1bf6	c.1.9.3	1e2a	a.7.2.1	1mug	c.18.1.2	2mhr	a.24.4.1
1bgv	c.58.1.1	1e7f	a.126.1.1	1nhk	d.58.6.1	2mnr	d.54.1.1
1bm9	a.4.5.7	1ebf	d.81.1.2	1npx	d.87.1.1	2nsy	c.26.2.1
1br6	d.165.1.1	1ecp	c.56.2.1	1one	c.1.11.1	2ovo	g.15.1.1
1bs0	c.67.1.4	1emy	a.1.1.2	1oyc	c.1.4.1	2pi1	c.25.1.2
1bs4	d.167.1.1	1eqo	d.58.30.1	1pbe	d.16.1.2	2pii	d.58.5.1
1bv1	d.129.3.1	1ert	c.47.1.1	1pfk	c.89.1.1	2pth	c.56.3.1
1bxo	b.50.1.2	1ew4	d.82.2.1	1php	c.86.1.1	2rn2	c.55.3.1
1bxy	d.59.1.1	1ex1	b.43.3.1	1plq	d.131.1.2	2sn3	g.3.7.1
1byf	d.169.1.1	1ex2	b.44.1.1	1poh	d.94.1.1	3chy	c.23.1.1
1c1k	a.120.1.1	1f4t	a.104.1.1	1qd9	d.79.1.1	3cla	c.43.1.1
1c3q	c.72.1.2	1f51	d.128.1.1	1qex	b.32.1.1	3lzm	d.2.1.3
1c7q	c.80.1.2	1f52	d.15.9.1	1qgx	e.7.1.1	3pm1	c.84.1.1
1c9h	d.26.1.1	1f8y	c.23.14.1	1qh4	a.83.1.1	3pm2	d.129.2.1
1cfz	c.56.1.1	1flj	b.74.1.1	1qh5	d.157.1.2	3pyp	d.110.3.1
1cmb	a.43.1.2	1flm	b.45.1.1	1qkj	c.87.1.1	3tdt	b.81.1.2
1cnu	d.109.1.2	1fp3	a.102.1.3	1qp1	c.1.17.1	4icb	a.39.1.1
1cnz	c.77.1.1	1fq0	c.1.10.1	1qp2	d.41.2.1	4tms	d.117.1.1
1coz	c.26.1.2	1fqt	b.33.1.1	1qto	d.32.1.2	5mdh	d.162.1.1
1ctj	a.3.1.1	1frb	c.1.7.1	1qtw	c.1.15.1	5pnt	c.44.1.1
1ctt	c.97.1.1	1ftr	d.58.33.1	1rb9	g.41.5.1	6ins	g.1.1.1
1cuo	b.6.1.1	1g3k	d.153.1.4	1rcf	c.23.5.1	7rsa	d.5.1.1
1czj	a.138.1.1	1g51	d.104.1.1	1tgx	g.7.1.1	7tim	c.1.1.1
1czp	d.15.4.1	1g52	d.74.4.1	1tib	c.69.1.17	8tl1	a.67.1.1
1d0d	g.8.1.2	1g71	e.16.1.1	1tr1	c.36.1.2	9rnt	d.1.1.1

Tableau 13 : Liste des 176 domaines. Chaque colonne contient le code PDB et le code SCOP.

3 - Matrices de contact

Nous avons déjà vu que les matrices de contact établies à l'aide des faces de contact des cellules de Voronoï permettent de visualiser aisément les proximités entre les AA. La Figure 93 qui représente une de ces matrices, a été établie selon la procédure classique de tessellation : la protéine est installée dans son environnement qui est relaxé neuf fois, les points de représentation sont les CGL (centres géométriques des chaînes latérales). Sur cette matrice, on constate la présence de points dans les coins supérieur gauche et inférieur droit.

Ces points sont les témoins de contacts entre d'une part des AA situés à l'extrémité N-terminale et d'autre part des AA à l'extrémité C-terminale, tel que celui représenté par la face de contact entre les deux cellules de la Figure 94.

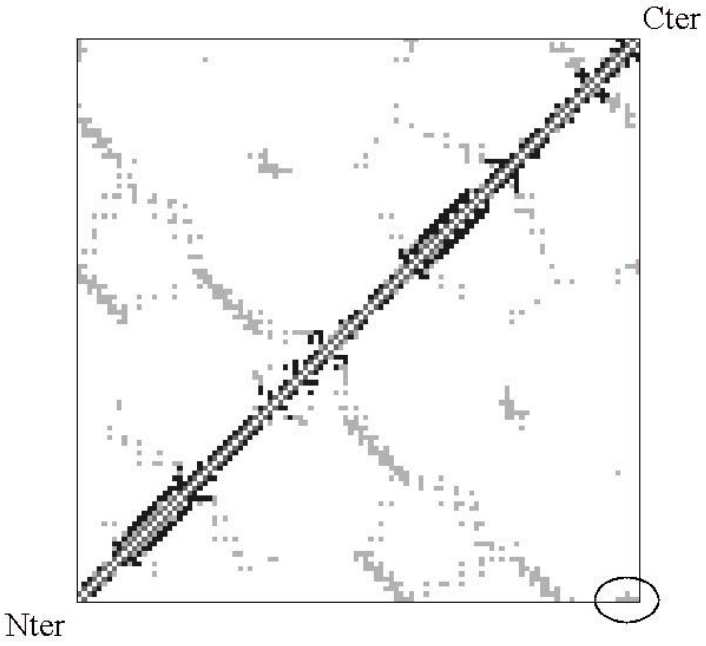


Figure 93 : Matrice de contact d'une protéine de transduction du signal d'*Escherichia coli* (code PDB : 2pii[98], 112 AA). L'ellipse signale les points symbolisant les contacts entre les deux extrémités de la chaîne polypeptidique.

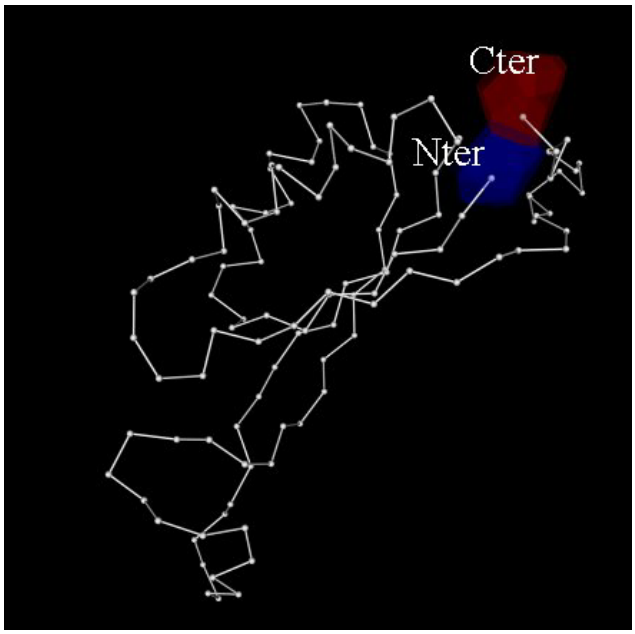


Figure 94 : Représentation des Cα de la même protéine que celle de la Figure 93. Les cellules du premier et du dernier résidu sont représentées, on constate qu'elles ont une face en commun donnant ainsi naissance à deux points symétriques sur la matrice (un des points signalés pas l'ellipse).

Pour tenter de comprendre ce qui se passait au niveau des extrémités des 177 chaînes protéiques de la banque, j'ai utilisé ces matrices en les additionnant toutes, afin d'obtenir une matrice globale. Cette opération simple *a priori* nécessite toutefois quelques précautions.

3.1 Pseudo-normalisation

Les 177 protéines de la banque ont chacune un nombre d'AA différent, ce qui implique nécessairement que les 177 matrices résultantes ont des tailles différentes. Ceci pose bien sûr un problème lorsque l'on veut procéder à leur addition. La solution consiste alors à normaliser les longueurs des séquences protéiques entre 0 et 100%. Avec cette convention, le premier AA de la chaîne protéique sera situé à 0% et le dernier sera situé à 100%. Cette opération que j'appelle une pseudo-normalisation, permet d'obtenir des matrices 100×100 qui ont donc toutes la même taille et avec lesquelles il devient alors possible de procéder à des additions terme à terme. Un point sur une de ces matrices signifie qu'il existe au moins un contact entre deux segments de la séquence protéique, chaque segment représentant 1 % de la totalité de la longueur de la séquence. La taille de ces segments dépend bien sûr du nombre d'AA de la protéine considérée, par exemple pour une protéine de 400 AA, un point sur la matrice pseudo-normalisée correspondante signifiera qu'il existe un contact entre deux segments de 4 AA. Avec cette méthode, le nombre de points présents sur la matrice ne sera plus proportionnel au véritable nombre de contacts présents dans la protéine.

3.2 Somme des matrices

L'addition des 177 matrices de contact donne la matrice 100×100, symétrique, présentée sur la Figure 95. Sur cette matrice le nombre de contacts est représenté par une couleur dont l'échelle va du blanc (pour un minimum de contacts) au rouge vif (pour un maximum) en passant par le noir. Pour alléger la représentation, les contacts triviaux dus à la proximités des AA le long de la structure primaire ne sont pas représentés (diagonale principale laissée en blanc). Malgré cela, on observe toujours, comme l'on pouvait s'y attendre, un très grand nombre de contacts au voisinage de cette diagonale, avec un certain nombre de taches rouge vif indiquant que ces contacts sont présents pour toutes ou quasiment toutes les protéines de la banque.

De manière beaucoup plus intéressante, la présence de pics noirs et rouges dans les coins supérieur gauche et inférieur droit de cette matrice semble indiquer que, dans notre banque, il existe un grand nombre de contacts entre les extrémités des chaînes

polypeptidiques. Cette observation nous a conduit à revoir la notion d'extrémités puisque les contacts observés ne se font pas entre les extrémités strictes des chaînes protéiques, c'est à dire entre les premiers et les derniers AA des séquences, mais entre leurs segments initiaux et terminaux. Il est facile de constater que typiquement les pics apparaissent entre le segment initial (0% et 15%) et le segment terminal (85% et 100%), dont trois îlots forts en (2%, 93%), (2%, 99%) et (8%, 96%). A l'opposé, la région (25%, 75%) est très pauvre en contact.

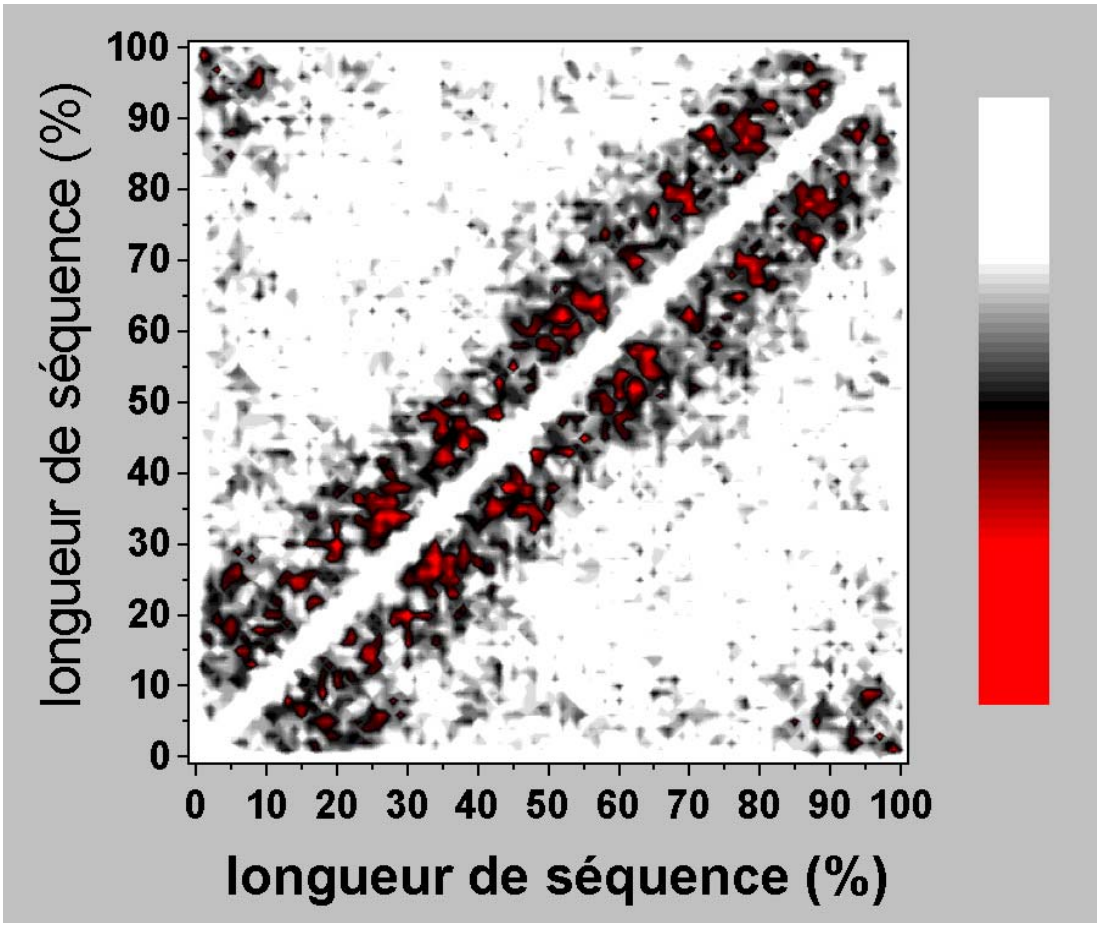


Figure 95 : Somme des 177 matrices de la banque de structures protéiques. L'échelle va du blanc au rouge vif pour un maximum de contacts en passant par le noir. Les contacts triviaux le long de la structure primaire et devant apparaître sur la diagonale principale ne sont pas indiqués pour alléger la représentation.

4 - Résultats quantitatifs

La suite logique de la démarche entreprise est donc d'essayer d'évaluer la proportion de structures protéiques faisant des contacts entre extrémités, mais comme nous venons de le voir, cette notion d'extrémité est un paramètre déterminant dont il est nécessaire de tenir compte de manière précise.

4.1 Notion d'écart en séquence normalisé

Cette notion est un complément à la normalisation de la séquence présentée au paragraphe précédent. Supposons que les cellules de Voronoï des résidus n°i et n°j ont une face en commun, le nombre d'AA séparant ces deux résidus est égal à | j – i + 1 |. L'écart en séquence normalisé que je noterai dans la suite Δ et qui s'exprime en % est égal à 100 × | j – i + 1 | / L, avec L représentant la longueur de la séquence étudiée (et non la longueur de la structure). Avec cette définition on obtient une valeur intuitive associée à l'écart considéré, par exemple si le premier AA d'une protéine est en contact avec le dernier, si on a donc un contact entre extrémités strictes, la valeur de Δ sera simplement de 100%.

4.2 Premiers résultats

C'est cet écart en séquence normalisé (Δ) que l'on retrouve en abscisse du graphe de la Figure 96. L'ordonnée de ce graphe représente le pourcentage de structures protéiques faisant au moins un contact avec un écart égal ou supérieur à Δ. Par exemple on peut constater qu'un peu moins de 10% (9.6%) des structures protéiques de la banque font au moins un contact avec un écart représentant 100% de la longueur totale de la séquence, en d'autres termes, 9.6% des chaînes de la banque ont un contact entre leurs extrémités strictes.

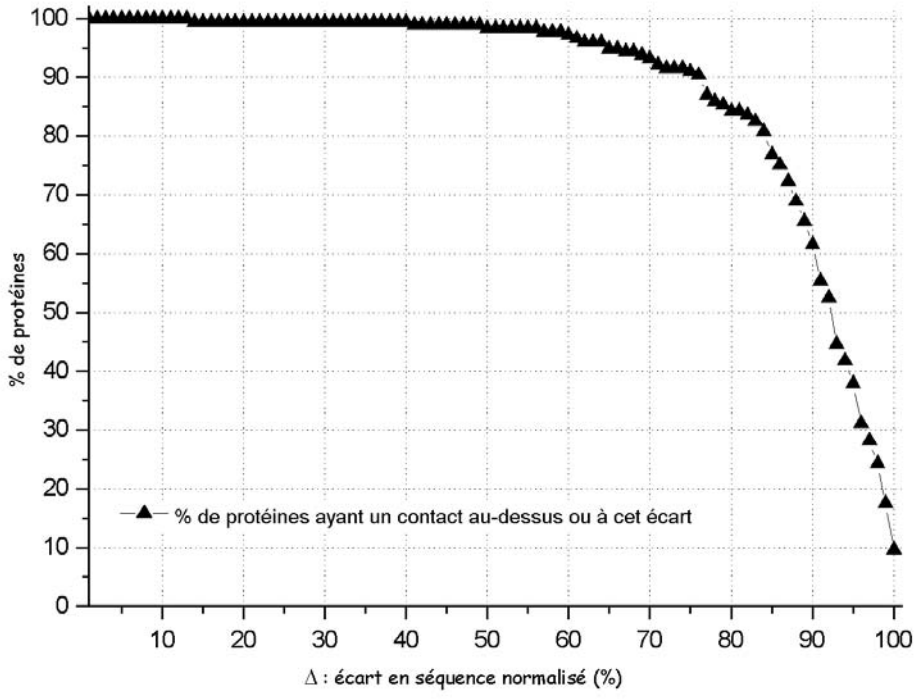


Figure 96 : Proportion de structures protéiques de la banque faisant un contact associé à un écart égal ou supérieur à Δ.

PDB	A	B	C	B	C	D	E	F
2bnh	456	Q	143	R	200	57	456	12.7
1qex	288	Q	170	K	283	113	288	39.6
1e7f	582	L	196	R	482	286	585	49.1
1trk	678	D	54	G	430	376	679	55.5
8tln	316	Y	84	Y	268	184	316	58.5
1a2o	347	I	102	V	310	208	349	59.9
3pmg	561	T	18	K	359	341	561	61.0
3tdt	274	D	274	Q	99	175	274	64.2
1dzr	183	F	64	L	181	117	183	64.5
1a8l	226	H	34	E	183	149	226	66.4
1qh4	380	H	19	S	192	173	256	68.0
1c8u	285	Q	32	T	227	195	285	68.8
1e2a	102	L	28	Y	100	72	105	69.5
1a73	162	P	29	P	142	113	163	69.9
1qgx	354	V	69	R	323	254	357	71.4
1whi	122	A	11	G	100	89	122	73.8
1cnu	134	V	4	K	106	102	137	75.2
1exm	403	A	84	G	390	306	406	75.6
1dlm	309	G	70	P	305	235	311	75.9
6ins	50	L	11	Y	48	37	50	76.0
1dhn	121	M	1	M	92	91	121	76.0
5mdh	333	I	4	W	257	253	333	76.3
1amm	174	G	40	D	172	132	174	76.4
1vpi	122	Y	27	E	120	93	122	77.0
1isu	62	G	1	Q	48	47	62	77.4
1dnp	469	Q	9	D	374	365	472	77.5
1azp	66	V	2	E	53	51	66	78.8
1dli	402	Y	10	R	327	317	402	79.1
1qh5	260	V	3	P	212	209	260	80.8
1ert	105	V	105	D	20	85	105	81.9
1a53	247	L	247	I	44	203	248	82.3
2pia	321	T	1	R	267	266	322	82.9
1qd9	124	I	21	V	123	102	124	83.1
1cqx	403	R	59	Y	393	334	403	83.1
1mka	171	K	24	Q	166	142	171	83.6
1ew4	106	I	17	F	105	88	106	84.0
1e15	496	R	2	N	421	419	499	84.2
1aoz	552	W	71	A	535	464	552	84.2
1gcb	452	A	452	G	70	382	454	84.4
2cga	245	G	2	T	208	206	245	84.5
1cuo	129	M	13	M	121	108	129	84.5
1a8p	257	G	34	E	251	217	257	84.8
2acy	98	A	1	V	85	84	100	85.0
1ctt	294	H	2	G	251	249	294	85.0
1uok	558	M	1	Y	479	478	558	85.8
1qpo	284	M	284	T	40	244	285	86.0
1br6	267	A	37	Q	267	230	268	86.2
1mng	203	L	16	V	190	174	203	86.2
3cla	213	R	13	H	196	183	213	86.4
1pbe	391	E	49	Y	390	341	394	86.8
1bm9	120	E	1	D	106	105	122	86.9
1qto	122	L	11	V	116	105	122	86.9
1feh	574	L	20	E	518	498	574	86.9
1pfk	319	M	1	G	279	278	320	87.2
1dxe	253	F	7	F	230	223	256	87.5
1coz	126	D	11	K	123	112	129	87.6
2sn3	65	K	1	T	57	56	65	87.7
1czp	98	E	10	E	95	85	98	87.8
1c9h	107	I	7	D	100	93	107	87.9

PDB	A	B	C	B	C	D	E	F
1c7q	442	I	2	F	393	391	445	88.1
1dmu	299	E	28	R	291	263	299	88.3
1aoe	192	L	13	F	182	169	192	88.5
1frb	315	W	20	C	298	278	315	88.6
1a44	185	V	2	T	166	164	186	88.7
1bs0	383	R	31	D	371	340	384	88.8
2rn2	155	N	16	V	153	137	155	89.0
2ovo	56	A	2	S	51	49	56	89.3
1b6t	157	P	11	L	152	141	159	89.3
1emy	153	G	1	L	137	136	153	89.5
1unk	87	I	7	F	84	77	87	89.7
1pjc	361	M	1	A	324	323	361	89.8
1b16	254	M	1	N	228	227	254	89.8
1plq	258	L	11	K	242	231	258	89.9
1bv1	159	N	159	A	16	143	160	90.0
1pii	452	Y	452	A	46	406	452	90.0
1bxo	323	D	14	K	304	290	323	90.1
1one	436	S	13	R	405	392	436	90.1
1mat	263	I	10	L	247	237	264	90.2
1a7v	125	K	9	R	121	112	125	90.4
2end	137	A	12	Y	136	124	138	90.6
1poh	85	V	6	A	82	76	85	90.6
1bgv	449	V	28	F	435	407	449	90.9
1ctj	89	A	2	D	82	80	89	91.0
1dqi	124	I	6	E	118	112	124	91.1
2mhr	118	W	2	G	109	107	118	91.5
1f8y	156	P	1	N	144	143	157	91.7
1eqo	158	W	158	P	14	144	158	91.8
1czj	110	Q	110	T	9	101	111	91.9
1drw	272	N	4	S	254	250	273	91.9
1dfx	125	V	6	H	120	114	125	92.0
1cnz	363	V	26	V	359	333	363	92.0
1fq0	213	P	16	A	211	195	213	92.0
1hcv	112	V	1	Q	104	103	113	92.0
1vfr	217	L	8	R	207	199	217	92.2
1mpg	282	W	14	Y	273	259	282	92.2
1mug	165	V	165	R	11	154	168	92.3
2pth	193	T	1	G	179	178	194	92.3
1dpt	117	F	2	K	109	107	117	92.3
2nsy	271	M	2	Y	252	250	271	92.6
1flj	259	Y	6	P	245	239	259	92.7
4tms	316	V	316	T	24	292	316	92.7
7rsa	124	A	4	V	118	114	124	92.7
1ugi	83	L	83	I	6	77	84	92.9
1lcl	141	S	1	T	132	131	141	93.6
1e0s	173	I	7	T	169	162	174	93.7
1ecp	237	D	14	D	236	222	238	93.7
1cmb	104	G	5	W	102	97	104	94.2
1g3k	173	P	173	G	10	163	174	94.3
3pyp	125	F	6	K	123	117	125	94.4
1nhk	143	R	4	S	139	135	144	94.4
1aor	605	K	30	I	601	571	605	94.5
1hoe	74	L	74	S	5	69	74	94.6
1aj2	282	S	13	K	279	266	282	94.7
1glq	209	F	8	G	205	197	209	94.7
1jac	133	D	5	Y	130	125	133	94.7
1mla	305	F	4	L	295	291	308	94.8
1tib	269	F	7	F	262	255	270	94.8
1a2z	220	K	2	E	210	208	220	95.0

PDB	A	B	C	B	C	D	E	F
1gpe	587	D	26	Y	583	557	587	95.1
1cfz	162	M	1	V	154	153	162	95.1
1c1k	217	M	1	V	207	206	217	95.4
1fqt	109	M	1	A	104	103	109	95.4
1dqn	230	V	5	I	224	219	230	95.7
1e19	313	K	2	L	303	301	314	96.2
1gdh	320	K	2	F	310	308	321	96.3
1bs4	168	Q	4	K	165	161	168	96.4
1f4t	367	M	1	N	355	354	368	96.5
1gvp	87	K	3	A	86	83	87	96.6
2mnr	357	R	10	L	356	346	359	96.7
1tgx	60	N	60	C	3	57	60	96.7
2abk	211	M	1	E	204	203	211	96.7
1flm	122	M	1	A	118	117	122	96.7
1iov	306	I	5	L	302	297	306	97.4
1a8b	318	T	5	G	314	309	318	97.5
1ee2	373	F	373	K	10	363	373	97.6
1g71	344	L	2	W	340	338	347	97.7
1qtw	285	M	1	Q	279	278	285	97.9
1xgs	295	T	3	I	291	288	295	98.0
2cev	298	K	1	F	293	292	299	98.0
1rb9	52	A	52	K	2	50	52	98.1
5pnt	157	H	157	A	4	153	157	98.1
1dea	266	R	2	N	262	260	266	98.1
1d0d	60	Y	1	C	59	58	60	98.3
7tim	247	F	4	R	246	242	247	98.4
1doz	309	K	3	L	307	304	310	98.4
1byf	123	D	123	D	1	122	125	98.4
1c3q	272	M	1	R	268	267	272	98.5
4icb	75	M	1	S	75	74	76	98.7

PDB	A	B	C	B	C	D	E	F
3lzm	164	M	1	K	162	161	164	98.8
1xva	292	S	3	T	291	288	292	99.0
9rnt	104	T	104	C	2	102	104	99.0
1gsa	314	I	2	L	313	311	315	99.0
1hnj	317	F	317	K	4	313	317	99.1
1zin	217	M	1	L	215	214	217	99.1
1duv	333	K	333	Y	4	329	333	99.1
1ypr	125	Y	125	S	1	124	126	99.2
1ifc	131	A	1	K	130	129	131	99.2
1kuh	132	G	132	V	2	130	132	99.2
1rcf	169	L	169	K	2	167	169	99.4
1oyc	399	S	1	D	397	396	399	99.5
2hgs	472	V	472	T	1	471	474	99.6
1bf6	291	Q	291	S	1	290	292	99.7
1ebf	358	L	358	S	1	357	359	99.7
1php	394	K	394	N	2	392	394	99.7
1fp3	402	M	1	L	401	400	402	99.8
1ako	268	R	268	M	1	267	268	100.0
1apx	249	A	249	G	1	248	249	100.0
1bxy	60	E	60	M	1	59	60	100.0
1ftr	296	F	296	M	1	295	296	100.0
1jcv	153	N	153	V	1	152	153	100.0
1lop	164	E	164	M	1	163	164	100.0
1qkj	351	L	351	M	1	350	351	100.0
1utg	70	M	70	G	1	69	70	100.0
2fdn	55	A	55	A	1	54	55	100.0
2hvm	273	V	273	G	1	272	273	100.0
2pii	112	I	112	M	1	111	112	100.0
3chy	128	M	128	A	1	127	128	100.0

Tableau 14 : Récapitulatif des résultats pour les structures protéiques et les plus lointains contacts directs, ordonnés dans le sens croissant. Colonne PDB : code PDB des structures protéiques ; A nombre d'AA présents dans la structure ; B code à une lettre de l'AA intervenant dans le contact ; C position le long de la séquence ; D différence entre les deux positions ; E nombre d'AA dans la séquence ; F écart en séquence normalisé (Δ).

