



HAL
open science

Analyse à grande échelle des textures des séquences protéiques via l'approche Hydrophobic Cluster Analysis (HCA).

Karine Albeau

► **To cite this version:**

Karine Albeau. Analyse à grande échelle des textures des séquences protéiques via l'approche Hydrophobic Cluster Analysis (HCA).. Autre [q-bio.OT]. Université de Versailles-Saint Quentin en Yvelines, 2005. Français. NNT: . tel-00011139

HAL Id: tel-00011139

<https://theses.hal.science/tel-00011139>

Submitted on 1 Dec 2005

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

MINISTERE DE LA JEUNESSE, DE L'EDUCATION NATIONALE
ET DE LA RECHERCHE

ECOLE PRATIQUE DES HAUTES ETUDES
UNIVERSITE VERSAILLES SAINT-QUENTIN-EN-YVELINES

THESE de DOCTORAT

présentée et soutenue publiquement
par

Karine ALBEAU

Pour l'obtention du grade de DOCTEUR de l'Ecole Pratique des Hautes Etudes

Discipline : Sciences de la Vie et de la Terre

Analyse à grande échelle des textures des séquences protéiques via
l'approche Hydrophobic Cluster Analysis (HCA).

Soutenance le 25 octobre 2005

devant le jury suivant :

Dr. Sophie Zinn-Justin	Rapporteur
Dr. Nathalie Colloc'h	Rapporteur
Dr. Stan Tomavo	Examineur
Dr. Jean-Paul Mornon	Examineur
Pr. Bernard Mignotte	Directeur de thèse

Laboratoire de génétique et biologie cellulaire – EPHE
Université Versailles Saint-Quentin-en-Yvelines
Bât Fermat, 45 avenue des Etats-Unis 78035 Versailles

Institut de Minéralogie et de Physique des Milieux Condensés (IMPMC)
CNRS UMR 7590, Universités Paris 6 et Paris 7, campus Boucicaut
140 rue de Lourmel 75015 Paris

Remerciements

Je remercie Bernard Capelle de m'avoir permis d'effectuer ce travail au sein du Laboratoire de Minéralogie - Cristallographie - Paris, puis en celui de l'Institut de Minéralogie et de Physique des Milieux Condensés (IMPMC) dont il est le directeur.

Je remercie Bernard Mignotte d'avoir accepté la direction de cette thèse.

Je remercie Jean-Paul Mornon qui a initié ce sujet et qui a suivi mes travaux tout au long de cette thèse.

Je remercie tout particulièrement Isabelle Callebaut pour avoir dirigé mes travaux et pour m'avoir plongée dans les subtilités d'HCA.

Je remercie Nathalie Colloc'h et Sophie Zinn-Justin d'avoir accepté d'être rapporteurs de cette thèse, ainsi que Stan Tomavo d'avoir bien voulu en être examinateur.

Je remercie François Coulier et Daniel Birnbaum du laboratoire de cancérologie expérimentale Inserm119 (Marseille) avec qui j'ai mené les études phylogénétiques du début de cette thèse, ainsi que Rémi, Alex, Claire, les deux Audrey, Elsa, Sébastien, Frédéric, André, Jean, Stéphane sans oublier Géraldo et tous ceux que j'oublie.

Une mention particulière est accordée à Magalie Leveugle qui, par sa sympathie et son aide, m'a permis de passer une excellente année à Marseille et de ramener dans mes bagages de bonnes compétences en phylogénie, mais aussi une excellente amie.

Je remercie aussi les anglais, Nathalie, Fabien et Christelle pour les discussions mémorables sur msn pendant mes heures tardives de rédaction.

Je remercie tous les membres de l'IMPMC ou ex-membres du LMCP et plus particulièrement les biologistes et ceux qui s'en rapprochent, Emilie, Franck, Jérôme, Quentin, Fériel, Marina, Nathalie, Guillaume, Eric, Jean-Christophe, Nicolas, Jacques, Jean, Annick, Sophie, Claire.

Je remercie également Mr Pierre Jollès pour sa sympathie et ses précieux conseils.

Merci enfin à Richard et Fabienne qui ont partagé toutes les étapes de cette thèse.

Merci aussi au grand Seb pour sa disponibilité, sa gentillesse et sa patience pour m'aider lors de mes débuts en programmation C.

Je remercie tous ceux qui ont cru en ce travail et qui m'ont soutenue par leurs encouragements ou leur présence.

Je remercie tout particulièrement, Benoît, qui m'a soutenue tout au long de ce travail et m'a souvent « débogué » à mes débuts.

Et bien sûr, je remercie ma mère et mon frère pour leur présence dans les moments difficiles.

A mon père,

Table des matières

Table des matières	1
Table des figures.....	5
Table des tableaux	9
Introduction générale	11
Première partie : Généralités sur l'analyse de la texture des séquences protéiques.....	15
Chapitre I Structure des protéines : généralités	17
1 Architecture des protéines	17
2 Structure secondaire des protéines.....	18
3 Structure tertiaire et repliement protéique	20
3.1 SCOP.....	22
3.2 CATH.....	23
4 Les domaines protéiques et régions non structurées	24
Chapitre II La méthode Hydrophobic Cluster Analysis	27
1 Introduction	27
2 Fondement de la méthode HCA.....	27
3 Représentation HCA : règles de segmentation de la séquence en amas	28
3.1 Représentation binaire des acides aminés	31
3.2 Un support hélicoïdal	32
3.3 Distance de connectivité.....	32
3.4 L'alphabet HCA.....	33
3.5 Un acide aminé interrupteur d'amas : la proline	33
4 Un amas correspond majoritairement à un type de structure secondaire	33
5 But et Perspectives.....	34
Chapitre III Analyse de la texture	37
1 Texture et analyse de texture : Généralités	37
1.1 Définition de la texture.....	37
1.2 Deux types de texture : aléatoire et structurée	37
2 Perception et analyse visuelle d'une texture	38
3 Quelques méthodes d'analyse de texture.....	39
3.1 Méthodes de premier ordre	39
3.2 Méthodes de second ordre : méthode de matrice de cooccurrence.....	40
3.3 Méthodes d'ordre supérieur : méthode des longueurs de plages de niveaux de gris (ou de sections).....	42
3.3.1 Paramètre SRE	43

3.3.2	Paramètre LRE	43
3.3.3	Paramètre GLD	44
3.3.4	Paramètre RLD.....	45
3.3.5	Paramètre RLP	45
4	<i>But de notre étude</i>	46
Deuxième partie : Développements méthodologiques pour analyser la texture dans les séquences protéiques.....		49
Préambule		51
Chapitre IV Exploitation des méthodes d'analyse de texture en imagerie.....		53
1	<i>Adaptation des méthodes d'analyse de texture en imagerie à notre système (tracé HCA)</i>	53
1.1	Transposition du tracé HCA dans une matrice.....	53
1.2	Choix d'une taille de fenêtre glissante	54
1.3	Choix d'un code	55
1.3.1	Code hydrophobe HCA (code01).....	55
1.3.2	Code à 4 groupes (code 1234).....	55
1.3.3	Code amas-non amas (code amas).....	56
1.4	Choix d'une direction (pour les calculs de matrice de cooccurrence et de longueurs de plage)	57
2	<i>Constitution des banques de protéines de référence</i>	57
3	<i>Résultats</i>	58
3.1	Méthode de premier ordre.....	58
3.1.1	Etude de la distribution du paramètre « pourcentage d'acides aminés hydrophobes ».....	58
3.1.2	Etude du profil d'hydrophobie	60
3.2	Méthode de cooccurrence.....	62
3.2.1	Etude des distributions des paramètres entropie et contraste	62
3.2.2	Etude de profils de cooccurrence	65
3.3	Méthode des longueurs de plage	66
3.3.1	Etude des distributions des paramètres SRE, LRE, GLD, RLD et RLP.....	66
3.3.2	Etude de profils des longueurs de plage	75
3.4	Distribution des amas hydrophobes dans les banques A, B, C et D.....	77
4	<i>Discussion et Conclusion</i>	78
Chapitre V DomHCA : un outil pour prédire les régions structurées.....		83
1	<i>Introduction : Travaux précédents</i>	83
1.1	Prédiction de domaines structuraux	83
1.2	Prédiction de « linkers » et de régions non structurées.....	84
1.3	Conclusion et but de l'étude	85
2	<i>Fondement de la procédure DomHCA et caractérisation des domaines globulaires</i>	85
2.1	Distribution des acides aminés hydrophobes V, I, L, F, M, Y et W	86
2.2	Distribution des amas hydrophobes	87
2.3	Distribution des tailles des domaines.....	88
3	<i>Principe de l'algorithme pour détecter les régions structurées</i>	90
4	<i>Ajustement des bornes des régions structurées</i>	91
4.1	Ajustement global	92
4.2	Ajustement des bornes de début et de fin de la région structurée	93

4.3	Correction apportée à la prédiction de régions structurées de petite taille entourées par de grandes régions charnières.....	93
5	<i>Score d'hydrophobie</i>	94
6	<i>Information déduite du score d'hydrophobie sur la présence éventuelle de passages membranaires dans les régions structurées</i>	95
6.1	Introduction.....	95
6.2	Indications quant à la présence éventuelle de passages membranaires dans le cadre de la prédiction DomHCA.....	96
6.3	Passages membranaires hélicoïdaux multiples	98
6.4	Passages membranaires hélicoïdaux isolés	100
6.5	Cas particuliers des porines	101
7	<i>Evaluation de la prédiction DomHCA</i>	103
7.1	Méthodologie	104
7.1.1	Constitution des échantillons tests	104
7.1.2	Cas des protéines monodomaines	105
7.1.3	Cas des protéines « pluridomaines »	106
7.2	Résultats	107
7.2.1	Hydrophobie et taille des régions prédites	107
7.2.2	Prédiction de régions structurées à partir de chaînes « monodomaines »	108
7.2.3	Prédiction des régions structurées à partir de chaînes pluri-domaines	112
7.2.4	Comparaison avec d'autres méthodes.....	119
7.3	Application aux séquences de <i>Plasmodium falciparum</i>	121
8	<i>Discussion et Conclusion</i>	122

Chapitre VI Autres développements et caractérisation de régions spécifiques dans les protéines

1	<i>Introduction</i>	127
2	<i>Identification de régions de répétition</i>	127
2.1	Méthode d'identification.....	127
2.1.1	Harmonique standard	127
2.1.2	Harmonique hydrophobe	130
2.1.3	Harmonique dégénérée.....	130
2.2	Caractérisation des harmoniques.....	131
2.2.1	Répétitions structurées, pseudo-structurées	131
2.2.2	Répétitions simples ou mixtes.....	131
3	<i>Identification de peptides de fusion</i>	132
3.1	Introduction.....	132
3.2	Méthode de détection.....	133
4	<i>Prédiction de la classe de repliement (A, B, C, D)</i>	134
4.1	Propension des acides pour un type de structure secondaire	134
4.2	Distribution des acides aminés au sein des différents types de repliements	135
4.3	Dictionnaire d'amas hydrophobes.....	136
4.3.1	Règles d'attribution des états A, B et ?	136
4.3.2	Prédiction de la tendance de repliement de régions structurées	136
4.4	Fréquence d'apparition des amas hydrophobes par rapport à l'aléatoire (Z-score) et corrélation aux structures secondaires	138
4.5	Distribution des amas en fonction de leur longueur et de leur nombre d'acides aminés hydrophobes	139

5	<i>Conclusion</i>	141
Chapitre VII Application à <i>Plasmodium falciparum</i> : Etude d'un génome particulier143		
1	<i>Introduction</i>	143
2	<i>Comparaison des génomes</i>	143
2.1	Distribution des bases et des codons dans les génomes.....	143
2.1.1	Introduction.....	143
2.1.2	Composition des génomes.....	144
2.2	Analyse de la composition en codons et acides aminés codés par les codons .	145
2.3	Analyse des protéines	149
2.4	Analyse des segments prédits structurés par DomHCA.....	150
2.5	Analyse des régions de répétitions	152
3	<i>Conclusion</i>	154
Conclusion générale et perspectives		157
Bibliographie.....		161
Annexe		171

Table des figures

Figure 1 : Représentation d'une chaîne polypeptidique.	18
Figure 2 : Hélice α et feuillet β .	19
Figure 3 : Exemples de repliement tout α (A) et tout β (B).	21
Figure 4 : Exemple de repliement α/β (A) et $\alpha+\beta$ (B).	22
Figure 5 : Répartition des protéines dans les différentes classes dans la base SCOP (version 1.65 de décembre 2003).	22
Figure 6 : Exemple d'un domaine discontinu dans la protéine 2MNR.	25
Figure 7 : Principe de la méthode HCA, illustré sur un segment de la séquence de l' α 1-antitrypsine (tiré de [CALLEBAUT et al., 1997]).	29
Figure 8 : Correspondance entre amas et structures secondaires illustrée pour un domaine globulaire complet (PDB 3CHY, Swiss-Prot CHEY_ECOLI).	30
Figure 9 : Exemple de textures structurées : A gauche, le « mur de briques » du catalogue de Brodatz et à droite, le « grillage » [BRODATZ, P. 1966].	37
Figure 10 : Exemple de textures aléatoires [BRODATZ, P. 1966].	38
Figure 11 : Illustration de la grossièreté sur deux textures de Brodatz différentes [BRODATZ, P. 1966].	38
Figure 12 : Illustration du contraste sur deux textures de Brodatz différentes [BRODATZ, P. 1966].	39
Figure 13 : Remplissage d'une matrice de cooccurrence.	41
Figure 14 : Remplissage d'une matrice de longueur de plages.	43
Figure 15 : Illustration des paramètres SRE et LRE.	44
Figure 16 : Illustration du paramètre GLD.	44
Figure 17 : Illustration du paramètre RLD.	45
Figure 18 : Illustration du paramètre RLP.	46
Figure 19: Représentation bidimensionnelle HCA de la séquence du prion humain.	47
Figure 20 : Représentation bidimensionnelle HCA et découpage en domaines de la protéine MAL8P1.111 de <i>Plasmodium Falciparum</i> .	48
Figure 21 : Transposition du tracé HCA dans une matrice.	53
Figure 22 : Illustration de la première et de la deuxième couronne HCA.	54
Figure 23 : Enfouissement moyen de chaque acide aminé, présenté en mode décroissant à partir des données brutes de Pintar et al [PINTAR, A. et al., 2003b].	56
Figure 24 : Principe de la fenêtre glissante.	58
Figure 25 : Distribution du pourcentage d'acides aminés hydrophobes, code 1234, fenêtre 17.	59
Figure 26 : Profil d'hydrophobie de différentes protéines issues des banques A, B, C, D, F et G.	62
Figure 27 : Distribution des valeurs du contraste dans chaque classe de notre banque en utilisant une fenêtre de 17 acides aminés.	64
Figure 28 : Distribution de l'entropie dans chaque classe avec la fenêtre de 17 acides aminés.	65
Figure 29 : Profil de cooccurrence de protéines issues de nos banques A, B, C, D, F et G.	66
Figure 30 : Définition de plages d'amas (à gauche) et de plages d'hydrophobes (à droite) dans une fenêtre.	67

Figure 31 : Distribution de SRE dans chaque classe en utilisant une fenêtre de 17 acides aminés.	69
Figure 32 : Distribution de LRE dans chaque banque avec la fenêtre de 17 acides aminés.	70
Figure 33 : Distribution de GLD dans chaque banque avec la fenêtre de 17 acides aminés.	71
Figure 34 : Distribution de RLD dans chaque classe en utilisant une fenêtre de 17 acides aminés.	73
Figure 35 : Distribution de RLP dans chaque classe en utilisant une fenêtre de 17 acides aminés.	74
Figure 36 : Profils des longueurs de plage de protéines issues de nos banques.	76
Figure 37 : Distribution du nombre d'acides aminés hydrophobes contenus dans les amas en fonction de leur longueur.	78
Figure 38 : Corrélation entre profil hydrophobe et représentation HCA.	80
Figure 39 : Distribution du pourcentage d'acides aminés hydrophobes (VILFMYW) sur un ensemble de séquences issues de SCOP (classe ABCDE, 7530 séquences).	86
Figure 40 : Distribution du pourcentage d'acides aminés hydrophobes (VILFMYW sur un ensemble de séquences issues de SCOP (classe ABCDE; 7530 séquences soit 2 472 934 fenêtres).	87
Figure 41 : Distribution des longueurs des domaines des séquences monodomaines et pluridomaines.	88
Figure 42 : Structure répertoriée dans CATH comme « monodomaine ».	89
Figure 43 : Schématisation de la procédure DomHCA pour la détection des régions structurées dans les séquences de protéines.	91
Figure 44 : Exemple de protéine « multidomaine », avec une région charnière séparant les deux domaines.	93
Figure 45 : Exemple de région intrinsèquement désordonnée, qui se replie après liaison avec un partenaire.	94
Figure 46 : Différentes configurations de protéines membranaires.	96
Figure 47 : Profils hydrophobes de protéines possédant des passages membranaires multiples.	97
Figure 48 : Passages membranaires de cinq protéines polytopiques.	99
Figure 49 : Tracé HCA de la protéine CFTR renfermant deux régions membranaires (MSD1 et MSD2) possédant chacune six hélices membranaires.	100
Figure 50 : Hélice transmembranaire détectée par DomHCA.	101
Figure 51 : Structure tridimensionnelle de la maltoporine (code PDB : 1AF6).	101
Figure 52 : Profil d'hydrophobie de la phosphoporine (code PDB : 1PHO).	102
Figure 53 : Exemple de passage membranaire bêta : la protéine porine 1PHO.	103
Figure 54 : Codification des différentes séquences protéiques étudiées.	105
Figure 55 : Paramètres pris en compte pour le calcul de score de validation.	105
Figure 56 : Profil hydrophobe des régions structurées prédites avec DomHCA.	107
Figure 57 : Distribution de la longueur des segments structurés prédits par DomHCA et comparaison avec les tailles des segments monodomaines définis à partir de notre échantillon « monodomaines » d'après CATH (cf. Figure 41).	108
Figure 58 : Distribution des scores DO obtenus avec DomHCA pour les jeux de données SLP et SEP.	109
Figure 59 : Distribution du nombre de prédictions correctes en fonction de l'imprécision autorisée.	109
Figure 60 : Alignement de la séquence PDB, de la séquence du domaine CATH et de la séquence de la région structurée prédite par DomHCA.	110
Figure 61 : Distribution de la valeur EDO.	111
Figure 62 : Distribution des valeurs n'.	111

Figure 63 : Linkers « 0 » et linkers « sans rupture ».	113
Figure 64 : Répartition des linkers totaux identifiés dans les chaînes pluridomaines.	113
Figure 65 : Plusieurs exemples de résultats positifs et négatifs obtenus avec DomHCA.	114
Figure 66 : Linkers correctement prédits par DomHCA.	116
Figure 67 : Distribution des longueurs des boucles « linkers » et des boucles « non-linkers » (boucles simples).	117
Figure 68 : Corrélation entre la composition en acides aminés des boucles « linkers » et des boucles « non-linkers ».	118
Figure 69 : Propensions des acides aminés pour les structures secondaires [CALLEBAUT, I. et al., 1997a].	118
Figure 70 : Composition en acides aminés des boucles non-linkers, des faux positifs et des faux négatifs de notre banque de séquences pluridomaines.	119
Figure 71 : Comparaison de méthodes de prédiction sur la protéine aconitase (1C96), chaîne A.	120
Figure 72 : Résultats des méthodes de prédiction testées dans CAFASP.	121
Figure 73 : Comparaison de la distribution des acides aminés dans les protéomes (protéines prédites) de <i>Plasmodium falciparum</i> , <i>Homo sapiens</i> , <i>Caenorhabditis elegans</i> et <i>Saccharomyces cerevisiae</i> (5334, 32035, 21629 et 6699 séquences, respectivement) [CALLEBAUT, I. et al., 2005].	121
Figure 74 : Exemple de découpage permettant une meilleure recherche automatique de domaines.	124
Figure 75 : Comparaison de deux harmoniques de longueurs et niveaux différents.	129
Figure 76 : Comparaison de deux harmoniques de niveaux différents.	129
Figure 77 : Exemple de deux harmoniques se chevauchant, celles-ci sont assemblées dans notre traitement.	129
Figure 78 : Exemple d'harmonique hydrophobe (PF11_0204).	130
Figure 79 : Exemple d'harmonique dégénérée (PFB0580w).	131
Figure 80 : Prédiction d'un peptide de fusion.	134
Figure 81 : Distribution des acides aminés dans les banques SCOP A, B, C et D.	135
Figure 82 : Distribution des Z-scores associés aux amas présents dans les banques A, B, C et D de SCOP (normalisation par le nombre d'amas de chaque banque).	139
Figure 83 : Distribution des tailles des amas dans les banques A, B, C, D, F et G de SCOP.	139
Figure 84 : Distribution du rapport S/L dans les banques A, B, C et D de SCOP.	140
Figure 85 : Séquence d'un fragment de gène issu de <i>Plasmodium falciparum</i> (24% GC).	144
Figure 86 : Usage du code génétique.	146
Figure 87 : Comparaison de la distribution des acides aminés dans les protéomes (protéines prédites) de <i>Plasmodium falciparum</i> , <i>Homo sapiens</i> , <i>Caenorhabditis elegans</i> et <i>Saccharomyces cerevisiae</i> (5334, 32035, 21629 et 6699 séquences, respectivement) [CALLEBAUT, I. et al., 2005].	147
Figure 88 : Comparaison de la distribution des acides aminés dans les protéomes de <i>Plasmodium falciparum</i> et <i>Dictyostelium discoideum</i> (5334 et 13574 séquences).	149
Figure 89 : Comparaison des tailles des protéines des protéomes de <i>Plasmodium falciparum</i> , <i>Homo sapiens</i> , <i>Caenorhabditis elegans</i> et <i>Saccharomyces cerevisiae</i> .	149
Figure 90 : Comparaison des tailles des régions structurées issues des protéomes de <i>Plasmodium falciparum</i> , <i>Homo sapiens</i> , <i>Caenorhabditis elegans</i> et <i>Saccharomyces cerevisiae</i> .	151
Figure 91 : Comparaison de la distribution des acides aminés dans les régions structurées prédites par DomHCA dans les protéomes (protéines prédites) de <i>Plasmodium falciparum</i> , <i>Homo sapiens</i> , <i>Caenorhabditis elegans</i> et <i>Saccharomyces cerevisiae</i> .	151

- Figure 92 : Comparaison des tailles des régions de répétition issues des protéomes de *Plasmodium falciparum*, *Homo sapiens*, *Caenorhabditis elegans* et *Saccharomyces cerevisiae*. 153
- Figure 93 : Comparaison de la distribution des acides aminés dans les régions de répétitions issues des protéomes (protéines prédites) de *Plasmodium falciparum*, *Homo sapiens*, *Caenorhabditis elegans* et *Saccharomyces cerevisiae*. 153
- Figure 94 : Distribution du nombre de régions de répétition contenues dans les protéines de *Plasmodium falciparum*. 154

Table des tableaux

Tableau 1 : Constitution des quatre groupes d'acides aminés suivant la valeur d'enfouissement moyen.	56
Tableau 2 : Effectifs des banques de protéines.	58
Tableau 3 : Méthode de premier ordre.	59
Tableau 4 : Méthode de cooccurrence.	63
Tableau 5 : Méthode des longueurs de plage, code Amas, fenêtre 17.	68
Tableau 6 : Méthode des longueurs de plage, code 01 et 1234, fenêtres 17 et 11.	68
Tableau 7 : Banque « Témoin » de <i>Plasmodium falciparum</i> .	92
Tableau 8 : Conversion des états DSSP en 3 états.	104
Tableau 9 : Résultats des prédictions de DomHCA.	115
Tableau 10 : Domaines identifiés dans MAL8P1.111 à partir d'une recherche de type HMMER dans les banques de domaines SMART et PFAM.	125
Tableau 11 : Présentation des quatre types d'harmoniques.	128
Tableau 12 : Exemples de chacun des trois types de régions répétitives.	131
Tableau 13 : Illustration d'harmoniques structurées simples et mixtes.	132
Tableau 14: Distribution des fréquences GC dans les séquences codantes (CDS) de plusieurs génomes.	145
Tableau 15 : Composition moyenne des protéines (site internet : (http://www.esil.inuvmrs.fr/~dgaut/Cours/biais.html)).	145
Tableau 16 : Bilan des caractéristiques des séquences incluses dans les protéomes de <i>Plasmodium falciparum</i> , <i>Homo sapiens</i> , <i>Caenorhabditis elegans</i> et <i>Saccharomyces cerevisiae</i> .	150
Tableau 17 : Bilan des régions structurées prédites par DomHCA dans les protéines des génomes de <i>Plasmodium falciparum</i> , <i>Homo sapiens</i> , <i>Caenorhabditis elegans</i> et <i>Saccharomyces cerevisiae</i> .	150
Tableau 18 : Bilan des régions de répétitions identifiées dans les protéines des génomes de <i>Plasmodium falciparum</i> , <i>Homo sapiens</i> , <i>Caenorhabditis elegans</i> et <i>Saccharomyces cerevisiae</i> .	152

Introduction générale

Le terme de « génome » a été introduit en 1920 par Hans Winkler pour désigner l'ensemble haploïde de des gènes d'un organisme. La nature du gène était encore inconnue. Puis, vers les années 1950, l'ADN a été identifié comme support du matériel génétique. La compréhension du génome implique une étude fonctionnelle afin d'identifier les informations génétiques qu'il contient, structurale pour connaître les différents niveaux d'organisation qui le compose et évolutive pour comprendre le génome actuel.

Les six premiers mois de cette thèse ont été consacrés à la poursuite d'une étude menée sur la présence d'évènements de duplications de gènes ou de portions de génomes entre la séparation des invertébrés et des cordés. En effet, des comparaisons de la quantité d'ADN chez des organismes modèles comme l'amphioxus et certains mammifères ont permis de proposer un modèle d'évolution des vertébrés impliquant deux évènements de duplications « 2R » [OHNO, S. 1970; OHNO, S. 1993]. La première duplication aurait eu lieu à la séparation des Céphalocordés et la seconde lors de l'apparition des vertébrés à mâchoires. Ainsi, les génomes des vertébrés tendraient à présenter souvent quatre copies de la même famille de gènes alors que les génomes invertébrés ne possèderaient qu'une seule copie. Pour vérifier cette hypothèse des « 2R », il faudrait pouvoir comparer les familles de gènes dupliqués chez les vertébrés avec le génome d'une espèce proche, mais située avant les évènements de duplication sur l'arbre phylogénétique des Chordés. La phylogénie de cet embranchement indiquait l'amphioxus comme un bon modèle car il faisait l'objet de recherches de synténies locales [ABI-RACHED, L. et al., 2002] et se plaçait juste avant les « 2R », mais son génome n'était pas encore séquencé. En 2002, le génome de *Ciona intestinalis*, qui se situe juste avant l'amphioxus dans l'arbre des Chordés, a été séquencé. Nous avons donc choisi d'étudier cette espèce pour voir si cet embranchement possédait les caractéristiques d'un génome non dupliqué à grande échelle, pour laquelle une seule copie de gène est présente en quatre exemplaires chez les vertébrés.

Cette étude, qui ne sera pas présentée dans ce manuscrit, est détaillée dans un article scientifique [LEVEUGLE, M. et al., 2004] (article inséré à la fin de la thèse). Ce travail a été mené conjointement avec Magalie Leveugle qui a présenté sa thèse intitulée « Evolution et duplications des génomes de vertébrés : Base de données et phylogénie » en novembre 2004 [LEVEUGLE, M. 2004].

Un défi important de l'ère post-génomique est de pouvoir déterminer les fonctions des protéines codées par les génomes et les interactions qu'elles opèrent entre elles. La prédiction et la caractérisation expérimentale des structures de ces protéines sont au cœur de cette problématique, puisque les fonctions des protéines sont très généralement tributaires de leurs conformations. Le traitement automatique des données brutes issues des génomes par les voies standards de la bioinformatique (recherche de similarités de séquence, reconnaissance de repliement, ...) laisse cependant bon nombre de séquences de protéines orphelines de familles structurales et fonctionnelles, ce qui rend leur caractérisation délicate. Ainsi, une part importante des protéomes, allant jusqu'à 60% des protéines prédites dans le génome de *Plasmodium falciparum* [GARDNER, M. J. et al., 2002], échappe à une prédiction fonctionnelle. Cependant, ces séquences pourraient souvent être reliées à d'autres, mais elles ont tellement divergé au cours du temps qu'elles ne partagent plus en commun que quelques résidus essentiels à leur repliement et à leur fonction, de sorte que leur parenté reste non significativement identifiée par les algorithmes classiques de comparaison de séquences. Par ailleurs, ces protéines sont souvent constituées de plusieurs domaines, dont la taille n'excède pas 400 acides aminés, et contiennent parfois des régions de faible complexité de séquence.

Un traitement efficace des protéomes nécessite donc de pré-découper les séquences en leurs différents domaines, qu'il est alors plus aisé de comparer aux banques de données, en évitant de nombreux biais, que ce soit par les recherches littérales classiques (BLAST, PSI-BLAST) ou par expertise à l'aide de méthodes de recherche de motifs ou à l'aide de la méthode Hydrophobic Cluster Analysis (HCA) [CALLEBAUT, I. et al., 1997a].

Dans ce contexte, le but de mon travail de thèse a été de mettre au point un programme de découpage systématique et automatique des séquences, utilisable à grande échelle sur des protéomes complets et basé sur la notion de texture 2D des séquences. Par rapport à d'autres méthodes [GEORGE, R. A. et al., 2002b; RIGDEN, D. J. 2002], cette démarche possède l'avantage de pouvoir être effectuée à partir d'une seule séquence, en l'absence de similitudes avec d'autres séquences, et permettrait de compléter d'autres approches de découpage « ab-initio », qui reposent sur l'utilisation de différents critères, comme l'entropie [GALZITSKAYA, O. V. et al., 2003] ou la prédiction de structures secondaires [MARSDEN, R. L. et al., 2002], [SUYAMA, M. et al., 2003].

La méthode HCA, développée au sein du département de Biologie Structurale de l'IMPMC et que je décrirai en détail dans la première partie de ce manuscrit, ne repose pas uniquement sur une analyse littérale (acide aminé par acide aminé) des séquences de

protéines, à la source même des limitations des méthodes d'analyse classique utilisées, mais sur une analyse de celles-ci au travers de leurs structures secondaires, éléments beaucoup plus stables face à la divergence des séquences considérées. Un second aspect de la méthode HCA est sa capacité à faire ressortir visuellement des régions présentant une texture particulière dans les séquences. En effet, la transposition bidimensionnelle HCA des séquences protéiques en acides aminés permet une étude de leur texture plus pertinente que celle qui pourrait être directement menée en une dimension (1D). La texture d'une protéine est souvent loin d'être uniforme et révèle la présence de régions de structures différentes, par exemple des domaines globulaires (caractérisés par une texture régulière d'amas hydrophobes de taille moyenne et couvrant environ un tiers de la surface), des domaines membranaires (caractérisés par de longs amas hydrophobes), des domaines non structurés présentant peu d'amas hydrophobes. La délimitation et la caractérisation de ces domaines, de quelques dizaines à quelques centaines d'acides aminés, sont cruciales pour décrypter efficacement les séquences dans l'optique d'une étude à grande échelle des génomes séquencés. Cette texture 2D des séquences, telle qu'offerte par HCA, pourrait être également mise à profit pour étudier plus avant les caractéristiques des séquences non globulaires, formant une part importante des protéomes et qui apparaissent intrinsèquement non structurées ou adopteraient des structures non globulaires dont la conformation reste inconnue [WRIGHT, P. E. et al., 1999; DUNKER, A. K. et al., 2001a; DUNKER, A. K. et al., 2001b]. Dans ce contexte, nous avons essayé d'extraire des régions de différentes textures dans les séquences protéiques en adaptant certaines méthodes d'analyse de texture utilisées en imagerie médicale.

Dans la première partie de ce travail, je présenterai une brève description des structures protéiques, de la méthode Hydrophobic Cluster Analysis et des méthodes d'analyse de texture. Dans une deuxième partie, je détaillerai notre utilisation des méthodes d'analyse de texture pour tenter de délimiter des régions caractéristiques au sein des protéines (chapitre IV). Je présenterai également DomHCA, un outil de prédécoupage automatique des séquences en domaines structurés, domaines non structurés et passages membranaires (chapitre V) et autres régions de répétitions (chapitre VI) que j'ai développé afin d'effectuer un premier découpage rapide, automatique et à grande échelle des séquences protéiques des génomes séquencés, comme par exemple celui de *Plasmodium falciparum* (chapitre VII).

Première partie :
Généralités sur l'analyse de la texture des
séquences protéiques

Chapitre I

Structure des protéines : généralités

1 *Architecture des protéines*

Les protéines sont un des éléments essentiels au maintien des processus nécessaires à la vie. Ce sont des polymères biologiques d'une grande variété fonctionnelle et structurale. Elles représentent une des classes les plus importantes des molécules biologiques en raison de leur capacité, entre autres, à catalyser spécifiquement une réaction, à s'auto-assembler (oligomère), à transporter des ions ou de petites molécules à travers différents milieux cellulaires, ou à réguler l'expression des gènes. Ce sont ces gènes eux-mêmes qui codent pour les protéines. Les régions codantes des gènes (ADN) sont transcrites en ARN messagers, puis ceux-ci sont traduits en séquences d'acides aminés par les ribosomes qui allongent les chaînes polypeptidiques par formation séquentielle de liaisons peptidiques (CO-NH) entre le groupement acide d'un acide aminé et le groupement amine de l'acide aminé suivant. Chaque acide aminé est constitué d'un carbone (appelé carbone C α), substitué par un groupement carboxyle (COOH), un groupement amine (NH₂), un atome d'hydrogène (H) et un radical R. Les radicaux R sont appelés chaînes latérales, alors que les atomes C α , N, C et O constituent la chaîne principale de la protéine ou squelette. La nature des radicaux confère au résidu des propriétés chimiques particulières (hydrophobie, charge, flexibilité, encombrement stérique). Une des plus importantes est l'hydrophobie, considérée comme le moteur du repliement protéique [KOSHI, J. M. et al., 1997; LADUNGA, I. et al., 1997].

La chaîne principale a une conformation définie pour chaque résidu par trois angles dièdres : ψ , ϕ et ω (Figure 1). L'angle phi (ϕ) est l'angle de rotation autour de la liaison N-C α , l'angle psi (ψ) est l'angle de rotation autour de la liaison C α -C'. La liaison peptidique est plane et en conformation trans ($\omega=0$) dans la majorité des cas, pour des raisons d'encombrement stérique (seule la proline est en conformation cis dans 10% des cas). Les deux angles dièdres ψ et ϕ permettent de définir les conformations énergétiques favorables de la chaîne polypeptidique dans l'espace et de préciser les interactions entre les différents groupements portés par les acides aminés [RAMAKRISHNAN, C. et al., 1965].

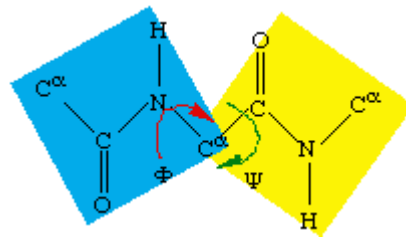


Figure 1 : Représentation d'une chaîne polypeptidique.

L'angle de rotation autour de la liaison N-C α est l'angle phi (ϕ) et celui autour de la liaison C α -C est l'angle psi (ψ). Figure reproduite de http://www.ujf-grenoble.fr/BIO/Gurvan/phi_psi.html.

Les contraintes stériques sont minimales pour la glycine ($-180^\circ < \Phi < 180^\circ$) et maximales pour la proline ($-90^\circ < \Phi < -40^\circ$). Tous les acides aminés hydrophobes ont normalement des angles Φ négatifs.

L'enchaînement des acides aminés le long de cette chaîne est la structure primaire ou séquence de la protéine. Cette séquence renferme l'ensemble des informations nécessaires au repliement de la protéine [ANFENSEN, C. B. 1973].

2 Structure secondaire des protéines

Les groupements amines et carboxyles des résidus de la chaîne principale sont capables de former des liaisons hydrogène, donnant naissance à des segments polypeptidiques caractérisés par des répétitions régulières locales. Ces régions appelées « structures secondaires régulières » minimisent les répulsions stériques entre les chaînes latérales des différents acides aminés, et maximisent le nombre de liaisons hydrogène : elles sont donc énergétiquement favorisées. Les angles possibles et les structures qu'ils engendrent le plus souvent sont représentés sur un diagramme de Ramachandran [RAMAKRISHNAN, C. et al., 1965].

L'hélice α (Figure 2 à gauche), l'une des deux structures secondaires régulières, se présente sous la forme d'un enroulement hélicoïdal de la chaîne d'acides aminés stabilisé par des interactions hydrogènes entre les résidus i et les résidus $i+4$, dont le pas est de 3.6 acides aminés par tour. Les hélices α sont plus ou moins longues, allant de 4 à 40 acides aminés, la longueur moyenne étant de 12. Certains acides aminés se retrouvent préférentiellement dans les hélices α . L'alanine, la leucine, l'acide glutamique et la méthionine sont des acides aminés formateurs d'hélice α . A l'inverse, on rencontre peu de glycine, de tyrosine, de sérine et de proline. De plus, une proline sera plus souvent au début d'une hélice α qu'à la fin [ALBERTS, B et al., 1997]. Si tous les résidus (ou la plupart) sont hydrophobes sur une face de l'hélice, l'autre face étant tapissée de résidus hydrophiles, l'hélice

α adopte un caractère amphiphile qui peut lui permettre de s'associer à d'autres faces hydrophobes (d'hélice α , de membrane, de feuillet β , etc...).

L'autre type de structure secondaire régulière est le brin β , dont l'association forme les feuillets β (le brin β isolé étant très peu stable). Contrairement à l'hélice α qui est une structure continue, le feuillet β est construit par l'assemblage de plusieurs régions polypeptidiques distinctes appelées « brins β » longs en général de 5 à 10 acides aminés (Figure 2, à droite). Ces brins sont adjacents, soit parallèlement (dans un feuillet β , tous les brins β sont alignés dans la même direction), soit antiparallèlement (les brins β ont des directions alternées).

La longueur des liaisons hydrogène est différente suivant l'arrangement des brins β au sein du feuillet: 0,31 nm pour un feuillet de brins parallèles et 0,29 nm pour un feuillet de brins antiparallèles. Les brins β sont reliés entre eux par des liaisons hydrogène formées entre le groupement CO d'un résidu d'un premier brin et la fonction NH d'un second brin adjacent. Certains acides aminés ont tendance à se retrouver préférentiellement dans les structures secondaires de type feuillet β : ce sont entre autres la valine, l'isoleucine et la thréonine. Le nombre de brins par feuillet varie de deux à une vingtaine, souvent il est compris entre 3 et 5.

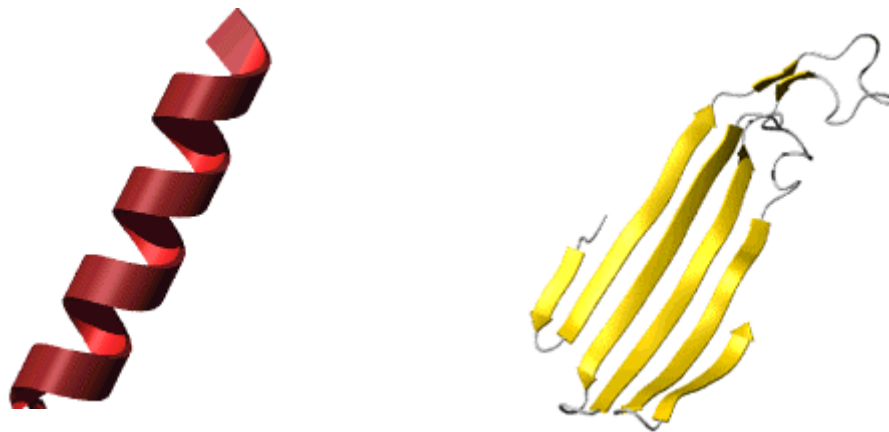


Figure 2 : Hélice a et feuillet b.

A gauche, Hélice a représentée en ruban, ce style de schématisation est très utilisé pour représenter les protéines, il permet de mieux appréhender l'architecture globale des structures. A droite, feuillet antiparallèle, le sens Nter vers Cter est indiqué sur les brins par une flèche (tiré du site internet: <http://www.ac-orleans-tours.fr/svt/mol3d/3d/module3/html/3page.html>).

Un grand nombre de protéines sont constituées par une alternance de structures secondaires (hélices α et/ou feuillet β) connectées par des boucles ou des tours, qui sont des structures dites non régulières ("coils"). On réserve généralement le nom de tour à la structure

qui connecte deux brins β antiparallèles permettant à la chaîne polypeptidique de faire un demi-tour. Les tours sont généralement courts (2 à 4 acides aminés en dehors des brins). Le terme de boucle est plutôt utilisé pour des séquences plus longues (plus de 4 résidus en général), qui peuvent alors prendre un plus grand nombre de conformations que les tours. Ces boucles connectent généralement des hélices α entre elles, ainsi que des hélices α avec des brins β , ou encore deux brins β spatialement distants. La longueur moyenne des boucles est de 6 à 10 acides aminés. La combinaison d'hélices et de feuillets confère à la protéine un cœur hydrophobe stable. Les boucles sont retrouvées à la surface des protéines et sont largement exposées au solvant. En conséquence, elles sont riches en acides aminés chargés et polaires.

3 Structure tertiaire et repliement protéique

La structure tertiaire est formée par l'agencement des structures secondaires entre elles. Elle est le résultat d'interactions diverses (liaisons hydrogène, hydrophobes, électrostatiques, covalentes comme les ponts disulfures...) entre acides aminés de la même chaîne principale, mais non voisins dans la séquence. Cette configuration stable et définie, qui permet à des acides aminés séquentiellement éloignés de se retrouver côte à côte, est primordiale pour l'activité biologique des protéines.

L'un des facteurs les plus importants qui gouverne le repliement d'une protéine est la distribution de ses chaînes latérales polaires et non polaires [BRANDEN, C et al., 1991]. Celui-ci s'effectue sans réarrangement des liaisons covalentes chimiques de la protéine à l'exception parfois de la formation de ponts disulfures entre des cystéines. Les nombreuses chaînes latérales hydrophobes d'une protéine ont tendance à être agglomérées à l'intérieur de la molécule, ce qui leur permet d'éviter le contact avec l'environnement aqueux. Au contraire, presque toutes les chaînes latérales polaires ont tendance à se placer près de l'extérieur de la molécule protéique, où elles peuvent interagir avec l'eau et avec d'autres molécules polaires. La stabilité du repliement provient des interactions faibles entre les atomes situés au cœur de la protéine et de ceux exposés au solvant. Ces interactions sont les forces de Van der Waals, les liaisons hydrogène et les liaisons ioniques. Le repliement d'une protéine s'effectue sous contrôle thermodynamique (la forme repliée possède normalement l'énergie la plus basse), dicté uniquement par l'information contenue dans sa séquence [ANFENSEN, C. B. 1973].

De manière remarquable, les structures protéiques ne définissent qu'un nombre limité et réduit de repliements distincts, un repliement (ou « fold ») étant caractérisé par une disposition unique, ou très proche, de structures secondaires régulières et par une connexion

(topologie) identique de ces structures secondaires. Les génomes procaryotes et eucaryotes codent pour plusieurs milliers de protéines de repliements divers. A l'heure actuelle nous connaissons au niveau structural plus de la moitié des types de repliements indépendants qu'utiliseraient les protéines. En effet, pour plus de 10^{12} séquences protéiques que contiendrait la biosphère terrestre, on estime que seul environ un millier de types de repliements tridimensionnels indépendants existeraient [WANG, Z. X. 1998; ZHANG, C. et al., 1998; GOVINDARAJAN, S. et al., 1999].

Les différentes topologies de repliement sont classées principalement en quatre familles : « tout α » ou « tout β » quand les hélices ou les brins représentent respectivement au moins 90% des structures secondaires de la protéine, « α/β » quand il y a une alternance de brins et d'hélices et « $\alpha+\beta$ » lorsqu'il y a ségrégation entre deux parties séparées de brins et d'hélices (Figure 3 et Figure 4). Il existe cependant plusieurs autres catégories regroupant les protéines multi-domaines, les protéines ayant peu de structures secondaires et les protéines membranaires.

Deux bases de données répertorient ces repliements protéiques : SCOP [MURZIN, A. G. et al., 1995] et CATH [ORENGO, C. A. et al., 1997; HADLEY, C. et al., 1999]. La banque de données SCOP regroupe les protéines de la Protein Data Bank (PDB) [BERMAN, H. M. et al., 2000] présentant une relation de similarité structurale et d'évolution. La classification de SCOP comprend les 5 classes de repliement définies ci-dessus alors que CATH n'en contient que 4 en regroupant les repliements α/β et $\alpha+\beta$ dans une classe unique alpha-bêta comme défini initialement par Levitt et Chothia [LEVITT, M. et al., 1976]).

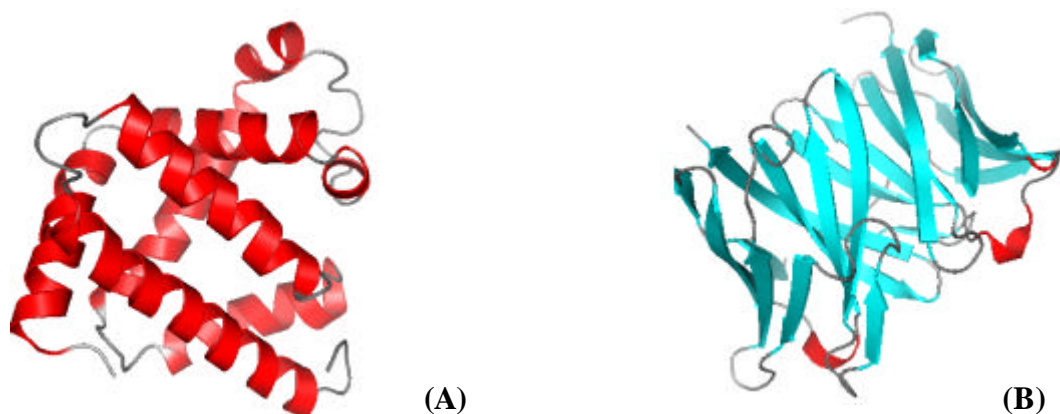


Figure 3 : Exemples de repliement tout α (A) et tout β (B).
A gauche, 2MM1 : myoglobine (*Homo sapiens*); à droite, 1MFA : immunoglobuline (*Mus musculus*).

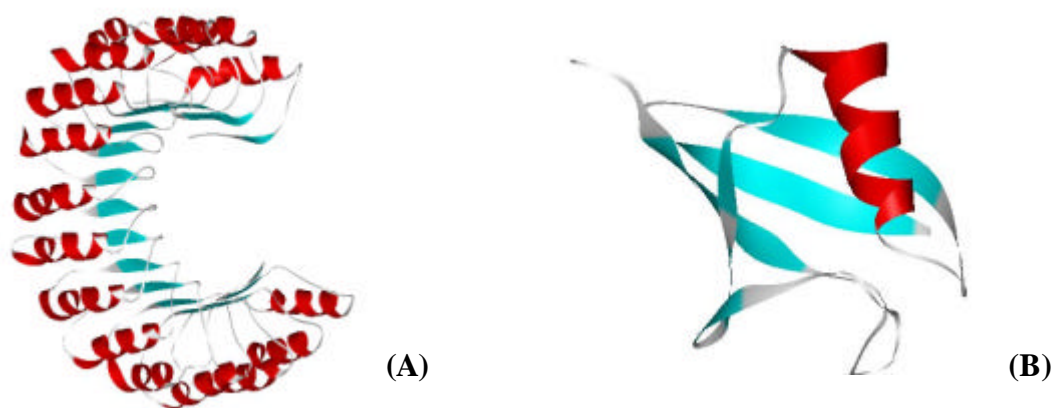


Figure 4 : Exemple de repliement α/β (A) et $\alpha+\beta$ (B).

A gauche, 2BNH : Inhibiteur de ribonucléase (*Sus Scrofa*); à droite, 1UBI : Ubiquitine (*Homo sapiens*).

3.1 SCOP

La classification SCOP (pour Structural Classification Of Proteins) comporte principalement cinq classes de repliements : A (tout α), B (tout β), C (α/β), D ($\alpha+\beta$), auxquelles se sont ajoutées rapidement les classes E (protéines multidomaines), F (protéines membranaires et de surface) et G (petites protéines). A ces classes se sont ajoutées récemment quatre classes moins bien définies (I, J, K et L) correspondant respectivement aux protéines « coiled_coil » (formant des tresses d'hélices), aux structures protéiques déterminées à faible résolution, aux peptides et aux modèles de protéines (Figure 5).

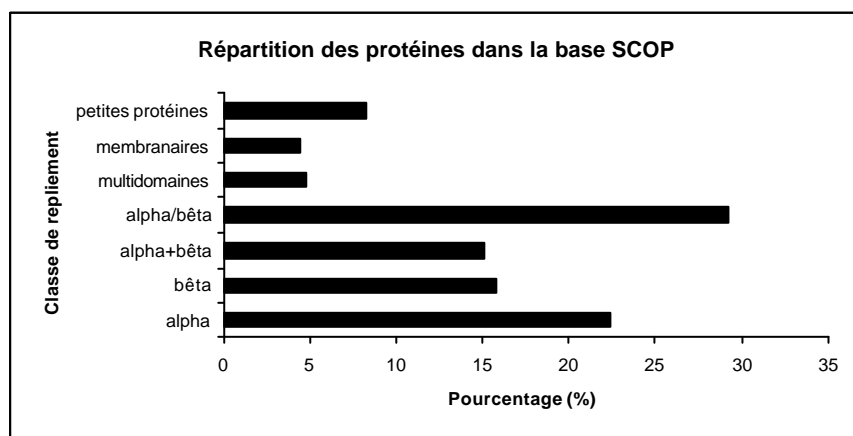


Figure 5 : Répartition des protéines dans les différentes classes dans la base SCOP (version 1.65 de décembre 2003).

Chaque classe de repliement comporte trois niveaux de hiérarchie :

- les familles regroupent des protéines ayant un lien de parenté. Généralement, elles présentent plus de 30% d'identité de séquence entre elles. Néanmoins, dans certains cas, leurs

fonctions et structures similaires suffisent à les inclure dans une même famille (par exemple, les globines forment une famille dont les membres présentent seulement environ 15% d'identité),

- les super familles rassemblent des protéines présentant une faible identité de séquence entre elles mais dont les repliements et fonctions semblent voisins et seraient probablement dû à une origine évolutive commune,
- le repliement regroupe les protéines partageant une majorité de structures secondaires dans le même arrangement et avec les mêmes connexions topologiques. Dans cette classe, on peut retrouver des protéines présentant des similarités structurales et n'ayant aucun lien de parenté entre elles.

3.2 CATH

La classification CATH (pour "Class Architecture Topology Homologous superfamily") est également une banque hiérarchique de structures protéiques. Les protéines sont désignées comme multi-domaines ou monodomaines, d'après trois algorithmes de reconnaissance de domaines (DETECTIVE [SWINDELLS, M. B. 1995], PUU [HOLM, L. et al., 1994] et DOMAK [SIDDIQUI, A. S. et al., 1995]). Ces domaines sont classés suivant quatre niveaux principaux de hiérarchie :

- C : La **classe** structurale est définie à partir de la composition en structures secondaires. Pour être qualifiées de tout α , les protéines doivent comporter plus de 50 % d'hélices α et moins de 5 % de brins β dans leur structure. De plus, elles doivent avoir plus de 50 % d'interactions ou de contacts $\alpha\alpha$ et moins de 5 % de contacts $\beta\beta$.

Pour être qualifiées de tout β , les protéines doivent comporter plus de 50 % de feuillets β et moins de 5 % d'hélices α dans leur structure. De plus, elles doivent avoir moins de 10 % d'interactions ou de contacts $\alpha\alpha$ et plus de 50 % de contacts $\beta\beta$.

Les repliements α/β et $\alpha+\beta$ sont regroupés dans une même classe de repliement. Celle-ci est définie par un ensemble d'hélices α et de feuillets β tel qu'il y ait entre 15 % et 55 % d'hélices α et entre 10 % et 45 % de feuillets β dans le repliement. Contrairement à $\alpha+\beta$, α/β contient beaucoup de feuillets β antiparallèles [MICHIE, A. D. et al., 1996].

Une dernière classe contient les domaines protéiques avec peu de structures secondaires.

- A : L'**architecture** décrit l'arrangement spatial des structures secondaires sans tenir compte de leurs connectivités. Dans ce niveau, sont retrouvées par exemple les protéines adoptant les architectures en tonneau ou en sandwich trois couches.

- T : La **topologie** rassemble des structures dont l'architecture et les connexions entre structures secondaires sont globalement proches [TAYLOR, W. R. et al., 1989].

- H : Les **familles** de protéines Homologues regroupent les domaines protéiques considérés comme partageant un ancêtre commun et pouvant ainsi être décrites comme homologues. Les similarités sont d'abord identifiées par comparaison de séquences puis par comparaison de structure. D'autres procédures utilisent les scores de SSAP pour relier des protéines avec un taux plus faible d'identité de séquence [TAYLOR, W. R. et al., 1989].

Il existe une dernière classe (S) qui regroupe les familles de séquences. Cette sous-classe de H contient les protéines avec des identités de séquence supérieures à 35%.

Un même type de repliement protéique peut donc être généré par un très grand nombre de séquences naturelles, souvent très différentes entre elles. Depuis la résolution de la première structure protéique en 1960 [KENDREW, J.C. et al., 1960], le nombre de structures résolues n'a cessé de croître. Ainsi, la Protein Data Bank (PDB) compte actuellement 31 059 entrées structurales (24 mai 2005) [BERMAN, H. M. et al., 2000]. On estime que le nombre de repliements différents dans la nature est de l'ordre du millier [WANG, Z. X. 1998; ZHANG, C. et al., 1998; GOVINDARAJAN, S. et al., 1999; WOLF, Y.I. et al., 2000] ce qui est très inférieur par rapport aux nombre de séquences connues (banque Non Redondante du NCBI, 2 543 432 séquences en juin 2005) [WOODSMALL, R. M. et al., 1993]. D'autre part, il apparaît que le nombre de séquences ne cesse de croître alors que la détection de repliements nouveaux tend à diminuer.

Des domaines protéiques présentant un même type de repliement ont souvent des séquences conservées et des fonctions voisines. Ainsi, lorsque deux séquences protéiques présentent plus de 30 % d'identité de séquence, elles adoptent, sauf exception rarissime, un même repliement [SANDER, C. et al., 1991; CHOTHIA, C. et al., 1997]. Cependant des protéines peuvent être souvent très divergentes (e.g. 10% d'identité de séquence) et être très similaires au niveau de leur structure tertiaire et de leur fonction.

4 Les domaines protéiques et régions non structurées

Les domaines protéiques sont des unités spatiales distinctes et compactes adoptant des repliements indépendants et qui ont souvent une fonction bien spécifique [HOLM, L. et al., 1995]. Les domaines structuraux sont la base de l'architecture des protéines, de leur

fonction et de leur évolution. Un domaine protéique peut être présent dans un grand nombre de contextes : les domaines peuvent s'associer avec d'autres domaines de nature très différente et selon des ordres très variés. Les domaines protéiques ont des longueurs comprises entre 20 et 400 acides aminés. Les domaines les plus petits (environ 40 acides aminés) sont généralement stabilisés par des ions métalliques comme le zinc ou par des ponts disulfures. Bien que les protéines puissent être très longues en séquence, actuellement peu de domaines possédant plus de 400 acides aminés ont été identifiés. Des distinctions peuvent être faites entre domaines protéiques continus et discontinus. Un domaine continu est formé d'une région peptidique continue qui se replie pour former un domaine indépendant et unique. Un domaine discontinu est composé de régions peptidiques non consécutives formant un domaine unique (Figure 6). L'évolution des protéines a en effet parfois entraîné des insertions de domaines dans d'autres domaines, créant des segments discontinus. Il a été mis en évidence sur les structures connues qu'un peu moins d'un tiers des domaines structuraux sont discontinus [JONES, S. et al., 1998].

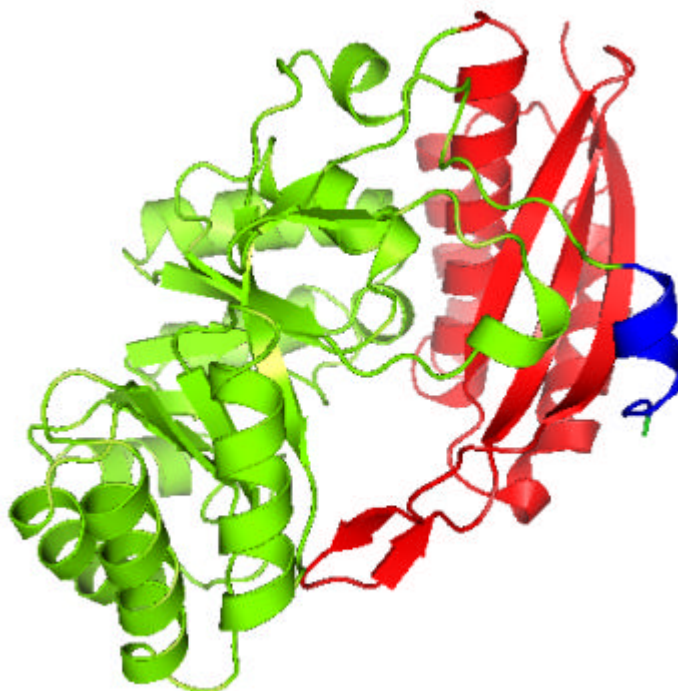


Figure 6 : Exemple d'un domaine discontinu dans la protéine 2MNR.

La Mandelate Racemase (PDB:2MNR) de *Pseudomonas putida* catalyse l'interconversion des énantiomères de l'acide mandélique. Elle se compose de deux domaines, un premier domaine continu (position 121 à 350) coloré en vert et un second domaine discontinu composé de deux fragments (position 3 à 120 et 351 à 358), colorés respectivement en rouge et en bleu.

Les domaines peuvent être globulaires et solubles, ou bien s'intégrer dans des membranes. Les domaines globulaires constitutifs d'une protéine sont souvent séparés par des

régions peu structurées, qui présentent peu ou pas de structures secondaires régulières. Il a été montré que ces régions non structurées permettent d'assurer un grand nombre de fonctions importantes [DUNKER, A. K. et al., 2002]. Ces régions non structurées semblent assurer leur fonction via une modulation d'affinité ou de spécificité à l'aide de petits peptides qu'elles contiennent alors que les domaines structurés assurent leur fonction en formant souvent des structures globulaires, parsemées de petites cavités (souvent le siège de sites actifs), qui peuvent aussi interagir grâce à leur grande surface de contact [COEYTAUX, K. 2004].

Les régions non structurées semblent avoir une signature bien particulière: une complexité (diversité d'usage des 20 acides aminés courants) de séquence faible et un biais dans la composition en acides aminés avec un faible taux en acides aminés hydrophobes (valine, leucine, isoleucine, méthionine, phénylalanine, tryptophane et tyrosine) et une forte proportion en acides aminés polaires et chargés (glutamine, sérine, proline, acide glutamique, lysine et parfois glycine et alanine) [ROMERO, P. et al., 2001; VUCETIC, S. et al., 2003]. Les régions non structurées peuvent être très fortement conservées en séquence entre espèces. Un certain nombre de segments désordonnés se replient en se liant à d'autres molécules biologiques alors que d'autres forment des linkers flexibles qui ont un rôle dans l'assemblage de grosses structures moléculaires [DYSON, H. J. et al., 2005].

Les caractéristiques d'une grande partie des domaines, obtenus par différentes expertises, sont regroupées dans des banques de domaines comme SMART [LETUNIC, I. et al., 2002], PRODOM [SERVANT, F. et al., 2002] ou PFAM [BATEMAN, A. et al., 2002].

Chapitre II

La méthode Hydrophobic Cluster Analysis

1 Introduction

Les protéines sont souvent constituées de régions aux propriétés physico-chimiques différentes, par exemple des domaines globulaires, des régions membranaires, des régions peu ou pas structurées. Nous avons voulu dans le cadre de ce travail tirer profit de la capacité qu'a l'approche HCA à résumer de nombreux détails de la séquence protéique en une série de motifs, à partir desquels il est possible de dégager des régions de textures différentes, correspondant notamment aux différents cas de figures décrits ci-dessus. Ainsi, la courbure de l'espace des séquences 1D dans l'espace tridimensionnel, telle qu'elle est utilisée dans la représentation HCA, permet de représenter la séquence d'une protéine de manière à prendre en compte les proximités spatiales locales des résidus pour mettre en évidence des signatures, ou amas hydrophobes, correspondant très majoritairement aux faces internes des structures secondaires régulières.