De la même façon on constate que 61.6% des structures de la banque ont au moins un contact avec un écart en séquence normalisé Δ égal ou supérieur à 90% de la longueur totale de la séquence. Précisons qu'un écart normalisé à 90%, de par sa définition même, peut représenter indifféremment des contacts entre un résidu à 0% (le premier AA) et un autre à 90% de la séquence mais également entre un résidu à 10% et 100% de la séquence (le dernier AA) ou bien encore entre un résidu à 5% et un autre à 95%. De plus, j'ai indiqué, dans le paragraphe décrivant la démarche que j'ai utilisée pour constituer cette banque, que pour certaines structures retenues, il pouvait manquer jusqu'à quatre résidus aux extrémités. Dans ces cas précis, même si le premier et le dernier résidu de la chaîne sont en contact, l'écart en séquence normalisé n'est jamais égal à 100%, la longueur totale retenue est celle de la

séquence et non pas celle de la structure correspondante. Le Tableau 14 donne pour chaque structure protéique de la banque le contact observé avec le plus grand écart en séquence normalisé.

4.3 Contacts du second ordre

Un des avantages des TdV est qu'elles permettent de s'affranchir de tout cut-off géométrique et en particulier de ceux liés aux distances entre résidus. Cependant il existe des cas pour lesquels la proximité tridimensionnelle de deux résidus ne se traduit pas par un contact, la Figure 97 en montre un exemple. Cette structure contient 129 AA et on peut constater que l'alanine n°1 (A1) et l'acide glutamique n°128 (E128) ($\Delta = 99.2\%$) sont très proches. Pourtant le contact avec le plus grand écart est celui intervenant entre les méthionines n°13 (M13) et n°121 (M121) ($\Delta = 84.5\%$). En fait, les cellules associées à l'alanine n°1 (A1) et l'acide glutamique n°128 (E128) ne partagent pas de face car la cellule associée à la proline n°23 (P23) vient s'insérer entre elles. Il nous a donc semblé judicieux de ne plus considérer uniquement des contacts directs entre cellules mais de considérer également ce que j'appellerai dans la suite des contacts du second ordre, c'est à dire des voisins de voisins, afin que ce type de proximités tridimensionnelles (et non plus de véritables contacts) soit détectable toujours sans utiliser de distances. Ainsi on pourra considérer que deux résidus sont proches si leurs cellules associées sont en contact direct avec une troisième et même cellule.

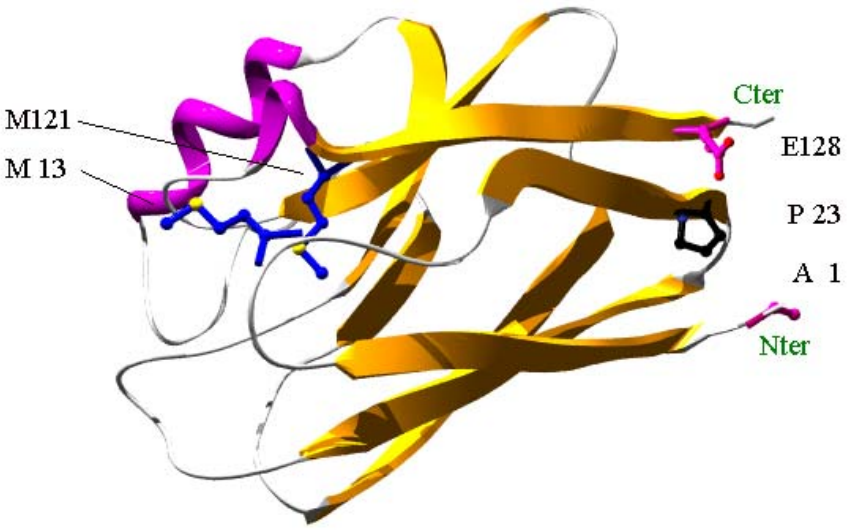


Figure 97 : Exemple de proximité entre extrémités (E128 et P23) non détectée par contact direct (1cuo[99], 129 AA).

4.4 Nouveaux résultats

Il est alors possible avec cette nouvelle définition de reprendre le graphe de la Figure 96, cette nouvelle courbe est donnée dans la Figure 98.

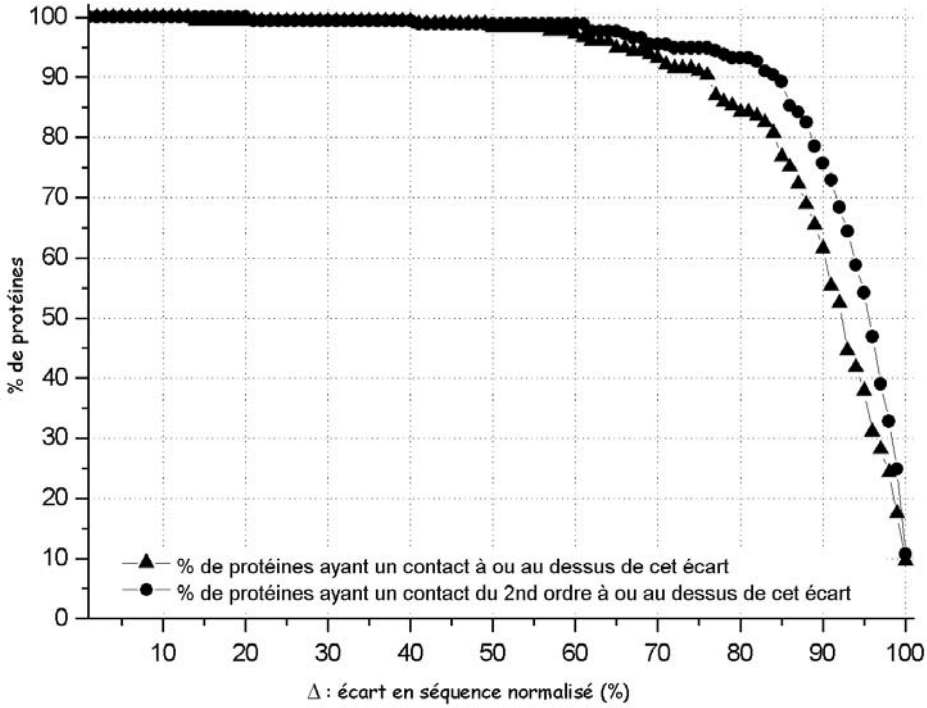


Figure 98 : Proportion de structures protéiques de la banque faisant un contact associé à un écart égal ou supérieur à Δ. Les contacts normaux sont représentés pas un triangle et reprennent les valeurs de la Figure 96, les contacts du 2nd ordre sont représentés par des disques.

On constate que lorsque l'on considère des contacts du second ordre, les proportions augmentent de manière non négligeable ; par exemple pour $\Delta = 90\%$, la proportion de protéines passe de 61.6% à 75.5%, soit un gain de 13.9%. Ce gain est maximum pour $\Delta = 93\%$ car dans ce cas on passe de 44.6% à 64.4% des protéines de la banque. Dans le cas des contacts directs, il faut considérer une valeur Δ de 92% pour que la proportion de structures protéiques faisant au moins un contact avec au moins un tel écart soit supérieure à 50%. Dans le cas des contacts du second ordre, cette proportion est déjà dépassée avec $\Delta = 95\%$.

4.5 Domaines protéiques

Dans les premières lignes du Tableau 14, on constate que certaines structures protéiques ont des scores qui semblent montrer que leurs extrémités sont relativement éloignées. Si l'on étudie par exemple la structure de code PDB 1a2o associée à une valeur Δ proche de 60% et

que l'on regarde sa structure il est facile de comprendre pourquoi ses extrémités ne sont pas proches. La structure de cette protéine est présentée Figure 99, l'extrémité N-terminale se situe à gauche prés de la lysine n°3 (K3) et l'extrémité C-terminale se situe quant à elle près de l'alanine n°346 (A346). Il est indéniable que ces deux extrémités sont on ne peut plus éloignées, le contact avec le plus grand écart normalisé de cette protéine est réalisé comme le précise le Tableau 14, entre l'isoleucine n°102 (I102) et la valine n°310 (V310) qui sont représentées en vert sur la figure. Il est facile de remarquer sur cette même figure que cette protéine est divisée en deux domaines structuraux, le premier à gauche est représenté en rouge et bleu, alors que le second à droite est représenté en jaune et violet. La frontière entre ces deux domaines se situe au niveau de la méthionine n°141 (M141 en noir sur la figure). Si l'on considère ces deux domaines séparément et que l'on cherche le contact avec le plus grand écart en séquence normalisé pour chacun d'eux on obtient les contacts suivants : pour le domaine n°1, entre la lysine n°3 (K3) et l'arginine n°132 (R132) représentées en orange et pour le domaine n°2, entre la sérine n°154 (S154) et l'alanine n°346 (A346). Pour ces deux contacts, les écarts en séquence normalisés sont respectivement d'à peu près 92% et 93% (en considérant la longueur totale de chaque domaine).

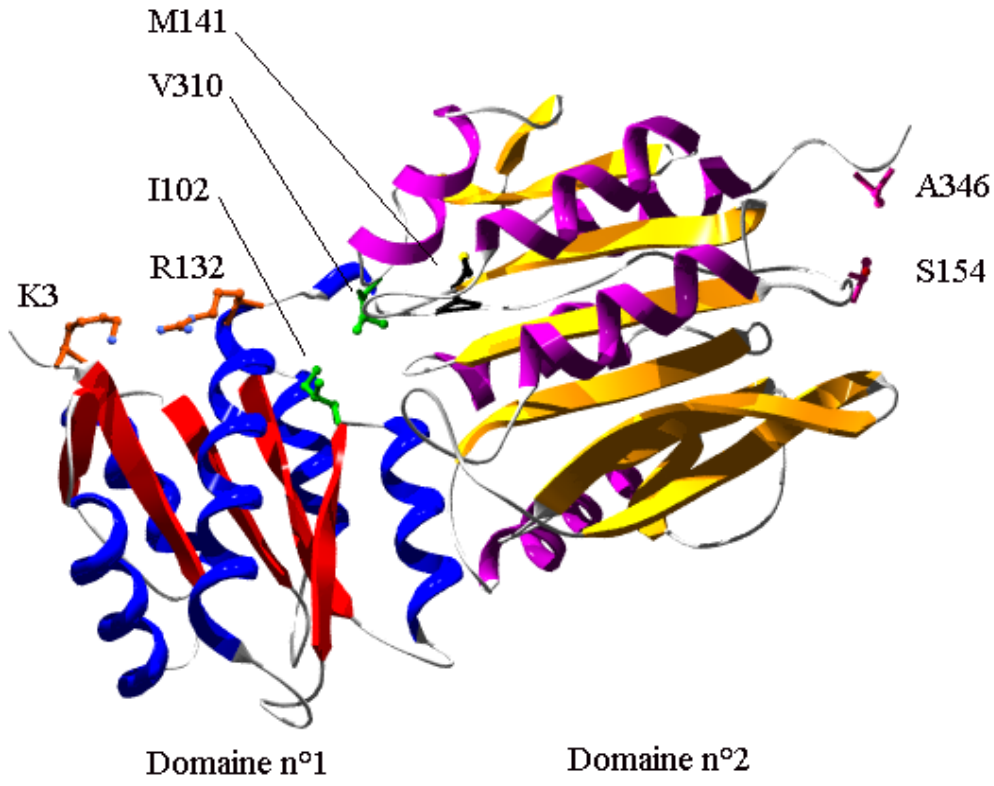


Figure 99 : Structure 1a2o[100], 347 AA. Ces structures présentent deux domaines : le domaine n°1 à gauche est représenté en rouge et bleu et le domaine n°2 est représenté en jaune et violet.

L'étude de cette structure indiquerait donc que la proximité des extrémités N- et C-terminales pourrait également intervenir au sein même des domaines protéiques et non uniquement à l'échelle des structures complètes.

4.6 Résultats pour les domaines

Nous avons donc utilisé la banque de 176 domaines protéiques décrite précédemment pour obtenir les graphes du type de ceux présentés Figure 96 et Figure 98. Ces courbes sont présentées Figure 100 pour les contacts directs et les contacts du second ordre.

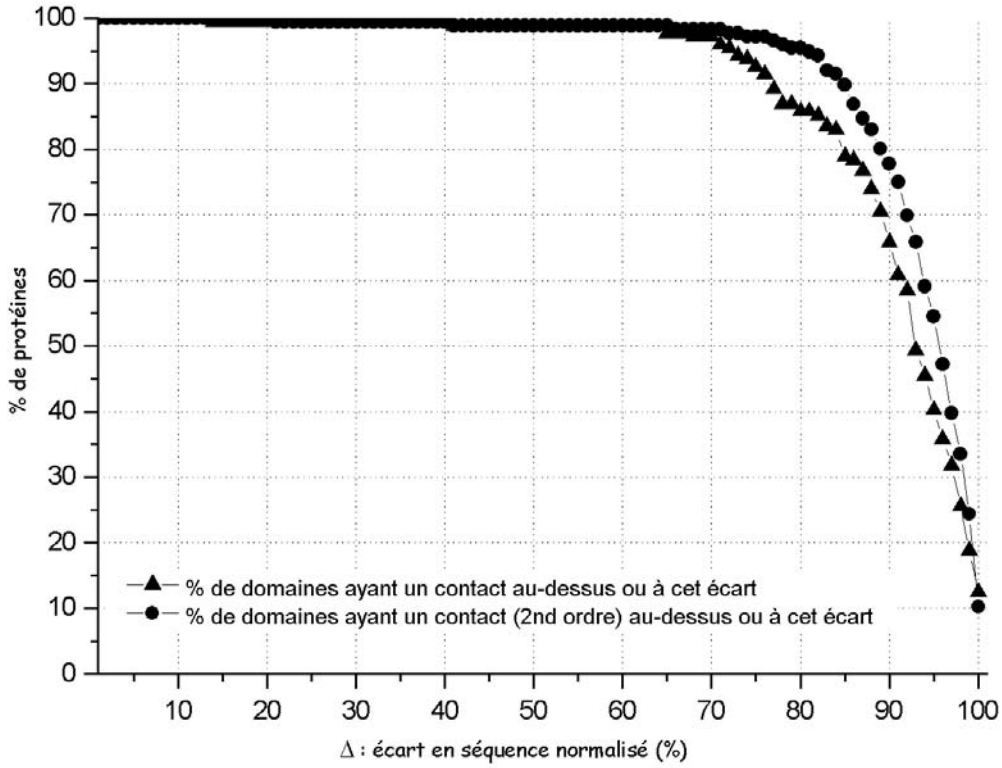


Figure 100 : Proportion de domaines protéiques de la banque faisant un contact (direct ou du second ordre) associé à un écart égal ou supérieur à Δ.

On constate que ce graphe présente le même aspect que celui de la Figure 98. Si l'on prend pour référence une valeur de Δ égale à 90%, on obtient alors pour les contacts directs une proportion de 65.9% (au lieu de 61.6% pour les protéines) et pour les contacts du second ordre, on obtient une proportion de 77.8% (au lieu de 75.7%). On voit donc bien que les scores obtenus sont meilleurs lorsque l'on considère des domaines et comme pour les structures entières, on peut constater que ce ne sont toujours pas les extrémités strictes qu'il faut considérer mais bien les segments initiaux et terminaux. La liste des contacts avec les plus grands écarts est présentée dans le Tableau 15 pour chaque domaine protéique.

PDB	A	B	A	B	C	D	E
2bnh	Q	143	R	200	57	456	12.7
1qex	Q	170	K	283	113	288	39.6
3tdt	D	274	Q	99	175	274	64.2
1dzr	F	64	L	181	117	183	64.5
1qh4	H	28	R	95	67	102	66.7
1e2a	L	28	Y	100	72	105	69.5
1a73	P	29	P	142	113	163	69.9
1qgx	V	69	R	323	254	357	71.4
1g51	F	4	M	122	118	166	71.7
1e7f	N	16	K	157	141	197	72.1
1bgv	K	38	T	142	104	143	73.4
1whi	A	11	G	100	89	122	73.8
1npx	K	33	E	125	92	126	73.8
1tr2	D	22	I	130	108	146	74.7
1cnu	V	4	K	106	102	137	75.2
1dlm	G	70	P	305	235	311	75.9
6ins	L	11	Y	48	37	50	76.0
1dhn	M	1	M	92	91	121	76.0
8tll	I	1	T	123	122	161	76.4
1dll	T	13	I	87	74	98	76.5
1vpi	Y	27	E	120	93	122	77.0
5mdh	N	31	E	168	137	179	77.1
1isu	G	1	Q	48	47	62	77.4
1azp	V	2	E	53	51	66	78.8
1ao1	E	42	P	354	312	395	79.2
1qh5	V	3	P	212	209	260	80.8
1glq	Q	131	Y	25	106	131	81.7
1ert	V	105	D	20	85	105	81.9
1a53	L	247	I	44	203	248	82.3
1qd9	I	21	V	123	102	124	83.1
1mka	K	24	Q	166	142	171	83.6
1f52	V	4	I	85	81	98	83.7
1gsa	E	2	T	163	161	193	83.9
1ew4	I	17	F	105	88	106	84.0
1gcb	A	452	G	70	382	454	84.4
2cga	G	2	T	208	206	245	84.5
1cuo	M	13	M	121	108	129	84.5
2acy	A	1	V	85	84	100	85.0
1br6	A	37	Q	267	230	268	86.2
3cla	R	13	H	196	183	213	86.4
1bm9	E	1	D	106	105	122	86.9
1qto	L	11	V	116	105	122	86.9
1qp2	A	10	L	109	99	115	87.0
2mnr	V	2	V	116	114	132	87.1
1pfk	M	1	G	279	278	320	87.2
1dxe	F	7	F	230	223	256	87.5
1coz	D	11	K	123	112	129	87.6
2sn3	K	1	T	57	56	65	87.7
1czp	E	10	E	95	85	98	87.8
1c9h	I	7	D	100	93	107	87.9
1c7q	I	2	F	393	391	445	88.1
1dmu	E	28	R	291	263	299	88.3
1aoe	L	13	F	182	169	192	88.5
1frb	W	20	C	298	278	315	88.6
1a44	V	2	T	166	164	186	88.7
1bs0	R	31	D	371	340	384	88.8
2rn2	N	16	V	153	137	155	89.0
2ovo	A	2	S	51	49	56	89.3
1b6t	P	11	L	152	141	159	89.3
1emy	G	1	L	137	136	153	89.5
1unk	I	7	F	84	77	87	89.7
1b16	M	1	N	228	227	254	89.8
1bv1	N	159	A	16	143	160	90.0

PDB	A	B	A	B	C	D	E
1ex1	P	8	S	97	89	100	90.0
1bxo	D	14	K	304	290	323	90.1
1mat	I	10	L	247	237	264	90.2
1mn1	L	9	T	91	82	92	90.2
1a7v	K	9	R	121	112	125	90.4
2end	A	12	Y	136	124	138	90.6
1poh	V	6	A	82	76	85	90.6
1mn2	G	2	D	102	100	111	91.0
1ctj	A	2	D	82	80	89	91.0
1dqi	I	6	E	118	112	124	91.1
2mhr	W	2	G	109	107	118	91.5
1f8y	P	1	N	144	143	157	91.7
1eqo	W	158	P	14	144	158	91.8
1czj	Q	110	T	9	101	111	91.9
1a8p	N	2	R	92	90	99	91.9
1cnz	V	26	V	359	333	363	92.0
1fq0	P	16	A	211	195	213	92.0
1hcv	V	1	Q	104	103	113	92.0
1az9	R	11	M	172	161	176	92.0
1vfr	L	8	R	207	199	217	92.2
1mug	V	165	R	11	154	168	92.3
2pth	T	1	G	179	178	194	92.3
1dpt	F	2	K	109	107	117	92.3
1f51	R	2	A	341	339	368	92.4
1plq	I	2	F	123	121	132	92.4
1dl2	Q	3	R	102	99	108	92.6
2nsy	M	2	Y	252	250	271	92.6
1flj	Y	6	P	245	239	259	92.7
4tms	V	316	T	24	292	316	92.7
7rsa	A	4	V	118	114	124	92.7
1ugi	L	83	I	6	77	84	92.9
1a2o	S	154	A	346	192	206	93.2
1lcl	S	1	T	132	131	141	93.6
1e0s	I	7	T	169	162	174	93.7
1ecp	D	14	D	236	222	238	93.7
1dnp	E	204	L	11	193	207	93.7
1tr1	Y	12	Q	196	184	197	93.9
3pm1	D	4	W	113	109	117	94.0
1cmb	G	5	W	102	97	104	94.2
1g3k	P	173	G	10	163	174	94.3
3pyp	F	6	K	123	117	125	94.4
1nhk	R	4	S	139	135	144	94.4
1hoe	L	74	S	5	69	74	94.6
1aj2	S	13	K	279	266	282	94.7
1jac	D	5	Y	130	125	133	94.7
1tib	F	7	F	262	255	270	94.8
1a2z	K	2	E	210	208	220	95.0
1cfz	M	1	V	154	153	162	95.1
1c1k	M	1	V	207	206	217	95.4
1fqt	M	1	A	104	103	109	95.4
1dqn	V	5	I	224	219	230	95.7
1mla	S	70	G	1	69	73	95.9
1e11	L	2	R	51	49	52	96.2
1e19	K	2	L	303	301	314	96.2
1bs4	Q	4	K	165	161	168	96.4
1qp1	I	1	L	163	162	169	96.4
1f4t	M	1	N	355	354	368	96.5
1ctt	A	115	Y	5	110	115	96.5
1gvp	K	3	A	86	83	87	96.6
1amm	G	1	M	86	85	89	96.6
1tgx	N	60	C	3	57	60	96.7
2abk	M	1	E	204	203	211	96.7
1flm	M	1	A	118	117	122	96.7

PDB	A	B	A	B	C	D	E
1pbe	P	102	L	1	101	105	97.1
1zin	G	1	Q	34	33	35	97.1
1ebf	P	1	G	186	185	191	97.4
1a8b	T	5	G	314	309	318	97.5
1uok	N	1	K	83	82	85	97.6
1g71	L	2	W	340	338	347	97.7
3pm2	T	141	F	4	137	141	97.9
1qtw	M	1	Q	279	278	285	97.9
2cev	K	1	F	293	292	299	98.0
1rb9	A	52	K	2	50	52	98.1
5pnt	H	157	A	4	153	157	98.1
1dea	R	2	N	262	260	266	98.1
1d0d	Y	1	C	59	58	60	98.3
1duv	Q	1	Q	180	179	183	98.4
7tim	F	4	R	246	242	247	98.4
1doz	K	3	L	307	304	310	98.4
1byf	D	123	D	1	122	125	98.4
1c3q	M	1	R	268	267	272	98.5
4icb	M	1	S	75	74	76	98.7
3lzm	M	1	K	162	161	164	98.8
1xva	S	3	T	291	288	292	99.0
9rnt	T	104	C	2	102	104	99.0
1ypr	Y	125	S	1	124	126	99.2
1ifc	A	1	K	130	129	131	99.2
1kuh	G	132	V	2	130	132	99.2

PDB	A	B	A	B	C	D	E
1rcf	L	169	K	2	167	169	99.4
1hnj	I	174	Y	2	172	174	99.4
1oyc	S	1	D	397	396	399	99.5
1bf6	Q	291	S	1	290	292	99.7
1one	L	295	P	2	293	295	99.7
1php	K	394	N	2	392	394	99.7
1fp3	M	1	L	401	400	402	99.8
1ako	R	268	M	1	267	268	100.0
1ao2	P	210	M	1	209	210	100.0
1apx	A	249	G	1	248	249	100.0
1bxy	E	60	M	1	59	60	100.0
1djn	H	156	R	1	155	156	100.0
1e12	Y	88	Y	1	87	88	100.0
1ex2	E	93	H	1	92	93	100.0
1ftr	A	150	F	1	149	150	100.0
1g52	K	120	F	1	119	120	100.0
1jcv	N	153	V	1	152	153	100.0
1lop	E	164	M	1	163	164	100.0
1qkj	L	351	M	1	350	351	100.0
1utg	M	70	G	1	69	70	100.0
2fdn	A	55	A	1	54	55	100.0
2hvm	V	273	G	1	272	273	100.0
2pi1	E	120	E	1	119	120	100.0
2pii	I	112	M	1	111	112	100.0
3chy	M	128	A	1	127	128	100.0

Tableau 15
Récapitulatif des résultats pour les domaines protéiques et les contacts directs.
Colonne PDB : code PDB des domaines protéiques ;
A code à une lettre de l'AA intervenant dans le contact ; B position le long de la séquence ;
C différence entre les deux positions ; D nombre d'AA dans la séquence ;
E écart en séquence normalisé (Δ).