Des domaines globulaires sont ainsi caractérisés par une distribution régulière d'amas hydrophobes de taille moyenne, couvrant environ un tiers de la surface, contrastant avec la texture de régions membranaires, riches en acides aminés hydrophobes et mimétiques, ou celle de régions non structurées, généralement pauvres en acides aminés hydrophobes.

2 Fondement de la méthode HCA

Environ les deux tiers des acides aminés dont l'hydrophobie est forte (Val, Ile, Leu, Phe, Met, Tyr, Trp) sont inclus dans des structures secondaires régulières (hélices α et brins β) [CALLEBAUT, I. et al., 1997a]. Ces acides aminés hydrophobes ont des propensions pour des structures secondaires régulières α et β largement supérieures aux propensions "coil" et peuvent ainsi être considérés comme les moteurs de constitution des éléments de structures secondaires régulières. Ce sont donc ces acides aminés hydrophobes qui sont mis en évidence, puisque ce sont eux qui vont créer, principalement par l'intermédiaire des faces internes des structures secondaires régulières, un réseau d'interaction nécessaire à l'obtention d'un repliement stable. Constituant le cœur hydrophobe de la protéine, ces acides aminés ne sont pas forcément voisins directs en séquence. Une chaîne polypeptidique linéaire ne correspond

cependant qu'exceptionnellement à une protéine fonctionnelle et il est nécessaire de la replier ou, en d'autres mots, de la courber pour établir des contacts à moyennes et longues distances.

Ainsi, la méthode HCA (Hydrophobic Cluster Analysis) [GABORIAUD, C. et al., 1987; CALLEBAUT, I. et al., 1997a] est fondée sur la détection et la comparaison d'amas d'acides aminés hydrophobes (notamment en taille et forme) par l'intermédiaire d'un support bidimensionnel. Quatre acides aminés particuliers (P, G, S et T) sont mis en évidence en utilisant des symboles : la proline (P) conférant une forte rigidité à la chaîne protéique est considérée comme interrupteur d'amas car elle a tendance à déformer ou interrompre les hélices et les brins. A l'opposé, la glycine (G), dont la chaîne latérale se limite à un atome d'hydrogène, confère une forte flexibilité à la chaîne polypeptidique et est très représentée dans les boucles. La sérine (S) souvent rencontrée dans les boucles peut se retrouver dans un environnement hydrophobe (son groupement hydroxyle formant des liaisons hydrogène avec un groupement carbonyle de la chaîne principale). Enfin, la thréonine (T) peut présenter le même comportement et, bien qu'elle soit modérément hydrophile, peut remplacer des résidus hydrophobes comme la valine dans les brins β .

3 Représentation HCA : règles de segmentation de la séquence en amas

La représentation 2D-HCA donne naissance à des amas hydrophobes de tailles et de formes très variées. Le principe du tracé HCA est illustré sur la Figure 7. La méthode HCA repose sur l'utilisation d'un tracé 2D [GABORIAUD, C. et al., 1987] obtenu en reportant la séquence de la protéine sur une hélice inscrite à la surface d'un cylindre. Ensuite, ce cylindre est ouvert le long d'une génératrice de manière à le dérouler dans le plan. Puis, il est dupliqué afin de retrouver l'environnement préexistant sur le cylindre avant coupure. Les acides aminés hydrophobes précédemment cités (V, I, L, F, M, Y et W) sont alors mis en évidence en les entourant et en rassemblant les contours adjacents. Ceux-ci forment ainsi des amas ou "*clusters*" hydrophobes.

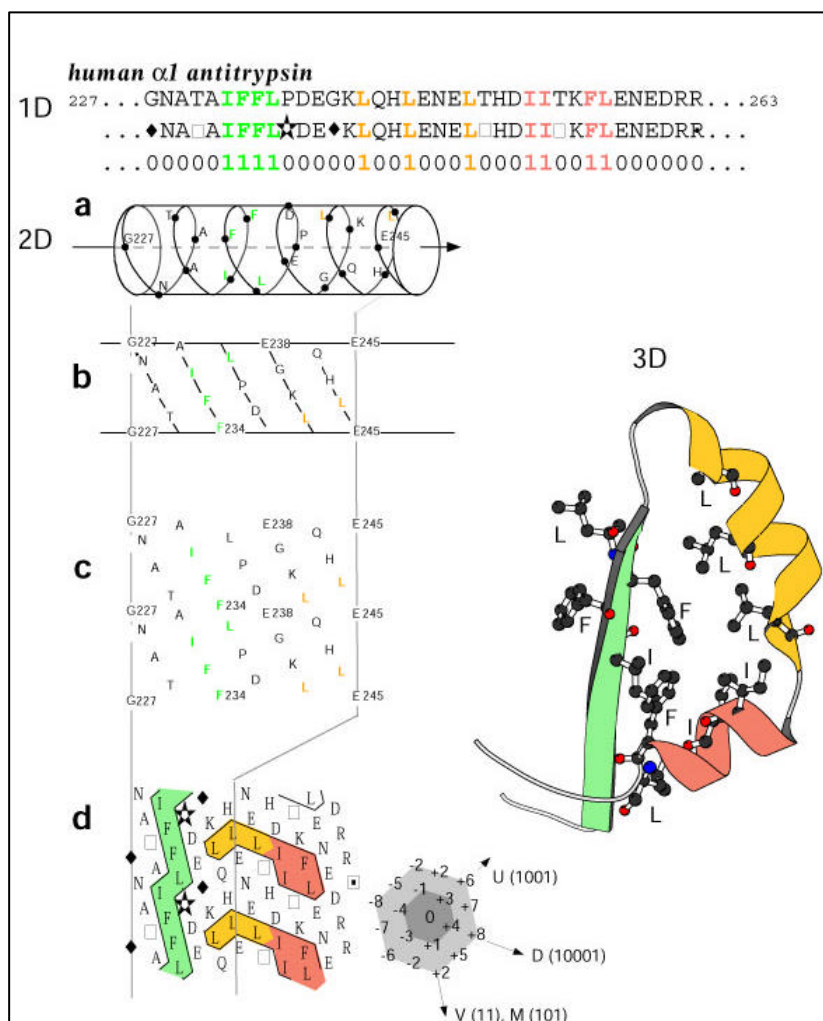


Figure 7 : Principe de la méthode HCA, illustré sur un segment de la séquence de l' $\alpha 1$ -antitrypsine (tiré de [CALLEBAUT et al., 1997]).

La séquence (1D), sur laquelle les acides aminés hydrophobes sont représentés en couleur, est inscrite sur une hélice **a** reporté sur un cylindre. Celui-ci, coupé suivant une génératrice est déplié et dupliqué afin de restaurer l'environnement de chaque acide aminé. Les acides aminés hydrophobes sont alors entourés et les contours adjacents réunis. Ces acides aminés ne se répartissent pas aléatoirement sur le diagramme (2D) mais forment des amas (« clusters ») dont la position correspond très majoritairement à celle des structures secondaires, hélices **a** et brins **b**, représentation sur laquelle sont visualisées les chaînes latérales des acides aminés hydrophobes). La forme des amas est de plus indicative de la nature de la structure secondaire qui lui est souvent associée : ainsi, l'amas vertical vert est associé à un brin **b**, tandis que l'amas horizontal (en jaune et rouge) est associé à des structures hélicoïdales (notons que la rupture d'orientation de l'amas horizontal est corrélée au découpage de l'hélice en deux éléments dans la structure tridimensionnelle). Les séquences séparant les amas correspondent très généralement à des régions de boucles reliant les structures secondaires entre elles.

Des symboles sont également utilisés pour représenter quatre acides aminés dont les propriétés structurales sont particulières : une étoile pour la proline (vient interrompre le réseau des liaisons hydrogène au sein des structures secondaires et est donc considérée comme un interrupteur d'amas), un losange pour la glycine (le plus petit des acides aminés, conférant une grande souplesse à la chaîne polypeptidique), un carré et un carré pointé respectivement pour la thréonine et la sérine (pouvant être enfouies ou exposées).

La formation de ces amas est essentiellement régie par une règle de segmentation (la distance de connectivité) et le choix d'un alphabet hydrophobe. Il a été montré que ces amas correspondent majoritairement aux faces internes des structures secondaires (hélices α et

brins β), alors que les séquences les séparant correspondent très majoritairement à des régions de boucles [WOODCOCK, S. et al., 1992]. On peut ainsi établir une correspondance entre amas et structures secondaires (Figure 8).

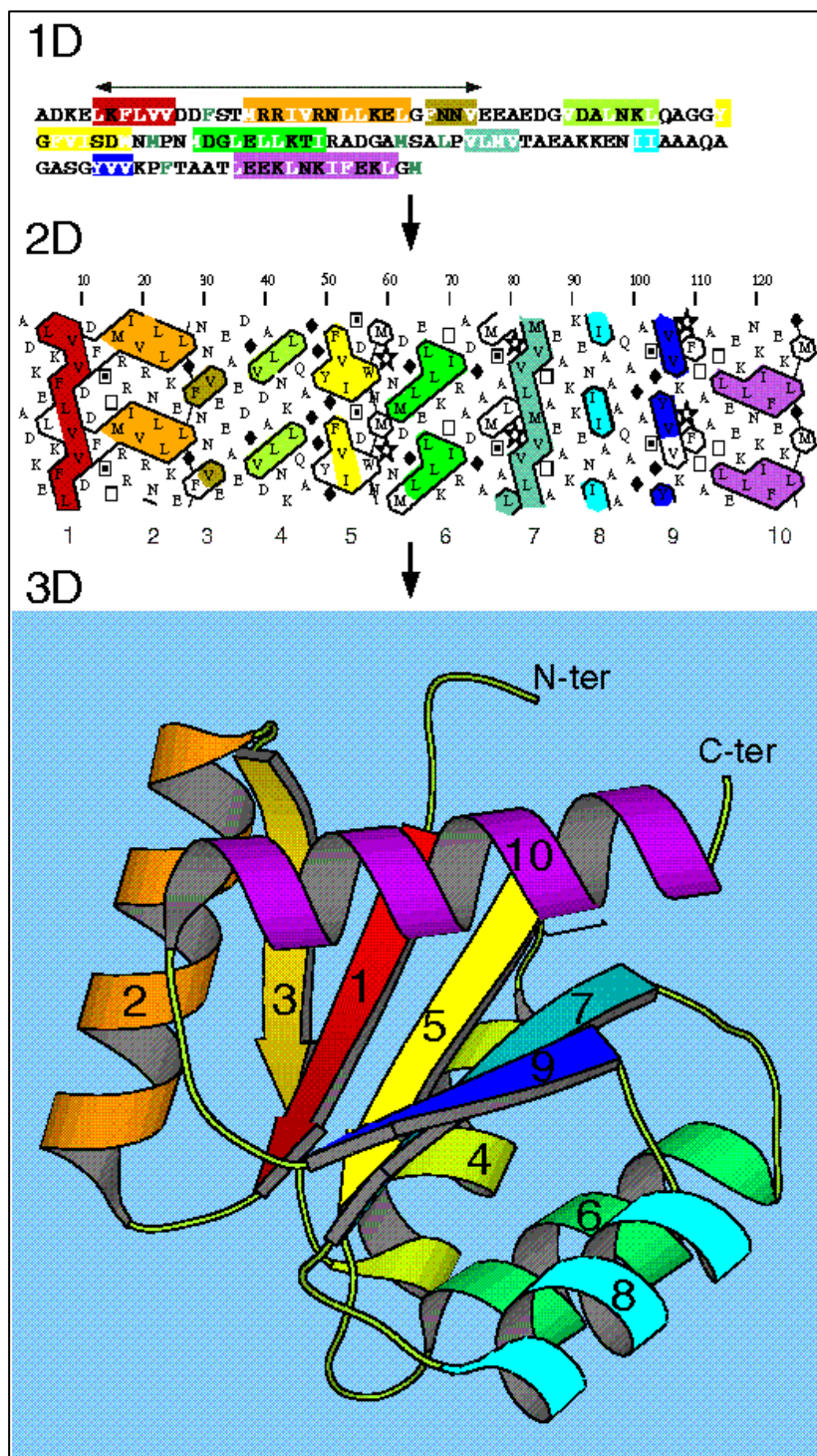


Figure 8 : Correspondance entre amas et structures secondaires illustrée pour un domaine globulaire complet (PDB 3CHY, Swiss-Prot CHEY_ECOLI). Les structures secondaires sont indiquées avec les mêmes colorations sur la séquence linéaire (1D), dans la moitié inférieure du diagramme HCA (2D) et sur la représentation de la structure tridimensionnelle (3D).

La plupart des acides aminés hydrophobes formant les amas (colorés dans la moitié supérieure du tracé HCA) participent aux structures secondaires régulières.

Comme illustré sur les Figures 7 et 8, la forme des amas peut généralement être corrélée à la nature même de la structure secondaire (hélice α ou brin β). Ainsi, des formes verticales -petit nombre d'acides aminés- peuvent être associées à des structures de brins β , des formes horizontales -plus grand nombre d'acides aminés- à des structures hélicoïdales. Les séquences séparant les amas hydrophobes correspondent très majoritairement à des régions de boucles reliant les structures secondaires entre elles.

Un amas hydrophobe au sens HCA est défini comme étant la séquence comprise entre le premier et le dernier acide aminé, nécessairement hydrophobes, de l'amas considéré et donc inclut dans la plupart des cas des acides aminés non hydrophobes. Cependant, certains amas hydrophobes ne renferment aucun acide aminé non hydrophobe. Les acides aminés hydrophobes sont généralement tournés vers le cœur de la structure alors que les acides aminés non hydrophobes sont exposés en surface.

3.1 Représentation binaire des acides aminés

L'approche HCA réduit la diversité des vingt acides aminés de la nature à une répartition binaire : hydrophobe ou non hydrophobe. Les acides aminés hydrophobes (valine, isoleucine, leucine, méthionine, phénylalanine, tyrosine et tryptophane) sont représentés par des 1 et les non hydrophobes par des 0. Avec ce codage binaire, l'amas VASFFGKW devient 10011001. Des études précédentes ont montré que ce type de simplification faisait ressortir des traits particuliers de la texture des protéines [CANARD, 1997].

Il existe plusieurs autres types de codage qui ont été explorés pour représenter les amas. Le « code décimal » consiste à traduire le résultat du codage binaire en son équivalent décimal. L'amas étant exprimé en code binaire, on obtient son Pcode en sommant les puissances de deux, indicées par la position de chaque chiffre dans le code binaire. Ainsi, 10011001 devient en code décimal $1+0+0+8+16+0+0+128$, soit 153.

Un autre codage est le Qcode. Il permet de représenter les amas sous forme de concaténation d'un petit nombre d'amas générateurs fondamentaux : s (Single), v (Vertical), m (Mosaic), u (Up) et d (Down) qui représentent les amas 1, 11, 101, 1001, et 10001 en code binaire. Tous les amas sont générables à partir de ces cinq amas fondamentaux. Il faut lire l'amas binaire de gauche à droite et dès qu'un des éléments fondamentaux est reconnu, on note la lettre correspondante en Qcode. Ensuite, la lecture est reprise à partir du dernier 1 du premier élément pris en compte pour trouver le second élément et ainsi de suite. L'amas 10011001 est uvu et l'amas 100111110001 est uvvvvd.

Un dernier codage repose sur l'utilisation de deux valeurs : H le nombre d'acides aminés hydrophobes de l'amas et L la longueur de l'amas. L'amas 10011001 est représenté par 4H8L (H correspond au nombre d'acides aminés hydrophobes de l'amas et L la longueur de l'amas). Ce code a un inconvénient majeur car il n'est pas univoque. En effet, tous les amas d'une même longueur et ayant le même nombre d'hydrophobes sont représentés par un même code. Ainsi, les amas 10101, 10011 et 11101 sont représentés par le même code 3H5L [CANARD, 1997].

Seul, le codage binaire sera utilisé dans ce manuscrit.

3.2 Un support hélicoïdal

Dans la représentation HCA, la séquence protéique est enroulée sur un cylindre, pour former l'équivalent d'une hélice α avec un pas d'hélice de 3,6 résidus par tour. Des études statistiques effectuées à partir de structures expérimentales ont montré que la meilleure segmentation était obtenue avec un support α -hélicoïdal, celui-ci conduisant en effet à un maximum de correspondance entre positions des amas et positions des structures secondaires, alors qu'avec d'autres configurations, cette correspondance était moindre [WOODCOCK et al., 1992]. Il a également été montré que l'on obtenait un minimum d'amas hydrophobes avec la configuration bidimensionnelle α -hélicoïdale, ou, en d'autres termes, une compaction maximale de l'information hydrophobe que pourrait sous-tendre la compacité naturelle des structures de protéines [CALLEBAUT, I. et al., 1997a]. Enfin, le support α -hélicoïdal paraît lui aussi plus apte que tout autre support à mettre en évidence les caractéristiques des séquences par rapport à l'aléatoire [HENNETIN, J. 2003].

3.3 Distance de connectivité

Un amas hydrophobe est constitué d'un groupe d'acides aminés hydrophobes proches entre eux sur la séquence. La proximité est définie par la distance de connectivité, qui fixe la distance maximale permettant encore de considérer deux positions comme proches et donc membre d'un même amas. La valeur 4 correspond à la trame d'une hélice α , tandis qu'une distance de connectivité de 3 et de 5 correspond respectivement à une hélice 3_0 et à une hélice π . Il a été montré que la distance de connectivité de 4 acides aminés offre la meilleure correspondance entre amas et structures secondaires régulières [WOODCOCK et al., 1992]. Dans la représentation hélicoïdale, les positions $i+1$, $i+3$ et $i+4$ sont proches de la position i . La proximité de $i+2$, qui n'est vraie que pour une structure en brin β , est ajoutée « artificiellement » pour former les règles complètes de la définition des amas HCA.

3.4 L'alphabet HCA

La méthode HCA utilise un alphabet hydrophobe composé des sept acides aminés suivants : Val, Ile, Leu, Phe, Met, Trp et Tyr. Ces acides aminés sont, en moyenne, les plus enfouis au sein des structures protéiques globulaires [SOYER et al., 2000] et ont des propensions pour des structures secondaires régulières α et β supérieures à leurs propensions "coil" [CALLEBAUT, I. et al., 1997a]. Même si la méthode HCA est basée sur un alphabet hydrophobe, les amas peuvent renfermer des acides aminés non hydrophobes. De même, certains acides aminés, comme la cystéine, l'alanine, la thréonine, peuvent parfois avantageusement intégrer l'alphabet hydrophobe, et jouer le rôle d'acide aminé mimétique.

3.5 Un acide aminé interrupteur d'amas : la proline

Un acide aminé est défini comme «interrupteur», c'est-à-dire dont la seule présence entraîne la terminaison d'un amas : la proline. En effet, la rigidité et la non disponibilité de son amine pour former des liaisons hydrogène convient mal aux structures secondaires régulières. De plus, la proline est souvent retrouvée dans les boucles. Puisque les amas hydrophobes ciblent les faces internes des structures secondaires, la présence d'un résidu proline signale très souvent la fin de cette structure et donc la fin de l'amas.

4 ***Un amas correspond majoritairement à un type de structure secondaire***

Des tendances très marquées pour des structures α ou β se dégagent pour certains amas. Ainsi, certains amas comme les amas 10011001, 10011011, 10011111, 10111011 ou 110011011 sont associés dans plus de 65% des cas avec des structures α alors que d'autres amas comme 1111, 101111 ou 111101 sont majoritairement associés à des structures β . Un dictionnaire des amas hydrophobes a été constitué, répertoriant les structures secondaires (dont les attributions ont été effectuées par P-SEA [LABESSE, G. et al., 1997]) associées à chacun des amas et calculé sur une banque PDB à 90% non redondante (version de septembre 2001, 4 704 séquences) [LE TUAN, 2003]. Nous avons utilisé le dictionnaire ainsi établi dans la deuxième partie de cette thèse, afin de donner une indication sur le type de repliement des domaines globulaires prédits, à partir de la seule information de séquence.

Par ailleurs, les amas hydrophobes peuvent être avantageusement exploités sur le plan fondamental pour caractériser les successions de structures secondaires utilisées par la nature pour construire des domaines globulaires stables. Il est en effet possible de comparer l'occurrence observée de chacun des amas par rapport à des banques de séquences aléatoires

et de constater ainsi que certains amas sont sur-représentés [HENNETIN, J. 2003] et par conséquent préférentiellement utilisés pour la constitution des domaines globulaires. Ces amas sur-représentés ont des longueurs et des contenus en acides aminés hydrophobes caractéristiques de la physico-chimie des domaines globulaires. D'autre part, ils constituent des "signatures" du repliement protéique. En effet, la présence d'amas hydrophobes dans une région de la séquence protéique sous-entend la présence d'un domaine globulaire alors que l'absence d'amas hydrophobes sous-entend une région charnière entre des domaines globulaires (voir deuxième partie).

La structure secondaire étant le plus souvent bien mieux conservée que ne l'est la séquence, il en résulte que des séquences même très divergentes auront généralement un profil d'amas hydrophobes dont la similitude pourra être reconnue ou suspectée, profil dont on peut se servir comme signature du repliement étudié. Ainsi, HCA se révèle particulièrement utile pour décrypter les apparentements entre les séquences de protéines en particulier à très haut niveau de divergence, conduisant à l'identification de nouveaux domaines (e.g. BRCT [CALLEBAUT, I. et al., 1997b], TUDOR [CALLEBAUT, I. et al., 1997c], BAH [CALLEBAUT, I. et al., 1999], RUN [CALLEBAUT, I. et al., 2001], et DENN [LEVIVIER, E. et al., 2001]), ou au rattachement de séquences à des familles déjà caractérisées sur le plan fonctionnel et/ou structural (e.g. RAG2 [CALLEBAUT, I. et al., 1998] et FERM [GIRAULT, J. A. et al., 1999]).

5 *But et Perspectives*

HCA est un outil puissant pour diverses étapes de l'analyse de séquences. Dans un premier temps, l'analyse spécifique de la « texture » des séquences permet rapidement de définir les limites de régions structurées, les passages membranaires, les régions peu ou pas structurées. Dans un second temps, la conservation des amas hydrophobes, véritables signatures des repliements des domaines globulaires, permet de détecter des apparentements entre séquences, même extrêmement divergentes.

Cependant, l'utilisation efficace de la méthode HCA repose toujours, pour une part importante, sur l'expérience et la culture de l'utilisateur.

Le but de notre étude est de réaliser une investigation des caractéristiques de texture des séquences protéiques, telles qu'elles peuvent être appréhendées par l'approche HCA et de proposer des outils permettant une analyse automatique, relativement indépendante de l'expertise de l'utilisateur. Le chapitre suivant présente les méthodes d'analyse de texture

utilisée en imagerie et que nous avons choisi d'explorer dans notre recherche de texture des séquences.

Chapitre III

Analyse de la texture

1 *Texture et analyse de texture : Généralités*

1.1 Définition de la texture

Le terme « texture » est employé dans de nombreux domaines : on peut parler de texture d'un cheveu, d'une sauce, d'un sol, d'un tableau, d'une image, etc... Intuitivement, la notion de texture nous paraît familière, mais en donner une définition précise devient plus complexe. Une texture est une information qui rend compte de l'état de surface ou de structure 3D (e.g. une éponge) d'un objet. Elle est caractérisée par l'arrangement plus ou moins régulier de motifs élémentaires.

1.2 Deux types de texture : aléatoire et structurée

On distingue deux types de texture :

- Les textures structurées : encore appelées « macroscopiques » ou « macrotextures », elles sont constituées par la répétition spatiale plus ou moins régulière, d'un motif de base (appelé « primitive » ou « texel »), dans différentes directions. Les textures périodiques constituent un sous-ensemble des textures structurées. L'exemple du mur de briques extrait du catalogue de textures de Brodatz [BRODATZ, 1966] illustre bien ce type de texture (Figure 9).

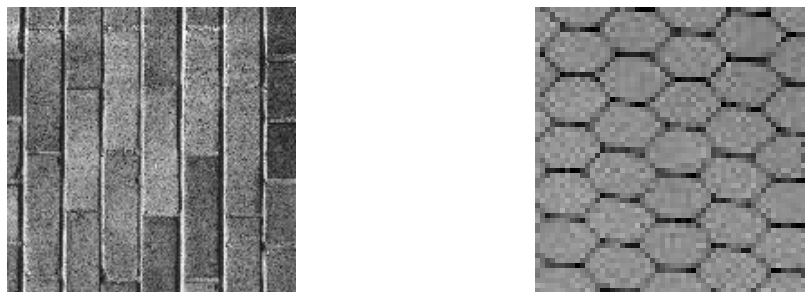


Figure 9 : Exemple de textures structurées : A gauche, le « mur de briques » du catalogue de Brodatz et à droite, le « grillage » [BRODATZ, P. 1966].

- Les textures aléatoires : dans ces textures, aucun motif particulier n'est localisable. Elles ont un aspect anarchique et désorganisé, tout en restant homogène, et ne répondent à aucune règle d'agencement particulière. La primitive est ramenée au niveau du pixel, ce qui vaut à ces textures le nom de textures « microscopiques » ou « microtextures » (Figure 10).



Figure 10 : Exemple de textures aléatoires [BRODATZ, P. 1966].

2 Perception et analyse visuelle d'une texture

Les textures donnent une information interne à une région. Dans des environnements flous ou perturbés, où l'information des contours n'est pas une donnée fiable, posséder une description de l'intérieur d'une région est un atout important. Les méthodes classiques d'analyse de textures sont très liées à la catégorie de textures à laquelle on s'intéresse. Pour les macrotextures, on utilisera de préférence des méthodes structurelles, tandis que l'analyse des microtextures se fera souvent par des méthodes de type statistiques.

L'œil humain distingue en moyenne 16 niveaux de gris différents du noir au blanc. Lors de l'analyse d'une image, l'observateur recherche des indices lui permettant de repérer les différentes textures de l'image. Il va instinctivement repérer, par une première analyse macroscopique, les limites (ou contours) entre les différentes régions homogènes plus ou moins importantes de l'image ; chaque région pourra être assimilée à une texture. Afin de transcrire en terme de texture ce qui est vu par l'observateur sur l'image, des caractéristiques de base ont été définies [AMADASUM et al., 1989]. Il s'agit par exemple du contraste, de la grossièreté, de la forme et de la direction.

La grossièreté : une texture grossière possède des primitives larges : il existe alors peu de variations entre l'intensité d'un pixel et celle de ses pixels voisins (Figure 11).

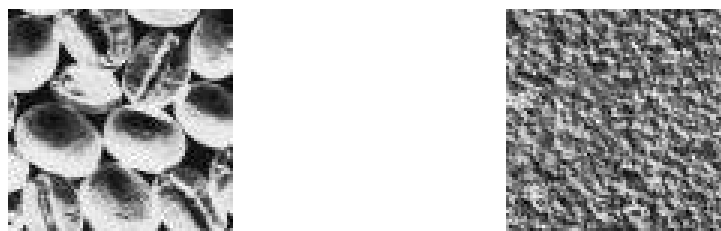


Figure 11 : Illustration de la grossièreté sur deux textures de Brodatz différentes [BRODATZ, P. 1966]. Le paramètre grossièreté est beaucoup plus élevé à gauche qu'à droite.

Le contraste : une texture possède un contraste élevé si les différences d'intensité entre primitives sont importantes [HERLIDOU, S. 1999] (Figure 12).



Figure 12 : Illustration du contraste sur deux textures de Brodatz différentes [BRODATZ, P. 1966]. A droite, le paramètre contraste est six fois plus grand qu'à gauche.

3 Quelques méthodes d'analyse de texture

L'extraction d'attributs caractéristiques a donné lieu à un certain nombre de méthodes d'analyse structurale et statistique. Elles sont utilisées actuellement pour l'indexation d'images et permettent de quantifier les différents niveaux de gris présents dans une image en terme d'intensité et de distribution. Comme il existe deux grandes classes de textures, deux approches sont nécessaires : les méthodes statistiques et les méthodes structurelles [HARALICK, 1979; CONNERS et al., 1980; VAN GOOL et al., 1985].

Les méthodes statistiques étudient les relations entre un pixel et ses voisins. Elles sont utilisées pour caractériser des structures fines, sans régularité apparente. Plus l'ordre de la statistique est élevé et plus le nombre de pixels (1 à n) mis en jeu est important.

Les méthodes structurelles permettent de décrire la texture en définissant les primitives et les « règles » d'arrangement qui les relient. En effet, les textures ordonnées possèdent des primitives qui se répètent dans les images en des positions suivant une certaine loi. Ces méthodes semblent plus adaptées à l'étude de textures périodiques ou régulières [HERLIDOU, S. 1999]. Nous ne présenterons ici que certaines méthodes statistiques que nous avons utilisées au cours de notre travail.

3.1 Méthodes de premier ordre

L'analyse par les méthodes de premier ordre se fait au niveau des pixels individuels d'une région d'intérêt de l'image appelée « ROI ». Les paramètres sont calculés à partir de l'histogramme des intensités. La moyenne, la variance et le « skewness » (mesurant le degré de symétrie d'une distribution) sont les paramètres les plus souvent utilisés pour caractériser une texture.

- La **moyenne** donne la valeur moyenne (ou intensité moyenne) des niveaux de gris appartenant à tous les pixels de la ROI. Les images qui ont une moyenne plus élevée apparaissent plus claires.

$$\text{MOY} = \frac{1}{N} \sum_{i,j} g(i,j)$$

où $g(i,j)$ représente la valeur du niveau de gris du pixel (i,j) , N est un facteur de normalisation qui correspond au nombre total de pixels.

- La **variance** correspond au moment d'ordre 2. Elle mesure la répartition des niveaux de gris autour de la valeur moyenne. Plus la variance est élevée et plus il y a d'écart importants entre les valeurs et la moyenne.

$$\text{VAR} = \frac{1}{N} \sum_{i,j} (g(i,j) - \text{MOY})^2$$

- Le « **skewness** » correspond au moment d'ordre 3 centré autour de la moyenne. Ce paramètre mesure la déviation de la distribution des niveaux de gris par rapport à une distribution symétrique. Pour une déviation vers les valeurs élevées, le skewness est positif ; alors que pour une déviation vers les valeurs basses, il est négatif.

$$\text{SKEW} = \frac{1}{N} \sum_{i,j} (g(i,j) - \text{MOY})^3$$

3.2 Méthodes de second ordre : méthode de matrice de cooccurrence

Dans les méthodes de premier ordre, qui correspondent à une description de l'histogramme des niveaux de gris, il n'y a pas d'informations sur la localisation du pixel. Il est nécessaire d'utiliser des méthodes d'ordre supérieur pour une analyse plus précise dans lesquelles la relation entre les niveaux de gris des pixels pris deux à deux intervient dans le calcul des paramètres. De nombreuses méthodes existent, nous ne présenterons ici que la méthode de matrice de cooccurrence [WESZKA et al., 1976; HE et al., 1988; PECKINPAUGH, 1991; DAVIS et al., 1979; GOTLIEB et al., 1990].