4.7 Monomères et multimères

Comme nous venons de le voir, le fait qu'une structure protéique soit composée de plusieurs domaines peut expliquer dans certains cas que les extrémités de la chaîne polypeptidique ne soient pas proches. Un autre argument peut être avancé pour expliquer ces observations. Si l'on considère par exemple la structure dont le code PDB est 1a73, la valeur Δ qui lui est associée est relativement mauvaise puisqu'elle est d'à peu près 69%. Cette structure monodomaine est présentée dans la Figure 101 sous sa forme biologiquement active, c'est à dire sous la forme d'un homodimère (sans l'ADN à laquelle cette protéine se lie). On constate effectivement que pour chaque chaîne les extrémités N- et C-terminales sont très éloignées les unes des autres mais il est par contre très intéressant de remarquer que

l'extrémité N-terminale de la chaîne A est relativement proche de l'extrémité C-terminale de la chaîne B. En effet il existe par exemple un contact entre la leucine n°9 de la chaîne A (A-L9) et l'arginine n°157 de la chaîne B (B-R157) et inversement, l'extrémité C-terminale de la chaîne A est proche de l'extrémité N-terminale de la chaîne B (contact entre A-R157 et B-L9). Ce phénomène de « swapping » ou d'échange, régulièrement observé dans les multimères[101-112], consiste à échanger des morceaux de structures (de quelques résidus à des domaines entiers) entre différentes chaînes tout en conservant globalement, pour chaque chaîne, la structure qu'elle adopte lorsqu'elle est seule. On voit donc ici que l'éloignement des extrémités de chaque chaîne est compensé par le rapprochement des extrémités entre chaînes. Dans le cas précis de la structure présentée ici, le « swapping » tend de plus à montrer que pour chaque chaîne on devrait observer une proximité des extrémités propre à chacun des deux éléments.

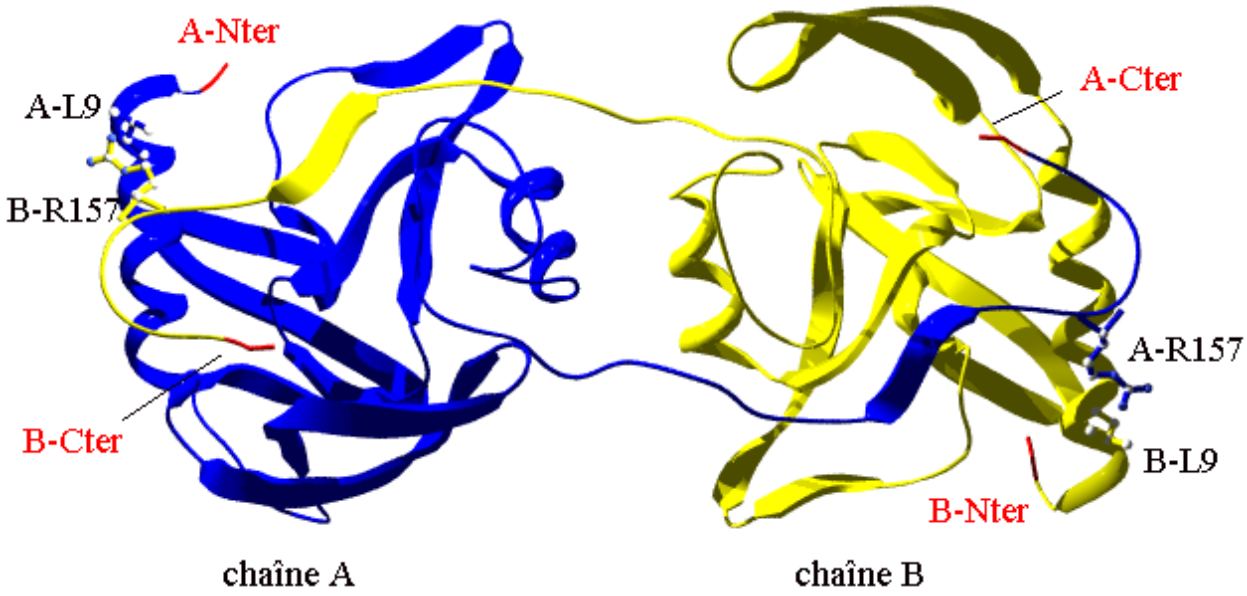


Figure 101 : Structure 1a73[113] sous forme dimérique. La chaîne A est en bleu et la chaîne B en jaune. Les extrémités N et C terminales sont en rouge.

Pour essayer d'aller un peu plus loin nous avons étudié ce qui se passait d'une part pour les protéines monomériques et d'autre part pour les protéines multimériques. On obtient ainsi la courbe présentée dans la Figure 102. On remarque sur ce graphe, pour les contacts directs ou du second ordre, et pour un écart en séquence normalisé Δ donné, que les proportions de protéines faisant au moins un contact avec un écart égal ou supérieur à cette valeur Δ sont

plus importantes pour les protéines monomériques que pour les protéines multimériques. Par exemple pour une valeur Δ représentant 90% de la longueur totale de la séquence la proportion pour les contacts directs sont de 58.9% pour les multimères alors que pour les monomères elle est de 74.7% soit 15.8% de plus. Pour les contacts du second ordre la proportion pour les multimères est de 74.4% alors qu'elle est de 83.1% pour les monomères soit une progression plus faible de 8.7%. On peut donc en conclure que dans notre banque les extrémités N- et C-terminales sont souvent plus proches pour les protéines monomériques que pour les protéines multimériques, pour certaines de ces dernières il semble que l'oligomérisation puisse suppléer à cet éloignement.

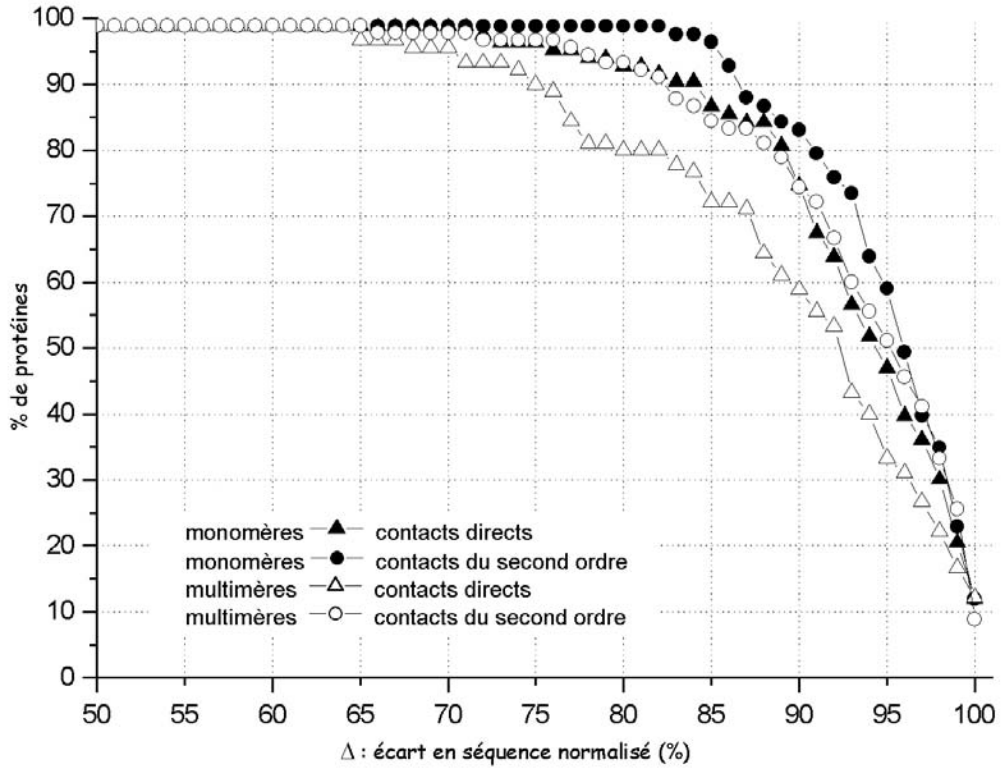
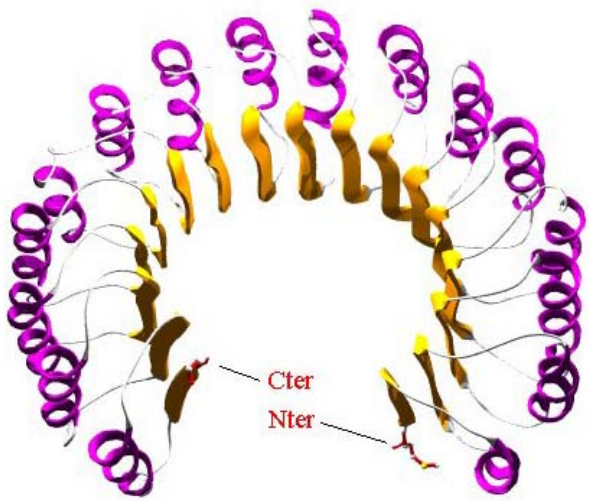


Figure 102 : Proportion des protéines de la banque faisant un contact associé à un écart égal ou supérieur à Δ pour les monomères (en noir) ou les multimères (en blanc).

5 - Exemples et détails

Dans ce paragraphe, je reprends en détail les résultats du Tableau 14 pour les protéines avec les plus mauvais scores afin d'essayer de mieux comprendre ce qui se passe lorsqu'il n'y a pas de proximité entre les extrémités.

5.1.1 2bnh[114]



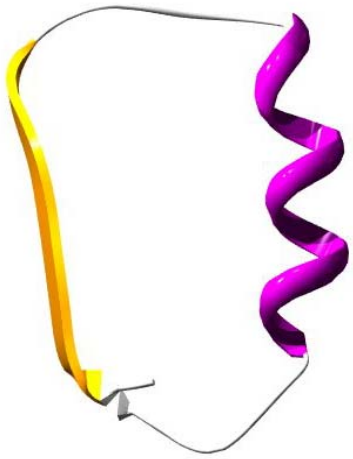


Figure 103 (à gauche) : Structure de la protéine de code PDB 2bnh (457 AA).
Figure 104 (à droite) : Elément unitaire de 2bnh entre la tyrosine n°25 (Y25) et la leucine n°53 (L53).

Cette structure en forme de fer à cheval a le plus mauvais score des protéines de notre banque ($\Delta = 12.7\%$) et selon les classifications de SCOP et de CATH, cette protéine ne contient qu'un seul domaine. Il est néanmoins évident que cette protéine peut être décomposée en seize domaines structuraux identiques d'à peu près 29 AA et composés d'un brin et d'une hélice tels que celui présenté sur la Figure 104. Cette sous-unité polypeptidique ne peut pas être considérée comme un véritable domaine indépendant puisqu'une telle structure ne serait pas stable seule, il est toutefois curieux de constater que les extrémités de ce segment protéique sont très proches l'une de l'autre. Ce type de segments est à rapprocher des « boucles de taille standard » appelées également TEF pour Tightened End Fragments[115-120] dont la taille est comprise entre 22 et 32 AA.

5.2 1qex[121]

Cette protéine est un homotrimère mais seules les extrémités N terminales sont proches les unes des autres, d'après SCOP cette protéine est monodomaine mais pour CATH cette structure est composée de trois domaines différents (Figure 105). Contrairement aux deux derniers domaines, le domaine n°1 ne semble pas pouvoir être stable s'il était indépendant. En ce qui concerne les extrémités de chaque domaine, on peut observer que celles du domaine n°1 sont très éloignées. Pour le domaine n°2, on constate que la méthionine n°167 (M167) et la tyrosine n°67 (Y67) sont à moins de 8Å l'une de l'autre avec une valeur Δ d'à peu près

98%. Pour le domaine n°3, la lysine n°283 (K283) et la glutamine n°170 (Q170) sont à moins de 6Å l'une de l'autre pour une valeur Δ de 95%.

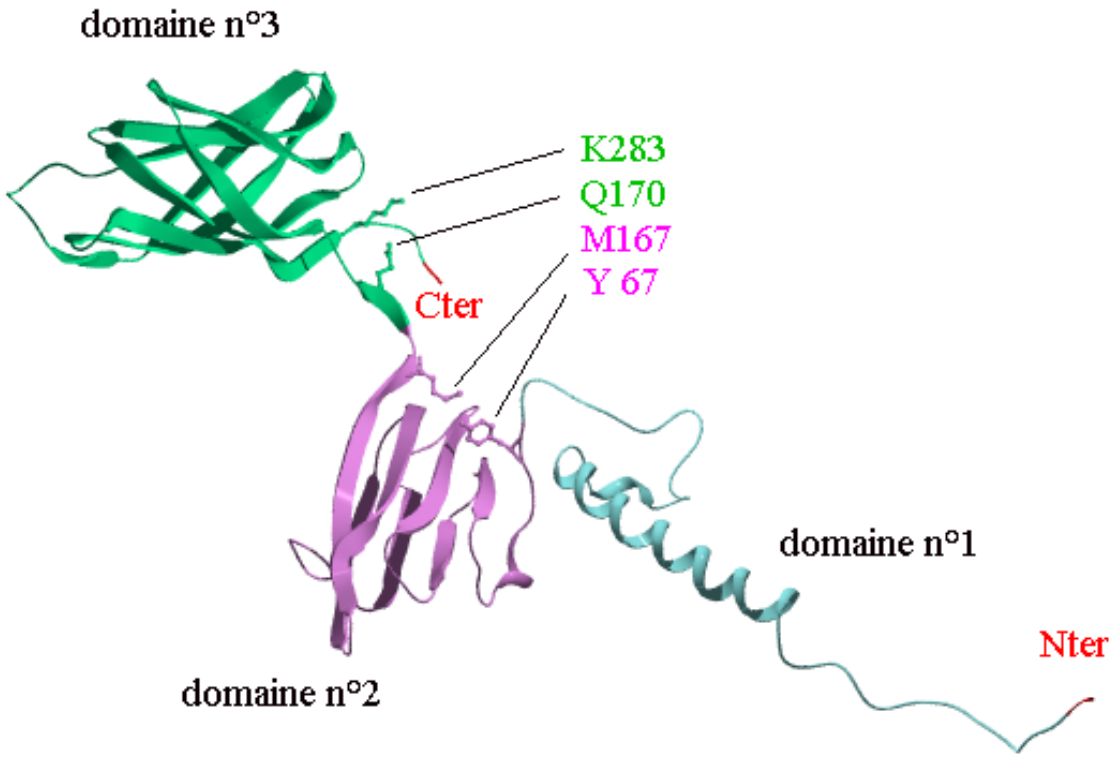


Figure 105 : Structure de la protéine 1qex (288 AA). D'après CATH cette protéine contient 3 domaines représentés en bleu (1-65), violet (66-168), vert(169-288).

5.3 1e7f[122]

Cette structure présente également un mauvais score puisque la valeur Δ qui lui est associée est de 49.1%. D'après SCOP, cette protéine se décompose entre trois domaines mais un examen visuel de cette structure montre que chacun de ces domaines est constitué de deux sous-domaines, c'est d'ailleurs la décomposition que propose CATH. Deux de ces sous-domaines sont représentés sur la Figure 106 (sous-domaines n°5 et n°6) ils constituent le dernier domaine détecté par SCOP. Il est facile de constater sur cette figure que les extrémités de ce domaine sont éloignées (l'extrémité du domaine n°5 est l'acide glutamique n°383 - E383), les extrémités des deux domaines CATH quant à elles sont beaucoup plus proches. Pour le domaine n°5, l'acide glutamique n°383 (E383) est en contact avec l'arginine n°485 (R485) ce qui donne une valeur Δ de 92%. Pour le domaine n°6, la phénylalanine n°502 (F502) est en contact avec la glutamine n°580 (Q580), ce qui fournit une valeur Δ un peu plus faible puisqu'elle est de 88%.

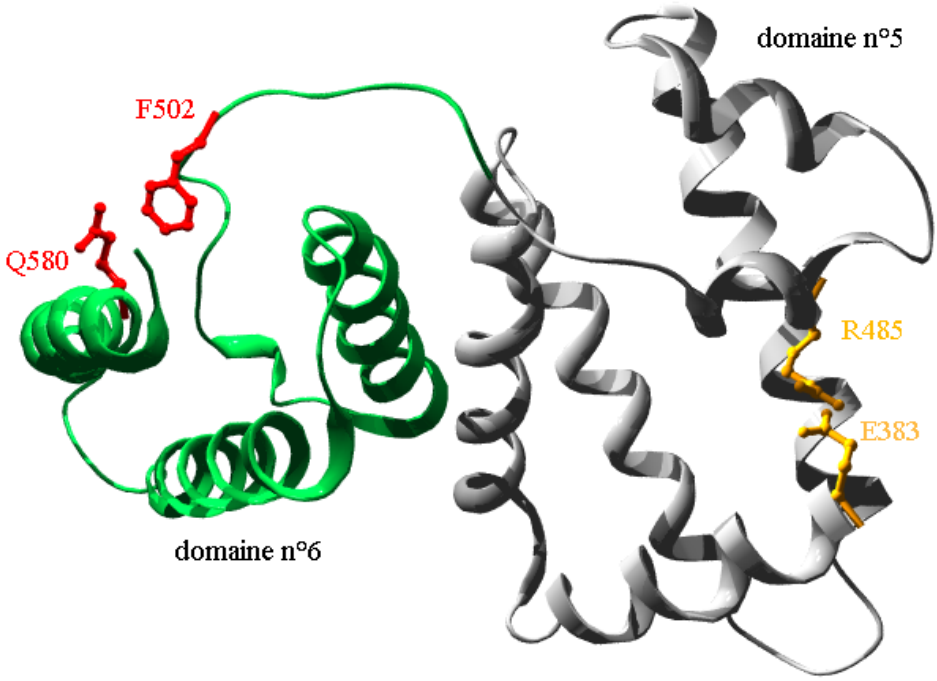


Figure 106 : Structure des sous-domaines n°5 (E383-D494) et n°6 (D495-G584) selon CATH, de la structure 1e7f (582 AA).

5.4 1trk[123]

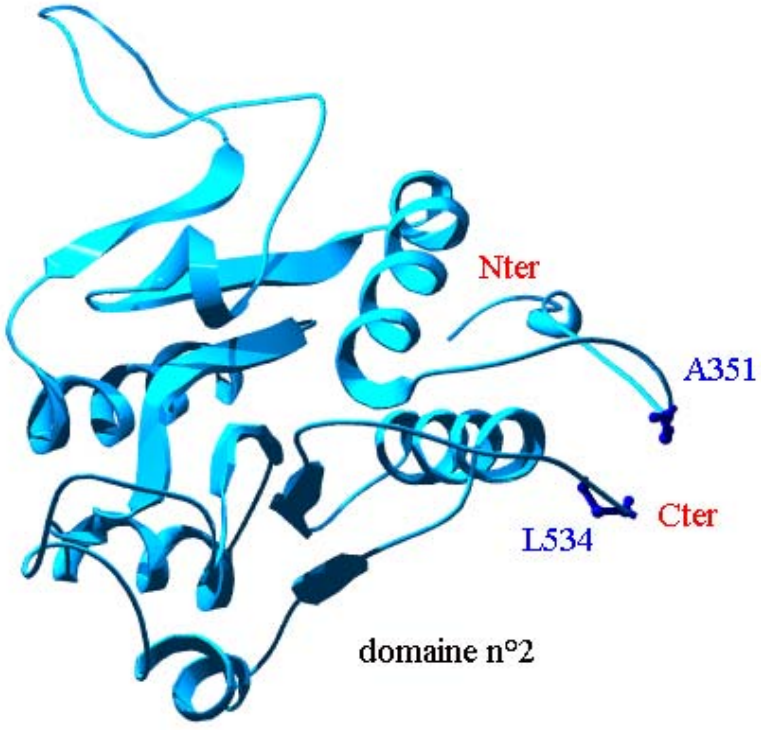


Figure 107 : Structure du domaine n°2 (L338-L534) selon SCOP.

Cette protéine est associée à une valeur Δ de 55.5%, elle est biologiquement fonctionnelle sous forme d'un homodimère mais seules les extrémités C-terminales des deux chaînes sont proches. Cette structure est divisée en trois domaines par SCOP et par CATH, avec des délimitations à peu près équivalentes. La Figure 107 représente la structure du domaine n°2 de cette protéine ; on peut constater que l'alanine n°351 (A351) et la leucine n°534 (L534) sont en contact avec un écart en séquence normalisé de 93%. Les valeurs Δ des deux autres domaines (non représentés) sont 86% pour le domaine n°1 et 74% pour le domaine n°3.

5.5 8tln[124]

Cette protéine dont la valeur Δ est de 58,5% est composée de deux domaines dont un est présent dans la banque de domaine (8tl1) avec une valeur Δ de 76.4%. On trouve la même valeur pour le second domaine soit un gain de près de 18%.

5.6 1a2o

Le cas de cette protéine a déjà été présenté Figure 99, nous avons vu que là aussi il s'agissait d'une structure composée de deux domaines dont les extrémités sont relativement proches avec des valeurs Δ supérieures à 90%.