La méthode de matrice de cooccurrence (ou méthode de dépendance spatiale des niveaux de gris) permet de déterminer la fréquence d'apparition d'un « motif » formé de deux

pixels séparés par une certaine distance d dans une direction particulière θ par rapport à l'horizontale. Afin de limiter le nombre de calculs, on prend généralement comme valeurs 0° , 45° , 90° , 135° , 180° et 1 pour la valeur d . La taille de la matrice est $N_g * N_g$, où N_g correspond au maximum des niveaux de gris de l'image. Pour ne pas avoir une taille de matrice trop élevée, on choisira le plus souvent $N_g=8$, 16 ou 32. A titre d'exemple, pour une image possédant 4 niveaux, la matrice de cooccurrence sera de taille 16 (Figure 13).

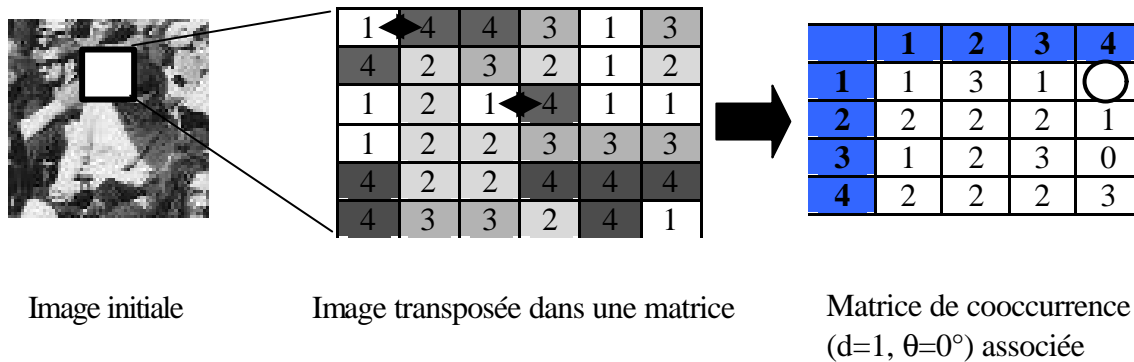


Figure 13 : Remplissage d'une matrice de cooccurrence.

A gauche, l'image initiale est intégrée dans une matrice dans laquelle chaque pixel est remplacé par une valeur caractérisant son niveau de gris. A droite, la matrice de cooccurrence va recenser les contacts entre chacun des pixels. Par exemple, nous comptons le nombre de fois où l'on retrouve le couple (1,4) dans la direction horizontale ($\theta=0^\circ$). Cette valeur (ici 2) est reportée dans la matrice de cooccurrence à l'intersection 1-4 (case entourée). La matrice de cooccurrence est toujours carrée (hauteur=longueur), sa taille correspond au nombre de niveaux de gris de l'image.

A chaque direction θ et pour chaque valeur de d correspond une matrice de cooccurrence $\phi(d,\theta)$. On définit généralement des matrices de cooccurrence symétriques. A partir de la matrice, il est possible d'extraire une quinzaine de paramètres. Ils contiennent des informations sur la finesse, la direction et la granularité de la texture. Pour une texture grossière, les valeurs de la matrice sont concentrées sur la diagonale principale. Au contraire, pour une texture fine, les valeurs de la matrice seront dispersées : en effet, pour une telle texture il existe beaucoup de transitions de niveaux de gris.

Nous présenterons ici quelques paramètres :

- La **moyenne**

$$MOY = \sum_{i,j} p(i,j)$$

où $p(i,j)$ correspond aux éléments de la matrice de cooccurrence; c'est à dire à la probabilité de passer d'un pixel de niveau de gris i à un pixel de niveau de gris j .

- La **variance** caractérise la distribution des niveaux de gris autour de la valeur moyenne M calculée précédemment.

$$\text{VAR} = \sum_{i,j} (i - \text{MOY})^2 p(i,j)$$

- Le **contraste** (=inertie) mesure les variations locales des niveaux de gris. Si elles sont importantes (c'est à dire s'il existe peu de régions homogènes), alors le contraste sera élevé.

$$\text{CONT} = \sum_{i,j} (i-j)^2 p(i,j)$$

- L'**entropie** mesure la complexité de l'image. Elle permet de caractériser le degré de granulation de l'image. Plus l'entropie est élevée et plus la granulation est grossière.

$$\text{ENT} = - \sum_{i,j} p(i,j) \log p(i,j)$$

3.3 Méthodes d'ordre supérieur : méthode des longueurs de plages de niveaux de gris (ou de sections)

Les méthodes d'ordre supérieur étudient les interactions entre plusieurs pixels. Le voisinage est de type mono ou bidimensionnel. La méthode des longueurs de plages de niveaux de gris est la plus souvent utilisée [CONNERS et al., 1980; CHU et al., 1990]. Elle consiste à compter le nombre de plages d'une certaine longueur j, de niveau de gris i dans une direction θ donnée (Figure 14). A chaque direction correspondra donc une matrice. Une plage de niveaux de gris (ou « section ») correspond à l'ensemble des pixels d'une image ayant la même valeur de niveau de gris. La longueur de la plage correspond au nombre de pixels appartenant à la plage. Nous utiliserons le terme de « plage » tout au long du manuscrit.

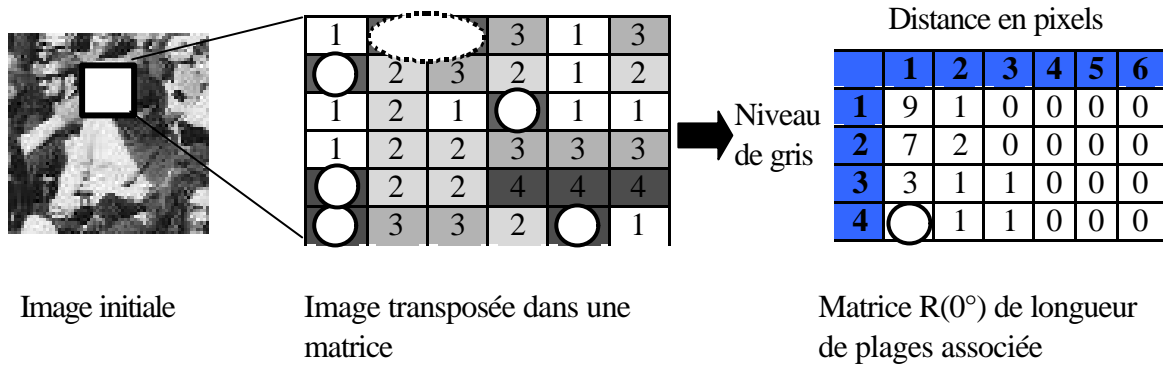


Figure 14 : Remplissage d’une matrice de longueur de plages.
 A gauche, l’image initiale est intégrée dans une matrice dans laquelle chaque pixel est remplacé par une valeur caractérisant son niveau de gris. A droite, la matrice de longueur de plage va recenser le nombre de plages d’une certaine longueur et d’un certain niveau de gris. Par exemple, nous comptons le nombre de plages de niveau 4 et de longueur 1 (région entourée en trait continu) dans la direction horizontale ($q=0^\circ$). Cette valeur (ici 5) est reportée dans la matrice de cooccurrence à l’intersection 41 (case entourée). Ensuite, nous comptabilisons les plages de niveau 4 et de longueur 2 (région entourée en pointillé) dans la direction horizontale ($q=0^\circ$), la valeur calculée est reportée dans la matrice de longueur de plage, etc.... La longueur de la matrice R est fonction de la distance maximale séparant deux pixels de l’image. La hauteur de la matrice R correspond au nombre de niveaux de gris de l’image.

Les paramètres dérivés des matrices de longueur de plages recouvrent une réalité perceptuelle plus évidente que les paramètres calculés à partir des matrices de cooccurrence.

3.3.1 Paramètre SRE

Le « poids » des plages courtes ou Short Run length Emphasis (SRE) rend compte de la finesse d’une texture. Plus l’image contient des plages de petite taille, plus le paramètre SRE est élevé (Figure 15, page suivante). Un paramètre SRE élevé indique que la texture est fine.

$$SRE = \frac{\sum_{l=1}^L \sum_{g=1}^G (P(l,g)/l^2)}{\sum_{l=1}^L \sum_{g=1}^G P(l,g)}$$

somme de toutes les valeurs de la matrice R de longueur

où $P(l,g)$ représente le nombre de plages de longueur l et de niveau de gris g

3.3.2 Paramètre LRE

Le « poids » des plages longues ou Long Run length Emphasis (LRE) rend compte de la grossièreté de la texture de l’image. Ce paramètre est l’inverse du paramètre SRE. Plus l’image contient des plages de grande taille, plus le paramètre LRE est élevé (Figure 15). Un paramètre LRE élevé indique que la texture est grossière.

$$\text{LRE} = \frac{\sum_{l=1}^L \sum_{g=1}^G P(l,g)^2}{\sum_{l=1}^L \sum_{g=1}^G P(l,g)}$$



Figure 15 : Illustration des paramètres SRE et LRE.

L'image de gauche présente une abondance de plages de petite taille contrairement à l'image de droite. Elle présente donc un SRE plus élevé que celle de droite et inversement un LRE plus faible.

3.3.3 Paramètre GLD

La **distribution des niveaux de gris** ou **Grey Level Distribution (GLD)** permet d'appréhender l'uniformité des plages dans les différents gris. Ce paramètre met en évidence une uniformité ou une hétérogénéité des plages partageant un même niveau de gris (Figure 16). Le paramètre GLD augmente avec le nombre de plages ayant des niveaux de gris identiques.

$$\text{GLD} = \frac{\sum_{g=1}^G \sum_{l=1}^L (P(l,g))^2}{\sum_{l=1}^L \sum_{g=1}^G P(l,g)}$$



Figure 16 : Illustration du paramètre GLD.

L'image de gauche contient de nombreuses plages de niveau 4 (gris foncé) indépendamment de la taille des plages. Plus la distribution des plages est uniforme (présence de beaucoup de plages de même niveau de gris), plus GLD est élevé.

3.3.4 Paramètre RLD

La **distribution des longueurs de plages** ou **Run Length Distribution (RLD)** mesure la non uniformité de la répartition des longueurs de plages, indépendamment de leur niveau de gris (Figure 17). Le paramètre RLD est minimum lorsque les plages sont également distribuées entre les longueurs.

$$RLD = \frac{\sum_{l=1}^L \sum_{g=1}^G (P(l,g))^2}{\sum_{l=1}^L \sum_{g=1}^G P(l,g)}$$

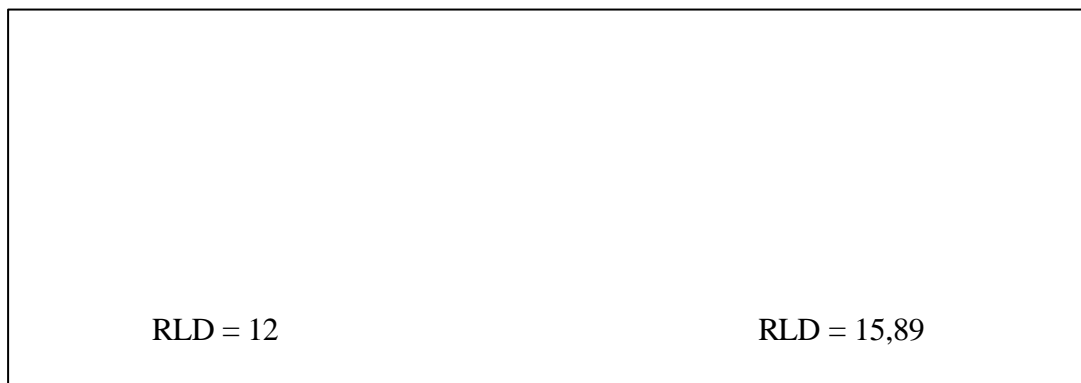


Figure 17 : Illustration du paramètre RLD.

L'image de gauche contient de nombreuses plages qui indépendamment de leur niveau de gris ont toutes la même longueur. Plus la distribution des longueurs des plages est non uniforme, plus RLD est élevé. L'image de gauche présente donc un RLD plus faible que celle de droite.

3.3.5 Paramètre RLP

Le pourcentage de plages ou Run Length Percentage (RLP) met en évidence la densité des plages dans la région d'intérêt de l'image (ROI) (Figure 18). Le paramètre RLP a une valeur d'autant plus faible que la texture est homogène et constituée de grandes plages.

$$RLP = \frac{\sum_{l=1}^L \sum_{g=1}^G P(l,g)}{A}$$

où A est le nombre total de pixels dans la région



Figure 18 : Illustration du paramètre RLP.

Les image de gauche et de droite contiennent respectivement 12 et 27 plages. Plus l'image contient de plages, plus son RLP est élevé. Ainsi, l'image de droite a un RLP de 0,75 plus élevé que celui de l'image de gauche.

4 But de notre étude

La notion de texture a été très peu utilisée dans le domaine des séquences nucléotidiques et protéiques. Quelques études ont été menées sur la texture de l'ADN avec la recherche de motifs répétitifs, de composition en nucléotides, l'existence d'isochores, etc... [NEVILL-MANNING, C. G. et al., 1998; KURTZ, S. et al., 1999; BARRETTE, I. et al., 2001; LERCHER, M. J. et al., 2002; KOLPAKOV, R. et al., 2003; LEFEBVRE, A. et al., 2003]. Au niveau des séquences protéiques, seules des études de composition sur les acides aminés ou d'identification de motifs ou de répétitions semblent avoir été faites ([ANDRADE, M. A. et al., 2000; SIGRIST, C. J. et al., 2002; SZKLARCZYK, R. et al., 2004; APOSTOLICO, A. et al., 2005; CHAKRABARTI, S. et al., 2005; GUTMAN, R. et al., 2005]). Par ailleurs, la notion de texture est utilisée dans des domaines variés tels que l'imagerie satellitaire, robotique ou médicale, l'industrie... ([BEZY-WENDLING, J. 1997; LORETTE, A. 1999; KAM, L. et al., 2000; BARBEAU, J. et al., 2002]). Elle offre l'avantage de fournir, de manière automatique, des paramètres quantifiés et objectifs, reflétant la texture vue par l'observateur. Ainsi, le radiologue réalise bien souvent de manière implicite sa propre analyse de texture. Celle-ci constitue une partie plus ou moins importante de l'étude de l'image radiologique, tâche complexe faisant intervenir à la fois l'œil et le cerveau, afin de mettre en place un diagnostic. Dans le cadre de l'analyse du tracé bidimensionnel HCA d'une protéine (Figure 19), l'utilisateur accomplit également une analyse visuelle de la texture. Le couple œil/cerveau permet à l'expert de faire abstraction des éléments non pertinents, de reconstruire intellectuellement un tracé HCA perçu par petites régions et de le comparer mentalement à des régions du tracé HCA ou au tracé HCA complet d'une autre protéine.

Cette analyse est expert-dépendante et nécessite une bonne culture biologique, une bonne connaissance des fonctions des protéines et des domaines les composant, une mémoire visuelle et une certaine intuition. De ce fait, cette analyse de texture est essentiellement qualitative et ne permet pas un traitement à grande échelle des séquences de protéines et des protéomes dont les génomes viennent d'être séquencés. Dans ce contexte, il nous est apparu envisageable de tenter de caractériser la «texture» d'une protéine à partir de sa représentation 2D établie avec HCA (Figure 19 ci-dessous), en transposant certaines méthodes d'analyse de texture à notre système. Cela offrirait l'avantage de fournir, de manière automatique, des paramètres quantifiés et objectifs, reflétant la texture vue par l'expérimentateur.

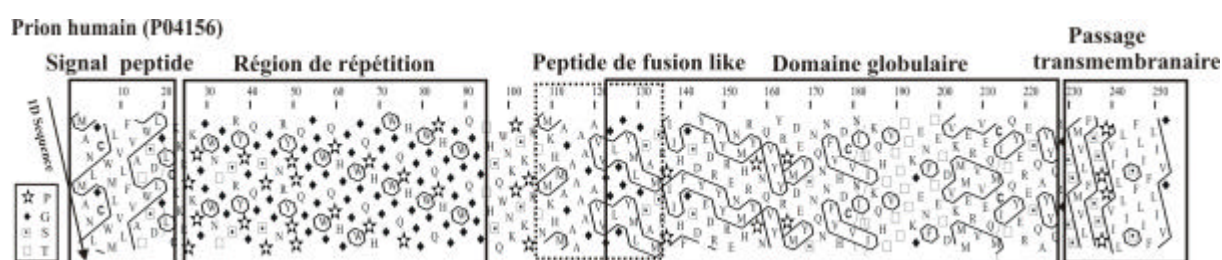


Figure 19: Représentation bidimensionnelle HCA de la séquence du prion humain.

Il est patent que la distribution des 20 acides aminés le long des séquences n'est pas uniforme et que celle-ci est fortement dépendante de la structuration qu'elles doivent localement adopter, sur des longueurs de séquence généralement comprises entre quelques acides aminés et trois à quatre centaines d'acides aminés. Par exemple, les domaines globulaires classiques bâtis sur l'enchaînement d'hélices **a** et/ou de brins **b** sont caractérisés par une proportion voisine de 1/3 d'acides aminés fortement hydrophobes (VILFMYW), avec une faible dispersion. Ceux-ci se regroupent en un nombre limité d'amas de taille moyenne (e.g. de 5 à 15 acides aminés). Les régions inter-domaines, intrinsèquement peu structurées, sont à l'inverse caractérisées par de faibles proportions d'acides aminés hydrophobes, qui de plus se regroupent en amas de petite taille. D'autres distributions typées caractérisent, entre autres, les régions de répétitions ou de pseudo-répétitions, ainsi que les diverses variétés de passages membranaires. Des textures différentes, visibles sur le tracé HCA, correspondant à différents domaines/régions décrites ci-dessus, caractérisent la séquence du prion.

Les méthodes d'analyse de texture appliquées aux protéines pourraient rendre compte de la nature des acides aminés présents dans leur séquence, de leur quantification et de leur répartition. Les méthodes de premier ordre, basées essentiellement sur des calculs de moyenne et de variance, de distribution des acides aminés hydrophobes, pourraient permettre de délimiter des régions dites « structurées » ou présentant un même type de texture. Ainsi, il semblerait possible de découper une séquence en domaines référencés (domaine structuré, région de répétition, passage membranaire, peptide de fusion, etc...) (Figure 20).

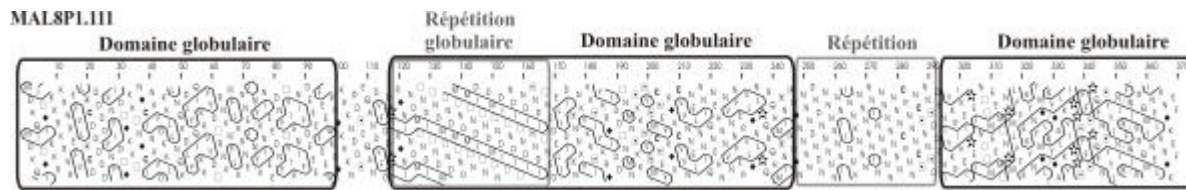


Figure 20 : Représentation bidimensionnelle HCA et découpage en domaines de la protéine MAL8P1.111 de *Plasmodium Falciparum*.

Dans cette étude, nous avons souhaité caractériser la texture des protéines visualisables par HCA en termes de « région contrastée », « région simple », « région complexe », « grossière » etc... autant de termes qui permettent de définir visuellement un objet. L'idée principale était une analyse à grande échelle des génomes visant à extraire des régions caractéristiques, les répertorier, les analyser et dans un second temps réaliser une synthèse de l'organisation en domaines de texture des séquences. En effet, une analyse sous cet angle et avec la sensibilité HCA n'a jamais été faite.

Au début de cette étude, nous avons exploité les concepts issus d'autres domaines scientifiques pour quantifier la texture dans nos séquences et en faire ressortir une information. Cette étude préliminaire qui s'est avérée peu concluante, nous a orienté vers un projet de découpage spécifique HCA des séquences en régions structurées/non structurées, en passages membranaires et en régions de répétition que nous avons ensuite appliqué à plusieurs génomes.

Deuxième partie :

Développements méthodologiques pour analyser la
texture dans les séquences protéiques

Préambule

Dans les chapitres précédents, nous avons pu constater que le tracé bidimensionnel HCA peut s'apparenter à une image et, qu'intuitivement, l'examineur va repérer, par une première analyse macroscopique, les limites (ou contours) entre les différentes régions homogènes plus ou moins importantes de l'image. Chaque région pourra être assimilée à une texture. Dans un premier temps (chapitre IV), nous avons alors exploré certaines méthodes d'analyse de texture utilisées en analyse d'images et tenté de les appliquer au tracé bidimensionnel HCA afin de définir des paramètres numériques spécifiques de régions de texture particulière. Dans un second temps (chapitre V et VI), nous avons développé plusieurs algorithmes pour détecter des régions particulières dans les séquences (segments structurés, peptides de fusion, régions de répétitions). Ces algorithmes sont basés sur les observations d'analyse de texture déduites de l'utilisation du tracé HCA, des propriétés de compacité des protéines et de propriétés physicochimiques. Enfin, une étude à grande échelle des caractéristiques précédemment mises en évidence a été réalisée au niveau des protéomes de différents génomes (*Homo sapiens*, *Plasmodium falciparum*, *Caenorhabditis elegans* et *Saccharomyces cerevisiae*) dans le but de les comparer (chapitre VII).

Chapitre IV

Exploitation des méthodes d'analyse de texture en imagerie

1 Adaptation des méthodes d'analyse de texture en imagerie à notre système (tracé HCA)

1.1 Transposition du tracé HCA dans une matrice

Nous avons inclus la séquence protéique sous sa représentation HCA dans une matrice de calcul similaire à celle utilisée dans les méthodes de cooccurrence et des longueurs de plages de niveaux de gris (cf. chapitre 3). Cette matrice est définie par une longueur et une hauteur. La hauteur et la longueur sont fixées à 5 acides aminés. Cette valeur correspond à la plus grande séquence d'acides aminés qu'il est possible de lire dans les directions verticale et diagonale du tracé HCA (Figure 21). Une matrice de 25 (longueur et hauteur de 5 acides aminés) correspond à une séquence de 17 acides aminés de long. En effet, par ses propriétés (duplication du tracé et connectivité de 4), le tracé HCA entraîne une duplication de certains acides aminés dans la matrice. Ainsi, les acides aminés FL (de la première colonne de la matrice ci-dessous) sont présents également dans la seconde colonne.

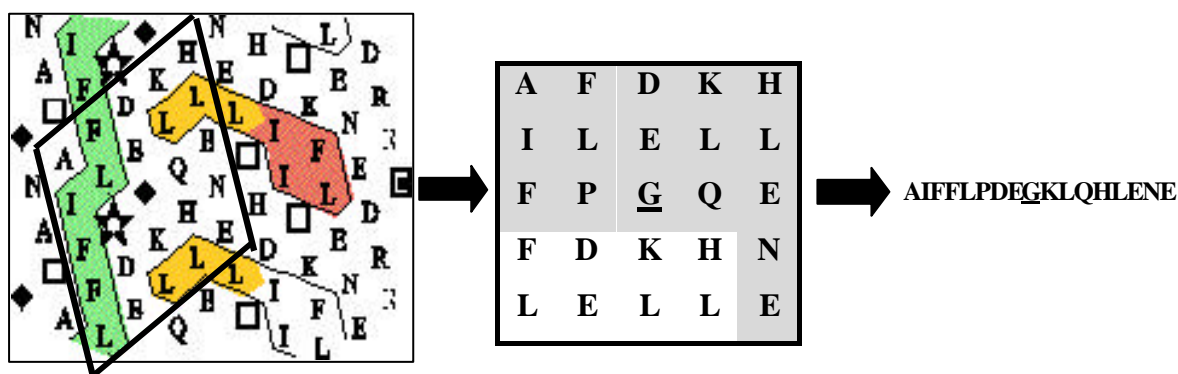


Figure 21 : Transposition du tracé HCA dans une matrice.

Dans le cas présent, cette matrice correspond à une fenêtre de longueur 17 acides aminés. La matrice est représentée par l'encadré noir sur le tracé HCA. La fenêtre correspond à la séquence à droite de la matrice. Le résidu central de cette fenêtre est ici la glycine (acide aminé souligné, losange noir dans la représentation HCA).

couronne correspond aux positions -2, +2, +6, +7, +8, +5, +2, -2, -6, -7, -8 et -5 sur le tracé HCA (résidus D, F, A, I, F, D, L, L, E, N, E et L).

Nous avons également testé une fenêtre de taille 31, mais elle s'avère moins sensible pour les calculs que la fenêtre de 17, qui a finalement été retenue pour les calculs de texture. En effet, cette longueur de fenêtre est trop importante et entraîne une moins bonne détection des changements de texture. Nous n'avons donc pas approfondi l'étude avec cette dernière taille de fenêtre.

1.3 Choix d'un code

Lors de l'analyse d'une image, les méthodes de calcul visent à rechercher des régions de même intensité de gris et de même répartition. Les pixels sont représentés par un indice attestant de leur niveau de gris (par exemple, de 1 à 4 allant du gris clair au gris foncé). Pour adapter ce système aux séquences de protéines, nous avons assimilé un acide aminé de la matrice à un pixel. Notre matrice peut contenir vingt types de pixels (les 20 acides aminés retrouvés dans la nature). L'indice de niveau de gris peut être remplacé par la nature de l'acide aminé, une valeur d'hydrophobie, une valeur d'enfouissement, une propension à être dans un certain type de structures secondaires, etc... Dans un premier temps, nous avons fixé comme indice une valeur d'hydrophobie (ou alphabet) et évalué plusieurs échelles de niveaux (2 et 4).

1.3.1 Code hydrophobe HCA (code01)

Nous avons séparé les acides aminés en deux groupes sur la base d'une dichotomie hydrophobe/hydrophile. Les acides aminés hydrophobes sont V, I, L, F, M Y et W; les autres sont réputés hydrophiles ou en tout cas non hydrophobes forts. Les hydrophobes sont codés 1 et les hydrophiles sont codés 0. Chaque acide aminé de la matrice est substitué par le code correspondant.

1.3.2 Code à 4 groupes (code 1234)

Les acides aminés peuvent également être classés suivant leur enfouissement moyen au sein d'un ensemble représentatif de structures protéiques (Figure 23) [PINTAR, A. et al., 2003b; PINTAR, A. et al., 2003a]. L'enfouissement correspond à la distance entre le résidu de la protéine et la plus proche molécule d'eau entourant la protéine. Ces valeurs d'enfouissement moyen ont été pondérées par les fréquences des effectifs de chaque acide aminé dans une banque Genpep utilisée au laboratoire [HENNETIN, J. 2003]. Nous nous

sommes très vite rendus compte qu'un code à 20 niveaux différents était pénalisant pour nos calculs car, entre autres, il sollicitait beaucoup de temps machine (matrice de calcul : 20 x 20). C'est pourquoi nous avons regroupé certains acides aminés de la manière décrite ci-après. A partir de la Figure 23 précédente, nous avons constitué quatre classes d'acides aminés partageant des valeurs d'enfouissement assez proches (code à 4 groupes).

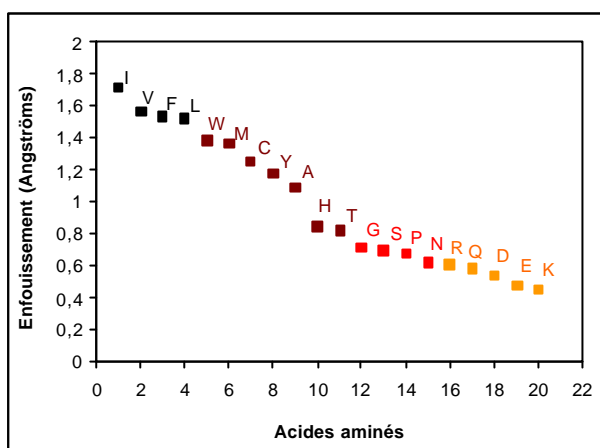


Figure 23 : Enfouissement moyen de chaque acide aminé, présenté en mode décroissant à partir des données brutes de Pintar et al [PINTAR, A. et al., 2003b].

Tableau 1 : Constitution des quatre groupes d'acides aminés suivant la valeur d'enfouissement moyen.

Groupe	Acides aminés	Enfouissement moyen (Angströms)
Groupe 1 (orange)	KEDQR	0,52
Groupe 2 (rouge)	NPSG	0,68
Groupe 3 (marron)	THAYCMW	1,07
Groupe 4 (noir)	LFVI	1,58

Pour chaque classe, nous avons déterminé un enfouissement moyen que nous appliquerons à chacun des acides aminés y appartenant.

1.3.3 Code amas-non amas (code amas)

Nous avons établi un dernier code pour séparer les acides aminés inclus dans des amas hydrophobes déterminés par la méthode HCA des acides aminés hors amas. Nous considérons les acides aminés (hydrophobes et non hydrophobes) inclus dans les amas comme codés par '1' et les autres par '0'. Ce code ne sera utilisé que dans la méthode des longueurs de plage.

Séq (AA)	L	T	P	K	L	L	P	K	R	A	L	A	A	L	A	V	G	G
Code binaire	1	0	0	0	1	1	0	0	0	0	1	0	0	1	0	1	0	0
Code amas-non amas	1	0	0	0	1	1	0	0	0	0	1	1	1	1	1	1	0	0

1.4 Choix d'une direction (pour les calculs de matrice de cooccurrence et de longueurs de plage)

La méthode de matrice de cooccurrence et la méthode des longueurs de plage consistent respectivement à déterminer la fréquence d'apparition d'un « motif » formé de deux pixels séparés par une certaine distance d ; et à compter le nombre de plages d'une certaine longueur j , de niveau de gris i . Dans les deux cas, le calcul est réalisé dans une direction particulière θ donné. Dans notre étude, nous avons fixé comme direction l'axe horizontal ($\theta = 0^\circ$). En effet, l'utilisation de l'axe vertical correspond à la séquence linéaire retrouvée dans le tracé HCA. Les autres directions ($45^\circ, 135^\circ$, etc...) nous semblaient moins facile à mettre en œuvre et moins adéquates à notre représentation HCA.