6 - Propensions

Nous nous sommes intéressés jusqu'ici aux fréquences des contacts entre extrémités, nous avons également cherché à savoir si les contacts observés entre les segments initiaux et terminaux étaient des contacts favorisés par rapport aux autres contacts observés. Pour répondre à cette question j'ai utilisé des propensions exprimées en fonction de l'écart en séquence normalisé :

$$\Pi(\Delta) = \frac{P(\Delta)}{\sum_{\Delta} P(\Delta)}$$

Avec Δ représentant l'écart en séquence normalisé et P(Δ) la proportion des contacts observés à l'écart normalisé Δ ($Nb_{observ}(\Delta)$) par rapport aux contacts possibles à ce même écart ($Nb_{poss}(\Delta)$).

$$P(\Delta) = \frac{Nb_{observ}(\Delta)}{Nb_{poss}(\Delta)}$$

Le nombre de contacts possibles à un écart en séquence normalisé donné (Δ) est la somme des contacts possibles aux écarts réels correspondant bien sûr à la valeur Δ. Par exemple pour une protéine de 100 AA le nombre de contacts possibles pour Δ = 100% est 1, alors que pour Δ = 99% il est de 2 (un contact entre l'AA n°1 et le n°99 et un contact entre l'AA n°2 et l'AA n°100) etc. Pour une protéine de 200 AA le nombre de contacts possibles à Δ = 99% est le nombre de contacts possibles aux écarts réels de 198 AA et 199 AA. On obtient donc un total de 3 + 2 = 5 contacts possibles. En fait, cette façon de procéder induit un biais dans le calcul du nombre de contacts possibles. En effet, les protéines sont des polymères, ce qui implique que les AA se suivent le long de la structure primaire. Les AA proches le long de cette structure sont donc nécessairement en contact et le nombre de contacts possibles correspondant à ces écarts se comporte plutôt comme un nombre de contacts observés voire obligatoires. Ceci n'est pas très important pour l'évolution de la courbe en fonction de Δ puisque ce biais intervient principalement pour les valeurs de Δ faibles alors que nous nous intéressons aux grandes valeurs. Mais il est important de noter que le nombre de ces contacts intervient également dans le compte total des contacts possibles. Aussi, pour supprimer ce biais nous n'avons pas pris en compte les contacts dont les AA sont séparés par moins de 4 AA. Une propension supérieure à un indique que les contacts à l'écart normalisé Δ considéré sont favorisés, si la propension est inférieure à un, on considèrera que les contacts à cet écart sont défavorisés.

Les résultats obtenus en fonction de l'écart normalisé en séquence sont présentés dans la Figure 108. On peut constater que la courbe correspondant aux protéines et celle correspondant aux domaines ont un profil à peu près équivalent, celle correspondant aux seuls domaines étant légèrement supérieure à celle des protéines. Cet écart entre les deux courbes augmente avec l'écart en séquence normalisé Δ. Les deux courbes présentent approximativement une symétrie autour de 50% avec des propensions supérieures à un aux deux extrémités. Les propensions pour les faibles écarts en séquence sont représentées mais n'ont pas de véritable sens puisque comme je l'ai expliqué plus haut, les contacts entre des AA séparés de moins de quatre AA n'ont pas été pris en considération dans le compte des contacts possibles, il est cependant intéressant de constater que les contacts sont favorisés jusqu'à une valeur Δ de plus de 15%. Ces courbes semblent donc confirmer que les contacts entre AA aux extrémités des domaines et/ou des chaînes polypeptidiques sont favorisés par rapport aux contacts entre AA dont l'écart en séquence normalisé se situe entre 20% et 90%.

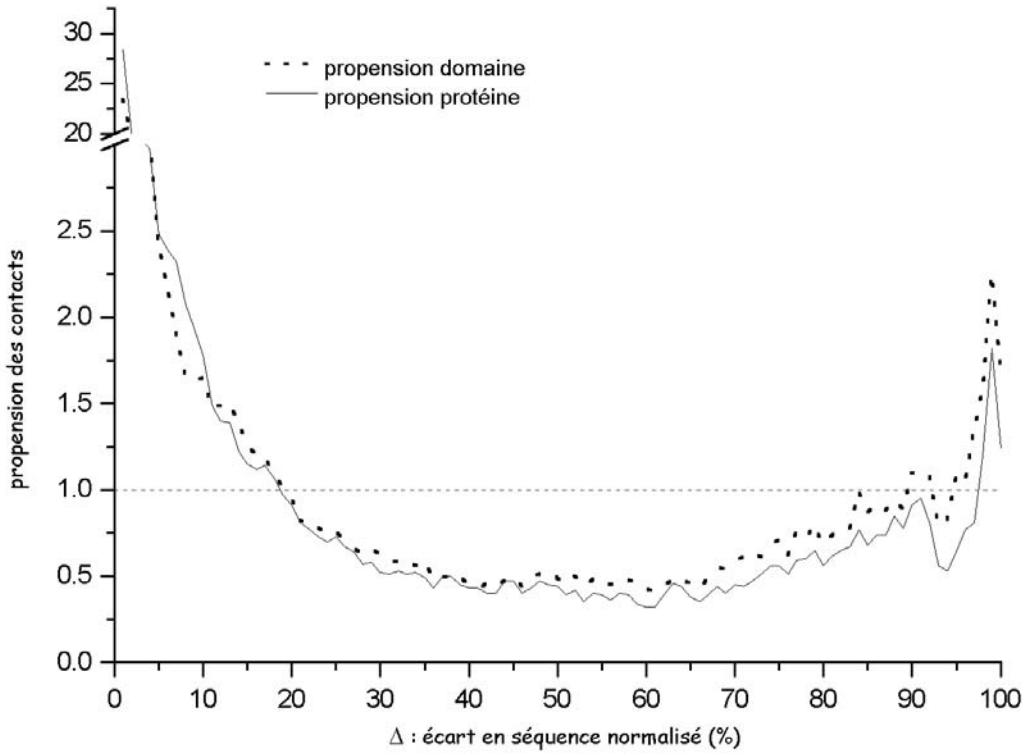


Figure 108 : Moyenne des propensions des contacts pour la banque de protéines et la banque de domaines.

7 - Conclusion

Les différents résultats exposés dans ce chapitre tendent à montrer que dans une majeure partie des structures de nos banques, les extrémités des protéines et plus particulièrement des domaines protéiques sont en contact ou voisines, confirmant ainsi les travaux de J. Thornton et B. L. Sibanda[92]. Cette observation n'est valable qu'en reconsidérant le terme d'extrémité et en l'élargissant afin de ne pas uniquement considérer les extrémités strictes ou des extrémités de longueurs fixées. Comme le suggère J. A. Christopher et T. O. Baldwin[93], la proximité des extrémités strictes est probablement une étape importante du processus de repliement (les deux extrémités pourraient être liées au ribosome), mais elle n'est pas la dernière et il est possible sinon très probable que ces extrémités se réorganisent après leur « libération » provoquant ainsi un éloignement. Leur étude montre ainsi qu'une faible réorganisation de petits fragments aux extrémités (environ 7 AA) permettrait d'obtenir des distances entre extrémités plus faibles que celles que l'on obtiendrait de manière aléatoire. L'élargissement de la notion d'extrémité par rapport aux études précédentes conduit à un grand nombre d'observations de leurs proximités qui pourrait être une trace « fossile » de

cette étape du repliement. L'approche présentée ici tire parti des avantages qu'apportent les TdV mais ne se soustrait pas à leurs inconvénients. Il serait intéressant de ne pas considérer uniquement des contacts (directs ou du 2nd ordre) mais également de considérer en complément des distances entre ces extrémités avec des seuils critiques (et les défauts qui leur sont associés) variant en fonction de la taille de la protéine. Par exemple, le cas de la structure 2bnh de la Figure 103 est particulièrement curieux. En effet, dans ce cas particulier, les TdV sont muettes, mais il est troublant de comparer la distance entre les deux segments extrêmes (20 Å) et le « diamètre » d'une telle structure (63 Å). Ce travail est actuellement mené au laboratoire par Guillaume Fourty et semble aboutir pour le moment aux mêmes conclusions. A travers les quelques mauvais cas rencontrés et détaillés plus haut il semble qu'une des raisons principales pour laquelle on n'observe pas de proximités réside dans le problème de la définition des domaines protéiques. Nous avons utilisé les définitions de SCOP qui semblent être moins fiables que celles de CATH mais qui, comme je l'ai déjà expliqué plus haut, conviennent mieux à nos besoins de traitement automatique. L'idéal aurait été bien sûr de redéfinir les domaines un par un mais le travail à effectuer aurait été d'une autre ampleur.

C h a p i t r e 6

Procédure d'attribution

1 - Introduction

Ce chapitre présente une application concrète des TdV puisqu'il décrit une nouvelle méthode d'attribution des structures secondaires. Ce travail sur les attributions est une suite naturelle de celui que j'ai effectué lors de mon stage de DEA. Il consistait en effet à mettre en place une banque ainsi que la procédure d'interrogation idoine permettant le criblage des structures secondaires significatives des différents types de repliements. L'élaboration de cette banque faisait entre autre déjà intervenir les TdV et l'algorithme résultant de ce stage a été utilisé afin d'identifier des candidats pouvant servir de support pour modéliser la forme pathogène du prion[105, 106]. Dans ce chapitre, j'expose tout d'abord les propriétés des matrices de contacts vues sous l'angle des différentes structures secondaires. Ces propriétés permettent à partir des résultats obtenus par diverses méthodes automatiques d'attribution (que je décris rapidement) d'introduire la notion d'empreinte. Cette notion est utilisée pour élaborer un nouveau programme d'attribution tirant parti des avantages spécifiques aux tessellations. Je présente en détail le principe de cet algorithme ainsi que les résultats obtenus, qui sont analysés et comparés à ceux des autres méthodes. J'étudie aussi l'influence de la résolution des structures et un exemple concret d'attribution est finalement présenté.

2 - Matrice de contact et structures secondaires

2.1 Matrices de distance et de contact

L'utilité des matrices de distance (distance matrices) que l'on appelle aussi cartes de distance (distance maps), pour décrire ou comparer des structures protéiques est reconnue depuis un certain nombre d'années[125-130].

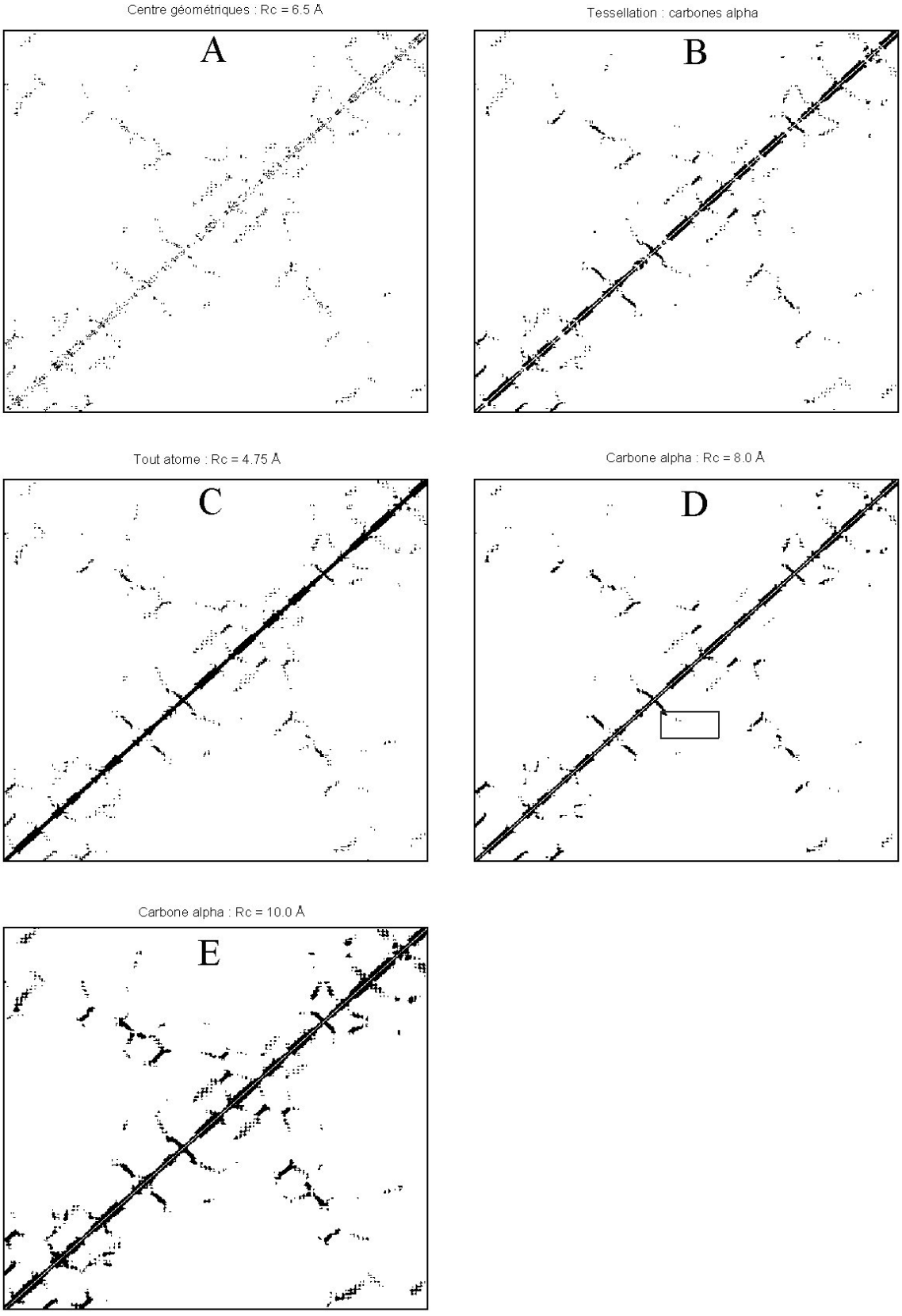


Figure 109 : Diverses matrices selon les définitions des contacts.

Ces matrices de distance sont des représentations 2D de structures 3D et à l'exception de la chiralité globale, elles contiennent toute l'information nécessaire pour reconstruire les structures 3D[131]. Les contacts entre les AA peuvent alors être définis en imposant diverses contraintes sur ces distances par l'intermédiaire de valeurs seuils ou « cut-off » ; on obtient ainsi des matrices de contact également appelées cartes de contact[132-137], dans lesquelles un point ou un '1' représente un contact entre deux résidus alors qu'un '0' ou un blanc indique qu'il n'y a pas de contact.

La Figure 109 montre plusieurs exemples de matrices de contact obtenues à partir de la protéine 3-isopropylmalate dehydrogenase provenant de *Thiobacillus ferrooxidans* (code PDB : 1a05, chaîne A, résolution de 2.00 Å). Chaque matrice est obtenue à partir d'une définition de contact différente. En haut à gauche (A) les contacts sont définis entre centres géométriques des différents AA, la valeur seuil Rc est de 6.5 Å[79], en haut à droite (B) se situe la matrice de contact obtenue par une tessellation effectuée sur les Cα. Au milieu à droite (D) les contacts sont définis entre Cα avec Rc = 8 Å[132], à gauche (C) les contacts sont définis entre tous les atomes lourds avec Rc = 4.75 Å[135]. En bas (E) les contacts sont définis entre Cα avec Rc = 10 Å[134].

Si les traits caractéristiques de ces matrices sont tous équivalents, il en va autrement pour leur aspect général ou leur « texture ». La matrice A et la E sont à ce titre en parfaite opposition : la matrice A favorise les contacts entre chaînes latérales et beaucoup moins les proximités des atomes du squelette des différents AA, d'où un aspect clairsemé même le long de la diagonale principale où apparaissent en général de manière flagrante les hélices α. La matrice E quant à elle est beaucoup plus tolérante et englobe un très grand nombre de contacts, cette tolérance présente bien sûr l'inconvénient d'une précision moindre. Cependant, la matrice D qui fonctionne également à partir des Cα montre qu'un cut-off plus faible (8Å au lieu des 10Å de la matrice E) ne permet pas de détecter de manière convenable tous les contacts (un exemple est donné par le rectangle). La matrice obtenue avec les TdV (B) semble un bon compromis entre le nombre de contacts détectés et leur précision. Avec la matrice A, cette matrice est la seule à avoir des zones sans contacts sur la diagonale principale. Ces quelques exemples (parmi d'autres) montrent la complexité d'un problème apparemment simple qui est de définir un contact entre deux résidus.

2.2 Les contacts dits forts

Les contacts triviaux dus à la proximité le long de la structure primaire sont très nombreux et ils apportent peu d'éléments informatifs, surtout pour des écarts inférieurs à 3 AA. Afin d'obtenir des matrices plus pertinentes, j'ai voulu enrichir l'information contenue dans les contacts et pour cela j'ai exploité un des principaux avantages des TdV. En effet comme nous l'avons vu, les contacts entre résidus sont définis par les faces des cellules de Voronoï, celles-ci peuvent être caractérisées par diverses valeurs comme leur périmètre ou leur aire.

J'ai ainsi utilisé une banque de 282 structures protéiques de résolution inférieure à 2.5 Å constituant ce que j'appellerai dans la suite la StatBank. Pour éviter toute redondance structurale au sein de cette banque je n'ai considéré que des structures de super-familles différentes, selon la classification établie dans SCOP[68]. Toutes ces structures ont été installées dans l'environnement habituel (relaxé neuf fois[63]) et les tessellations non pondérées ont été effectuées sur les Cα. Ce choix s'est imposé car il permet de détecter la régularité des hélices et des brins et, comme nous le verrons, il permet d'effectuer les attributions des structures déterminées à basse résolution et/ou n'étant constituées que de la trace de la chaîne principale.

Tous les contacts entre résidus dont l'écart en séquence était inférieur ou égal à 6 AA ont été répertoriés et classés selon la nature du couple d'AA et leur écart en séquence. Pour les 210 (21×20/2) couples d'AA possibles, chacun des six écarts a été considéré afin d'obtenir finalement 1260 (210×6) moyennes d'aire. Pour une structure particulière, préparée dans les mêmes conditions que celles de la StatBank, il devient alors possible de comparer les aires des faces de contact avec les moyennes correspondantes déjà calculées. Si l'aire d'une face est supérieure à la moyenne qui lui correspond augmentée de 2 $Å^2$, j'ai qualifié le contact associé de fort et dans la matrice, le '1' a été remplacé par un '2'. La valeur de 2 $Å^2$ a été déterminée afin de diminuer le nombre total de désaccords entre la méthode présentée ici et les procédures déjà existantes.

La Figure 110 représente la carte des contacts obtenue pour la protéine 3-isopropylmalate dehydrogenase provenant de *Thiobacillus ferrooxidans* (code PDB 1a05[66], 357 AA, chaîne A). Les contacts sont symbolisés par des points gris, les points noirs représentent les contacts dits forts. L'extrémité N-terminale se trouve en bas à gauche et l'extrémité C-terminale en haut à droite. Sur les côtés et sous la carte sont indiqués les éléments de structures secondaires définis par PROMOTIF[138], les hélices sont symbolisées

par des rectangles noirs repérés en lettres capitales de A à R, les brins sont symbolisés par des rectangles blancs repérés en lettres minuscules de a à l.

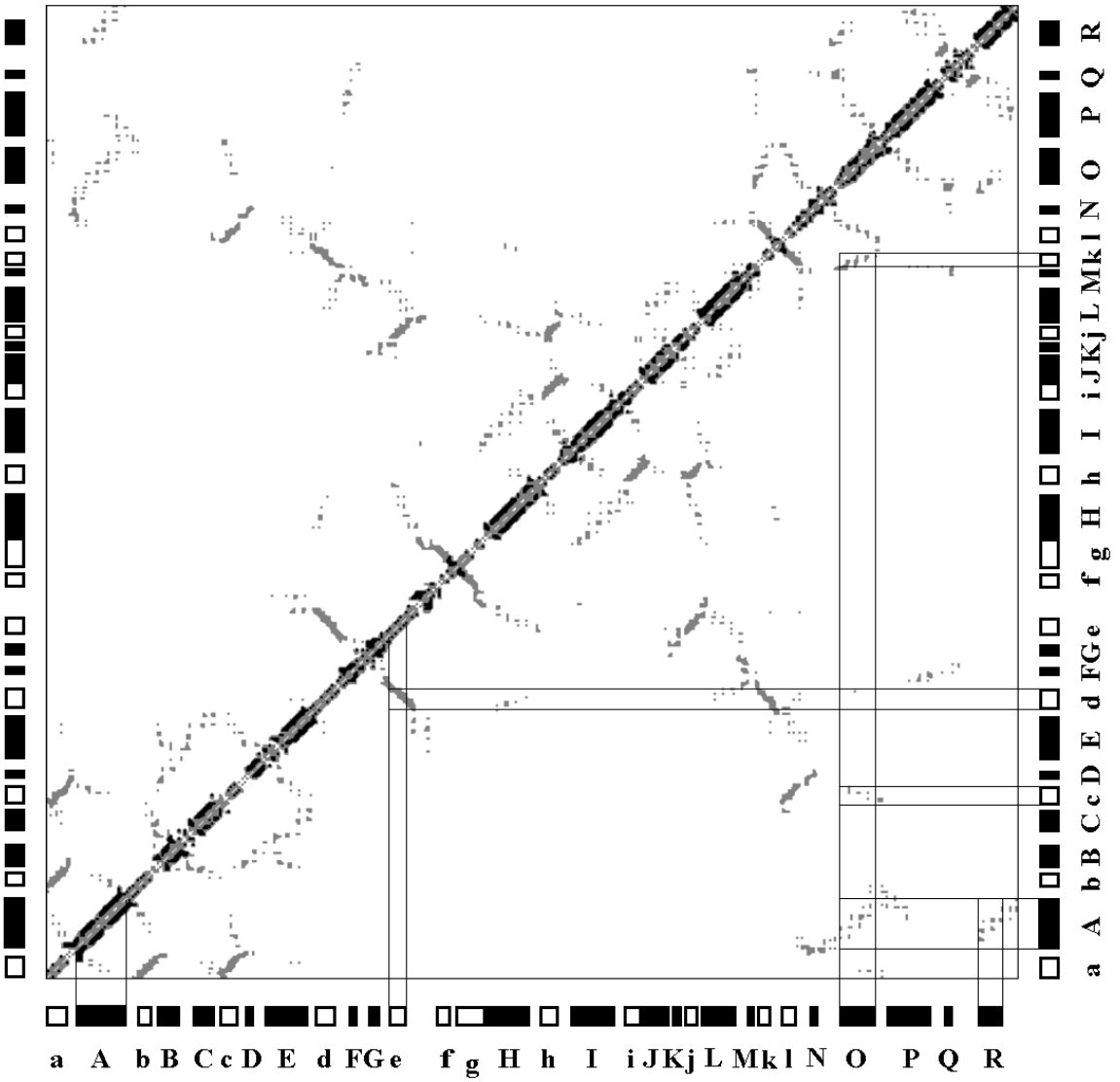


Figure 110 : Matrice de contact déduite de la protéine 3-isopropylmalate dehydrogenase de *Thiobacillus ferrooxidans* (code PDB 1a05). L'extrémité N-terminale se trouve dans le coin inférieur gauche. Les contacts sont représentés en gris ou spécialement en noir pour les contacts dits forts avec ΔAA ≤ 6. Sur les côtés et en dessous de la matrice, les rectangles noirs représentent les hélices repérées de A à R, les rectangles blancs représentent les brins de a à l.

2.3 Propriétés des matrices de contact

Une grande partie des contacts forme une diagonale reliant les extrémités N- et C-terminales. Ces contacts sont souvent internes aux éléments de structures secondaires et les caractéristiques géométriques de la diagonale sont représentatives de la nature de la structure

secondaire à laquelle ils sont associés. Ainsi, il est facile de constater que les hélices sont représentées par une diagonale plus épaisse dont les bords sont continûment noirs (et non pas gris). La topologie de l'hélice α favorise les contacts entre l'AA n° i et l'AA n° i+4 ou n° i-4, c'est la raison pour laquelle les aires des faces de ces contacts sont généralement supérieures à la moyenne. Les contacts avec des écarts plus faibles, dus à la proximité le long de la structure primaire sont également présents et certains contacts entre l'AA n° i et n° i±3 sont aussi des contacts forts. Les contacts avec un écart supérieur à 4 AA étant extrêmement rares pour les hélices, on obtient finalement une diagonale aux bords noirs d'à peu près 9 AA d'épaisseur. La Figure 111 représente l'hélice (T7-C16) de la toxine beta-purothionine provenant de *Triticum aestivum* (code PDB : 1bhp[65]). Les rectangles blancs mettent en évidence une des caractéristiques importantes des cellules liées aux hélices α, la présence des contacts entre l'AA n° i et l'AA n° i±2, principalement due à la proximité séquentielle car la géométrie hélicoïdale ne tend pas à favoriser ces contacts. Ceci se traduit concrètement par la présence (quasiment systématique) de faces triangulaires relativement petites entre ces résidus.