2 ***Constitution des banques de protéines de référence***

Nous avons récupéré la banque ASTRAL version 1.63 à 95% non redondante (8373 séquences partageant au plus 95% d'identité de séquence) [BRENNER, S. E. et al., 2000]. Nous avons séparé les séquences en plusieurs classes rassemblant des protéines partageant le même type de repliement selon la classification SCOP (Tableau 2) :

A : protéines présentant un repliement tout alpha

B : protéines présentant un repliement tout bêta

C : protéines présentant un repliement alpha/bêta

D : protéines présentant un repliement alpha+bêta

E : protéines multi-domaines

F : protéines membranaires et de surface

G : petites protéines ayant peu de structures secondaires.

Les classes A, B, C, D, F et G ont été utilisées séparément dans les méthodes d'analyse de texture pour calibrer le système. Les classes A, B, C, D et E ont été groupées pour constituer un jeu de données de 7530 séquences (classe « ABCDE_Scop »)

caractéristiques des domaines globulaires des protéines. Cette classe sera utilisée dans la caractérisation des domaines globulaires des protéines (cf. chapitre V).

Tableau 2 : Effectifs des banques de protéines.

Banque	Effectifs totaux (séquences)	Effectifs totaux (Acides aminés)	Effectifs utilisés avec une fenêtre de 17	Effectifs utilisés avec une fenêtre de 11
A	1 521	219 642	195 306	204 432
B	2 262	335 447	299 255	312 827
C	1 798	460 121	431 353	442 141
D	1 794	281 923	253 219	263 983
E	155	64 303	/	/
F	155	31 516	29 036	29 966
G	688	38 228	27 220	31 348
Total	8 373	1 431 180	1 235 389	1 284 697

L'évaluation des paramètres dans les méthodes d'analyse de texture se fait en calculant la valeur du paramètre pour chaque fenêtre de la séquence protéique. Une fenêtre de 17 acides aminés ne tient pas compte des 8 premiers et 8 derniers résidus, une fenêtre de 11 des 5 premiers et 5 derniers résidus. Le nombre d'acides aminés d'une protéine utilisés dans le calcul ne prend pas en compte les quelques acides aminés du début et de fin de la séquence correspondant respectivement à la première et à la dernière fenêtre de la séquence (Figure 24). Les effectifs de chaque banque sont différents suivant la taille de fenêtre utilisée pour le calcul du paramètre (Tableau 2).



Figure 24 : Principe de la fenêtre glissante.

La première fenêtre de 17 a pour résidu central L, la seconde G et la troisième P etc.... Les huit premiers acides aminés (grisés, **GSLRHSGP**) ne sont pas pris en compte dans les calculs, de même que les huit derniers acides aminés de la séquence (**SPDSPKGS**).

3 Résultats

3.1 Méthode de premier ordre

3.1.1 Etude de la distribution du paramètre « pourcentage d'acides aminés hydrophobes »

Nous avons étudié le paramètre « pourcentage d'acides aminés hydrophobes » dans la méthode de premier ordre afin d'avoir une vision du profil d'hydrophobie des séquences

protéiques des différentes classes de notre banque et de mettre en évidence éventuellement des similarités de profils. Nous avons évalué ce paramètre en utilisant les deux tailles de fenêtre 17 et 11 acides aminés. Le code utilisé a peu d'importance puisqu'il n'influe pas sur le nombre d'acides aminés hydrophobes de la fenêtre. Nous avons utilisé le code 01. Le Tableau 3 présente les valeurs moyennes du pourcentage d'acides aminés hydrophobes calculées pour chaque classe de notre banque.

Tableau 3 : Méthode de premier ordre.

		classe SCOP	min	%VILFMYW (ó)	max
code 01	fen 17	A	0,00	32,36 +/- 9,87	64,71
		B	0,00	31,74 +/- 9,25	64,71
		C	0,00	33,16 +/- 9,12	64,71
		D	0,00	32,59 +/- 9,28	64,71
		F	0,00	38,59 +/- 13,97	82,35
		G	0,00	23,56 +/- 9,94	52,94
		fen 11	A	0,00	32,39 +/- 12,55
	B		0,00	31,74 +/- 11,90	81,81
	C		0,00	33,19 +/- 11,85	81,81
	D		0,00	32,61 +/- 11,90	90,91
	F		0,00	38,61 +/- 16,35	100,00
	G		0,00	23,33 +/- 12,18	63,64

Les valeurs moyennes de chaque classe sont identiques pour les deux tailles de fenêtres. Nous avons également calculé les distributions des pourcentages d'acides aminés hydrophobes pour les fenêtres 17 et 11. Nous avons représenté ici la distribution obtenue en utilisant la fenêtre de 17 (Figure 25).

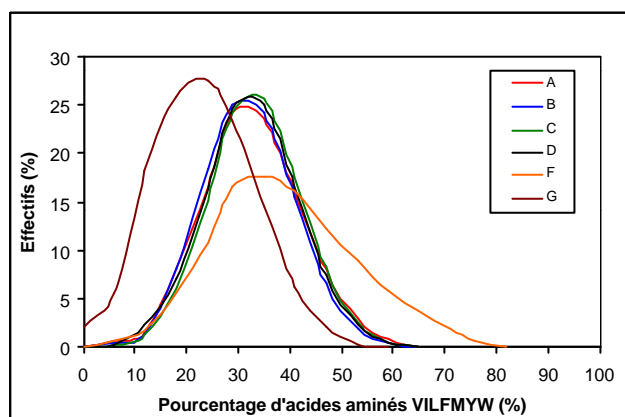


Figure 25 : Distribution du pourcentage d'acides aminés hydrophobes, code 1234, fenêtre 17.

Les courbes représentatives des quatre classes A, B, C et D, correspondant respectivement aux repliements tout α , tout β , α/β et $\alpha+\beta$, sont superposées ; il n'y a pas de différence significative du pourcentage d'acides aminés hydrophobes. Le pourcentage moyen d'acides aminés hydrophobes calculé par fenêtre est de 32,5%. Le nombre minimal d'acides aminés hydrophobes par fenêtre est de 0 résidus et le nombre maximal est de 11 résidus.

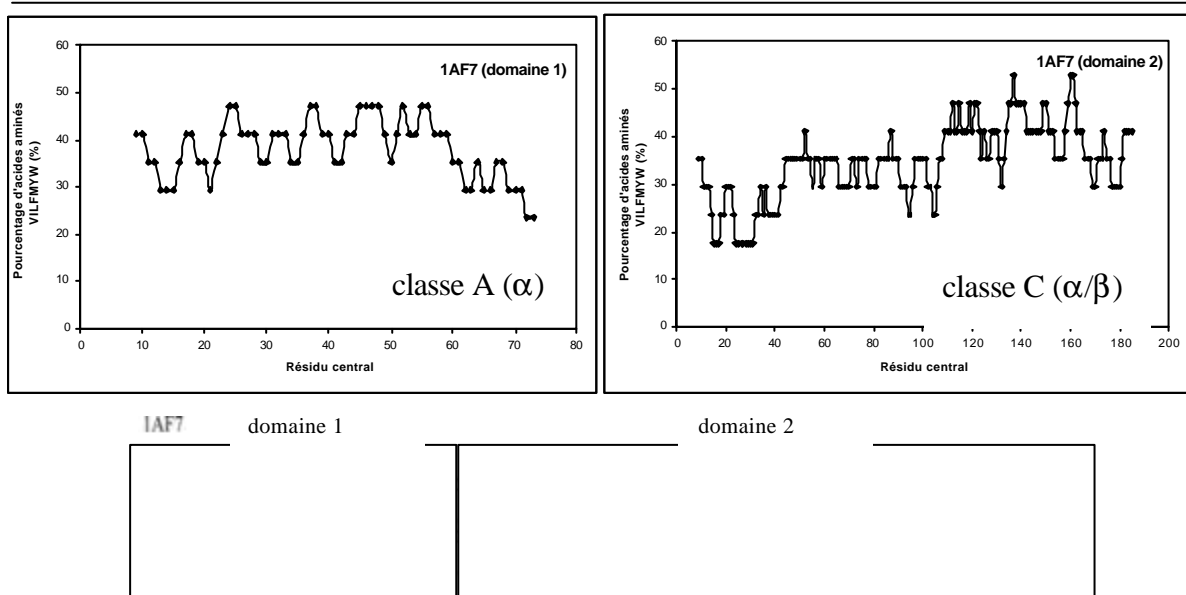
La banque F présente une distribution plus aplatie que les autres courbes (kurtosis plus faible par rapport aux courbes A, B, C et D). Le pourcentage moyen d'acides aminés hydrophobes (38,59%) diffère significativement par rapport au autres valeurs moyennes calculées (rejet de l'hypothèse H_0 dans le test statistique de comparaison de deux moyennes d'après [GRAIS, B 1992]). Cette banque, correspondant aux protéines membranaires, contient de nombreux amas hydrophobes plus longs et plus riches en acides aminés hydrophobes. Le nombre minimal d'acides aminés hydrophobes par fenêtre est de 0 résidus et le nombre maximal est de 14 résidus.

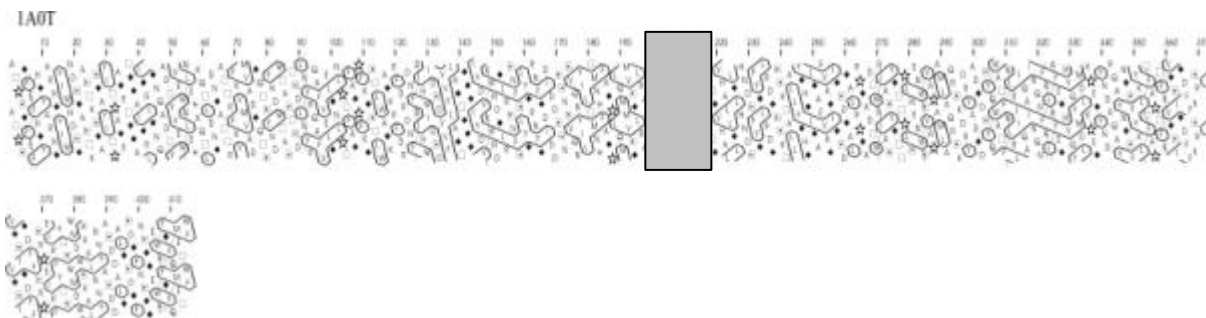
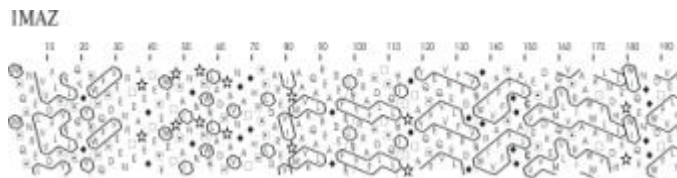
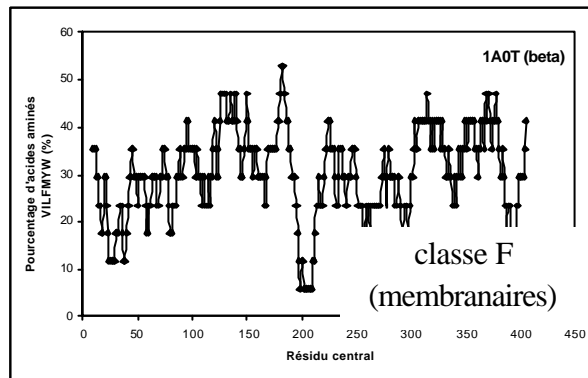
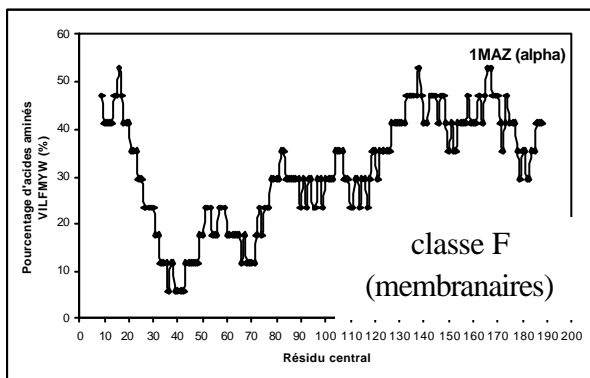
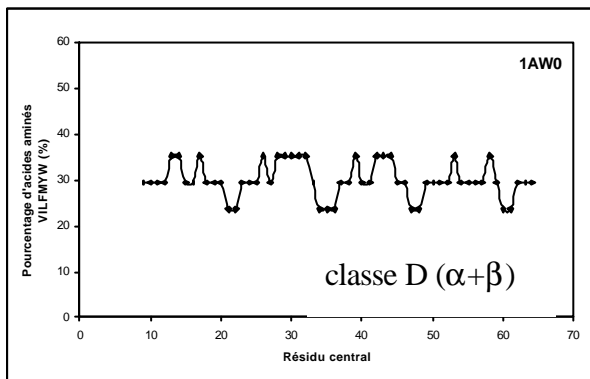
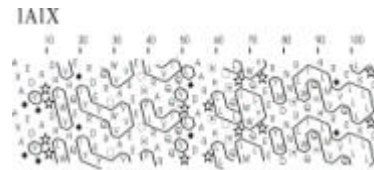
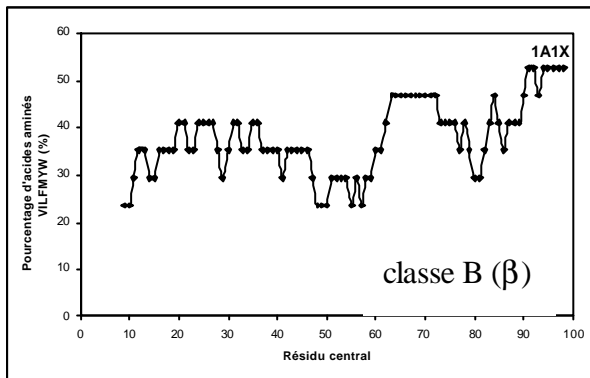
La banque G, contenant les petites protéines, présente une distribution déplacée vers la gauche par rapport aux autres courbes (skewness négatif par rapport aux courbes A, B, C et D). Le pourcentage moyen d'acides aminés hydrophobes (23,56%) diffère significativement par rapport aux autres valeurs moyennes calculées (rejet de l'hypothèse H_0 dans le test statistique de comparaison de deux moyennes d'après [GRAIS, B 1992]). Le nombre minimal d'acides aminés hydrophobes par fenêtre est de 0 résidus et le nombre maximal est de 9 résidus.

Nous avons constaté que dans les calculs avec la fenêtre de taille 11, les écarts-types sont logiquement plus élevés et les valeurs sont plus dispersées par rapport à la valeur moyenne de la banque.

3.1.2 Etude du profil d'hydrophobie

Nous avons étudié le profil d'hydrophobie de différentes séquences représentatives de chaque classe structurale (Figure 26).





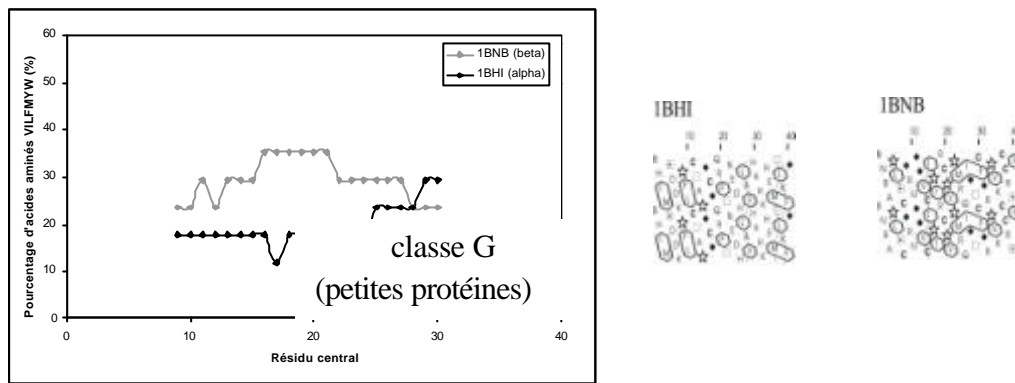


Figure 26 : Profil d'hydrophobie de différentes protéines issues des banques A, B, C, D, F et G.
Les chaînes protéiques 1AF7 (domaine 1), 1A1X, 1AF7 (domaine 2), 1AW0, 1MAZ, 1A0T, 1BHI et 1BNB correspondent respectivement à des protéines tout alpha, tout bêta, alpha/bêta, alpha+bêta, membranaires (hélices transmembranaires), membranaires (brins bêta formant un pore), à des protéines de petites tailles avec une structure majoritaire en hélice et à de protéines de petites tailles avec une structure majoritaire en brins.

Nous constatons à partir des séquences étudiées que les valeurs de pourcentage d'acides aminés hydrophobes très faibles (5-10%) correspondent à des régions de la séquence présentant peu ou pas d'amas hydrophobes (encadrées sur les tracés HCA) et inversement des régions riches en amas hydrophobes présentent des pourcentages d'acides aminés hydrophobes entre 60 et 80% (comme par exemple dans les protéines 1A0T ou 1MAZ).

3.2 Méthode de cooccurrence

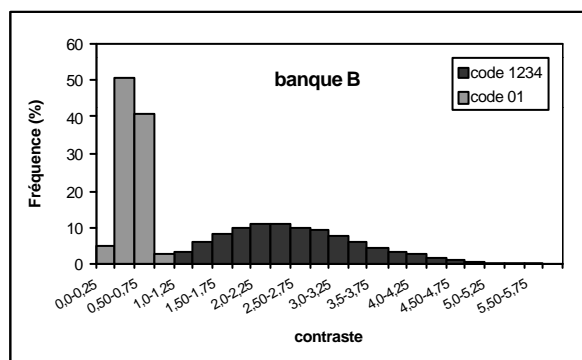
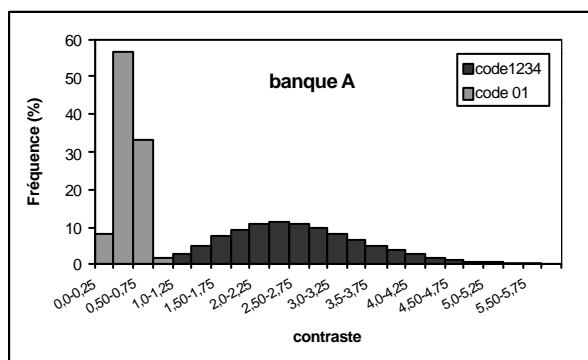
3.2.1 Etude des distributions des paramètres entropie et contraste

Les deux paramètres nous semblant pertinents sont l'entropie et le contraste. L'entropie mesure la complexité de la séquence. Elle permet de caractériser le degré de granulation de la séquence sous sa représentation HCA. Le contraste mesure les variations locales des niveaux d'enfouissement ou d'hydrophobie de la séquence. Nous avons déterminé l'entropie et le contraste moyen pour les classes A, B, C, D, F et G de notre banque en utilisant les deux tailles de fenêtres 17 et 11 et deux codages différents (code 01 et code 1234 correspondant respectivement au code hydrophobe HCA et au code à 4 groupes) (Tableau 4, page suivante).

Tableau 4 : Méthode de cooccurrence.

		min	entropie (ó)	max	min	contraste (ó)	max		
code 1234	fen 17	A	0,24	0,70 +/- 0,06	0,86	0,30	2,61 +/- 0,90	7,55	
		B	0,00	0,71 +/- 0,06	0,86	0,00	2,53 +/- 0,91	7,40	
		C	0,00	0,71 +/- 0,06	0,86	0,00	2,60 +/- 0,91	7,40	
		D	0,07	0,71 +/- 0,06	0,86	0,20	2,66 +/- 0,93	7,05	
		F	0,28	0,69 +/- 0,07	0,86	0,20	2,24 +/- 0,91	7,00	
		G	0,26	0,70 +/- 0,06	0,86	0,30	2,26 +/- 0,86	6,15	
		fen 11	A	0,14	0,76 +/- 0,09	0,88	0,00	2,62 +/- 1,18	8,50
	B		0,00	0,76 +/- 0,09	0,88	0,00	2,54 +/- 1,17	9,00	
	C		0,00	0,76 +/- 0,08	0,88	0,00	2,60 +/- 1,17	9,00	
	D		0,00	0,76 +/- 0,08	0,88	0,00	2,67 +/- 1,21	8,20	
	F		0,14	0,74 +/- 0,10	0,88	0,10	2,25 +/- 1,15	7,70	
	G		0,14	0,76 +/- 0,09	0,88	0,10	2,26 +/- 1,09	8,10	
	code 01		fen 17	A	0,00	0,38 +/- 0,07	0,46	0,00	0,42 +/- 0,14
		B		0,00	0,38 +/- 0,06	0,46	0,00	0,45 +/- 0,14	1,00
C		0,00		0,39 +/- 0,06	0,46	0,00	0,45 +/- 0,14	1,00	
D		0,00		0,39 +/- 0,06	0,46	0,00	0,45 +/- 0,14	0,95	
F		0,00		0,39 +/- 0,06	0,46	0,00	0,46 +/- 0,15	1,00	
G		0,00		0,32 +/- 0,09	0,46	0,00	0,37 +/- 0,16	0,90	
fen 11		A		0,00	0,46 +/- 0,11	0,59	0,00	0,42 +/- 0,19	1,00
		B	0,00	0,45 +/- 0,11	0,59	0,00	0,46 +/- 0,19	1,00	
		C	0,00	0,46 +/- 0,11	0,59	0,00	0,45 +/- 0,19	1,00	
		D	0,00	0,46 +/- 0,11	0,59	0,00	0,45 +/- 0,19	1,00	
		F	0,00	0,47 +/- 0,11	0,59	0,00	0,46 +/- 0,19	1,00	
		G	0,00	0,38 +/- 0,15	0,59	0,00	0,37 +/- 0,20	1,00	

Nous ne présentons ici que les distributions obtenues en utilisant une taille de fenêtre de 17. En effet, nous avons constaté que l'évaluation des paramètres de cooccurrence et de longueur de plage, faisant intervenir l'utilisation d'une matrice de calcul, n'est pas intéressante quand la matrice est de petite taille (pour une fenêtre de 11 acides aminés, nous utilisons une matrice de taille 15 avec une longueur de 3 et une hauteur de 5). Dans la méthode des longueurs de plage, une longueur de 3 résidus dans la direction horizontale ne permet en effet pas de faire ressortir une information d'uniformité ou d'homogénéité dans la distribution des plages recherchées. Les Figure 27 et 28 montrent les distributions calculées pour les paramètres entropie et contraste avec une fenêtre de 17 acides aminés et les codes 1234 et 01.



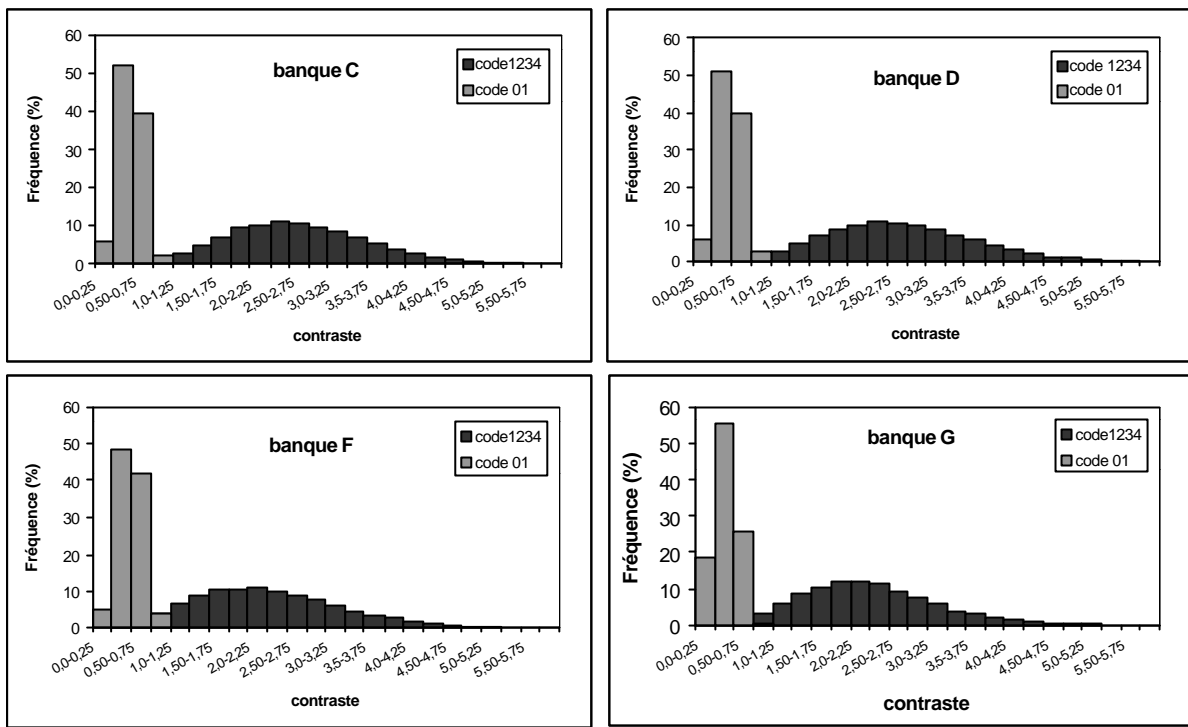
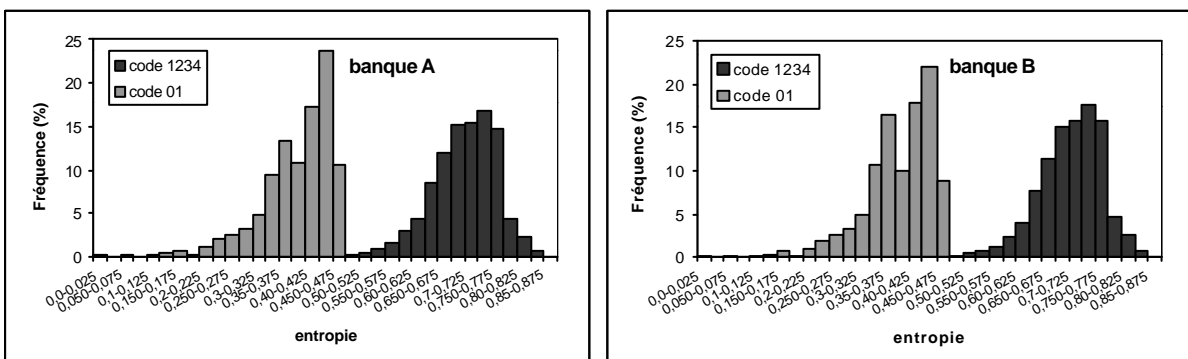


Figure 27 : Distribution des valeurs du contraste dans chaque classe de notre banque en utilisant une fenêtre de 17 acides aminés.

Pour le contraste, toutes les classes de notre banque présentent un aspect global similaire au sein d'un même code (Figure 27). Les modes sont respectivement compris dans les intervalles 0,5-0,75 et 2,25-2,50, pour les codes 01 et 1234. L'étendue (différence entre la plus grande et la plus petite valeur) est identique pour toutes les classes au sein d'un même code. Seule, la distribution de la banque G semble différente avec le code 01. Les valeurs de contraste moyen des banques F et G sont plus faibles que celles des autres banques.

Pour l'entropie, nous avons à nouveau des distributions très similaires, hormis pour la banque G (Figure 28).



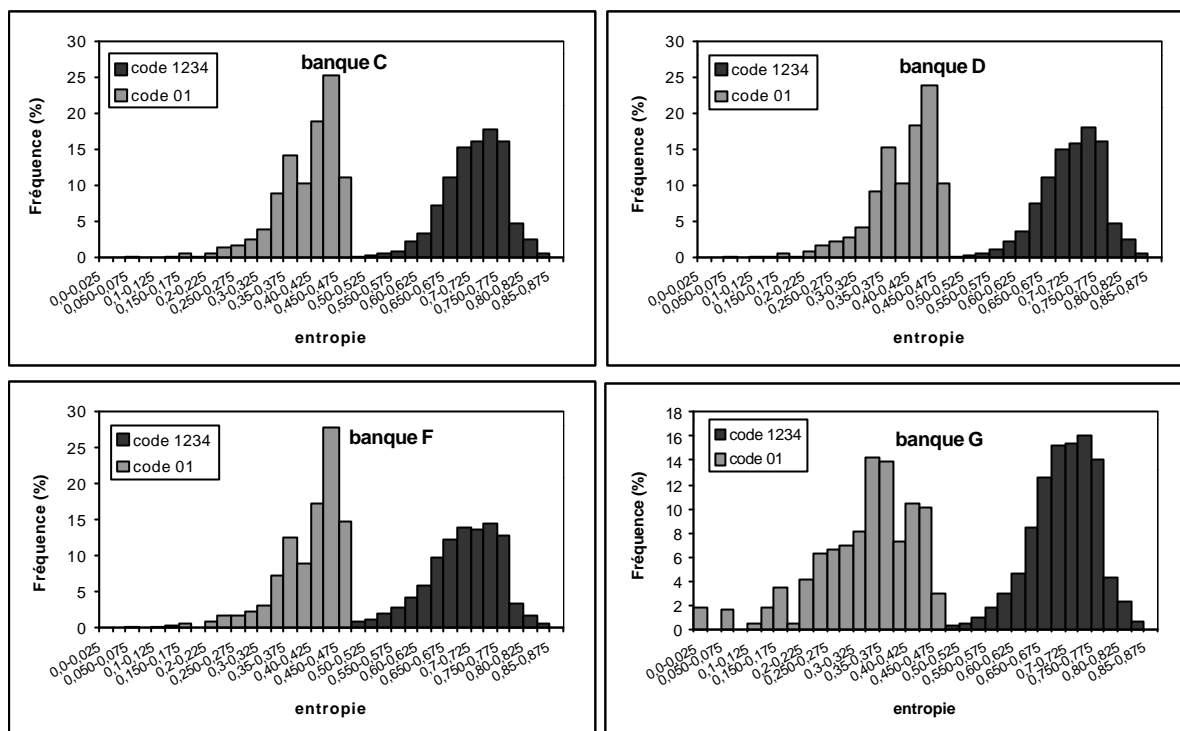
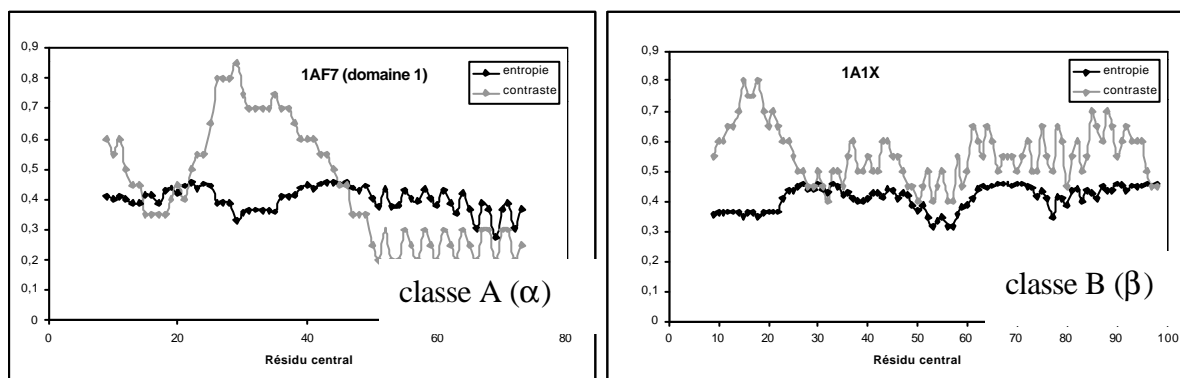


Figure 28 : Distribution de l'entropie dans chaque classe avec la fenêtre de 17 acides aminés.