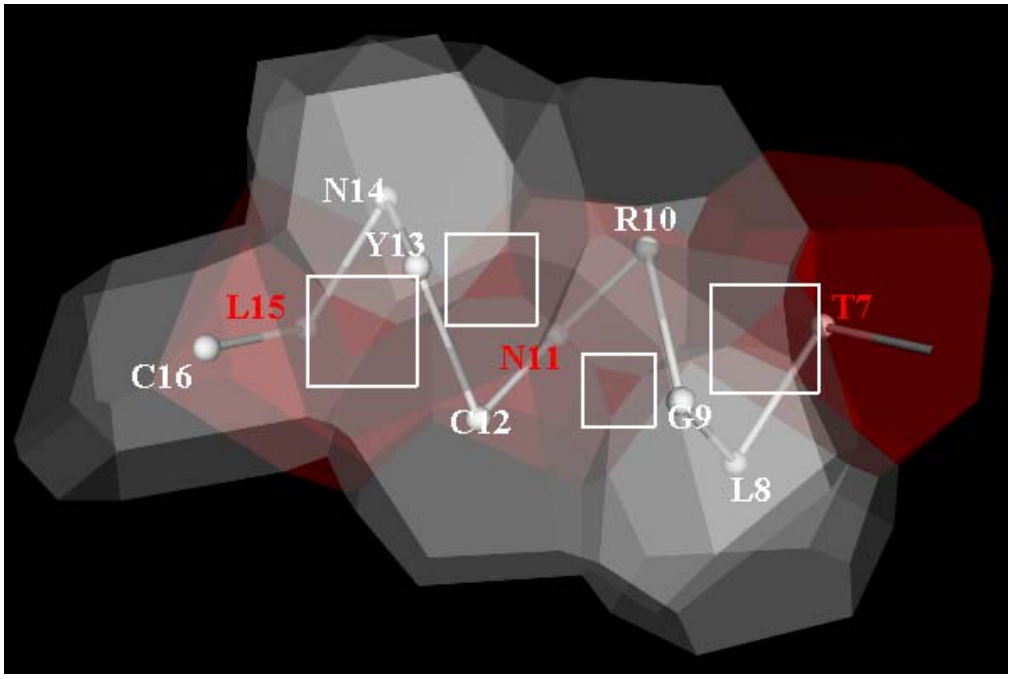


Figure 111 : Hélice α (T7-C16) de la toxine beta-purothionine provenant de *Triticum aestivum* (code PDB : 1bhp). Les cellules correspondant à T7, N11 et L15 sont en rouge, celles correspondant à L8, G9, R10, C12, Y13, N14 et C16 sont en blanc.

Les brins sont représentés par une diagonale plus mince bordée de manière plus ou moins continue en noir. La diagonale est plus fine car les contacts entre l'AA n° i et l'AA n° i±3 ou plus, sont rarement observés à l'intérieur des brins. L'irrégularité des bords

s'explique par le fait que les contacts forts entre l'AA n° i et l'AA n° i±2 ne sont pas systématiques mais restent fréquents contrairement aux hélices pour lesquelles ils sont extrêmement rares. La Figure 112 montre le brin (I18-N24) de l'inhibiteur de la trypsine du pancréas bovin (code PDB : 1bpi). Aucun contact avec un écart égal ou supérieur à 3 AA n'est présent. Les aires des contacts avec un écart de 2 AA peuvent être très variables, ceci est illustré dans la Figure 112 par les deux faces mises en valeur par des rectangles. La face entre l'isoleucine n° 18 (I18) et l'arginine n° 20 (R20), signalée par un rectangle blanc, est relativement importante (4.45 $Å^2$), surtout si on la compare à la face repérée par le rectangle rouge entre l'arginine n° 20 (R20) et la phénylalanine n° 22 (F22) qui est quarante fois plus petite (0.11 $Å^2$). L'écart entre chaque cellule représentée en rouge est de 2 AA, il n'existe donc aucun contact entre ces cellules, contrairement aux cellules en blanc qui comme nous venons de le voir se touchent. Cette irrégularité des contacts à ±2 AA explique celle que l'on retrouve le long de la diagonale pour les brins.

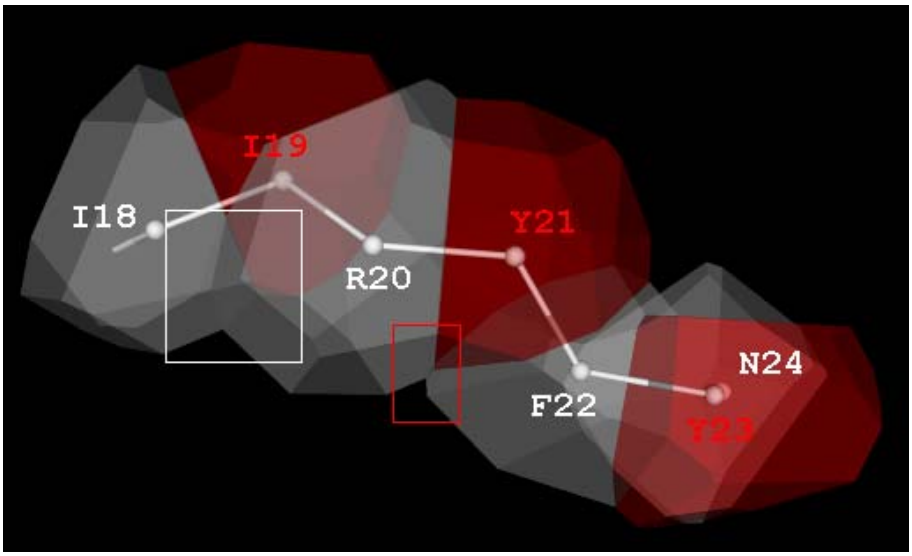


Figure 112
Brin (I18-N24) de la protéine de code PDB 1bpi. Les cellules correspondant à I18, R20, F22 et N24 sont représentées en blanc ; celles correspondant à I19, Y21 et Y23 sont représentées en rouge.

Les boucles reliant les structures secondaires quant à elles ne semblent pas avoir de caractéristiques spécifiques détectables.

Les contacts qui ne sont pas repérés pas des points le long de la diagonale et qui mettent donc en relation des AA plus éloignés les uns des autres le long de la structure primaire, ne sont pas répartis uniformément sur la carte. La plupart d'entre eux forment des lignes dont l'aspect et la direction donnent des informations importantes sur les structures secondaires

régulières ainsi mises en relation. Les lignes parallèles ou perpendiculaires à la diagonale représentent des contacts entre structures secondaires parallèles ou anti-parallèles de même nature. Ceci s'explique simplement par le fait que quand l'AA n° i et l'AA n° j appartenant à deux structures différentes sont en contact, l'AA n° i+δ est en contact avec l'AA n° j+δ dans le cas parallèle ou l'AA n° j-δ dans le cas anti-parallèle, δ pouvant être considéré comme le pas de chaque type de structure, typiquement 4 pour les hélices α et 1 pour les brins. La Figure 113 montre les hélices G13-D28 et L293-H305 de la protéine 3-isopropylmalate déhydrogénase (code PDB : 1a05) dont la carte des contacts est présentée Figure 110 (hélices A et O). L'écart entre les AA dont les cellules sont représentées de la même couleur est de 4 AA et on constate que la continuité des contacts entre les deux structures est assurée le long des deux structures, ce qui se traduit sur la carte des contacts par une ligne discontinue (voir Figure 110). On constate de plus que si un AA n° i fait un contact avec l'AA n° j, non seulement l'AA n° i+4 fait un contact avec l'AA n° j+4 mais l'AA n° i peut faire aussi un contact avec l'AA n° j+4 ou j-4 selon les cas. Ainsi Figure 113, la valine n° 299 (V299) voit la leucine n° 24 (L24) et l'alanine n° 20 (A20), la thréonine n° 295 (T295) voit l'alanine n° 20 (A20) et l'isoleucine n° 16 (I16) etc. Cette observation explique l'élargissement des lignes de contact entre hélices.

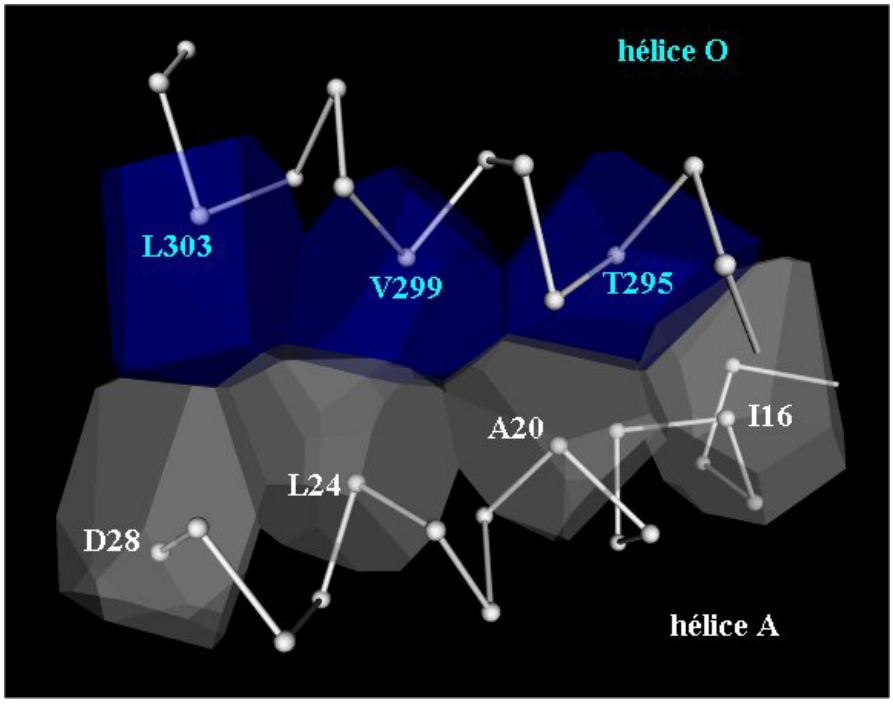


Figure 113 : Représentation de deux hélices de 1a05. En bas l'hélice A (G13-D28), les cellules en blanc sont celles de I16, A20, L24 et D28. En haut l'hélice O (L293-H305), les cellules en bleu sont celles de T295, V299 et L303.

Pour les brins, on retrouve la même caractéristique mais avec un pas de 1. La Figure 114 montre deux brins issus de 1a05 : brin e (D128-E134) et brin d (A102-Q108). L'écart entre les AA dont les cellules sont représentées de la même couleur est maintenant de 1, les contacts sont donc continus et si l'AA n° i voit l'AA n° j et les AA n° j+1 et n° j-1, il ne voit pas au-delà comme c'est le cas pour les hélices, ceci explique à la fois que les lignes soient plus fines et continues.

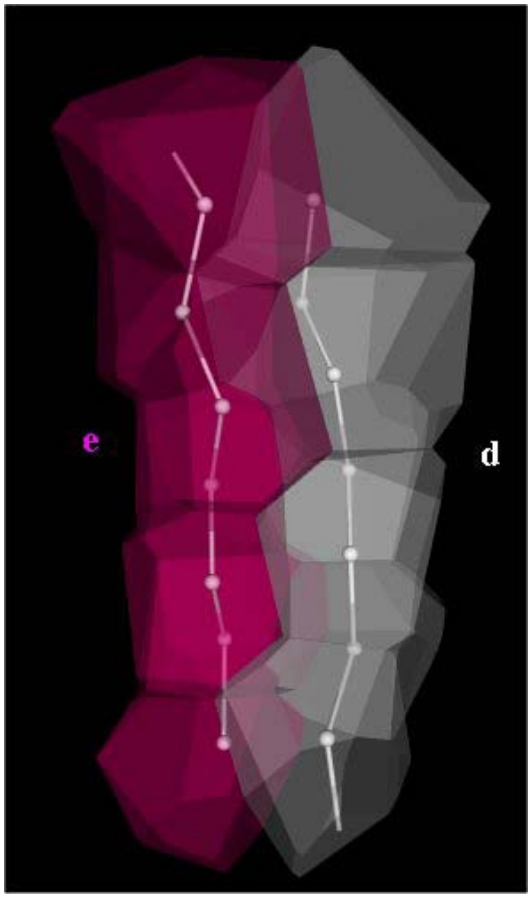
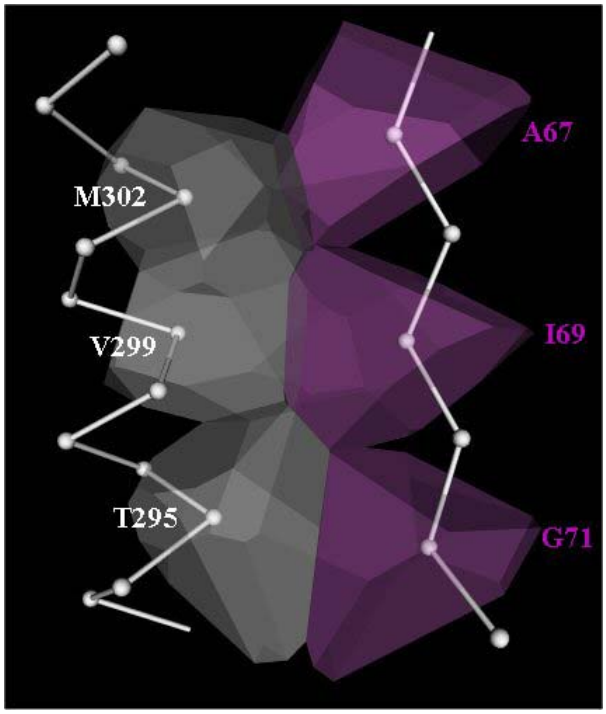


Figure 114
Représentation de deux brins de 1a05. En rose le brin e (D128-E134), en blanc le brin d (A102-Q108).

Dans la matrice, les lignes qui ne sont ni parallèles ni perpendiculaires à la diagonale représentent des contacts entre des structures secondaires régulières de natures différentes, c'est à dire entre des brins et des hélices. Quand l'AA n° i et n° j sont en contact, l'AA n° $i+\delta_1$ et n° $j+\delta_2$ le sont aussi, δ_1 et δ_2 variant selon chaque cas par exemple en fonction de l'orientation des structures en présence l'une par rapport à l'autre. Puisque ces contacts font toujours intervenir une hélice, il est logique que ces lignes soient toujours discontinues. La Figure 115 représente ce type de contacts entre l'hélice O et le brin c. Les cellules représentées sont celles des seuls AA en contact entre les deux structure (3 AA pour chaque),

le pas associé au brin est de 2, par contre celui de l'hélice n'est pas régulier puisque les cellules visibles sont celles des AA n° 295, n° 299 (ΔAA = 4) et n° 302 (ΔAA = 3). On a donc bien à faire à une ligne discontinue mais dont l'épaisseur peut être moins régulière que pour des contacts entre hélices ou entre brins. Ceci est confirmé par la Figure 116 qui représente également l'hélice O et le brin k de la même protéine. Ici aussi les cellules représentées sont celles des seuls AA en contact entre les deux structures. Dans ce cas la régularité des pas est inversée, en effet le pas de 4 est respecté pour l'hélice alors que celui du brin ne l'est pas, puisqu'un des AA (S266) ne fait pas de contact avec l'hélice.



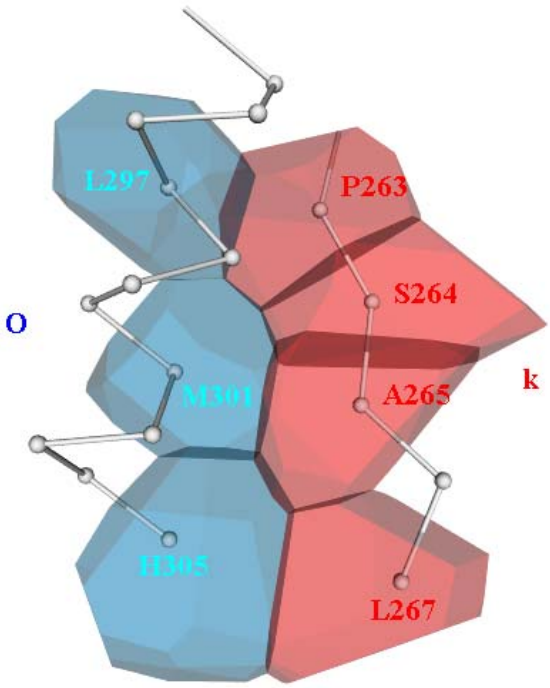


Figure 115 (à gauche) :
Représentation d'une hélice et d'un brin de 1a05. A gauche l'hélice O (L293-H305), les cellules représentées en blanc sont celles de T295, V299 et M302. A droite le brin c (A67-A72), les cellules représentées en violet sont celles de A67, I69 et G71. Seules les cellules en contact sont représentées.
Figure 116 (à droite) :
Représentation d'une hélice et d'un brin de 1a05. A gauche l'hélice O (I294-H305), les cellules représentées en bleu sont celles de L297, M301 et H305. A droite le brin k (P263-L267), toutes les cellules du brin (sauf celle de S266 qui n'est pas représentée) sont en rouge. Seules les cellules en contact sont représentées.

3 - Programme d'attribution

A partir de ces observations, nous nous sommes demandés s'il n'était pas possible de tirer parti des avantages qu'apportent les TdV pour concevoir une nouvelle méthode d'attribution des structures secondaires. Depuis la prédiction des hélices α et des hélices π[139]

et des feuillets β[140], différentes méthodes d'attribution des structures secondaires ont été développées. Initialement, les cristallographes devaient procéder à une inspection visuelle des structures pour effectuer ces attributions. Depuis, de nombreux auteurs ont conçu différents algorithmes pour attribuer automatiquement certains ou tous les types de structures secondaires[141-148]. Aujourd'hui, les programmes d'attribution automatique les plus employés sont DSSP[74], DEFINE[149], P-Curve[150] et STRIDE[151]. Une brève description de chacun de ces algorithmes permettra de mieux comprendre les problèmes en jeu, je décris plus longuement DSSP qui est le précurseur de ces algorithmes et qui a pu notamment servir de référence pour les suivants.

3.1 DSSP : Dictionary of Secondary Structure of Proteins, 1983[74]

L'attribution visuelle des structures secondaires par les cristallographes posait un problème de cohérence puisque les experts n'étaient pas toujours d'accord entre eux. Ceci présentait un inconvénient lorsqu'il s'agissait par exemple de vérifier la qualité des prédictions de structures secondaires faites à partir des seules séquences. Le but original de Kabsch et Sander était donc d'automatiser les attributions des structures secondaires afin d'améliorer leur prédiction. Il est d'ailleurs curieux de constater que DSSP est resté une référence des programmes d'attribution alors que le travail de Kabsch et Sander sur la prédiction n'a, à ce jour, toujours pas été publié[152]. Néanmoins DSSP est un des plus anciens et reste probablement le plus utilisé des algorithmes d'attribution automatique.

Il repose sur l'étude du réseau de liaisons hydrogène entre l'atome d'oxygène d'une unité peptidique i et l'atome d'hydrogène lié à l'atome d'azote d'une autre unité peptidique j. Les atomes d'hydrogène étant rarement présents dans les fichiers de coordonnées de structures protéiques, DSSP procède donc à une première approximation afin de positionner ces atomes. A partir d'une deuxième approximation, celle de l'énergie de la liaison hydrogène, l'énergie de liaison est calculée, si la valeur déterminée est inférieure à - 0.55 kcal/mole, DSSP considère que la liaison hydrogène (i - j) est bien présente.

L'attribution des hélices α débute lorsque deux AA ont des liaisons hydrogène du type (i - i+4), elle se termine de la même façon avec deux AA formant des liaisons hydrogène du type (i-4 - i). Cette procédure est également suivie pour attribuer les hélices 3_{10} avec des liaisons du type (i - i+3) et pour les hélices π avec des liaisons du type (i - i+5). Les résidus

attribués en brin β sont définis comme ayant deux liaisons hydrogène dans le feuillet ou étant entourés par deux liaisons hydrogène dans le feuillet (Figure 117).

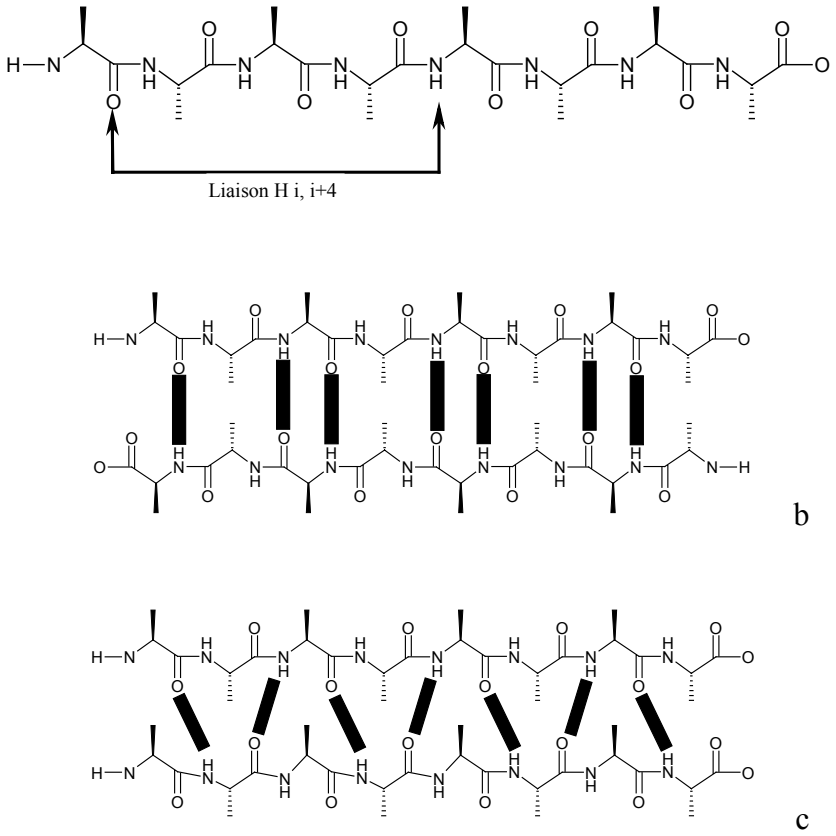


Figure 117 : a - Type de liaison hydrogène entraînant l'attribution en hélice α.
b – Liaisons hydrogène dans un feuillet anti-parallèle.
c – Liaisons hydrogène dans un feuillet parallèle.

3.2 STRIDE : STRuctural IDEntificaton[151]

Cet algorithme utilise également des réseaux de liaisons hydrogène fondés sur une énergie de liaison déterminée par l'analyse de données expérimentales. A ceci s'ajoute l'utilisation de propensions calculées statistiquement à partir des angles de torsion pour les hélices α et les brins β. Les différents paramètres sont optimisés pour correspondre au mieux aux attributions déterminées visuellement par les cristallographes. L'alliance de ces deux paramètres permet une certaine souplesse d'attribution, par exemple un ensemble médiocre de liaisons hydrogène pourra donner lieu à une attribution si les angles de torsion sont particulièrement favorables. STRIDE comme DSSP, duquel il est relativement proche, attribue les différentes hélices (α, 3_{10}, π) mais ne distingue pas les feuillets parallèles et anti-parallèles.

3.3 DEFINE[149]

Cette méthode n'utilise que les coordonnées des Cα et procède en établissant des matrices de distance. Des masques de distances représentant des structures secondaires idéales sont alors utilisés et recherchés dans la matrice. Lorsqu'un accord est trouvé, la structure secondaire est alors progressivement allongée de manière à tolérer des irrégularités ou des courbures modérées. DEFINE attribue les hélices α et également les hélices 3_{10} lorsqu'elles ne sont pas isolées mais ne fait pas la distinction entre les feuillets parallèles et anti-parallèles. De plus, le mode d'attribution permet d'attribuer des brins isolés (ne faisant pas partie d'un feuillet). DEFINE attribue également les boucles Ω.

3.4 P-Curve[150]

Cette procédure est fondée sur une analyse mathématique de la torsion de la chaîne principale. Elle utilise une fonction décrivant la déviation par rapport à une symétrie hélicoïdale parfaite, en terme de courbure de l'axe décrivant le polypeptide et en terme de changement dans la position des différents monomères par rapport à cet axe et les uns par rapport aux autres. Les valeurs de ces différents paramètres sont comparées à des valeurs de référence pour effectuer les attributions des résidus. Au moment de la réalisation de ce travail, le code permettant d'utiliser cette procédure n'était pas disponible, les résultats que je présente dans la suite ne tiennent donc pas compte des attributions effectuées par cette méthode.

3.5 P-SEA : Protein Secondary Element Assignment[153]

Ce programme est fondé sur une double attribution, la première s'appuie sur des valeurs d'angles entre AA proches le long de la structure primaire, la seconde est effectuée en fonction de distances entre AA également proches en séquence. Le résultat final est une attribution consensuelle entre ces deux approches. Les différents paramètres sont déterminés pour optimiser les résultats par rapport à ceux de DSSP, DEFINE et P-Curve. P-SEA procède à une attribution à trois états, hélice, brin et boucle.

En 1993, Colloc'h et coll[154] ont comparé les attributions de DSSP, P-Curve et DEFINE sur une banque non redondante de 152 structures protéiques. En analysant résidu par résidu les différents résultats, ils trouvèrent que le pourcentage d'accord des trois méthodes était seulement de 63 %. Si les trois programmes donnent le même nombre de résidus dans chacun

des trois états, il existe des différences sur le nombre d'hélices et de brins, impliquant donc un grand désaccord sur la longueur des différents éléments. La distribution des longueurs des hélices et des brins montrait également qu'il existe des artefacts propres à chacun des algorithmes. Par exemple, DSSP a tendance à attribuer beaucoup d'hélices de 4 AA (tendance déjà observée par les auteurs du programme eux-mêmes), DEFINE a tendance à attribuer des brins de 4 AA également (il coupe en fait des brins plus longs en morceaux de 4 AA), P-Curve tend à attribuer des structures trop longues. Pour remédier à ces inconvénients, Colloc'h et coll proposèrent une méthode d'attribution consistant en un consensus entre les trois programmes. Un résidu était attribué dans l'état déterminé par au moins deux des programmes étudiés. Avec cette méthode, les effets des artefacts étaient diminués, il était de plus montré que les résidus attribués dans le même état par les trois méthodes étaient de manière générale associés à de meilleures prédictions des structures secondaires que les autres. Les résultats de cette étude montraient donc que l'utilisation combinée des méthodes d'attribution existantes donne de meilleurs résultats que chacune des méthodes prise séparément. Je me suis donc inspiré de cette observation pour établir une méthode d'attribution fondée sur ce principe et sur les TdV.

4 - La méthode

4.1 Constitution des tables Tab_Cen et Tab_Ext

Les attributions des 282 structures protéiques constituant la StatBank (Tableau 16) ont été déterminées à l'aide des quatre programmes suivants : DSSP, STRIDE, DEFINE et P-SEA. Les résultats de ces programmes ont été traités de manière à obtenir une attribution à trois états : a pour les hélices (α, 3_{10} et π), b pour les brins (parallèles, anti-parallèles et les "β-bridges") et c pour tout le reste.

Comme nous l'avons vu chaque protéine de la StatBank a été installée dans son environnement (relaxé neuf fois) puis une tessellation non pondérée finale a été effectuée sur les Cα afin d'obtenir la matrice de contact correspondante. Pour l'AA n° i de chaque matrice, j'ai extrait les valeurs de a_{ii-6} à a_{ii-2} et de a_{ii+2} à a_{ii+6}, a_{ij} étant la valeur du point d'abscisse i et d'ordonnée j dans la matrice. Le triplet central [a_{ii-1}, a_{ii}, a_{ii+1}] n'a pas été pris en considération puisqu'il est presque toujours égal à [1,0,1]. On obtient donc ainsi une matrice linéaire de dix

éléments qui représente le voisinage séquentiel du résidu considéré, cette matrice sera appelée dans la suite l'empreinte de l'AA n° i.

Cette empreinte a été ensuite associée avec les quatre quintuplets (un pour chaque méthode d'attribution) composés des attributions (a, b et c) des résidus i-2 à i+2. Cette opération a été reproduite pour tous les résidus de la StatBank exceptés les six premiers et les six derniers de chaque chaîne protéique. En effet, pour ces cas précis, de telles empreintes ne convenaient pas puisque tous les résidus n'étaient pas présents pour les établir. En fait, pour ces AA la distinction n'a plus été faite entre contacts forts et normaux mais entre contacts réalisables et irréalisables. Pour les empreintes de ces résidus, '0' signifie toujours qu'il n'y a pas de contact, '1' qu'il y a un contact (fort ou normal) mais '2' signifie qu'un contact est impossible. Par exemple pour le premier résidu à l'extrémité N-terminale de la chaîne l'empreinte commence par (22222xxxxx, x pouvant prendre les valeurs 0 ou 1). Afin d'obtenir des quintuplets pour ces empreintes, les attributions existantes ont été complétées avec des attributions boucle (c).

StatBank															
1a05	1abe	1aoe	1b0x	1bm9	1c5e	1cq3	1dd3	1dxj	1eq6	1flj	1imb	1nhk	1qf9	1ugi	2kau
1a28	1abr	1aoh	1b33	1bo4	1c76	1cqd	1dd6	1dzr	1esc	1fqt	1jdw	1noa	1qhw	1utg	2mhr
1a2z	1ad1	1apx	1b3a	1bov	1c9h	1cqy	1dif	1e19	1eur	1fs1	1jpc	1opc	1qjb	1vhh	2nsy
1a34	1ad6	1aqb	1b3t	1bpl	1c9o	1cvw	1dk0	1e20	1euv	1fup	1knb	1oun	1qje	1vhr	2pth
1a3a	1ae1	1ars	1b4b	1bs4	1c9s	1cxy	1dk7	1e2a	1eyv	1fxo	1kp6	1oyc	1qk2	1vid	2sic
1a44	1aew	1aun	1b5e	1bs9	1cbj	1d0b	1dlm	1e6t	1ez3	1fzd	1kpf	1p35	1qsd	1wgj	2wrp
1a4m	1ag9	1ava	1b8o	1btn	1cbk	1d0i	1dlw	1eay	1f05	1g24	1kuh	1pbw	1qtn	1whi	3cla
1a4y	1agi	1avb	1b93	1bu7	1cc8	1d0q	1dm9	1ed1	1f0c	1g71	1lau	1pdo	1rav	1who	3daa
1a58	1agj	1aw8	1bb9	1bue	1cfy	1d1j	1doz	1edy	1f3u	1gak	1lbe	1php	1ris	1ycq	3eip
1a6o	1ah7	1awc	1bdo	1bx4	1cfz	1d1p	1dqi	1ei7	1f3v	1gci	1ldt	1pml	1rop	256b	3pyp
1a7w	1aho	1awd	1bhd	1bx7	1ciq	1d3v	1dqo	1ej1	1f3z	1gen	1lki	1poc	1rpx	2a0b	3tdt
1a80	1ail	1ay7	1bhp	1byf	1cjd	1d4t	1ds7	1ejf	1f41	1gnk	1luc	1poh	1sfp	2acy	4aah
1a8b	1air	1ayx	1bj1	1bzy	1ck4	1d6r	1dsz	1ekg	1f5m	1got	1mat	1pud	1tbg	2afg	4icb
1a99	1aj8	1azp	1bjp	1c1k	1cku	1d8d	1dtd	1ekj	1f60	1gym	1mka	1pyt	1tfe	2ahj	4pah
1aa7	1ako	1azz	1bk5	1c26	1cl8	1d9t	1dun	1el6	1f8y	1hfe	1mml	1qau	1tig	2bbk	4ubp
1aac	1amk	1b00	1blu	1c2t	1cmb	1dan	1dwv	1em9	1f94	1hoe	1mro	1qb0	1toa	2ctc	6ins
1aap	1amw	1b0n	1bm8	1c3m	1coz	1dce	1dxe	1emv	1fd3	1hyp	1msk	1qb2	1tup	2e2c	6prc
1aaz		1b0w		1c52		1dcp		1enh		1icf		1qc7	1tyf	2end	7cei

Tableau 16 : Liste des 282 codes PDB des protéines de StatBank.

Les 243 (3^5) quintuplets possibles ont été répertoriés pour chaque type d'empreinte et la fréquence de chaque type de quintuplet a été notée N(xxxxx). Par exemple, l'empreinte (0012112100) a été rencontrée 4519 fois dans la StatBank dont 4425 fois (97.92 %) avec le

quintuplet (aaaaa) (pour cette empreinte N(aaaaa) = 4425) et quelques fois avec l'empreinte (caaaa) ou (aaaac). Il était alors possible d'associer chaque empreinte avec la fréquence d'observation de chaque type d'attribution pour chaque position au sein du quintuplet. Ceci peut être résumé dans un tableau (3x5) comme celui représenté en gris dans la Figure 118, dans lequel chaque fréquence est notée $I_{i+p}^{q}(i)$ avec p représentant la position dans le quintuplet centré sur i, q une des trois attributions possibles A, B ou C et i la position dans la séquence de l'AA considéré. Par exemple, pour l'empreinte associée à l'AA n° i, $I_{i+2}^{A}(i)$ est la fréquence d'apparition de l'AA n° i+2 attribué a et elle peut être calculée de la manière suivante : $I_{i+2}^{A}(i) = \sum_{x=a,b,c} N(xxxxa) \Big/ \sum_{x=a,b,c} N(xxxxx)$. A partir de cette définition on voit que $I_{i+2}^{A}(i) + I_{i+2}^{B}(i) + I_{i+2}^{C}(i) = 1$, c'est la raison pour laquelle $I_{i+2}^{A}(i)$ a été considérée comme la probabilité pour le dernier résidu du quintuplet d'être dans la conformation hélice pour l'empreinte associée à l'AA n° i. Les 3718 empreintes trouvées et les tableaux correspondants ont été enregistrés dans une table nommée Tab_Cen. Pour les résidus aux extrémités N- et C-terminales, les 388 empreintes ont été enregistrées dans une table nommée Tab_Ext.

4.2 Attribution

Pour attribuer chaque résidu de n'importe quel fichier PDB, la procédure commence par sélectionner les Cα, l'environnement est ensuite ajouté et relaxé neuf fois, la tessellation finale est alors effectuée afin d'obtenir la matrice des contacts. La procédure d'attribution se décompose en deux parties.

La première partie utilise l'information se trouvant le long de la diagonale principale de la matrice. L'empreinte de chaque résidu est déterminée et associée au tableau correspondant de Tab_Cen ou Tab_Ext selon les cas. Pour chacun des AA, le même processus est appliqué et est représenté dans la Figure 118. Pour l'AA n° i, les tableaux de l'AA n° i-2 à l'AA n° i+2 sont considérés et les trois probabilités moyennes sont calculées : $P_i^X = \frac{1}{5}\sum_{j=-2}^{2} I_i^X(i+j)$ où X peut être A, B ou C. La plus grande moyenne donne l'attribution temporaire de l'AA n° i et à la fin de cette partie de la procédure, une succession de structures secondaires temporaire est obtenue. Si une ou plusieurs empreintes ne sont pas trouvées dans Tab_Cen ou Tab_Ext, la moyenne est calculée sur les autres probabilités.

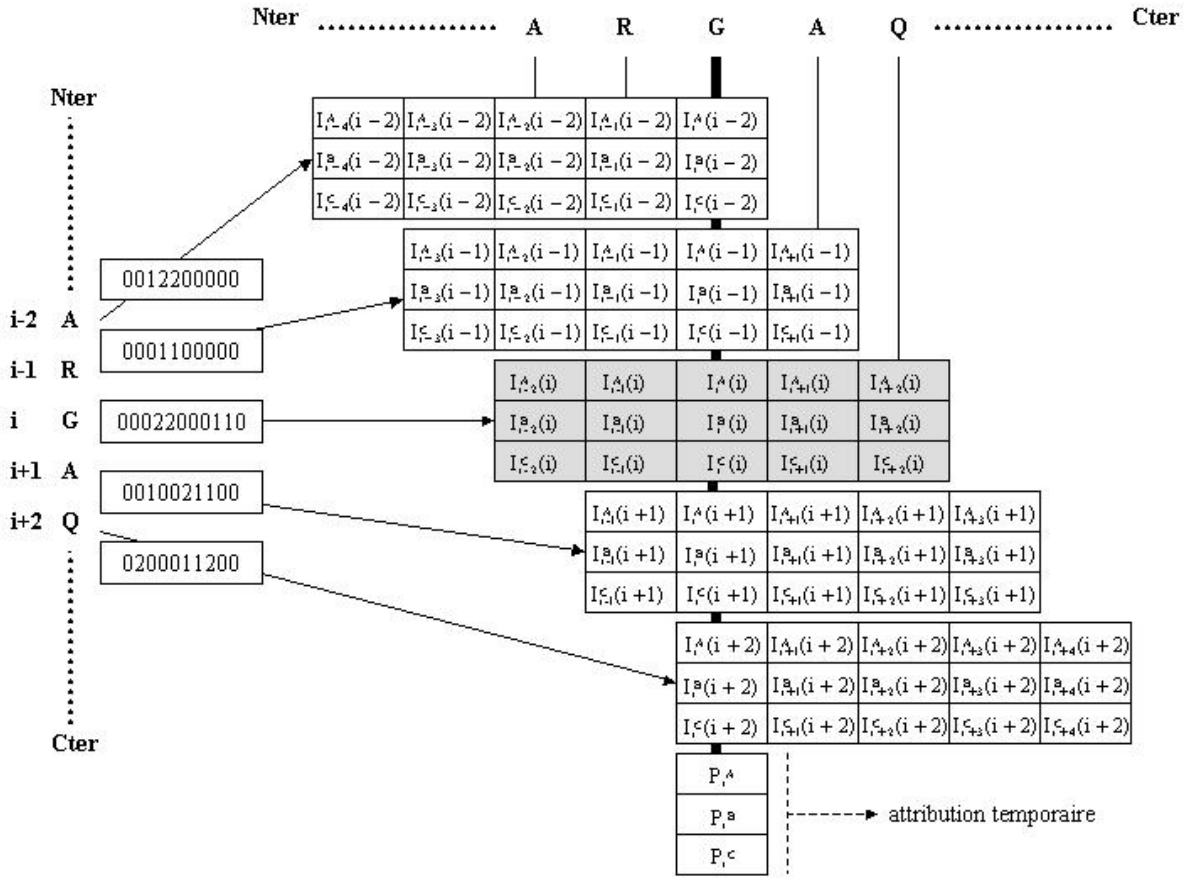


Figure 118
Ce schéma représente une partie du processus pour l'AA n° i. De gauche à droite : pour chaque AA du n° i-2 au n° i+2, les empreintes correspondantes sont extraites de la matrice de contact. Chacune de ces empreintes est associée au tableau correspondant tiré de Tab_Cen (ou Tab_Ext pour les 6 premiers ou les 6 derniers résidus de la chaîne). Une colonne de chaque tableau (de la dernière pour l'AA n° i-2 jusqu'à la première pour l'AA n° i+2) donne une probabilité pour chaque conformation (a, b ou c) pour l'AA n° i. Pour chaque conformation la probabilité moyenne P_i^X (X pour A, B ou C) est calculée.

Les critères appliqués dans la seconde partie ont été affinés de manière à minimiser le nombre de désaccords entre mes résultats et ceux conjugués de DSSP, PSEA, DEFINE et STRIDE. Cette partie du processus d'attribution utilise l'information contenue dans les contacts entre les brins. Les contacts entre les brins temporaires sont ainsi recherchés et analysés alors que les éléments de structures secondaires trop courts (un ou deux AA) sont supprimés. Les différentes étapes de ce processus sont effectuées dans l'ordre suivant :

Quand un AA n° i attribué en brin avec $P_i^B > 0.5$ est détecté, le programme recherche des contacts avec chaque résidu attribué brin ou boucle avec $P_j^C < 0.6$ et $|i-j| > 6$. Quand un de ces contacts est trouvé, le contact entre d'une part les AA n° i+d ou n° i-d et d'autre part les AA n° j+d ou n° j-d avec d commençant à la valeur 1 est recherché avec les mêmes conditions. Si ce contact existe, la recherche est répétée pour d+1 et ainsi de suite jusqu'à une

valeur D. Si D > 3 alors tous les résidus entre j ± (D-1) sont attribués en brin. Ces nouvelles attributions ne sont pas prises en considération pour la prochaine étape qui suit le même principe mais avec des conditions différentes. Quand un résidu n° i en conformation brin ou boucle avec P_i^C<0.6 est détecté le programme recherche des contacts avec le résidu n° j attribué brin ou boucle avec P_j^C < 0.6 et |i-j| > 6. Quand un de ces contacts est détecté, les contacts entre l'AA n° i+d ou n° i-d et n° j+d ou n° j-d avec d allant de 1 à D sont recherchés avec les mêmes conditions pour l'AA n° i±d, mais pour l'AA n° j±d la seule condition est que l'attribution temporaire doit être brin. Si tous ces contacts existent et si D > 3 alors tous les résidus entre l'AA n° i ± (D-1) sont attribués en brin. Dans les étapes suivantes, les nouvelles attributions sont utilisées mais les hélices et les brins constitués de un ou deux résidus sont simplement supprimés et attribués en boucle. Les premiers stades de cette partie du processus d'attribution ont été créés afin d'essayer d'allonger les brins, la suite de la procédure a pour tâche principale de vérifier que les brins sont finalement correctement attribués.

Pour ce faire, les contacts entre les résidus attribués brins sont détectés et leur voisinage séquentiel est vérifié. Tous les contacts entre deux résidus n° i et n° j sont répertoriés. Pour chacun d'eux, la condition suivante doit être remplie au moins une fois. Pour l'AA n° i, si le résidu suivant ou précédent est aussi attribué en brin alors le résidu suivant ou précédent l'AA n° j doit être aussi attribué brin. Si ces conditions ne sont jamais remplies et si P_i^B<0.6 alors le résidu est finalement attribué en boucle. La dernière étape de la procédure consiste à supprimer encore une fois les brins de moins de trois résidus.

5 - Résultats

CheckBank														
1a17	1ax0	1bix	1bxs	1cyn	1dp4	1efv	1fas	1gpr	1kpt	1pbv	1qtw	1tml	2asr	2plc
1a1x	1ayf	1bkp	1bxy	1cyo	1dpj	1egp	1fkj	1gux	1lts	1pfk	1reg	1ukz	2ay1	2poo
1a6j	1b16	1bkr	1byq	1d1g	1dqe	1egw	1fle	1hcb	1mjc	1phr	1rfs	1unk	2bnh	2prd
1ah4	1b1i	1blx	1byr	1d2z	1dv8	1ej8	1flm	1hcq	1mnm	1ppf	1rkd	1uro	2bop	2prg
1ahs	1b4f	1bn8	1c08	1d4o	1dyn	1eo9	1flt	1hle	1moq	1ptf	1rvv	1ute	2cev	2rn2
1aj2	1b66	1bou	1cex	1d7d	1dz3	1ep0	1fof	1htp	1mwp	1puc	1sei	1vcc	2cpg	2tnf
1al3	1b67	1bpi	1chd	1dea	1e6i	1eqo	1fq0	1hxn	1mzm	1qd9	1sml	1vmo	2cro	2trc
1amx	1b7d	1bqk	1cjb	1dfu	1e79	1ert	1fua	1ibr	1nba	1qex	1spp	1wap	2dor	3fap
1ann	1bbp	1br9	1cjw	1dio	1eai	1ew4	1fxd	1ida	1ndd	1qgh	1srv	1wba	2erl	3fib
1apa	1bbz	1bu5	1cnu	1dj8	1ecm	1eyq	1g31	1ido	1nec	1qhv	1stm	1xxa	2izh	4fgf
1arb	1bd8	1buo	1co6	1djr	1ecp	1f2k	1g43	1inp	1nnc	1qip	1sup	1ycs	2lis	4mon
1arv	1beo	1bv1	1ctf	1dk8	1ecy	1f2l	1g6g	1iro	1npk	1qkj	1tcd	2abk	2mta	4nos
1aug	1bf4	1bvy	1cv8	1dly	1edm	1f7d	1gdo	1jac	1otf	1qq8	1tif	2arc	2pii	

Tableau 17 : Liste des 194 codes PDB des protéines de CheckBank.

Pour vérifier la validité et étudier le comportement de mon algorithme (nommé VoTAP pour Voronoï Tessellation Assignment Procedure), j'ai utilisé une seconde banque de structures protéiques que j'ai appelée CheckBank (Tableau 17). La méthode de construction a été la même que celle employée pour StatBank mais les structures retenues sont bien sûr différentes afin d'éviter tout biais dû au fait que CheckBank a été utilisée pour établir Tab_Cen et Tab_Ext. J'ai comparé mes résultats avec les attributions de DSSP, PSEA, DEFINE et STRIDE déterminées sur les 194 structures de la nouvelle banque.

Puisque mon programme effectue une attribution à trois états, j'ai préalablement converti les différents états de chacune des méthodes en trois classes en suivant les mêmes conventions que celles utilisées précédemment. Les comparaisons entre les différentes méthodes sont données pour chaque type de structure secondaire dans le Tableau 18. Dans celui-ci, la dernière colonne correspond à un consensus inspiré du travail de Colloc'h et coll dans lequel chaque résidu est attribué dans la conformation déterminée par au moins deux des trois méthodes suivantes : DSSP, PSEA, DEFINE. Le programme STRIDE n'a pas été pris en compte pour éviter un biais dû à la ressemblance des attributions de DSSP et de STRIDE (> 90% voir ci-dessous).

VoTAP structures	Comparaison avec				
secondaires	DSSP (%)	PSEA (%)	STRIDE (%)	DEFINE (%)	CONSENSUS (%)
hélice	93.0	92.7	96.7	96.7	95.6
brin	77.3	79.7	79.1	73.1	82.1
boucle	79.3	83.1	78.3	64.2	82.7
total	83.2	85.3	84.4	76.9	86.7

Tableau 18
Comparaison des résultats des différentes méthodes d'attribution pour chaque type de structure. Le pourcentage d'accord est établi AA par AA pour chacun des trois types d'éléments (hélice, brin et boucle). Mes résultats sont comparés à ceux de DSSP, PSEA, DEFINE, STRIDE et à un consensus fondé sur ces algorithmes.

L'accord avec les cinq méthodes testées ici (> 92 %) montre que les hélices sont les structures secondaires les mieux attribuées ; ce résultat avait déjà été constaté par Colloc'h et coll. Ceci s'explique par deux facteurs combinés : la régularité de la géométrie de l'hélice et le nombre de résidus se trouvant dans cette conformation. Dans le cas présent, ceci se traduit concrètement par un grand nombre de résidus auquel correspond un petit nombre d'empreintes. Par exemple, les dix empreintes les plus représentées, sur un total de 3718 (soit 0.3 %), correspondent à 6325 AA (sur un total de 48911 AA présents dans la StatBank soit

12.9 %) attribués en hélice par une des quatre méthodes. Pour l'attribution des brins, l'accord est plus faible ; ceci peut s'expliquer d'une part par la ressemblance des empreintes caractérisant les brins et certaines boucles et d'autre part par le fait que les brins sont des structures relativement courtes (en nombre d'AA) comparées aux hélices, ce qui a pour conséquence de rendre plus difficile la détection de leur régularité. Ceci confirme le fait déjà observé[154] que les brins sont plus difficiles à attribuer que les hélices et explique pourquoi VoTAP procède à une attribution des brins en deux étapes qui allie les contacts internes aux structures et ceux existant entre les structures puis à une vérification suivie d'une éventuelle correction.

	DSSP (%)	PSEA (%)	STRIDE (%)	DEFINE (%)	CONSENSUS (%)
DSSP	-	80.2	95.8	73.0	87.8
PSEA	80.2	-	81.4	77.2	92.0
STRIDE	95.8	81.4	-	74.4	87.7
DEFINE	73.0	77.2	74.4	-	84.8
CONSENSUS	87.8	92.0	87.7	84.8	-
VoTAP	83.2	85.3	84.4	76.9	86.7

Tableau 19
Comparaison des résultats des différentes méthodes d'attribution. Le pourcentage d'accord est établi AA par AA. Le consensus est le même que celui présenté dans le Tableau 18.