Avec le code 01, la distribution de la courbe G est beaucoup plus étendue que celle des autres courbes. Avec le code 1234, elle ressemble au contraire plus à une courbe gaussienne avec des valeurs plus élevées que les autres courbes. Les valeurs moyennes sont identiques pour le code 1234 alors qu'elles diffèrent avec le code 01, l'entropie moyenne étant beaucoup plus faible dans la banque G que dans les autres banques.

3.2.2 Etude de profils de cooccurrence

Nous avons tracé les profils de contraste et d'entropie des huit fragments protéiques utilisés précédemment (Figure 29).



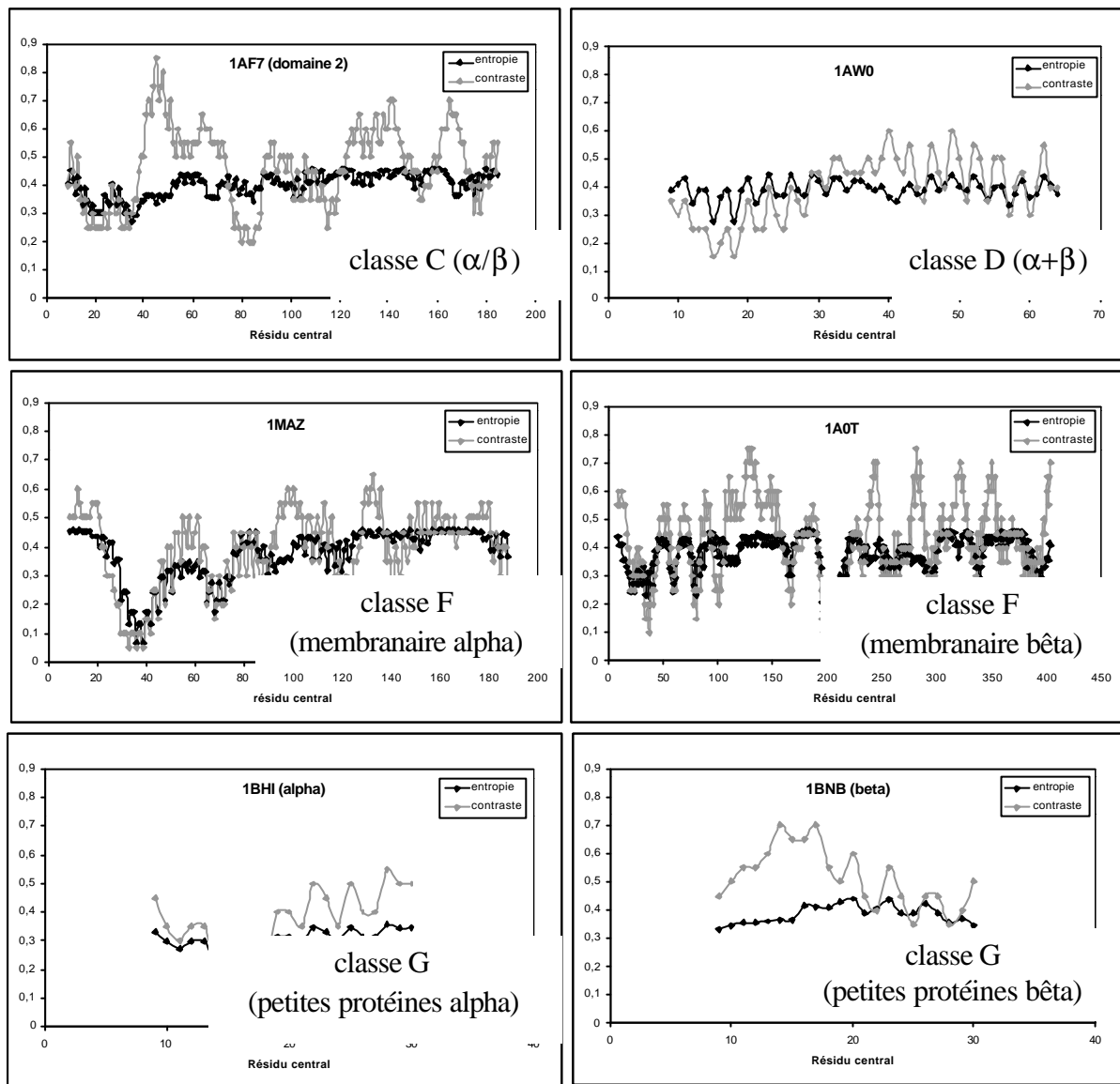


Figure 29 : Profil de cooccurrence de protéines issues de nos banques A, B, C, D, F et G.

Nous avons constaté que dans l'ensemble, l'entropie varie peu au niveau de nos séquences, sauf lorsque sont présentes des régions possédant peu d'acides aminés hydrophobes (comme par exemple les régions 200 à 215 de la protéine 1A0T et 35 à 70 de la protéine 1MAZ). Le contraste semble être plus sensible cependant, on constate qu'il fluctue également avec la présence/absence d'amas hydrophobes dans la séquence (Figure 29).

3.3 Méthode des longueurs de plage

3.3.1 Etude des distributions des paramètres SRE, LRE, GLD, RLD et RLP

Pour la mise en place de la méthode de longueurs de plage, nous avons redéfini plusieurs paramètres en accord avec le système HCA:

- SRE : paramètre « poids » des plages courtes (une plage correspondant à l'ensemble des acides aminés d'une séquence ayant la même valeur d'hydrophobie ou une autre grandeur suivant l'échelle choisie).
- LRE : paramètre « poids » des plages longues
- GLD : paramètre « distribution des niveaux d'hydrophobie ou d'enfouissement », qui mesure l'uniformité de la distribution des niveaux d'hydrophobie
- RLD : paramètre « distribution des plages »
- RLP : paramètre qui correspond au pourcentage de plages.

Nous avons appliqué les deux codages (code hydrophobe HCA et code à 4 groupes) en utilisant la fenêtre de 17 acides aminés. Nous avons testé également le code spécial amas-non amas (acide aminé inclus dans un amas et acide aminé non inclus dans amas) qui devrait nous permettre de cerner la répartition des amas les uns par rapport aux autres dans la séquence, leur densité dans la séquence et leur type (long ou court).

L'utilisation du code amas transpose les amas comme des plages appelées «plages d'amas» et l'utilisation du code 01 transpose les successions d'acides aminés hydrophobes en plages d'hydrophobes. Nous précisons ici la différence entre plage d'amas et plage d'hydrophobes (Figure 30).

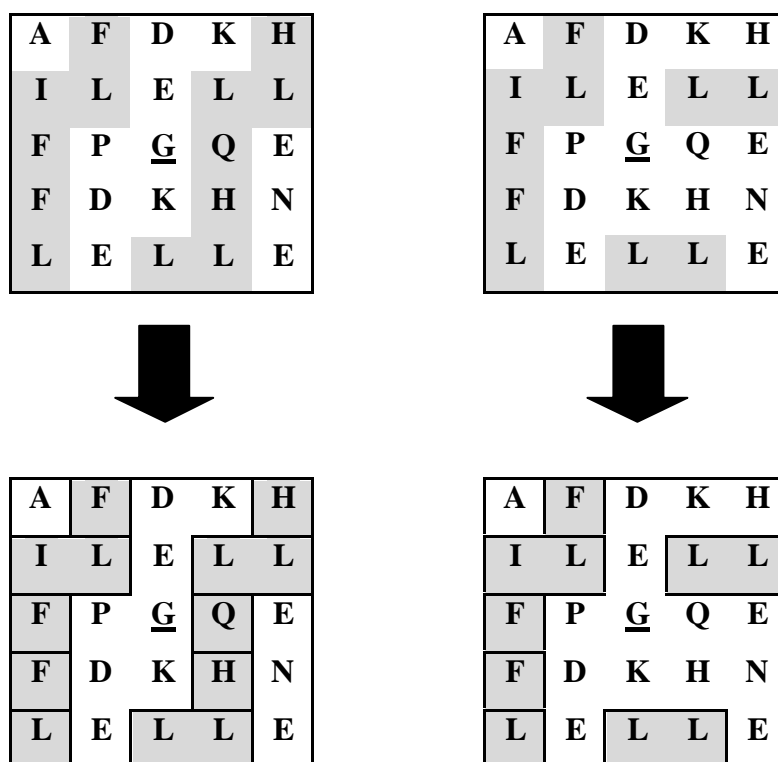


Figure 30 : Définition de plages d'amas (à gauche) et de plages d'hydrophobes (à droite) dans une fenêtre.

Une plage est une région de la matrice constituée par un ou plusieurs acides aminés lus dans la direction horizontale. A gauche, tous les acides aminés inclus dans des amas sont grisés. Les plages d'amas sont encadrées. Nous dénombrons 7 plages d'amas de longueur 1 (F, F, L, F, Q, H et H) et 3 plages d'amas de longueur 2 (IL, LL et LL). Une plage d'amas peut contenir n'importe quel acide aminé, qu'il soit hydrophobe fort ou non hydrophobe fort. A droite, seuls les acides aminés hydrophobes (V, I, L, F, M, Y et W) sont grisés. Les plages d'hydrophobes (encadrées) correspondent à la plus grande succession d'acides aminés hydrophobes (V, I, L, F, M, Y et W). Nous dénombrons 4 plages d'hydrophobes de longueur 1 (F, F, L et F) et 3 plages d'amas de longueur 2 (IL, LL et LL). Le résidu central de la fenêtre est souligné.

Les paramètres moyens calculés avec les codes amas, 01 et 1234 sont présentés

Tableau 5 et Tableau 6.

Tableau 5 : Méthode des longueurs de plage, code Amas, fenêtre 17.

	classe SCOP	SRE (°)			LRE (°)			GLD (°)			RLD (°)			RLP (°)			
		min		max	min		max	min		max	min		max	min		max	
code amas	fen 17	A	0,04	0,49 +/- 0,19	0,97	1,13	7,07 +/- 5,94	25,00	4,14	6,58 +/- 1,56	12,08	1,75	5,35 +/- 2,43	22,08	0,29	0,73 +/- 0,21	1,41
		B	0,04	0,54 +/- 0,17	1,00	1,00	5,51 +/- 4,47	25,00	4,14	5,51 +/- 1,57	12,52	1,75	5,99 +/- 2,67	25,00	0,29	0,80 +/- 0,20	1,47
		C	0,04	0,52 +/- 0,18	1,00	1,00	6,29 +/- 5,40	25,00	4,14	6,84 +/- 1,59	12,52	1,75	5,72 +/- 2,60	25,00	0,29	0,76 +/- 0,21	1,47
		D	0,04	0,52 +/- 0,18	0,97	1,13	6,18 +/- 5,23	25,00	4,14	6,84 +/- 1,59	12,08	1,75	5,72 +/- 2,58	22,08	0,29	0,76 +/- 0,21	1,41
		F	0,04	0,44 +/- 0,213	0,96	1,26	9,02 +/- 7,45	25,00	4,14	6,32 +/- 1,59	11,70	1,75	5,10 +/- 2,37	21,09	0,29	0,67 +/- 0,24	1,35
		G	0,04	0,53 +/- 0,17	0,96	1,26	6,23 +/- 4,63	25,00	4,14	6,79 +/- 1,52	11,52	1,75	5,58 +/- 2,58	21,09	0,29	0,76 +/- 0,20	1,35

Tableau 6 : Méthode des longueurs de plage, code 01 et 1234, fenêtres 17 et 11.

		classe	SRE (°)			LRE (°)			GLD (°)			RLD (°)			RLP (°)			
			min		max	min		max	min		max	min		max	min		max	
code 1234	fen 17	A	0,24	0,82 +/- 0,10	1,00	1,00	2,06 +/- 0,80	15,57	2,50	5,63 +/- 0,82	10,27	1,75	13,04 +/- 4,29	25,00	0,41	1,15 +/- 0,15	1,47	
		B	0,04	0,83 +/- 0,09	1,00	1,00	1,92 +/- 0,72	25,00	2,14	5,73 +/- 0,81	11,09	1,89	13,88 +/- 4,35	25,00	0,29	1,18 +/- 0,14	1,47	
		C	0,04	0,82 +/- 0,09	1,00	1,00	1,99 +/- 0,75	25,00	1,67	5,64 +/- 0,80	11,52	1,75	13,44 +/- 4,31	25,00	0,29	1,17 +/- 0,15	1,47	
		D	0,20	0,83 +/- 0,09	1,00	1,00	1,97 +/- 0,74	19,50	2,25	5,68 +/- 0,80	10,09	1,75	13,58 +/- 4,36	25,00	0,35	1,17 +/- 0,15	1,47	
		F	0,34	0,82 +/- 0,10	1,00	1,00	2,10 +/- 0,87	10,33	2,80	5,79 +/- 0,87	10,52	2,11	13,01 +/- 4,36	25,00	0,53	1,15 +/- 0,15	1,47	
		G	0,39	0,83 +/- 0,09	1,00	1,00	1,98 +/- 0,88	11,00	2,82	5,76 +/- 0,82	10,52	3,00	13,52 +/- 4,31	25,00	0,53	1,17 +/- 0,15	1,47	
		fen 11	A	0,11	0,84 +/- 0,11	1,00	1,00	1,81 +/- 0,69	9,00	1,67	3,81 +/- 0,63	7,53	2,43	8,79 +/- 3,13	15,00	0,46	1,12 +/- 0,15	1,36
	B		0,11	0,85 +/- 0,11	1,00	1,00	1,71 +/- 0,63	9,00	1,67	3,84 +/- 0,62	7,53	2,43	9,28 +/- 3,14	15,00	0,46	1,14 +/- 0,15	1,36	
	C		0,11	0,85 +/- 0,11	1,00	1,00	1,76 +/- 0,66	9,00	1,67	3,79 +/- 0,61	7,53	2,43	9,02 +/- 3,12	15,00	0,46	1,13 +/- 0,15	1,36	
	D		0,11	0,85 +/- 0,11	1,00	1,00	1,75 +/- 0,66	9,00	1,67	3,81 +/- 0,62	7,53	2,43	9,09 +/- 3,15	15,00	0,46	1,13 +/- 0,15	1,36	
	F		0,28	0,84 +/- 0,12	1,00	1,00	1,84 +/- 0,73	6,83	1,67	3,92 +/- 0,66	7,14	2,43	8,77 +/- 3,17	15,00	0,55	1,11 +/- 0,16	1,36	
	G		0,28	0,85 +/- 0,11	1,00	1,00	1,76 +/- 0,65	6,83	1,67	3,87 +/- 0,63	7,00	2,43	9,04 +/- 3,13	15,00	0,55	1,13 +/- 0,15	1,36	
	code 01		fen 17	A	0,04	0,59 +/- 0,14	0,97	1,13	5,34 +/- 2,77	25,00	3,00	6,93 +/- 1,34	12,08	1,75	5,66 +/- 2,62	22,08	0,29	0,79 +/- 0,17
		B		0,04	0,62 +/- 0,13	1,00	1,00	4,85 +/- 2,54	25,00	2,60	7,25 +/- 1,36	12,52	1,75	6,26 +/- 2,82	25,00	0,29	0,83 +/- 0,17	1,47
C		0,04		0,61 +/- 0,13	1,00	1,00	4,92 +/- 2,53	25,00	2,60	7,16 +/- 1,36	12,52	1,75	6,09 +/- 2,75	25,00	0,29	0,82 +/- 0,17	1,47	
D		0,04		0,61 +/- 0,13	0,97	1,13	4,93 +/- 2,56	25,00	2,60	7,18 +/- 1,36	12,08	1,75	6,13 +/- 2,75	22,08	0,29	0,82 +/- 0,17	1,41	
F		0,04		0,62 +/- 0,14	1,00	1,00	4,74 +/- 2,58	25,00	3,00	7,28 +/- 1,40	12,52	1,75	6,42 +/- 2,99	25,00	0,29	0,84 +/- 0,18	1,47	
G		0,04		0,56 +/- 0,16	0,95	1,26	6,85 +/- 4,32	25,00	3,33	6,64 +/- 1,35	11,52	1,75	5,18 +/- 2,47	19,35	0,29	0,73 +/- 0,19	1,35	
fen 11		A		0,11	0,61 +/- 0,17	1,00	1,00	3,55 +/- 1,49	9,00	2,60	4,92 +/- 0,79	7,53	2,43	4,56 +/- 2,02	15,00	0,46	0,84 +/- 0,17	1,36
		B	0,11	0,64 +/- 0,17	1,00	1,00	3,32 +/- 1,44	9,00	2,60	5,08 +/- 0,80	7,53	2,43	4,90 +/- 2,22	15,00	0,46	0,87 +/- 0,18	1,36	
		C	0,11	0,63 +/- 0,17	1,00	1,00	3,36 +/- 1,44	9,00	2,60	5,01 +/- 0,81	7,53	2,43	4,80 +/- 2,16	15,00	0,46	0,87 +/- 0,17	1,36	
		D	0,11	0,64 +/- 0,17	1,00	1,00	3,36 +/- 1,45	9,00	2,60	5,03 +/- 0,81	7,53	2,43	4,82 +/- 2,17	15,00	0,46	0,86 +/- 0,17	1,36	
		F	0,11	0,65 +/- 0,17	1,00	1,00	3,28 +/- 1,43	9,00	2,60	5,07 +/- 0,82	7,53	2,43	4,99 +/- 2,18	15,00	0,46	0,87 +/- 0,18	1,36	
		G	0,11	0,57 +/- 0,20	1,00	1,00	4,13 +/- 1,89	9,00	2,60	4,95 +/- 0,70	7,53	2,43	4,36 +/- 1,88	15,00	0,46	0,79 +/- 0,19	1,36	

Nous avons tracé les distributions de chaque paramètre avec les trois codes 01, 1234 et amas pour une taille de fenêtre 17 (cf. Figure 31, 32, 33, 34 et 35 pages suivantes). Toutes les distributions sont semblables, hormis celles se rapportant aux classes F et G, en utilisant les codes amas et 01. Les plages d'amas et d'hydrophobes permettent en effet d'obtenir une information de texture différente, de nature hydrophobe, sur la fenêtre considérée selon les classes étudiées.

- Paramètre SRE

Le paramètre SRE (ou « poids » des plages courtes) permet de mettre en évidence l'abondance de plages de petites tailles (Figure 31).

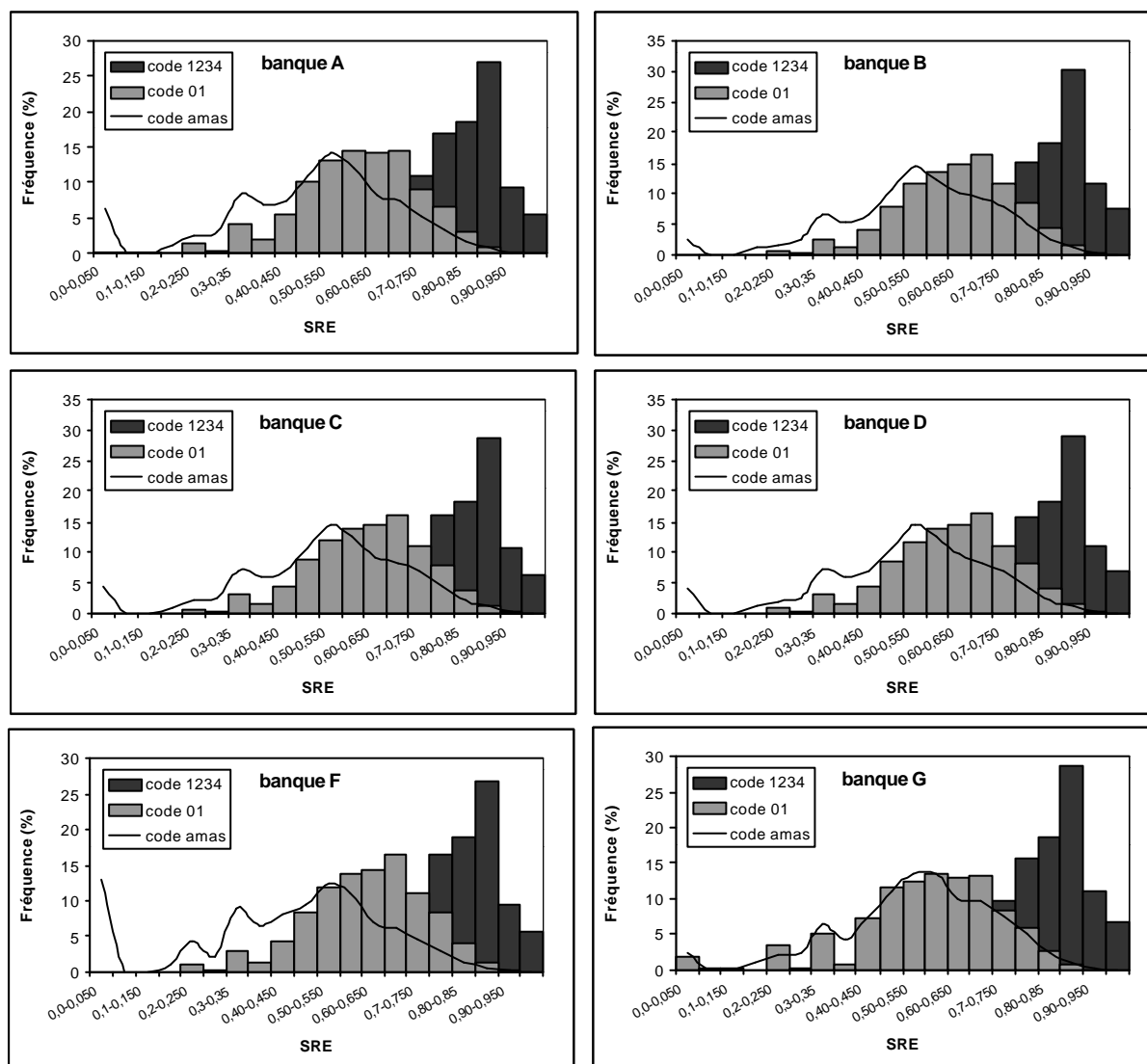


Figure 31 : Distribution de SRE dans chaque classe en utilisant une fenêtre de 17 acides aminés.

Toutes les distributions calculées avec le code 1234 et le code 01 semblent similaires. Par contre, les distributions calculées avec le code amas montrent quelques différences. Les banques A, B, C et D semblent identiques dans l'ensemble et présentent deux pics (un premier pour une valeur de SRE de 0,3 comprenant presque 15% des effectifs et un second pour une valeur de 0,5 comprenant 7% des effectifs). La distribution F présente également deux pics situés aux mêmes valeurs de SRE que précédemment (l'intensité du premier pic est cependant plus forte (environ 9% des effectifs) et l'intensité du second plus faible (environ 12% des effectifs)) ainsi qu'un troisième pic (comprenant environ 5% des

effectifs). La courbe F correspond aux protéines membranaires et de surface. Ces protéines contiennent peu d'amas hydrophobes de petites tailles. Il est cohérent d'avoir des valeurs plus faibles de SRE pour cette banque. A l'inverse la banque G qui correspond aux protéines de petites tailles, renferme des amas de petite taille et présente des valeurs SRE plus élevées (le second pic est plus élargi que dans les autres banques et déplacé vers la droite).

Nous constatons que globalement dans les classes A, B, C et D, il n'y a pas de différence au niveau du nombre d'amas de petites tailles.

- Paramètre LRE

Le paramètre LRE rend compte, à l'inverse du SRE, du « poids » des pages longues. Nous obtenons un résultat en accord avec celui obtenu précédemment (Figure 32).

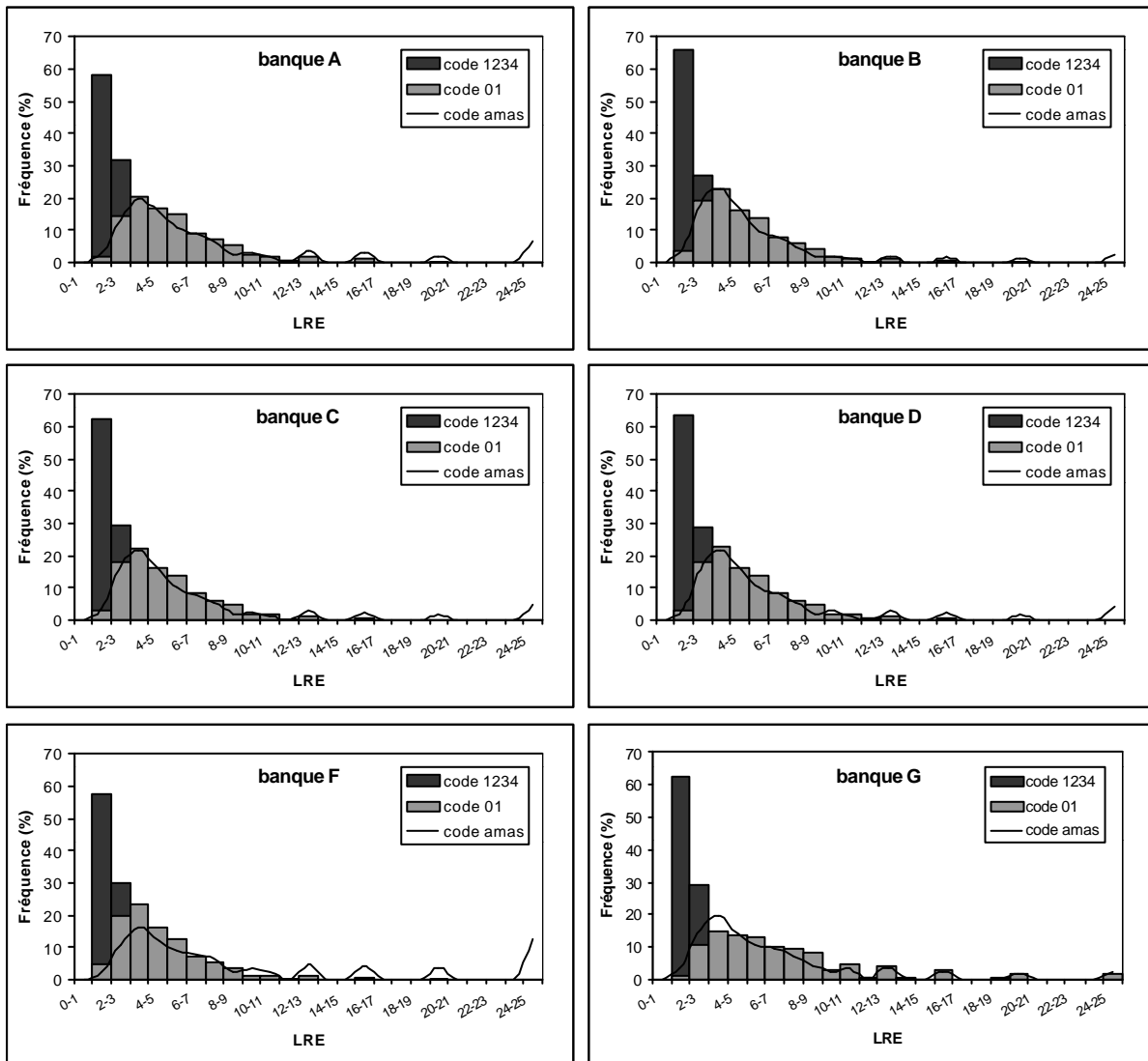


Figure 32 : Distribution de LRE dans chaque banque avec la fenêtre de 17 acides aminés.

Toutes les distributions sont semblables hormis, en utilisant le code amas et le code 01, pour lesquels les classes F et G se distinguent légèrement. De façon surprenante, avec le code 01, la banque G (petites protéines) présente dans l'ensemble des valeurs de LRE légèrement plus élevées que dans les autres banques. Avec le code amas par contre et logiquement, un nombre important de plages d'amas de grandes tailles peut être mis en évidence dans la classe F (protéines membranaires).

- Paramètre GLD

Le paramètre GLD correspond à la distribution des niveaux d'hydrophobie de la région étudiée. Il permet de mettre en évidence une uniformité de la distribution des plages. Comme précédemment, les distributions A, B, C et D sont similaires (Figure 33).

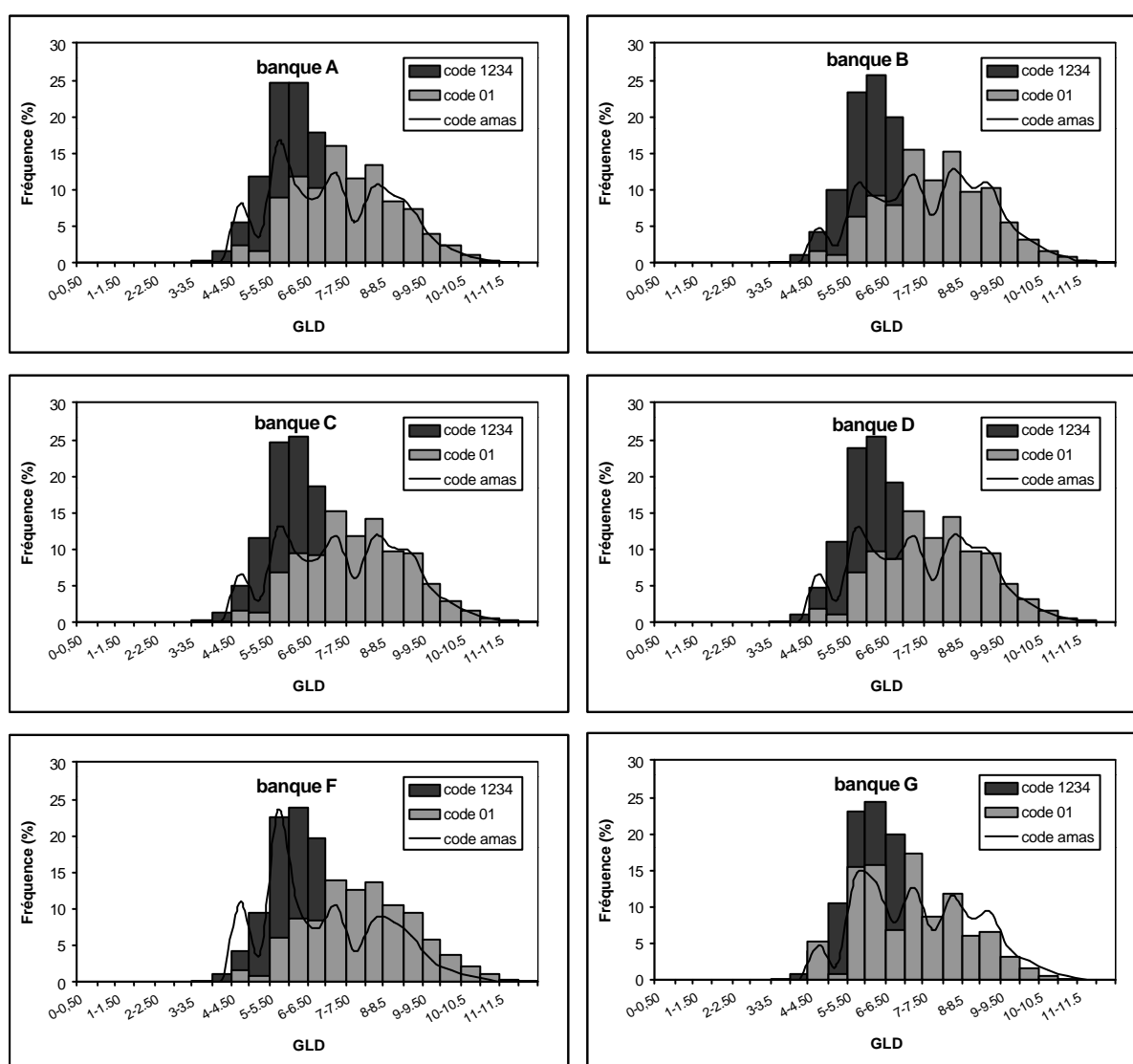
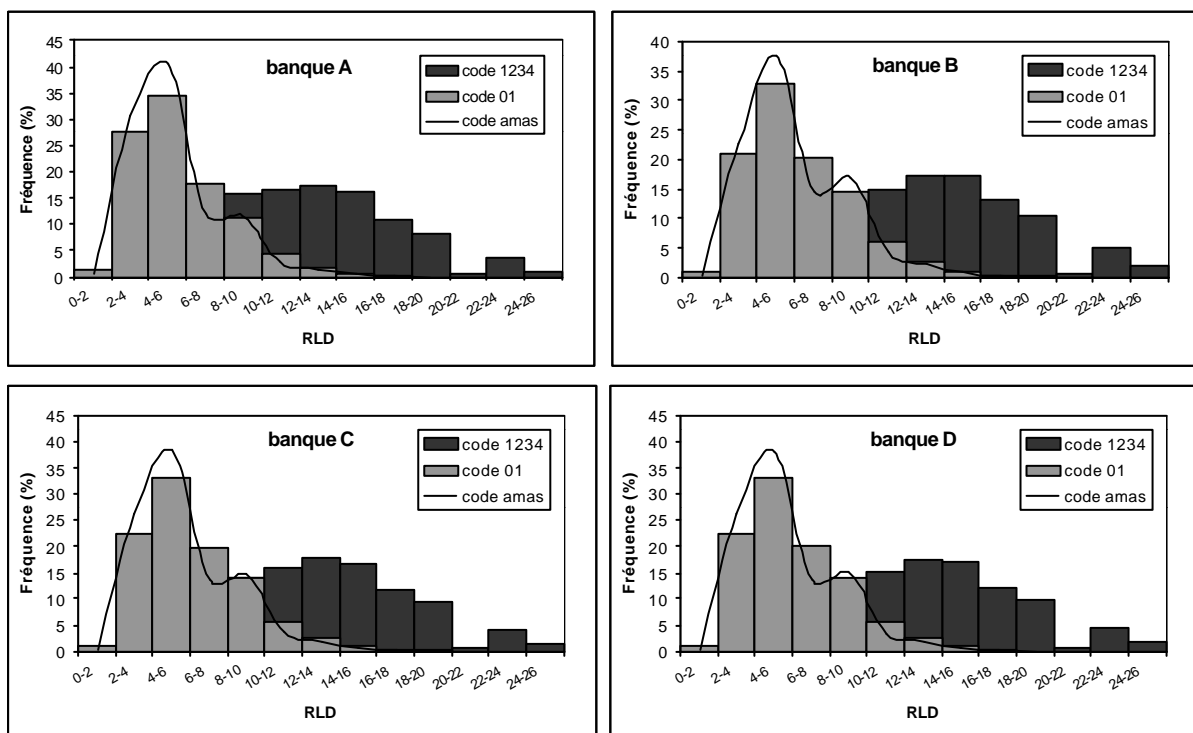


Figure 33 : Distribution de GLD dans chaque banque avec la fenêtre de 17 acides aminés.

Seules diffèrent les distributions F et G, en utilisant le code amas. Nous constatons de nombreux pics dans la distribution F avec un pic en 5,50 qui présente environ 25% des effectifs, ce qui nous indique que près d'un quart des fenêtres issues de la banque F ont des valeurs faibles de GLD. Ces valeurs faibles de GLD montrent la non uniformité de la distribution des « niveaux de gris ». Dans le cas du code amas, deux niveaux d'« amas » sont retrouvés (acide aminé inclus dans un amas et acide aminé non inclus dans un amas). Il semble que la banque F présente une répartition non uniforme des plages d'amas dans ces fenêtres. La banque F aux protéines membranaires et de surface. Ces protéines contiennent de nombreux amas hydrophobes de grande taille. Cependant, les passages membranaires riches en acides aminés hydrophobes sont souvent connectés par des boucles beaucoup moins riches en acides aminés hydrophobes dans le cas de passages membranaires multiples ou sont isolés dans la séquence. Cela pourrait expliquer cette distribution hétérogène des plages. A l'inverse, la courbe G, correspondant aux petites protéines, semble avoir des valeurs GLD plus élevées et présente dans l'ensemble une uniformité dans la distribution des plages d'amas.

- Paramètre RLD

Le paramètre RLD met en évidence le nombre de plages de même longueur. Le paramètre RLD augmente si le nombre de plages de même longueur diminue. La distribution avec le code amas montre une valeur modale de 5 caractérisée par une amplitude très forte (près de 40% des fenêtres) (Figure 34).



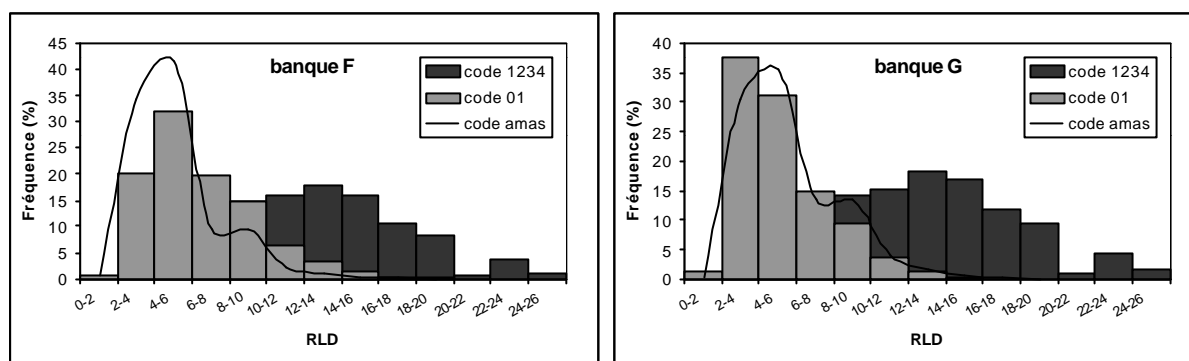


Figure 34 : Distribution de RLD dans chaque classe en utilisant une fenêtre de 17 acides aminés.

Ainsi, avec le code amas, près de la moitié des fenêtres de nos banques A, B, C, D, F et G présentent un RLD faible montrant une homogénéité des tailles des plages d'amas au sein des fenêtres. Entre 10 et 15% des fenêtres présentent un RLD beaucoup plus élevé (entre 8 et 10) ce qui indique la présence de plages de taille différente. La banque F (protéines membranaires et de surface) se caractérise par un pic moins élevé au niveau des valeurs 8 et 10 de RLD et par une décroissance de ce pic moins marquée que dans les autres banques. Ces protéines contiennent de nombreux amas hydrophobes de grandes tailles et peu d'amas hydrophobes de petites tailles. Il semble ainsi cohérent de trouver un plus grand nombre de plages de longueur similaire (Figure 34).

L'utilisation du code 01 met en évidence non pas des plages d'amas, mais des plages hydrophobes. Les distributions dans les classes A, B, C, D et F semblent similaires. La distribution de la banque G (petites protéines) présente un pic contenant près de 38% des effectifs pour des valeurs moyennes de RLD aux environs de 3, ce qui montre que la distribution des tailles des plages d'acides aminés hydrophobes tend à être uniforme dans les fenêtres.

- Paramètre RLP

Le paramètre RLP correspond au pourcentage de plages (dans notre cas, avec le code amas, RLP indique le pourcentage de plages d'amas). Les distributions calculées avec le code 01 et le code amas sont similaires (les acides aminés hydrophobes étant principalement retrouvés dans les amas). La courbe correspondant au code 1234 semble avoir une distribution différente selon la classe étudiée (Figure 35).

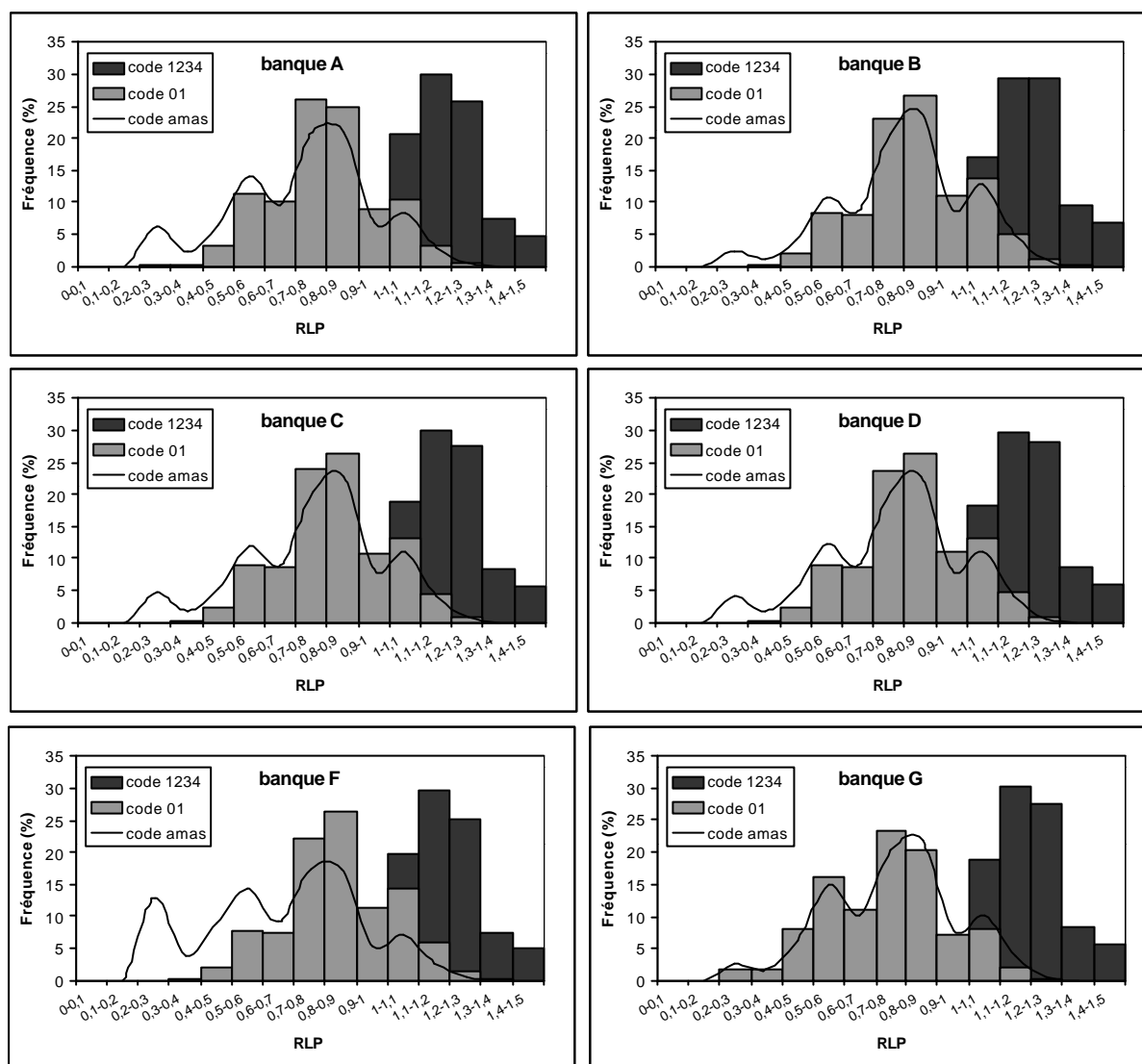


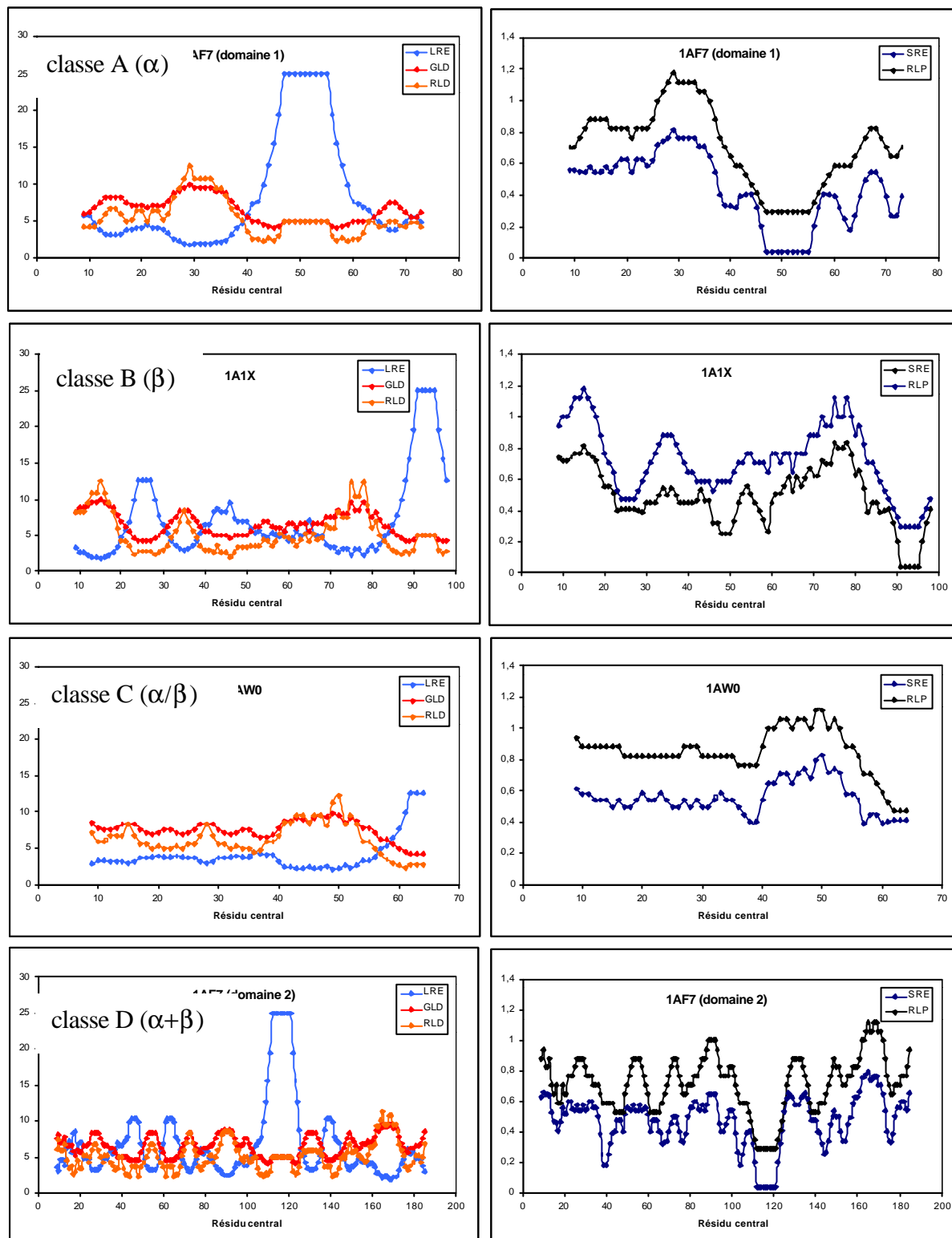
Figure 35 : Distribution de RLP dans chaque classe en utilisant une fenêtre de 17 acides aminés.

Dans cette étude, le code amas est particulièrement intéressant. En effet, appliqué au calcul de RLP, il fait ressortir le nombre d'acides aminés inclus dans les amas. Les plages dénombrées sont alors les plages d'amas. La distribution du RLP est une courbe sinusoïdale qui semble montrer différents types de texture d'amas dans les fenêtres. Quatre densités de plages d'amas sont identifiées parmi les fenêtres (RLP valant respectivement 0,25 ; 0,55 ; 0,75 et 1,05), allant d'un faible nombre d'acides aminés de la fenêtre de calcul inclus dans des amas à une grande majorité des acides aminés constituant des amas. Nous remarquons que dans les banques A, B, C, D et G, 40 à 50% des fenêtres ont une valeur de RLP compris entre 0,7 et 0,9. Un RLP élevé indique une texture assez hétérogène. Il semble donc que l'ensemble des fenêtres de nos banques contient une texture hétérogène en amas et à peu près équivalente. Une fraction plus petite (15% des fenêtres) a des valeurs de RLP plus faibles avoisinant 0,5-0,6. Dans la courbe F, une dernière fraction correspondant à des RLP de 0,2-

0,3 (5 à 10% des fenêtres) apparaît. Ce pic correspond sans doute à des régions renfermant des amas hydrophobes de grande taille, composant les passages membranaires.

3.3.2 Etude de profils des longueurs de plage

Nous avons tracé les profils des longueurs de plage pour nos huit fragments protéiques étudiés précédemment (Figure 36).



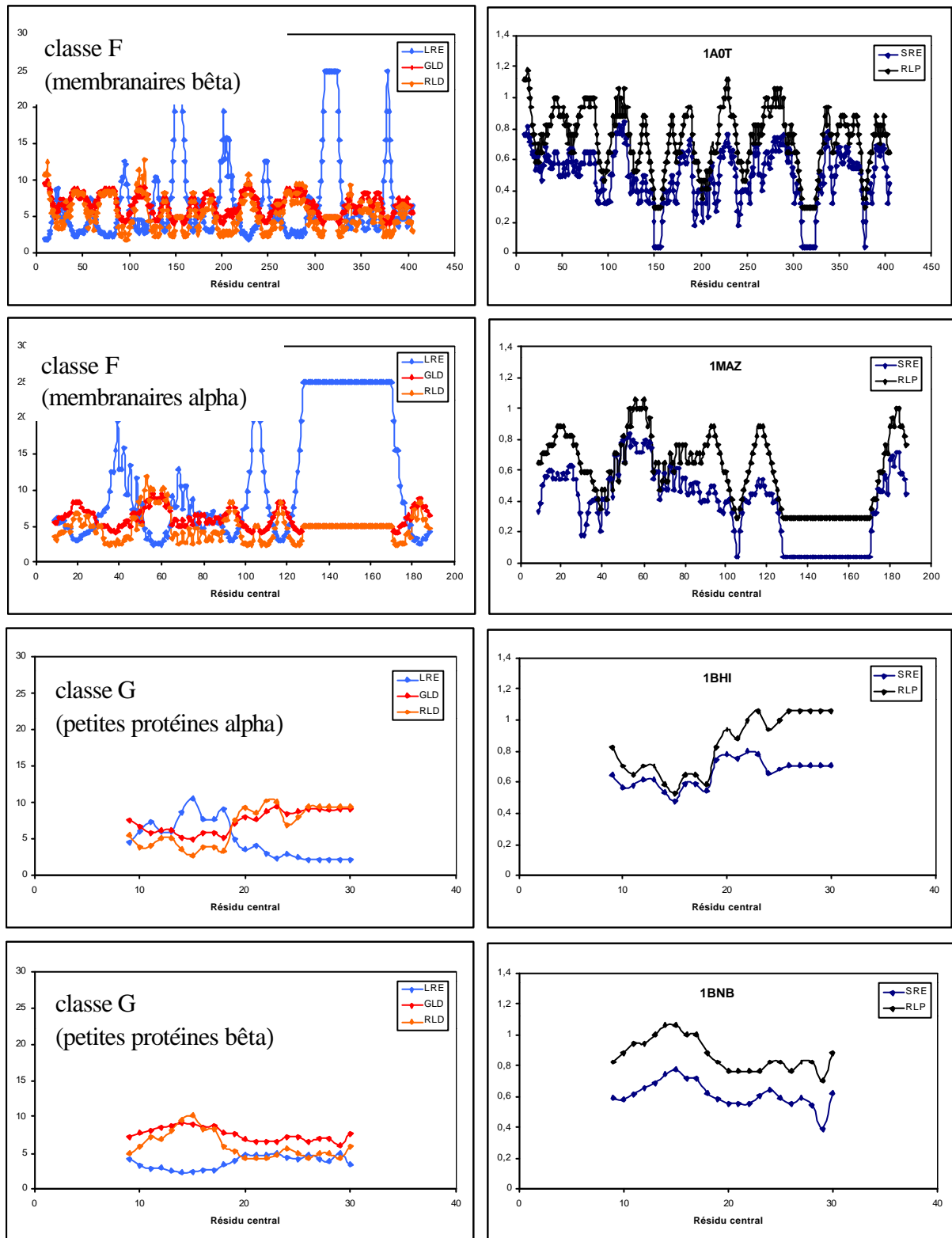


Figure 36 : Profils des longueurs de plage de protéines issues de nos banques.

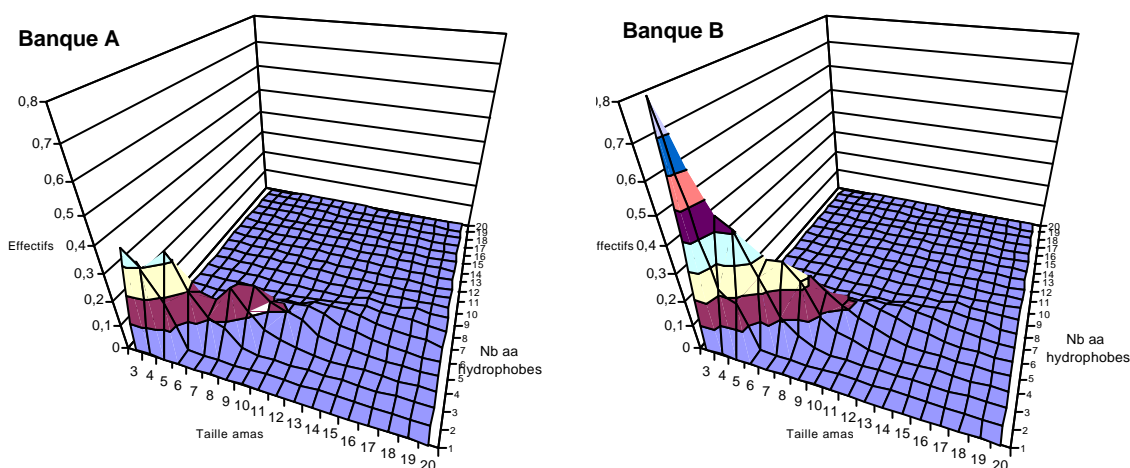
Nous avons séparé les cinq paramètres de longueur de plage en deux groupes : d'une part, SRE et RLP qui semblent fournir la même information de texture. Ils montrent l'importance des plages d'amas courtes et la présence dans ce cas de textures fines ; et d'autre

part LRE, GLD et RLD qui nous donnent une information sur l'uniformité de la répartition des plages et de l'importance des plages longues.

L'utilisation d'un code amas-non amas impose une évolution similaire des paramètres GLD et RLD dans une séquence. Comme dans les profils précédents (hydrophobie et cooccurrence), les différents paramètres ne varient que lors d'un changement de niveau d'hydrophobie très important dans la séquence. Il faut noter que le paramètre LRE présente des valeurs très élevées en présence de plages d'amas longues dans la séquence, ce qui coïncide avec la présence de régions de séquence riches en amas de grande taille (par exemple la région 120 à 180 de la protéine 1MAZ). Comme nous n'avons pas mis en évidence de caractéristiques particulières dans la répartition des amas dans chacune des banques A, B, C et D, qui présentent des paramètres moyens de texture identiques (cf. pourcentage moyen d'hydrophobie, contraste, entropie, SRE, LRE, GLD, RLD et RLP), nous avons souhaité vérifier la composition des amas les constituant. Cette étude n'a pas été faite sur les banques F et G étant donné que nous avons déjà mis en évidence une abondance d'amas de grande taille dans la banque F (protéines membranaires) et inversement une pauvreté en amas et plus spécifiquement en amas de grande taille dans la banque G (petites protéines).

3.4 Distribution des amas hydrophobes dans les banques A, B, C et D

Nous avons souhaité évaluer la répartition des tailles des amas dans les protéines présentant un repliement tout α , tout β , α/β et $\alpha+\beta$ et estimer dans quelle proportion les acides aminés hydrophobes (V, I, L, F, M, Y et W) sont répartis dans les différents amas hydrophobes. Nous avons ainsi tracé la distribution de la taille des amas hydrophobes en fonction du nombre d'acides aminés hydrophobes les composant au sein des protéines contenues dans nos classes A, B, C et D (Figure 37).



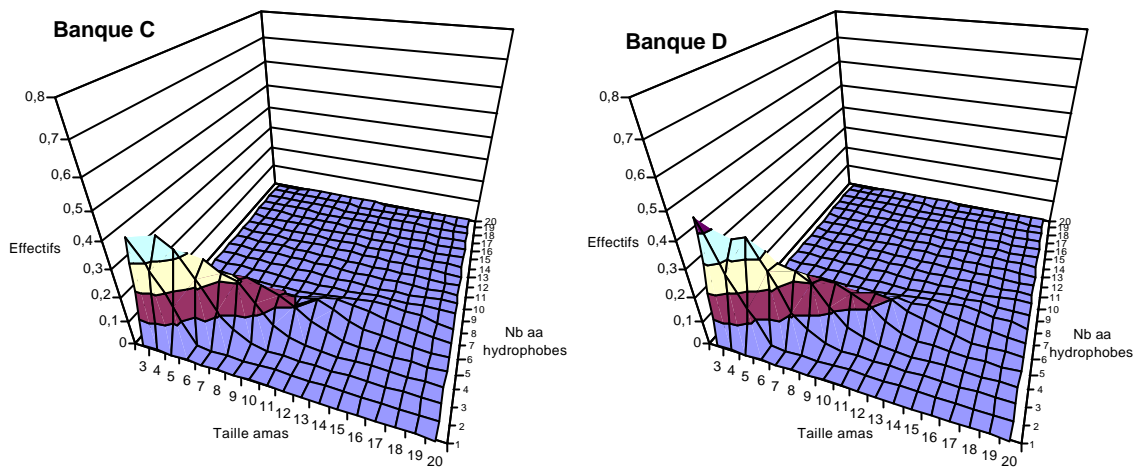


Figure 37 : Distribution du nombre d'acides aminés hydrophobes contenus dans les amas en fonction de leur longueur.
Seules les amas hydrophobes ayant une taille comprise entre 3 et 20 acides aminés ont été considérés.

Les amas de taille 1 et 2 (1 et 11), très répandus dans les séquences protéiques, ne sont généralement pas associés à des structures secondaires régulières [HENNETIN, J. 2003] (ils ne sont donc pas représentés dans la distribution). La distribution de la Figure 37 est limitée aux amas de 20 acides aminés. Bien que moins nombreux, les amas de grandes tailles sont présents dans les domaines globulaires dans les classes A, B, C et D de SCOP, mais correspondent souvent à des amas associés à plusieurs structures secondaires régulières dits amas multiples.

Pour ces quatre courbes, le nombre d'acides aminés hydrophobes constituant chaque taille d'amas semblent constant et avoisine le tiers. Le rapport taille/nombre d'hydrophobes est linéaire. Les banques C et D présentent des distributions similaires avec une abondance d'amas de taille 4 à 6 et des amas de grande taille beaucoup moins nombreux. La banque B contient beaucoup d'amas hydrophobes de petite taille et très peu d'amas de grande taille à l'inverse de la banque A qui présente comme attendu un grand nombre d'amas de grande taille. Les amas hydrophobes de ces quatre classes contiennent au minimum un tiers d'acides aminés hydrophobes (V, I, L, F, M, Y et W). Plus ils sont longs, plus ils contiennent des acides aminés hydrophobes.