Comme nous l'avons vu, DEFINE est un programme fondé sur les matrices de distance entre Cα. Celles-ci sont comparées à des structures idéales et le domaine de tolérance entre la réalité et ces structures théoriques est fonction du paramètre « d'erreurs cumulées » ε. Nous avons utilisé une valeur de 0.75 Å pour les hélices et de 0.5 Å pour les brins au lieu de la valeur par défaut de 1 Å qui produit un excès de structures secondaires[155, 156]. Néanmoins, comme le montre le Tableau 19, les pourcentages d'accord de cette méthode avec DSSP, PSEA et STRIDE montrent que ce paramètre pourtant corrigé ne donne pas entière satisfaction. Une des principales conséquences de ce problème reste la trop grande proportion de structures secondaires attribuées par DEFINE (35.5 + 28 = 63.5 %, Tableau 20) par rapport aux autres programmes. De plus, malgré le fait que DEFINE participe à l'élaboration du consensus, le pourcentage d'accord avec ce dernier est plus faible que celui de VoTAP. On remarque que DSSP est en meilleur accord avec le consensus que VoTAP mais avec un avantage de 1 %, ce qui est peu si l'on considère que DSSP contribue aussi à l'élaboration du consensus. Puisque STRIDE et DSSP (> 95 %) sont très proches l'un de l'autre, il n'est pas

étonnant de constater que leur accord est à peu près équivalent. PSEA est le seul programme qui dépasse largement les performances de VoTAP. Outre le fait qu'il intervient dans le consensus, il est intéressant de noter que PSEA comme VoTAP utilise seulement les coordonnées des Cα et qu'il est fondé également sur un consensus puisqu'il utilise simultanément des critères angulaires et de distance.

méthode	état	% d'AA dans chaque état	nb moyen d'AA par structure
DSSP	hélice	30.1	10.9
	brin	23.1	5.4
	boucle	46.8	6.2
PSEA	hélice	30.6	11.7
	brin	26.6	6.3
	boucle	42.8	6.1
STRIDE	hélice	31.5	12.1
	brin	23.9	5.5
	boucle	44.6	6.0
DEFINE	hélice	35.5	13.0
	brin	28.0	5.5
	boucle	36.5	5.1
VoTAP	hélice	32.1	11.7
	brin	25.5	6.1
	boucle	42.4	5.8

Tableau 20 : Proportions et compositions des structures secondaires de CheckBank en fonction des différentes méthodes d'attribution.

Le Tableau 20 montre que les proportions et les longueurs des structures secondaires régulières attribuées par VoTAP se situent entre les différentes valeurs extrêmes des autres programmes d'attribution. Les longueurs des hélices et des brins sont à peu près les mêmes que celles de PSEA mais les proportions sont légèrement différentes en étant plus proches des moyennes des quatre programmes (31.9 AA pour les hélices et 25.4 AA pour les brins). Au vu de ces résultats, VoTAP semble donc être un bon compromis entre les différents programmes.

Pour huit résidus sur les 33193 (soit 0.02 %) de CheckBank, l'attribution n'a pas été possible car les cinq empreintes consécutives n'étaient présentes ni dans Tab_Cen ni dans Tab_Ext. Par défaut, ces résidus ont été attribués en boucle et dans les huit cas, cette attribution était en accord avec les autres méthodes. Ceci s'explique simplement par le fait que l'irrégularité des boucles conduit à la création d'empreintes peu représentées dans les deux tables, la succession de plusieurs de ces empreintes semble suffisante en soi pour induire l'attribution en boucle.

VoTAP est fondé sur des statistiques tirées d'une nouvelle forme de consensus qui présente plusieurs avantages. Avec le consensus proposé par Colloc'h et coll, l'attribution finale d'un résidu particulier est généralement déterminée par au moins deux des trois attributions proposées par DSSP, P-Curve et DEFINE. Ceci implique que l'information portée par l'attribution non retenue est perdue. De plus, une structure secondaire régulière est principalement définie par les relations géométriques que chacun de ses éléments constitutifs entretient avec son voisinage. La méthode que je propose présente l'avantage de prendre en considération la totalité des quatre attributions du résidu concerné mais également les quatre attributions des deux résidus le suivant et le précédant. Le bruit que constituent les artefacts propres à chaque algorithme est ainsi dissout dans l'information utile dont la pertinence résulte de l'accord entre les différentes méthodes. Il aurait été très intéressant d'utiliser P-Curve qui est fondé sur des critères très différents des autres méthodes pour enrichir cette information utile mais ce programme n'était pas disponible à l'époque de ce travail. VoTAP est indépendant de tout « cut-off » puisqu'il a l'avantage de ne pas considérer de critères géométriques particuliers. La seule donnée utilisée est le voisinage local qui est fourni par la TdV. Pour toute structure, cette information est absolue, unique et contrairement aux critères numériques, elle peut s'adapter à diverses distorsions.

6 - Influence de la résolution des structures

Ceci peut se vérifier en étudiant l'influence de la résolution sur le comportement des différents algorithmes. Pour y parvenir, j'ai utilisé 26 structures qui ont été remplacées dans la PDB par des structures de meilleures résolutions. Ces structures sont archivées dans la banque PDBObs (PDB Obsolete http://pdbobs.sdsc.edu/index.cgi) qui est accessible à partir du serveur de la PDB. La liste des 52 structures utilisées est présentée dans le Tableau 21, les structures de basses résolutions sont regroupées dans la banque BR (résolution moyenne 3.0 Å) et celles de hautes résolutions dans la banque HR (résolution moyenne 1.9 Å).

Les attributions de ces structures ont été effectuées par les différentes méthodes et leurs résultats comparés. Pour certaines structures de la banque BR, seules les coordonnées des Cα étaient disponibles, c'est pourquoi DSSP et STRIDE n'ont pas été capables d'attribuer les résidus de cette banque. J'ai aussi constaté que pour plusieurs protéines, DEFINE n'a pas été en mesure d'attribuer tous leurs résidus. Les résultats sont présentés dans le Tableau 22. Ce tableau met en valeur les difficultés de DEFINE à rester cohérent entre les deux banques

puisque l'accord entre les deux est seulement de 86.8 % mais pour seulement 77.5 % de la banque BR.

Le meilleur accord est observé pour STRIDE mais celui-ci n'attribue que 74.4 % de la banque, ce qui est également le cas de DSSP avec un score comparable à PSEA. Ce dernier est avec VoTAP le seul qui attribue 100 % de la banque BR et les deux scores sont proches avec un avantage de 2 % pour PSEA. Quand les données étaient présentes dans les fichiers PDB, nous avons utilisé les attributions des cristallographes comme références pour comparer les attributions des différentes méthodes et l'évolution entre les basses et les plus hautes résolutions. Malgré un désaccord entre les deux banques proche de 13 %, l'accord entre les attributions de DEFINE et celles des cristallographes n'évolue pas avec l'augmentation de la résolution et reste le moins bon des cinq méthodes.

banque BR		banque HR	
code PDB	R (Å)	code PDB	R (Å)
14ps	2.6	1qjb	2.0
151c	2.4	351c	1.6
1abk	2.0	2abk	1.9
1abp	2.4	1abe	1.7
1abx †	3.5	2abx	2.5
1act	2.8	2act	1.7
1afn	2.6	2afn	2.0
1alp	2.8	2alp	1.7
1baa	2.8	2baa	1.8
1bjl	3.0	3bjl	2.3
3cha ‡	2.8	5cha ‡	1.7
4cna	2.9	5cna	2.0
1cpp †	2.6	2cpp	1.6
2dpv ‡	3.3	4dpv ‡	2.9
1erl	1.6	2erl ‡	1.0
2fnr	3.0	1fnd	1.7
1fxb	2.3	1iqz ‡	0.9
1grs † ‡	3.0	3grs	1.5
1mhr †	5.5	2mhr	1.7
5pfk †	7.0	6pfk	2.6
2psi	2.9	1qlp	2.0
2rhn	3.5	1ayn	2.9
1scp †	3.0	2scp	2.0
1sdh	2.4	3sdh	1.4
1sga	2.8	2sga	1.5
1trc	3.6	1fw4	1.7

Tableau 21 : Liste des codes PDB des 26 protéines de la banque basse résolution (BR) et des protéines correspondantes dans la banque de haute résolution (HR). Les résolutions de chaque structure sont indiquées en Å. † signifie que seules les Cα étaient présents dans le fichier PDB. ‡ indique que DEFINE n'a pas pu attribuer une partie ou tous les résidus de la protéine.

Les autres écarts entre les deux banques varient de 1.6 % pour STRIDE à 3.1 % pour DSSP. C'est à ces deux méthodes que reviennent les meilleurs accords avec les cristallographes mais il est important de rappeler, encore une fois, que ces deux méthodes ne sont pas capables d'effectuer les attributions pour les plus basses résolutions (5phk, 1mhr par exemple). Pour les deux seules méthodes qui attribuent tous les résidus, les écarts sont de 3.0 % pour PSEA et 2.7 % pour VoTAP. Avec 78.6 % d'accord avec les cristallographes pour la banque BR et 81.3 % pour la banque HR, notre méthode a donc un petit avantage sur PSEA malgré le fait que ce dernier soit plus cohérent d'une banque à l'autre. Pour la protéine 5pfk (résolution de 7 Å), l'accord de notre programme avec les attributions des cristallographes de 6pfk (résolution de 2.6 Å) est de 86.2 %, l'accord de PSEA est de 81.5 %, celui de DEFINE est de 82.4 %.

méthode	proportion d'AA attribués dans BR (%)	accord entre BR et HR (%)	accord entre BR et les cristallographes (%)	accord entre HR et les cristallographes (%)
VoTAP	100.0	88.2	78.6	81.3
DSSP	74.4	90.2	81.5	84.7
PSEA	100.0	90.1	76.9	79.9
STRIDE	74.4	91.4	81.0	82.6
DEFINE	77.5	86.8	73.0	73.8

Tableau 22 : Comparaison des 5 méthodes d'attribution. Les deux dernières colonnes indiquent l'accord de chaque méthode avec les attributions présentes dans les fichiers PDB de la protéine de la banque HR.

7 - Un exemple concret

La Figure 119 et la Figure 120 illustrent les différentes attributions telles qu'elles sont réalisées par les différents programmes étudiés sur le cytochrome bovin (code PDB 1cyo, R = 1.50 Å). A partir de cet exemple particulier, deux remarques valables de manière plus générale peuvent être faites. On constate tout d'abord que dans l'ensemble, les différents programmes sont en accord avec les cristallographes et entre eux, mais il existe tout de même des divergences concernant l'attribution de structures particulières. Par exemple PSEA n'attribue pas les brins B3, B4, B5 ni les hélices A2 et A6. Plus précisément PSEA est la seule méthode à ne pas attribuer B3 alors que A2 n'est attribuée par aucune méthode à part VoTAP, le brin B5 n'est pas détecté non plus par DEFINE ainsi que l'hélice A6. A l'inverse, DSSP et STRIDE attribuent un brin supplémentaire entre A1 et B2 et DEFINE fait de même à l'extrémité C-terminale. Ces divergences montrent que si sur l'ensemble les diverses

méthodes s'accordent, certaines structures sont problématiques comme le sont A2, B4, B5 et A6. Les critères utilisés et les « cut-off » associés se montrent ainsi ou trop stricts ou trop tolérants quand se présentent certaines irrégularités dans les structures. De plus, la définition propre à chaque programme de ce qu'est une structure pose quelques problèmes de cohérence entre les différentes méthodes, par exemple le brin B4 n'est détecté que par DSSP et STRIDE mais que par un seul résidu. Sachant que PSEA n'attribue pas de brins en dessous de trois résidus, il est possible que les critères de PSEA aient détecté ce brin mais avec un nombre insuffisant de résidus. Ceci est le cas pour VoTAP, qui détecte correctement les deux premiers résidus du brin, mais considérant, comme PSEA, que l'existence d'un brin n'est réelle qu'à partir de trois résidus, notre méthode n'attribue finalement aucune structure. Il est important de noter par ailleurs que DSSP et STRIDE ne considèrent pas B4 comme un brin non plus mais comme un « β bridge ». La nécessité de comparer les différentes méthodes entre elles impose de traduire les attributions spécifiques à chaque algorithme dans un langage commun (ici une attribution à trois états a, b et c) dont les conventions peuvent bien sûr influencer les résultats. Ceci est également valable pour le brin que détecte DSSP et STRIDE entre A1 et B2 et pour le brin final de DEFINE, deux brins qui ne comprennent que deux résidus.

La seconde observation principale que l'on peut faire concernant la cohérence entre les divers programmes d'attribution a déjà été notée par Colloc'h et coll au sujet des limites des structures secondaires et donc de leurs longueurs. L'exemple proposé montre à quel point les programmes divergent sur ce point puisque la plupart des structures détectées présente des limites différentes. Là encore, on retrouve les problèmes d'artefacts dus aux critères numériques et aux « cut-off » associés et aussi le problème plus profond de la définition même d'une hélice et d'un brin sur laquelle les différents programmes semblent diverger.

Figure 119 : Attributions de la protéine 1cyo selon les différentes méthodes. Les structures secondaires sont repérées en fonction de l'attribution des cristallographes, B pour les brins (de 1 à 5) et A pour les hélices (de 1 à 6) et soulignées. Les caractères gras correspondent aux attributions communes avec celles des cristallographes, les capitales aux attributions communes à toutes les méthodes. Les divergences de longueurs sont en italique.

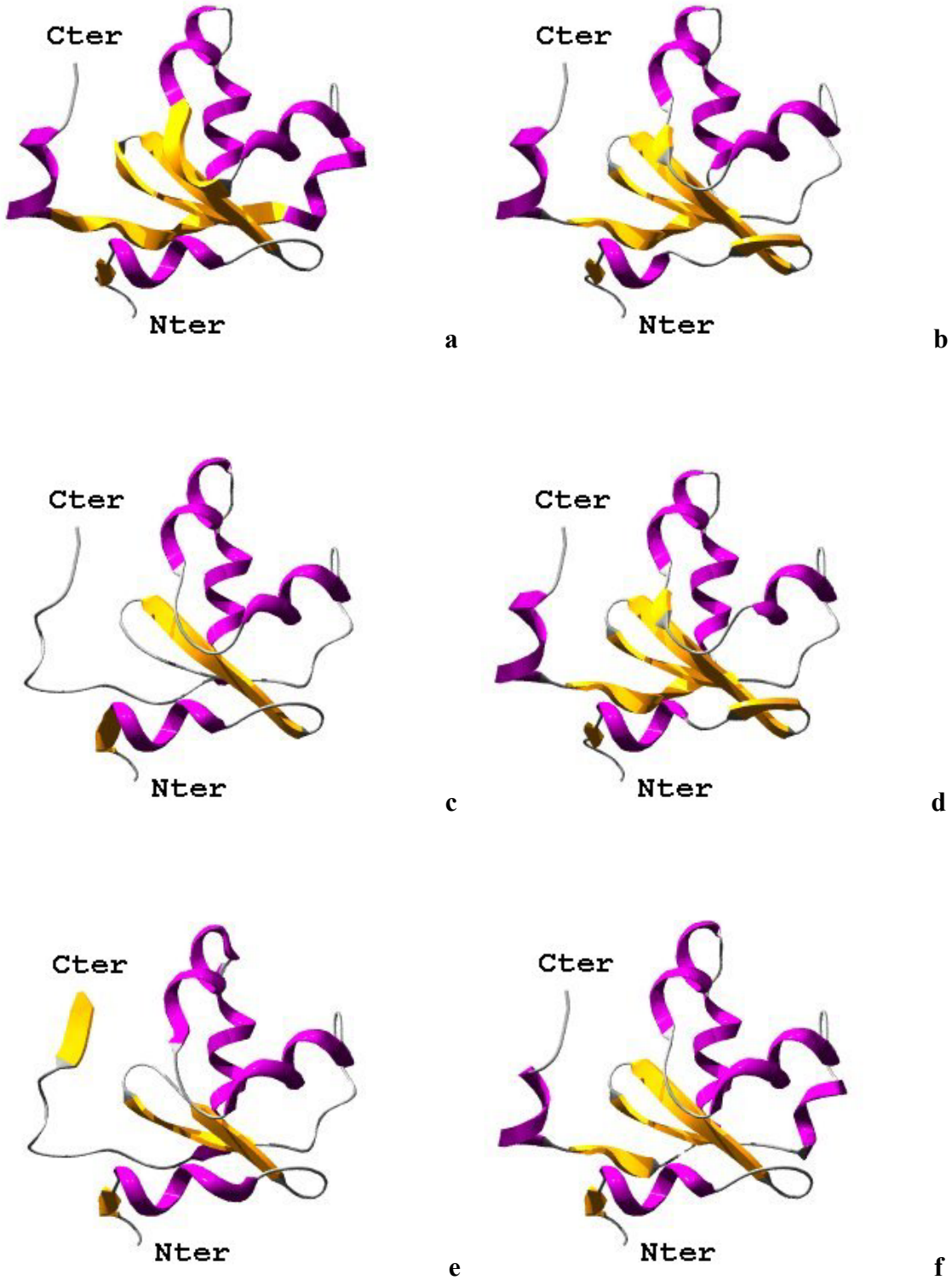


Figure 120 : Les différentes attributions de 1cyo. a - cristallographes, b - DSSP, c - PSEA, d - STRIDE, e - DEFINE et f - VoTAP. Les hélices sont en violet et les brins en jaune.

8 - Conclusion

Dans ce chapitre nous avons vu que les TdV permettent d'élaborer un outil très efficace pour attribuer les structures secondaires de n'importe quelle structure protéique à partir de la position des Cα. Puisque le seul critère utilisé est le voisinage de chaque résidu, et que notre méthode n'est fondée sur aucune définition figée, les valeurs numériques telles que les distances ou les angles et les « cut-off » qui leur sont associés s'avèrent inutiles. Les artefacts dus à ces « cut-off » sont ainsi estompés, ce qui a pour principale conséquence de régulariser les limites des structures. De plus, la souplesse des TdV permet d'obtenir de bons résultats même pour les basses résolutions. Les qualités des consensus qui avaient été énoncées par Colloc'h et coll se trouvent ici confirmés, il est intéressant de noter par ailleurs que ces qualités ne sont pas spécifiques au problème de l'attribution puisqu'il a été montré[157] qu'il était préférable de combiner les méthodes de prédiction de structures secondaires plutôt que de les utiliser seules. Cette étude a par ailleurs fait l'objet d'un article qui paraîtra prochainement dans la revue « Proteins, Structure, Function and Genetics ».

Conclusion générale

Le travail de thèse présenté ici m'a permis d'aborder la structure des protéines selon un axe relativement nouveau. En effet si l'approche par TdV n'est pas récente, elle demeure curieusement peu explorée et assez peu utilisée. Notre façon d'appliquer les TdV reste originale et les résultats, ainsi que les applications présentées, semblent prometteurs. J'ai montré en effet que les TdV pouvaient servir à attribuer les structures secondaires régulières à partir de structures expérimentales et qu'elles permettaient de définir une nouvelle accessibilité à l'environnement/solvant. En ce qui concerne ces deux applications, les résultats obtenus sont comparables aux méthodes actuellement utilisées. De plus, comme je l'ai souvent évoqué, un des atouts majeurs des TdV est qu'elles permettent de définir des contacts de manière absolue sans l'utilisation de cut-off. J'ai également conçu d'autres applications à l'aide des TdV ; par exemple, j'ai élaboré un programme permettant de faire des recherches de motifs de structures secondaires en incluant des structures « décoratives » c'est à dire non indispensable au repliement. Cet outil a été utilisé afin de rechercher des supports potentiels pour modéliser la forme pathogène du prion dont la structure reste à ce jour inconnue[105, 106]. J'ai également créé une procédure permettant d'établir une banque d'hélices internes (c'est à dire totalement enfouies) qui a aidé à caractériser la composition en AA des peptides de fusion[158]. Mon travail de thèse a donc permis de mieux connaître les TdV construites au niveau des AA sur les protéines et a montré à travers quelques applications qu'elles pouvaient être exploitées à des fins utiles. Ces applications, quoique modestes, ne montrent pas moins que les TdV peuvent constituer un outil efficace et pertinent dont il reste néanmoins à améliorer les performances, notamment en terme de temps de calcul. D'autres études concernant les TdV sont actuellement en cours : Jean-François SADOC du Laboratoire de Physique des Solides à Orsay, parvient entre autre, à diminuer le nombres de petites faces (à trois et quatre côtés) de manière à obtenir un ensemble de cellules aux formes plus régulières, le but idéal étant d'associer à chaque catégorie d'AA une cellule type ou un ensemble de cellules types permettant la reconstruction de protéines à la manière d'un « lego ». Il parvient également par un système de pondération plus complexe que celui présenté ici à obtenir des

cellules dont le volume, non plus moyen mais réel est égal à celui déterminé par Pontius. Anne Poupon et son étudiante en thèse Julie Bernauer du Laboratoire d'Enzymologie et Biochimie Structurales utilisent les TdV pour modéliser les surfaces des protéines afin de mettre au point un programme de « docking », leur permettant d'estimer la possibilité pour deux structures de pouvoir s'associer (s'emboîter), dans le but de mieux comprendre le fonctionnement des protéines. Dans notre laboratoire, les TdV seront associées à une autre thématique très largement développée dans notre équipe : HCA[69, 159, 160] (Hydrophobic Cluster Analysis). Cette approche permet, à partir d'un codage binaire des AA fondé sur la dichotomie hydrophile/hydrophobe, de regrouper les AA relativement proches le long de la séquence en amas hydrophobes dont les formes peuvent être associées aux brins β ou aux hélices α. Les TdV permettront peut être de caractériser ces amas dans l'espace à trois dimensions et également d'étudier les relations de voisinages qu'entretiennent ces différents amas entre eux.

Mon travail sur les TdV a donné lieu à plusieurs publications :

- Protein secondary structure assignment through Voronoï tessellation. Dupuis F, Sadoc JF, Mornon JP. Proteins, 2003.Sous presse.
- Structural features of prions explored by sequence analysis I. Sequence data. Mornon JP, Prat K, Dupuis F, Callebaut I. Cell Mol Life Sci 2002; 59 : 1366-1376.
- Structural features of prions explored by sequence analysis. II. A PrP(Sc) model. Mornon JP, Prat K, Dupuis F, Boisset N, Callebaut I. Cell Mol Life Sci 2002; 59 : 2144-2154.
- Viral fusion peptides and identification of membrane-interacting segments. Del Angel VD, Dupuis F, Mornon JP, Callebaut I. Biochem Biophys Res Commun 2002; 293 : 1153-1160.

Deux autres articles sont en cours de préparation : un concernant la proximité des extrémités N- et C-terminales et l'autre concernant le logiciel Voro3D.

Bibliographie

Ouvrage de référence :

Spatial Tessellations: Concepts and Applications of Voronoï Diagrams.

Okabe A, Boots B, Sugihara K, Chiu S N. Chichester: John Wiley, 2000.

Références :

1 **Zuckerkandl E, Pauling L**. Molecules as documents of evolutionary history. J Theor Biol: 1965; 8: *357-366.*

2 **Fitch WM, Margoliash E**. Construction of phylogenetic trees. Science: 1967; 155: *279-284.*

3 **Dayhoff MO**. Computer analysis of protein evolution. Sci Am: 1969; 221: *86-95.*

4 **Luscombe NM, Greenbaum D, Gerstein M**. What is bioinformatics? A proposed definition and overview of the field. Methods Inf Med: 2001; 40: *346-358.*

5 **Richards FM**. The interpretation of protein structures: total volume, group volume distributions and packing density. J Mol Biol: 1974; 82: *1-14.*

6 **Harpaz Y, Gerstein M, Chothia C**. Volume changes on protein folding. Structure: 1994; 2: *641-649.*

7 **Gerstein M, Tsai J, Levitt M**. The volume of atoms on the protein surface: calculated from simulation, using Voronoi polyhedra. J Mol Biol: 1995; 249: *955-966.*

8 **Tsai J, Gerstein M**. Calculations of protein volumes: sensitivity analysis and parameter database. Bioinformatics: 2002; 18: *985-995.*

9 **Liang J, Dill KA**. Are proteins well-packed? Biophys J: 2001; 81: *751-766.*

10 **Liang J, Edelsbrunner H, Fu P, Sudhakar PV, Subramaniam S**. Analytical shape computation of macromolecules: I. Molecular area and volume through alpha shape. Proteins: 1998; 33: *1-17.*

11 **Kussell E, Shimada J, Shakhnovich EI**. Excluded volume in protein side-chain packing. J Mol Biol: 2001; 311: *183-193.*

12 **Andersson KM, Hovmoller S**. The protein content in crystals and packing coefficients in different space groups. Acta Crystallogr D Biol Crystallogr: 2000; 56: *789-790.*

13 **Fleming PJ, Richards FM**. Protein packing: dependence on protein size, secondary structure and amino acid composition. J Mol Biol: 2000; 299: *487-498.*

14 **Gerstein M, Sonnhammer EL, Chothia C**. Volume changes in protein evolution. J Mol Biol: 1994; 236: *1067-1078.*

15 **Tsai J, Taylor R, Chothia C, Gerstein M**. The packing density in proteins: standard radii and volumes. J Mol Biol: 1999; 290: *253-266.*

16 **Edelsbrunner H, Koehl P**. The weighted-volume derivative of a space-filling diagram. Proc Natl Acad Sci U S A: 2003; 100: *2203-2208.*

17 **Finney JL**. Volume occupation, environment and accessibility in proteins. The problem of the protein surface. J Mol Biol: 1975; 96: *721-732.*

18 **Tsai J, Voss N, Gerstein M**. Determining the minimum number of types necessary to represent the sizes of protein atoms. Bioinformatics: 2001; 17: *949-956.*

19 **Richards FM**. Calculation of molecular volumes and areas for structures of known geometry. Methods Enzymol: 1985; 115: *440-464*.

20 **Quillin ML, Matthews BW**. Accurate calculation of the density of proteins. Acta Crystallogr D Biol Crystallogr: 2000; 56: *791-794.*

21 **Paci E, Marchi M**. Intrinsic compressibility and volume compression in solvated proteins by molecular dynamics simulation at high pressure. Proc Natl Acad Sci U S A: 1996; 93: *11609-11614.*

22 **Pontius J, Richelle J, Wodak SJ**. Deviations from standard atomic volumes as a quality measure for protein crystal structures. J Mol Biol: 1996; 264: *121-136*.

23 **Chakravarty S, Bhinge A, Varadarajan R**. A procedure for detection and quantitation of cavity volumes proteins. Application to measure the strength of the hydrophobic driving force in protein folding. J Biol Chem: 2002; 277: *31345-31353.*

24 **Liang J, Edelsbrunner H, Fu P, Sudhakar PV, Subramaniam S**. Analytical shape computation of macromolecules: II. Inaccessible cavities in proteins. Proteins: 1998; 33: *18-29.*

25 **Wernisch L, Hunting M, Wodak SJ**. Identification of structural domains in proteins by a graph heuristic. Proteins: 1999; 35: *338-352.*

26 **Chelli R, Gervasio FL, Procacci P, Schettino V**. Stacking and T-shape competition in aromatic-aromatic amino acid interactions. J Am Chem Soc: 2002; 124: *6133-6143.*

27 **Casari G, Sippl MJ**. Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds. J Mol Biol: 1992; 224: *725-732.*

28 **Zimmer R, Wohler M, Thiele R**. New scoring schemes for protein fold recognition based on Voronoi contacts. Bioinformatics: 1998; 14: *295-308*.

29 **McConkey BJ, Sobolev V, Edelman M**. Quantification of protein surfaces, volumes and atom-atom contacts using a constrained Voronoi procedure. Bioinformatics: 2002; 18: *1365-1373.*

30 **Lo Conte L, Chothia C, Janin J**. The atomic structure of protein-protein recognition sites. J Mol Biol: 1999; 285: *2177-2198.*

31 **Nadassy K, Wodak SJ, Janin J**. Structural features of protein-nucleic acid recognition sites. Biochemistry: 1999; 38: *1999-2017.*

32 **Nadassy K, Tomas-Oliveira I, Alberts I, Janin J, Wodak SJ**. Standard atomic volumes in double-stranded DNA and packing in protein-- DNA interfaces. Nucleic Acids Res: 2001; 29: *3362-3376.*

33 **Wako H, Yamato T**. Novel method to detect a motif of local structures in different protein conformations. Protein Eng: 1998; 11: *981-990.*

34 **Singh RK, Tropsha A, Vaisman, II**. Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues. J Comput Biol: 1996; 3: *213-221.*

35 **Zheng W, Cho SJ, Vaisman, II, Tropsha A**. A new approach to protein fold recognition based on Delaunay tessellation of protein structure. Pac Symp Biocomput: 1997;: *486-497.*

36 **Munson PJ, Singh RK**. Statistical significance of hierarchical multi-body potentials based on Delaunay tessellation and their application in sequence-structure alignment. Protein Sci: 1997; 6: *1467-1481.*

37 **Gan HH, Tropsha A, Schlick T**. Lattice protein folding with two and four-body statistical potentials. Proteins: 2001; 43: *161-174.*

38 **Adamian L, Liang J**. Helix-helix packing and interfacial pairwise interactions of residues in membrane proteins. J Mol Biol: 2001; 311: *891-907.*

39 **Carter CW, Jr., LeFebvre BC, Cammer SA, Tropsha A, Edgell MH**. Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. J Mol Biol: 2001; 311: *625-638.*

40 **Fleischmann RD*, et al.*** Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science: 1995; 269: *496-512.*

41 **Lander ES*, et al.*** Initial sequencing and analysis of the human genome. Nature: 2001; 409: *860-921.*

42 **Venter JC*, et al.*** The sequence of the human genome. Science: 2001; 291: *1304-1351.*

43 **Check E**. Mouse genome: The real deal. Nature: 2002; 420: *457.*

44 **Bairoch A, Apweiler R**. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res: 2000; 28: *45-48.*

45 **Kendrew JC, Dickerson RE**. Structure of myoglobine. A three dimensional Fourier synthesis at Å of resolution. Nature: 1960; 185: *422-427.*

46 Würtrich K, *NMR of proteins and nucleic acids*. 1986, New York: Wiley.

47 **Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE**. The Protein Data Bank. Nucleic Acids Res: 2000; 28: *235-242.*

48 **Wolf YI, Grishin NV, Koonin EV**. Estimating the number of protein folds and families from complete genome data. J Mol Biol: 2000; 299: *897-905.*

49 **Wang ZX**. A re-estimation for the total numbers of protein folds and superfamilies. Protein Eng: 1998; 11: *621-626.*

50 **Zhang C, DeLisi C**. Estimating the number of protein folds. J Mol Biol: 1998; 284: *1301-1305.*

51 **Govindarajan S, Recabarren R, Goldstein RA**. Estimating the total number of protein folds. Proteins: 1999; 35: *408-414.*

52 **Koshi JM, Goldstein RA**. Mutation matrices and physical-chemical properties: correlations and implications. Proteins: 1997; 27: *336-344.*

53 **Ladunga I, Smith RF**. Amino acid substitutions preserve protein folding by conserving steric and hydrophobicity properties. Protein Eng: 1997; 10: *187-196.*

54 **Anfinsen CB**. Principles that govern the folding of protein chains. Science: 1973; 181: *223-230.*

55 **Holm L, Sander C**. Dictionary of recurrent domains in protein structures. Proteins: 1998; 33: *88-96.*

56 **Holm L, Sander C**. 3-D lookup: fast protein structure database searches at 90% reliability. Proc Int Conf Intell Syst Mol Biol: 1995; 3: *179-187.*

57 **Islam SA, Luo J, Sternberg MJ**. Identification and analysis of domains in proteins. Protein Eng: 1995; 8: *513-525.*

58 **Sowdhamini R, Rufino SD, Blundell TL**. A database of globular protein structural domains: clustering of representative family members into similar folds. Fold Des: 1996; 1: *209-220.*

59 **Taylor WR**. Protein structural domain identification. Protein Eng: 1999; 12: *203-216.*

60 **Hong L, Koelsch G, Lin X, Wu S, Terzyan S, Ghosh AK, Zhang XC, Tang J**. Structure of the protease domain of memapsin 2 (beta-secretase) complexed with inhibitor. Science: 2000; 290: *150-153.*

61 **Parsiegla G, Reverbel-Leroy C, Tardif C, Belaich JP, Driguez H, Haser R**. Crystal structures of the cellulase Cel48F in complex with inhibitors and substrates give insights into its processive action. Biochemistry: 2000; 39: *11238-11246.*

62 **Xu Z, Horwich AL, Sigler PB**. The crystal structure of the asymmetric GroEL-GroES-(ADP)7 chaperonin complex. Nature: 1997; 388: *741-750.*

63 **Angelov B, Sadoc JF, Jullien R, Soyer A, Mornon JP, Chomilier J**. Nonatomic solvent-driven Voronoi tessellation of proteins: an open tool to analyze protein folds. Proteins: 2002; 49: *446-456.*

64 **Frolow F, Harel M, Sussman JL, Mevarech M, Shoham M**. Insights into protein adaptation to a saturated salt environment from the crystal structure of a halophilic 2Fe-2S ferredoxin. Nat Struct Biol: 1996; 3: *452-458.*

65 **Stec B, Rao U, Teeter MM**. Refinement of purothionins reveals solute particles important for lattice formation and toxicity. Part 2: structure of beta -purothionin at 1.7 Å resolution. Acta Crystallogr D Biol Crystallogr: 1995; 51: *914-924*.

66 **Imada K, Inagaki K, Matsunami H, Kawaguchi H, Tanaka H, Tanaka N, Namba K**. Structure of 3-isopropylmalate dehydrogenase in complex with 3- isopropylmalate at 2.0 A resolution: the role of Glu88 in the unique substrate-recognition mechanism. Structure: 1998; 6: *971-982.*

67 **Boeckmann B, *et al.*** The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res: 2003; 31: *365-370.*

68 **Murzin AG, Brenner SE, Hubbard T, Chothia C**. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol: 1995; 247: *536-540.*

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

70 **Hennetin J, Le TK, Canard L, Colloc'h N, Mornon JP, Callebaut I**. Non-intertwined binary patterns of hydrophobic/nonhydrophobic amino acids are considerably better markers of regular secondary structures than nonconstrained patterns. Proteins: 2003; 51: *236-244.*

71 **Sadoc JF, Jullien R, Rivier N**. The Laguerre polyhedral decomposition: application to protein folds. Eur. Phys. J. B: 2003; 33: *355-363*.

72 **Soyer A, Chomilier J, Mornon JP, Jullien R, Sadoc JF**. Voronoi tessellation reveals the condensed matter character of folded proteins. Phys Rev Lett: 2000; 85: *3532-3535.*

73 Sadoc JF, Mosseri R, *Geometrical Frustration*. 1999, Cambridge: Cambridge University Press.

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

75 **Vendruscolo M, Najmanovich R, Domany E**. Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading? Proteins: 2000; 38: *134-148.*

76 **Hinds DA, Levitt M**. Exploring conformational space with a simple lattice model for protein structure. J Mol Biol: 1994; 243: *668-682.*

77 **Mirny L, Domany E**. Protein fold recognition and dynamics in the space of contact maps. Proteins: 1996; 26: *391-410.*

78 **Mirny LA, Shakhnovich EI**. How to derive a protein folding potential? A new approach to an old problem. J Mol Biol: 1996; 264: *1164-1179.*

79 **Miyazawa S, Jernigan RL**. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol: 1996; 256: *623-644.*

80 **Skolnick J, Jaroszewski L, Kolinski A, Godzik A**. Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? Protein Sci: 1997; 6: *676-688.*

81 **Thomas PD, Dill KA**. An iterative method for extracting energy-like quantities from protein structures. Proc Natl Acad Sci U S A: 1996; 93: *11628-11633.*

82 **Shrake A, Rupley JA**. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. J Mol Biol: 1973; 79: *351-371.*

83 **Gerstein M, Chothia C**. Packing at the protein-water interface. Proc Natl Acad Sci U S A: 1996; 93: *10167-10172.*

84 **Fedorov AN, Friguet B, Djavadi-Ohaniance L, Alakhov YB, Goldberg ME**. Folding on the ribosome of Escherichia coli tryptophan synthase beta subunit nascent chains probed with a conformation-dependent monoclonal antibody. J Mol Biol: 1992; 228: *351-358.*

85 **Fedorov AN, Baldwin TO**. Contribution of cotranslational folding to the rate of formation of native protein structure. Proc Natl Acad Sci U S A: 1995; 92: *1227-1231.*

86 **Lietzow MA, Jamin M, Jane Dyson HJ, Wright PE**. Mapping long-range contacts in a highly unfolded protein. J Mol Biol: 2002; 322: *655-662.*

87 **Roder H, Elove GA, Englander SW**. Structural characterization of folding intermediates in cytochrome c by H-exchange labelling and proton NMR. Nature: 1988; 335: *700-704.*

88 **Ptitsyn OB**. Protein folding : general physical model. FEBS Lett: 1981; 131: *197-202*.

89 **Taylor WR, Lin K**. Protein knots: A tangled problem. Nature: 2003; 421: *25.*

90 **Taylor WR**. A deeply knotted protein structure and how it might fold. Nature: 2000; 406: *916-919.*

91 **Mansfield ML**. Are there knots in proteins? Nat Struct Biol: 1994; 1: *213-214.*

92 **Thornton JM, Sibanda BL**. Amino and carboxy-terminal regions in globular proteins. J Mol Biol: 1983; 167: *443-460.*

93 **Christopher JA, Baldwin TO**. Implications of N and C-terminal proximity for protein folding. J Mol Biol: 1996; 257: *175-187.*

94 **Baker W, van den Broek A, Camon E, Hingamp P, Sterk P, Stoesser G, Tuli MA**. The EMBL nucleotide sequence database. Nucleic Acids Res: 2000; 28: *19-23.*

95 **Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL**. GenBank. Nucleic Acids Res: 2000; 28: *15-18.*

96 **Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ**. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res: 1997; 25: *3389-3402.*

97 **Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM**. CATH--a hierarchic classification of protein domain structures. Structure: 1997; 5: *1093-1108.*

98 **Cheah E, Carr PD, Suffolk PM, Vasudevan SG, Dixon NE, Ollis DL**. Structure of the Escherichia coli signal transducing protein PII. Structure: 1994; 2: *981-990.*

99 **Inoue T, Nishio N, Kanamoto K, Suzuki S, Yamaguchi K, Kataoka K, Tobari J, Kai Y**. Crystallization and preliminary X-ray study of iso-2 azurin from the methylotrophic bacterium, Methylomonas J. Acta Crystallogr D Biol Crystallogr: 1999; 55: *307-309.*

100 **Djordjevic S, Goudreau PN, Xu Q, Stock AM, West AH**. Structural basis for methylesterase CheB regulation by a phosphorylation-activated domain. Proc Natl Acad Sci U S A: 1998; 95: *1381-1386.*

101 **Rousseau F, Schymkowitz JW, Itzhaki LS**. The unfolding story of three-dimensional domain swapping. Structure (Camb): 2003; 11: *243-251.*

102 **Bennett MJ, Choe S, Eisenberg D**. Domain swapping: entangling alliances between proteins. Proc Natl Acad Sci U S A: 1994; 91: *3127-3131.*

103 **Bennett MJ, Schlunegger MP, Eisenberg D**. 3D domain swapping: a mechanism for oligomer assembly. Protein Sci: 1995; 4: *2455-2468.*

104 **Liu Y, Eisenberg D**. 3D domain swapping: as domains continue to swap. Protein Sci: 2002; 11: *1285-1299.*

105 **Mornon JP, Prat K, Dupuis F, Boisset N, Callebaut I**. Structural features of prions explored by sequence analysis. II. A PrP(Sc) model. Cell Mol Life Sci: 2002; 59: *2144-2154.*

106 **Mornon JP, Prat K, Dupuis F, Callebaut I**. Structural features of prions explored by sequence analysis I. Sequence data. Cell Mol Life Sci: 2002; 59: *1366-1376.*

107 **Newcomer ME**. Trading places. Nat Struct Biol: 2001; 8: *282-284.*

108 **Janowski R, Kozak M, Jankowska E, Grzonka Z, Grubb A, Abrahamson M, Jaskolski M**. Human cystatin C, an amyloidogenic protein, dimerizes through three-dimensional domain swapping. Nat Struct Biol: 2001; 8: *316-320.*

109 **Hakansson M, Svensson A, Fast J, Linse S**. An extended hydrophobic core induces EF-hand swapping. Protein Sci: 2001; 10: *927-933.*

110 **Sinha N, Tsai CJ, Nussinov R**. A proposed structural model for amyloid fibril elongation: domain swapping forms an interdigitating beta-structure polymer. Protein Eng: 2001; 14: *93-103.*

111 **Liu Y, Gotte G, Libonati M, Eisenberg D**. A domain-swapped RNase A dimer with implications for amyloid formation. Nat Struct Biol: 2001; 8: *211-214.*

112 **Staniforth RA, Giannini S, Higgins LD, Conroy MJ, Hounslow AM, Jerala R, Craven CJ, Waltho JP**. Three-dimensional domain swapping in the folded and molten-globule states of cystatins, an amyloid-forming structural superfamily. Embo J: 2001; 20: *4774-4781.*

113 **Flick KE, Jurica MS, Monnat RJ, Jr., Stoddard BL**. DNA binding and cleavage by the nuclear intron-encoded homing endonuclease I-PpoI. Nature: 1998; 394: *96-101.*

114 **Kobe B, Deisenhofer J**. Mechanism of ribonuclease inhibition by ribonuclease inhibitor protein based on the crystal structure of its complex with ribonuclease A. J Mol Biol: 1996; 264: *1028-1043.*

115 **Lamarine M, Mornon JP, Berezovsky N, Chomilier J**. Distribution of tightened end fragments of globular proteins statistically matches that of topohydrophobic positions: towards an efficient punctuation of protein folding? Cell Mol Life Sci: 2001; 58: *492-498.*

116 **Berezovsky IN, Trifonov EN**. Loop fold structure of proteins: resolution of Levinthas paradox. J Biomol Struct Dyn: 2002; 20: *5-6.*

117 **Berezovsky IN, Trifonov EN**. Loop fold nature of globular proteins. Protein Eng: 2001; 14: *403-407.*

118 **Berezovsky IN, Kirzhner VM, Kirzhner A, Trifonov EN**. Protein folding: looping from hydrophobic nuclei. Proteins: 2001; 45: *346-350.*

119 **Berezovsky IN, Kirzhner VM, Kirzhner A, Rosenfeld VR, Trifonov EN**. Closed loops: persistence of the protein chain returns. Protein Eng: 2002; 15: *955-957.*

120 **Berezovsky IN, Grosberg AY, Trifonov EN**. Closed loops of nearly standard size: common basic element of protein structure. FEBS Lett: 2000; 466: *283-286.*
121 **Kostyuchenko VA, Navruzbekov GA, Kurochkina LP, Strelkov SV, Mesyanzhinov VV, Rossmann MG**. The structure of bacteriophage T4 gene product 9: the trigger for tail contraction. Structure Fold Des: 1999; 7: *1213-1222.*
122 **Bhattacharya AA, Grune T, Curry S**. Crystallographic analysis reveals common modes of binding of medium and long-chain fatty acids to human serum albumin. J Mol Biol: 2000; 303: *721-732.*
123 **Nikkola M, Lindqvist Y, Schneider G**. Refined structure of transketolase from Saccharomyces cerevisiae at 2.0 A resolution. J Mol Biol: 1994; 238: *387-404.*
124 **Holland DR, Tronrud DE, Pley HW, Flaherty KM, Stark W, Jansonius JN, McKay DB, Matthews BW**. Structural comparison suggests that thermolysin and related neutral proteases undergo hinge-bending motion during catalysis. Biochemistry: 1992; 31: *11310-11316.*
125 **Rossmann M, Liljas A**. Recognition of Structural Domains in Globular Proteins. J Mol Biol: 1974; 85: *177-181*.
126 **Sippl MJ**. On the problem of comparing protein structures. Development and applications of a new method for the assessment of structural similarities of polypeptide conformations. J Mol Biol: 1982; 156: *359-388.*
127 **Nishikawa K, Ooi T**. Comparison of homologous tertiary structures of proteins. J Theor Biol: 1974; 43: *351-374.*
128 **Phillips DC**. The development of crystallographic enzymology. Biochem Soc Symp: 1970; 30: *11-28*.
129 **Go M**. Correlation of DNA exonic regions with protein structural units in haemoglobin. Nature: 1981; 291: *90-92.*
130 **Holm L, Sander C**. Protein structure comparison by alignment of distance matrices. J Mol Biol: 1993; 233: *123-138.*
131 **Havel TF, Kuntz ID, Crippen GM**. The theory and practice of distance geometry. Bull. Math. Biol.: 1983; 45: *665-720*.
132 **Saitoh S, Nakai T, Nishikawa K**. A geometrical constraint approach for reproducing the native backbone conformation of a protein. Proteins: 1993; 15: *191-204.*
133 **Selbig J**. Contact pattern-induced pair potentials for protein fold recognition. Protein Eng: 1995; 8: *339-351.*
134 **Galaktionov S, Nikiforovich GV, Marshall GR**. Ab initio modeling of small, medium, and large loops in proteins. Biopolymers: 2001; 60: *153-168*.
135 **Singer MS, Vriend G, Bywater RP**. Prediction of protein residue contacts with a PDB-derived likelihood matrix. Protein Eng: 2002; 15: *721-725.*
136 **Kim MK, Jernigan RL, Chirikjian GS**. Efficient generation of feasible pathways for protein conformational transitions. Biophys J: 2002; 83: *1620-1630.*
137 **Goel NS, Ycas M**. On the Computation of the Tertiary Structure of Globular Proteins II. J Theor Biol: 1979; 77: *253-305*.
138 **Hutchinson EG, Thornton JM**. PROMOTIF--a program to identify and analyze structural motifs in proteins. Protein Sci: 1996; 5: *212-220.*
139 **Pauling L, Corey RB, Branson HR**. The structure of proteins: Two hydrogen-bonded helical configurations of the polypetptide chain. Proc Natl Acad Sci U S A: 1951; 37: *205-234*.
140 **Pauling L, Corey RB**. Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets. Proc Natl Acad Sci U S A: 1951; 37: *729-740*.

141 **Lewis PN, Momany FA, Scheraga HA**. Folding of polypeptide chains in proteins: a proposed mechanism for folding. Proc Natl Acad Sci U S A: 1971; 68: *2293-2297.*

142 **Kuntz ID**. Protein folding. J Am Chem Soc: 1972; 94: *4009-4012.*

143 **Levitt M, Greer J**. Automatic identification of secondary structure in globular proteins. J Mol Biol: 1977; 114: *181-239.*

144 **Crawford JL, Lipscomb WN, Schellman CG**. The reverse turn as a polypeptide conformation in globular proteins. Proc Natl Acad Sci U S A: 1973; 70: *538-542.*

145 **Rose GD, Seltzer JP**. A new algorithm for finding the peptide chain turns in a globular protein. J Mol Biol: 1977; 113: *153-164.*

146 **Chou PY, Fasman GD**. Beta-turns in proteins. J Mol Biol: 1977; 115: *135-175.*

147 **Kolaskar AS, Ramabrahmam V, Soman KV**. Reversals of polypeptide chain in globular proteins. Int J Pept Protein Res: 1980; 16: *1-11.*

148 **Ramakrishnan C, Soman KV**. Identification of secondary structures in globular proteins--a new algorithm. Int J Pept Protein Res: 1982; 20: *218-237.*

149 **Richards FM, Kundrot CE**. Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. Proteins: 1988; 3: *71-84*.

150 **Sklenar H, Etchebest C, Lavery R**. Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis. Proteins: 1989; 6: *46-60*.

151 **Frishman D, Argos P**. Knowledge-based protein secondary structure assignment. Proteins: 1995; 23: *566-579.*

152 Andersen CA, Rost B, *Secondary Structure Assignment*, in *Structural Bioinformatics*, P. Bourne and H. Weissig, Editors. 2002: Wiley.

153 **Labesse G, Colloc'h N, Pothier J, Mornon JP**. P-SEA: a new efficient assignment of secondary structure from C alpha trace of proteins. Comput Appl Biosci: 1997; 13: *291-295.*

154 **Colloc'h N, Etchebest C, Thoreau E, Henrissat B, Mornon JP**. Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. Protein Eng: 1993; 6: *377-382.*

155 **Colloc'h N, Cohen FE**. Beta-breakers: an aperiodic secondary structure. J Mol Biol: 1991; 221: *603-613.*

156 **Presnell SR, Cohen BI, Cohen FE**. A segment-based approach to protein secondary structure prediction. Biochemistry: 1992; 31: *983-993.*

157 **King RD, Ouali M, Strong AT, Aly A, Elmaghraby A, Kantardzic M, Page D**. Is it better to combine predictions? Protein Eng: 2000; 13: *15-19.*

158 **Del Angel VD, Dupuis F, Mornon JP, Callebaut I**. Viral fusion peptides and identification of membrane-interacting segments. Biochem Biophys Res Commun: 2002; 293: *1153-1160.*

159 **Gaboriaud C, Bissery V, Benchetrit T, Mornon JP**. Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences. FEBS Lett: 1987; 224: *149-155.*

160 **Lemesle-Varloot L, Henrissat B, Gaboriaud C, Bissery V, Morgat A, Mornon JP**. Hydrophobic cluster analysis: procedures to derive structural and functional information from 2-D-representation of protein sequences. Biochimie: 1990; 72: *555-574.*