4 Discussion et Conclusion

En adaptant certaines méthodes d'analyse d'image au tracé bidimensionnel HCA, nous avons souhaité pouvoir « quantifier » et caractériser des régions de texture différente dans les séquences protéiques. Pour calibrer nos paramètres (pourcentage d'acides aminés hydrophobe, contraste, entropie, SRE, LRE, GLD, RLD et RLP) et observer leur variation,

nous avons utilisé six banques de repliements protéiques issues de SCOP (A, B, C, D, F, et G), deux tailles de fenêtres de travail (17 et 11 acides aminés), plusieurs codes (01, 1234 et amas). Nous avons tracé les distributions de ces paramètres pour chacune des banques. Les résultats obtenus lors des comparaisons de moyennes deux à deux sur des banques identiques évaluées avec des tailles de fenêtre différentes sont très variables. Par exemple, dans le cas de calcul d'entropie, de LRE, GLD, RLD et RLP, la taille de la fenêtre semble jouer un rôle sur les valeurs moyennes calculées alors que dans le calcul de contraste ou de SRE, les tailles de fenêtres 17 et 11 nous donnent des résultats voisins. Ces données nous indiquent que le signal de texture extrait des paramètres ne ressort pas clairement en utilisant l'une ou l'autre fenêtre. De plus, pour les calculs utilisant la fenêtre de taille 11, les écarts-types sont plus élevés et les valeurs sont plus dispersées par rapport à la valeur moyenne de la banque. Dans ce contexte, pour l'analyse des paramètres de texture, nous avons choisi par la suite d'utiliser la fenêtre 17. Pour le calcul des paramètres de la méthode des longueurs de plage avec le codage amas-non amas, nous avons choisi de travailler uniquement sur une fenêtre de 17 car 11 correspond à une fenêtre trop petite pour rendre compte de la distance entre amas.

L'utilisation des codes 01 et 1234 ne s'est pas révélée informative. En effet, le code 01 utilisé dans les paramètres entropie, contraste, et dans les cinq paramètres de la méthode des longueurs de plage nous fournit une information sur la répartition des acides aminés hydrophobes dans la fenêtre au travers de calculs assez complexes. Or, cette information est directement obtenue en calculant le profil d'hydrophobie de la protéine sur une fenêtre de 17 acides aminés. Le code amas, utilisé uniquement dans la méthode des longueurs de plage, est particulièrement intéressant car il rend compte de la taille et de la répartition uniforme des amas hydrophobes dans la séquence. Dans le cas du calcul du LRE, cette propriété nous a permis de mettre en évidence la présence de nombreux amas de grande taille dans les protéines membranaires (banque F) et inversement la pauvreté en amas des protéines de petites tailles (banque G), comme attendu.

Dans l'ensemble, les valeurs moyennes de chacun des paramètres calculés dans les six banques ne diffèrent pas sauf pour les banques F et G qui possèdent des compositions bien particulières. La banque F contient des protéines membranaires et de surface et la banque G des protéines de petites tailles. Les banques A, B, C et D correspondent aux domaines globulaires des protéines et présentent des variations identiques des paramètres de texture. Il n'est donc pas possible de différencier les repliements alpha des repliements bêta à partir de ces valeurs moyennes. Néanmoins, l'observation de « profils » d'hydrophobie, de cooccurrence et de longueur a permis de mettre en évidence la sensibilité des paramètres en

fonction de la texture en amas de la séquence. Ainsi, des régions pauvres en amas sont souvent corrélées à des valeurs extrêmes des paramètres de texture que nous avons définis, alors que des régions présentant une répartition uniforme d'amas hydrophobes n'entraînent que peu de variations de ces paramètres.

Les informations de texture que nous avons extraites à partir de ces paramètres mettent en évidence la répartition uniforme des amas dans les séquences. Cependant, nous avons observé qu'avec le seul profil d'hydrophobie, nous obtenions déjà cette information. De plus, sur ce profil, nous pouvons identifier les limites des régions homogènes en amas des régions présentant peu d'amas hydrophobes (Figure 38).

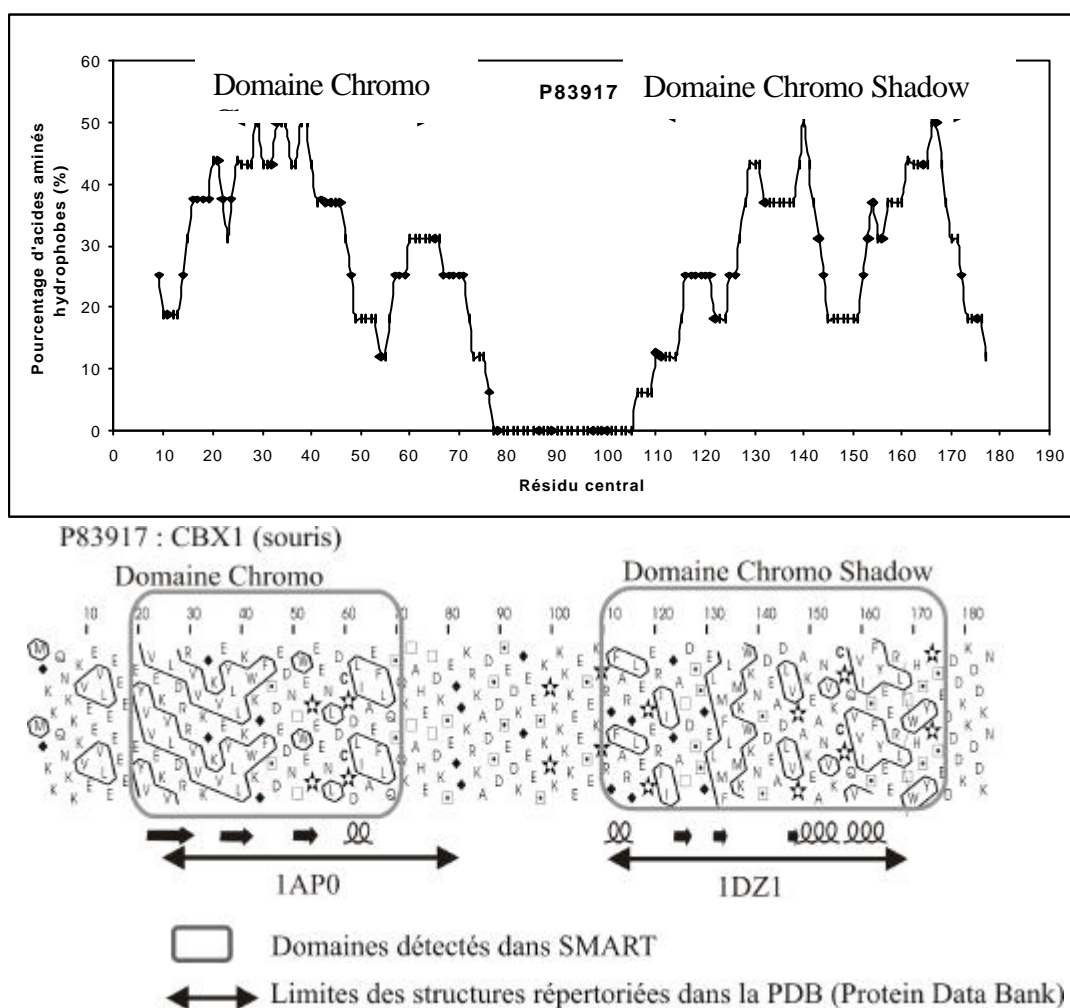


Figure 38 : Corrélation entre profil hydrophobe et représentation HCA.

Une analyse visuelle du tracé HCA permet de délimiter facilement, les domaines globulaires structurés comportant des acides aminés regroupés en amas. Les régions non structurées ou peu structurées contiennent moins d'acides aminés hydrophobes, constituant souvent de petits amas. L'analyse du profil hydrophobe montre que ces régions non

structurées ou peu structurées correspondent à des diminutions très fortes du pourcentage d'acides aminés hydrophobes (voir pour exemple positions 70 à 110 de la protéine HP1 de souris : P83917, Figure 38). Ainsi, la protéine HP1 peut être divisée en deux domaines globulaires distincts. Ceux-ci correspondent respectivement aux domaines CHROMO et CHROMO SHADOW [AASLAND, R. et al., 1995; YE, Q. et al., 1997].

En conclusion de cette approche, nous avons constaté l'inadéquation des méthodes d'analyse de texture classiquement utilisées en imagerie pour extraire des informations différenciées et pertinentes des séquences protéiques transposées sous le format HCA. En effet, ces méthodes utilisent souvent de nombreuses images qui renferment des millions de pixels pour extraire des paramètres significatifs. Notre image HCA composée de pixels « acides aminés » est de loin beaucoup moins riche et ne contient que quelques dizaines à quelques centaines de pixels, nombre manifestement insuffisant pour transposer l'approche classique d'analyse de texture 2D à notre problématique.

Néanmoins, cette étude nous a conforté dans la recherche d'une procédure de délimitation des régions structurées des séquences, basée sur leur profil hydrophobe, ce que nous présentons ci-après.

Chapitre V

DomHCA : un outil pour prédire les régions structurées

1 **Introduction : Travaux précédents**

1.1 Prédiction de domaines structuraux

Les domaines sont considérés comme les unités de base de repliement, de fonction et d'évolution [HOLM, L. et al., 1994]. La connaissance de l'organisation en domaines d'une protéine est souvent le point de départ crucial pour la compréhension de sa structure et de sa fonction. La prédiction *in silico* des limites de domaines peut être réalisée en se focalisant soit sur les domaines eux-mêmes, soit sur des régions dites « linkers » entourant les domaines. L'étude et la caractérisation de familles de protéines ont permis de développer des méthodes de prédictions de domaines au moyen de comparaisons de séquences (e.g. DIVCLUS [PARK, J. et al., 1998], DOMAINATION [GEORGE, R. A. et al., 2002a], CHOP [LIU, J. et al., 2004],...) et de constituer des bases de données de domaines potentiels (ProDOM [CORPET, F. et al., 2000], DOMO [GRACY, J. et al., 1998], ...). Les méthodes proposées sont généralement efficaces quand les similarités de séquence sont élevées mais ne peuvent s'appliquer aux séquences orphelines. Ces séquences orphelines représentent néanmoins une part importante des séquences issues des génomes (entre 20 et 30% des cadres de lecture ouverts (ORFs) de la plupart des génomes séquencés [SIEW, N. et al., 2004] et jusqu'à 60% du nombre total de protéines prédites du génome de *Plasmodium falciparum* [GARDNER, M. J. et al., 2002]). Pour notamment surmonter cette difficulté, des approches *ab initio*, basées uniquement sur l'information de séquence, ont été développées. Les premières méthodes reposent sur des propriétés physicochimiques [BUSETTA, B. et al., 1984; VONDERVISZT, F. et al., 1986; KIKUCHI, T. et al., 1988]. Des méthodes plus récentes ont pour bases des analyses statistiques des longueurs des domaines (Domain Guess by Size [WHEELAN, S. J. et al., 2000]), des calculs d'hydrophobie et de charge (FoldIndex, [UVERSKY, V. N. et al., 2000]), des simulations *ab initio* de structures tridimensionnelles (SnapDragon, [GEORGE, R. A. et al., 2002b]), la comparaison des structures secondaires prédites avec celles observées dans des chaînes ayant des domaines dont les limites sont connues (DomSSEA, [MARSDEN, R. L. et al., 2002]), des corrélations entre l'entropie de la chaîne latérale et les interactions

fortes entre résidus [GALZITSKAYA, O. V. et al., 2003], les propensions des acides aminés à être dans un état ordonné ou dans un état désordonné (GlobPlot, [LINDING, R. et al., 2003b]) ou encore des réseaux de neurones combinant des informations d'évolution, de composition en acides aminés, de prédiction de structures secondaires et d'accessibilité au solvant (CHOPnet, [LIU, J. et al., 2004]).

Un autre moyen pour délimiter les domaines d'une séquence consiste à caractériser les régions de « linkers », qui sont le plus souvent situées entre domaines structurés.

1.2 Prédiction de « linkers » et de régions non structurées

Des analyses statistiques de la composition en acides aminés de « linkers » ont été réalisées [GEORGE, R. A. et al., 2003; SUYAMA, M. et al., 2003; TANAKA, T. et al., 2003] permettant, dans certains cas, le développement de méthodes de prédictions de linkers comme DomCut [SUYAMA, M. et al., 2003]. Les régions désordonnées ou non structurées peuvent aussi être considérées comme des séquences linkers, mais se distinguent des linkers courts car elles sont généralement conservées au cours de l'évolution et adoptent parfois des structures secondaires régulières en se liant à d'autres molécules [DYSON, H. J. et al., 2005; FINK, A. L. 2005]. Plusieurs méthodes peuvent être utilisées pour prédire les régions désordonnées. Certaines sont basées sur la détection de régions de faible complexité, c'est à dire renfermant des acides aminés souvent répétés (pas forcément de manière régulière), essentiellement proline, glutamine, sérine et thréonine (SEG [WOOTTON, J. C. 1994] ou CAST [PROMPONAS, V. J. et al., 2000]) ou de régions où peu ou pas de structures secondaires régulières sont prédites (NORSp [LIU, J. et al., 2003]). Il faut noter cependant que les régions de faible complexité ne sont pas forcément des régions désordonnées et vice versa [DUNKER, A. K. et al., 2002]. Des méthodes plus spécifiques ont été développées. Elles sont principalement basées sur des caractéristiques discriminant les régions désordonnées des domaines globulaires ([BRACKEN, C. et al., 2004], FoldIndex [UVERSKY, V. N. et al., 2000], GlobPlot [LINDING, R. et al., 2003b], PONDR [ROMERO, P. et al., 2001; VUCETIC, S. et al., 2003], DisEMBL [LINDING, R. et al., 2003a], DISOPRED ou DISOPRED2 [JONES, D. T. et al., 2003; WARD, J. J. et al., 2004a; WARD, J. J. et al., 2004b]).

Récemment, une nouvelle méthode de détection de linkers a été proposée. Elle repose sur la comparaison d'un jeu de données de séquences structurées et de séquences non structurées, et sur la distance entre deux amas hydrophobes (tels que définis par la méthode

HCA). Cette méthode nommée PreLink, permet de calculer la probabilité d'une séquence à être non structurée et de prédire certains linkers [COEYTAUX, K. et al., 2005].

1.3 Conclusion et but de l'étude

Les méthodes présentées ci-dessus utilisent soit des alignements multiples et sont alors tributaires de la connaissance d'autres membres de la famille protéique étudiée, soit des jeux de données (apprentissage) et sont efficaces sur des protéines «standards». Cependant, elles obtiennent souvent de mauvais résultats sur des séquences présentant des textures particulières et n'ayant pas ou peu d'identité de séquences avec une protéine connue.

La méthode Hydrophobic Cluster Analysis (HCA), que nous avons précédemment présentée, repose sur les principes physico-chimiques et topologiques gouvernant le repliement des domaines globulaires [GABORIAUD, C. et al., 1987; CALLEBAUT, I. et al., 1997a]. Elle permet directement, à partir d'une seule séquence, l'accès à la texture et au découpage en domaines (voir pour exemple [MORNON, J. P. et al., 2002a; MORNON, J. P. et al., 2002b]). En effet, les régions structurées contiennent typiquement des amas hydrophobes de longueur proche de celles des structures secondaires régulières alors que les régions non structurées sont appauvries en amas hydrophobes ou ne renferment que des amas de petites tailles. L'utilisation d'une représentation en deux dimensions de la séquence, permet beaucoup mieux qu'à l'aide d'une séquence linéaire de souligner les changements de texture dans les séquences protéiques et se révèle être un outil efficace pour identifier les régions structurées.

La méthode HCA, dans son ensemble, a démontré son efficacité dans l'identification des limites de régions structurées, permettant notamment l'identification de nouveaux domaines (e.g. [CALLEBAUT, I. et al., 1997b; CALLEBAUT, I. et al., 2001]), mais elle reste fortement dépendante de l'expertise humaine, car elle n'est pas aisément automatisable. Le but de notre étude a été dans ce contexte, de développer un outil spécifique, DomHCA, permettant le prédécoupage automatique d'une séquence protéique en domaines structurés et non structurés.

2 *Fondement de la procédure DomHCA et caractérisation des domaines globulaires*

Nous avons calibré la méthode DomHCA sur un jeu de domaines globulaires issus de données connues (classe ABCDE de SCOP, 7530 séquences, cf. chapitre IV) avec des

paramètres définis, issus de la pratique HCA. Nous avons examiné la représentativité des acides aminés dans les domaines globulaires de cette banque.

2.1 Distribution des acides aminés hydrophobes V, I, L, F, M, Y et W

Comme référence, nous avons calculé la distribution du pourcentage d'acides aminés (V, I, L, F, M, Y et W) dans les domaines globulaires de notre banque ABCDE de SCOP [MURZIN, A. G. et al., 1995].

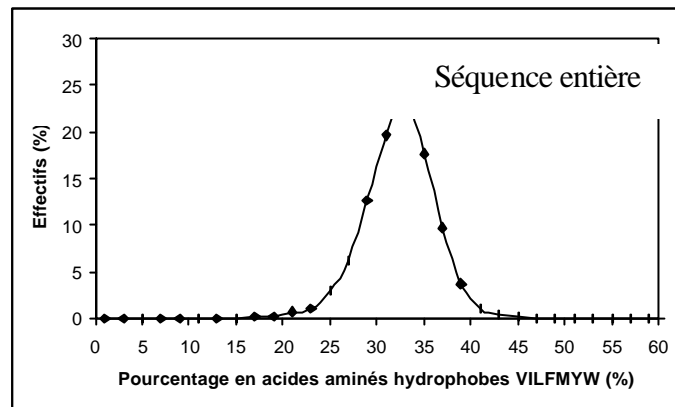


Figure 39 : Distribution du pourcentage d'acides aminés hydrophobes (VILFMYW) sur un ensemble de séquences issues de SCOP (classe ABCDE, 7530 séquences). Un intervalle de 2% a été utilisé pour le calcul du pourcentage d'acides aminés hydrophobes (VILFMYW) sur la protéine entière.

Nous avons observé que les domaines globulaires sont caractérisés par un pourcentage moyen d'acides aminés hydrophobes de 32,32 % (Figure 39). Ceux-ci sont groupés en amas hydrophobes (selon la définition HCA), qui correspondent majoritairement aux faces internes des structures secondaires régulières [WOODCOCK, S. et al., 1992; HENNETIN, J. et al., 2003]. Seuls les plus petits amas (codés 1 et 11, un ou deux acides aminés hydrophobes contigus) sont indifféremment associés à des structures secondaires régulières ou à des régions de boucles.

Inspirés par les méthodes d'analyse de texture, nous avons utilisé une fenêtre de 17 acides aminés glissant le long de la séquence pour calculer la distribution des fréquences des acides aminés hydrophobes (V, I, L, F, M, Y, W) sur notre jeu de domaines globulaires (Figure 40). La valeur correspondante est attribuée au résidu central de la fenêtre. Le pourcentage moyen d'hydrophobie observé sur les fenêtres de 17 acides aminés est de 32.56%. Aucune fenêtre ne présente un pourcentage moyen d'hydrophobie supérieur à 80.

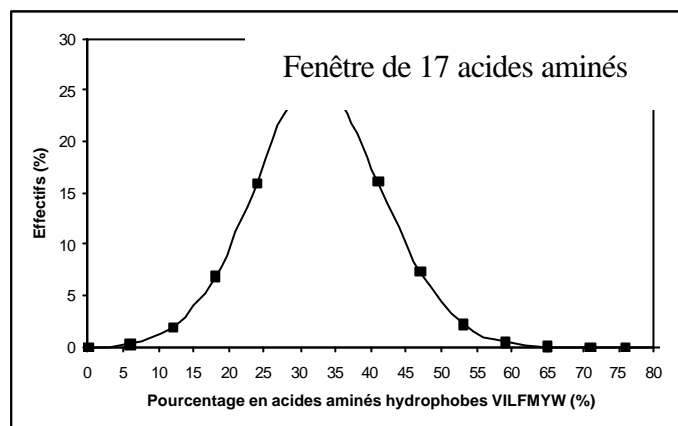


Figure 40 : Distribution du pourcentage d'acides aminés hydrophobes (VILFMYW sur un ensemble de séquences issues de SCOP (classe ABCDE; 7530 séquences soit 2 472 934 fenêtres).

Un intervalle de 6% a été utilisé pour le calcul de ce pourcentage. Les différentes valeurs ainsi étudiées à intervalles de 6%, correspondent au nombre de résidus hydrophobes possibles dans une fenêtre (1/17 6%, 2/17 12%, etc...).

Nous avons vérifié que l'utilisation d'une fenêtre de 17 acides aminés pour calculer le pourcentage d'acides aminés hydrophobes nous donne un pourcentage moyen d'hydrophobie similaire. Le découpage à l'aide de fenêtre de 17 acides aminés ne tient pas compte des huit premiers et derniers acides aminés des domaines considérés dans le calcul du pourcentage d'hydrophobie et conduit à l'obtention d'une moyenne d'hydrophobie légèrement plus élevée. Cette idée de fenêtre glissante, utilisée pour l'analyse de texture précédente (chapitre 4), peut être également exploitée dans l'optique d'une détection de domaines globulaires au sein des séquences protéiques. En effet, elle permet de rendre compte de l'évolution du pourcentage local d'acides aminés hydrophobes de la séquence protéique, contrairement à une valeur d'hydrophobie globale calculée sur la séquence entière.

2.2 Distribution des amas hydrophobes

Nous avons constaté que les domaines globulaires renferment environ 33% d'acides aminés V, I, L, F, M, Y, W, retrouvés dans les amas hydrophobes. Nous avons souhaité évaluer dans quelle proportion ces acides aminés hydrophobes sont répartis dans les différents amas hydrophobes. Nous avons tracé précédemment la distribution de la taille des amas hydrophobes en fonction du nombre d'acides aminés hydrophobes les composant sur nos quatre classes A, B C et D (cf. chapitre 3, pages 68 et 69). Nous avons conclu que dans l'ensemble des classes A, B, C et D, les amas sont bâtis sur une balance en acides aminés hydrophobes, « ni trop ni trop peu » et renferment généralement un tiers d'acides aminés

hydrophobes. Nous avons vérifié que le même résultat était retrouvé dans la classe E, correspondant aux protéines multidomaines (distribution non montrée).

2.3 Distribution des tailles des domaines

Nous nous sommes plus particulièrement intéressés à la taille des domaines globulaires des séquences protéiques lorsqu'ils sont seuls ou multiples dans une séquence. Comme décrit précédemment, les classes A, B, C, D et E de SCOP contiennent des fragments de séquences correspondant à différents types de repliements. Nous avons utilisé également la banque CATH (version 2.5.1 de janvier 2004) [ORENGO, C. A. et al., 1997]), qui, issue de la PDB et très similaire à SCOP, nous fournit une information supplémentaire sur la composition en domaines des séquences (domaine unique ou pluridomaine) et sur la nature de ces domaines (domaines continus ou discontinus). Nous avons ainsi collecté 2820 chaînes protéiques structurées à partir d'une version non redondante de la Protein Data Bank (PDB, 25% d'identité de séquence entre deux séquences au maximum, résolution maximale de 2,5 Angströms, longueur minimale de 50 acides aminés). Les séquences non référencées dans CATH ou renfermant des domaines discontinus ont été retirées de notre échantillon test pour constituer un effectif de 1403 séquences. Ces chaînes sont triées en séquences monodomaines et pluridomaines en utilisant la classification CATH : 1113 fichiers PDB correspondent à des « monodomaines » et 290 à des « multidomaines », nous employons des guillemets pour tenir compte du fait qu'un certain nombre de « monodomaines » ainsi répertoriés sont en fait, après une analyse plus minutieuse, constitués de plusieurs domaines distincts. Nous avons tracé la distribution des longueurs des protéines monodomaines et pluridomaines issues du traitement ci-dessus (Figure 41).

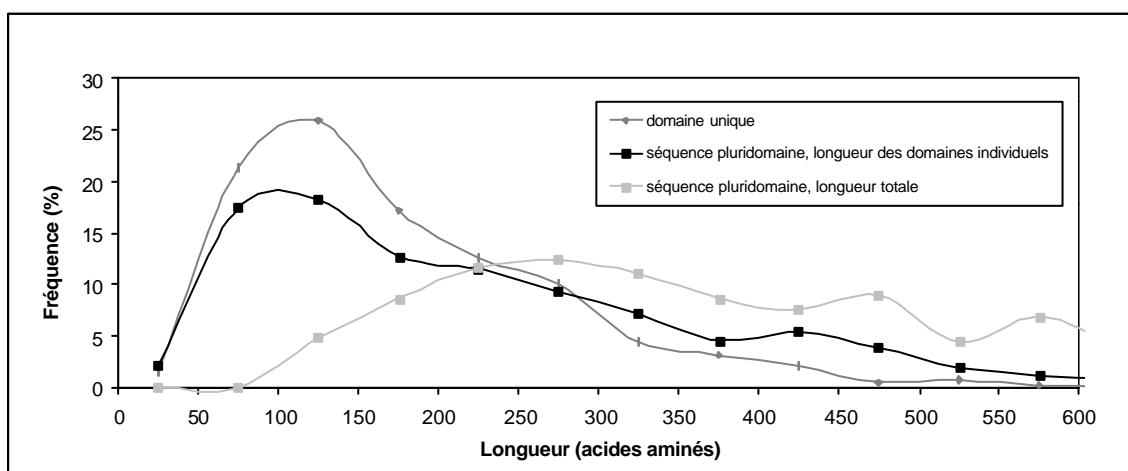


Figure 41 : Distribution des longueurs des domaines des séquences monodomaines et pluridomaines.

Les longueurs totales des chaînes « monodomaines » et « pluridomaines » diffèrent logiquement. La longueur moyenne des domaines individuels des protéines pluridomaines et celle des monodomaines sont identiques à celle observée dans la littérature (environ 100 à 150 acides aminés) [ISLAM, S.A. et al., 1995; SOWDHAMINI, R. et al., 1996; JONES, S. et al., 1998; WHEELAN, S. J. et al., 2000; MARSDEN, R. L. et al., 2002]. Il faut remarquer cependant que certaines chaînes dites « monodomaines » sont particulièrement longues. Celles-ci contiennent en réalité plusieurs domaines et ont été mal attribuées par la classification CATH. En conséquence, le taux de séquences « multidomaines » dans notre banque est vraisemblablement sous-estimé. Les chaînes « multidomaines » ne sont souvent pas découpées pour préserver l'intégrité fonctionnelle, spécialement lorsqu'elles forment des interfaces compactes comme dans la glucuronidase (Figure 42). L'utilisation de plusieurs méthodes combinées devrait améliorer la fiabilité des attributions [SAINI, H. K. et al., 2005].

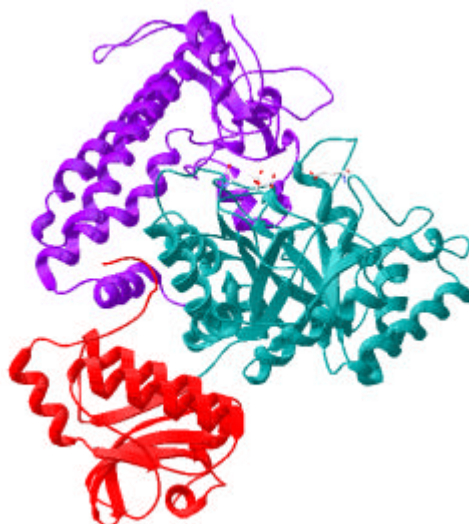


Figure 42 : Structure répertoriée dans CATH comme « monodomaine ».

Il s'agit de la cellulose α -glucuronidase de *Pseudomonas* (code PDB : 1GQI). Cette chaîne de 708 acides aminés comporte en réalité trois domaines (colorés en rouge, violet et bleu) [NURIZZO, D. et al., 2002].

Après avoir exploré plusieurs critères (pourcentage moyen d'acides aminés hydrophobes (V, I, L, F, M, Y et W) voisin de 33%, utilisation d'une fenêtre de 17 acides aminés) permettant de caractériser les domaines globulaires, nous avons développé une procédure de prédiction de régions structurées (DomHCA). Dans ce contexte, celle-ci sera essentiellement basée sur le profil hydrophobe calculé le long de la séquence protéique, associé à la distribution des amas hydrophobes.

3 Principe de l'algorithme pour détecter les régions structurées

Nous avons vu que les domaines globulaires des protéines contiennent environ un tiers d'acides aminés hydrophobes (Figure 39). Des pourcentages similaires sont également observés lorsque l'on parcourt ces domaines globulaires avec des fenêtres de taille moyenne (longueur 17, Figure 40). La longueur 17 a été choisie en accord avec les observations faites lors de l'analyse de texture précédemment réalisée. Nous avons tenté de définir les limites des domaines globulaires en balayant les séquences avec des fenêtres glissantes de 17 résidus, dans lesquelles le pourcentage d'acides aminés hydrophobes est calculé. En suivant ainsi l'évolution du pourcentage d'acides aminés hydrophobes le long de la séquence, il est possible de détecter des points de rupture et ainsi de proposer une première délimitation des régions dites « structurées ».

Nous considérons que si le pourcentage d'acides aminés hydrophobes d'un bloc est inférieur à 15% (cf. Figure 39), le résidu central du bloc devrait se trouver dans une région non structurée. A l'inverse, si le pourcentage d'acides aminés hydrophobes est supérieur à 30%, le résidu central serait dans une région structurée. Si le pourcentage d'acides aminés hydrophobes est compris entre 15% et 30%, le résidu central pourrait être dans une région structurée. Il est possible d'appuyer cette hypothèse par la présence dans ce segment d'amas hydrophobes, marqueurs de structures secondaires le plus souvent retrouvés dans les régions structurées.

Si le résidu central de notre fenêtre correspond à une position incluse dans un amas hydrophobe, celui-ci intègre la région structurée. Par contre, si le résidu central n'est pas inclus dans un amas, nous considérons son environnement direct. Ainsi, la distance entre les deux amas encadrant le résidu central est calculée. Si celle-ci est inférieure ou égale à 8 acides aminés, le résidu central est considéré comme appartenant à une région structurée. Cette valeur de 8 acides aminés correspond au nombre de résidus situés de part et d'autre du résidu central dans une fenêtre de 17 acides aminés. Les régions potentiellement structurées sont ensuite réunies, en considérant une distance de séparation maximale de 15 acides aminés entre les résidus centraux des fenêtres bordant deux régions différentes. Cette distance a été fixée après une analyse minutieuse des tracés HCA de plusieurs protéines de *Plasmodium falciparum*. De cette façon, les séquences protéiques sont grossièrement prédécoupées en régions structurées.

Deux types de régions structurées peuvent être définies en considérant leur profil hydrophobe et leur taille: celles qui sont typiquement des “régions structurées” et celles que nous appellerons “pseudo-structurées”, car elles sont généralement de plus petite taille (souvent entre 11 et 25 acides aminés) et présentent moins d’acides aminés hydrophobes (15 à 30 % d’acides aminés hydrophobes). La Figure 43 présente les principales étapes de la méthode de découpage en régions structurées par DomHCA.

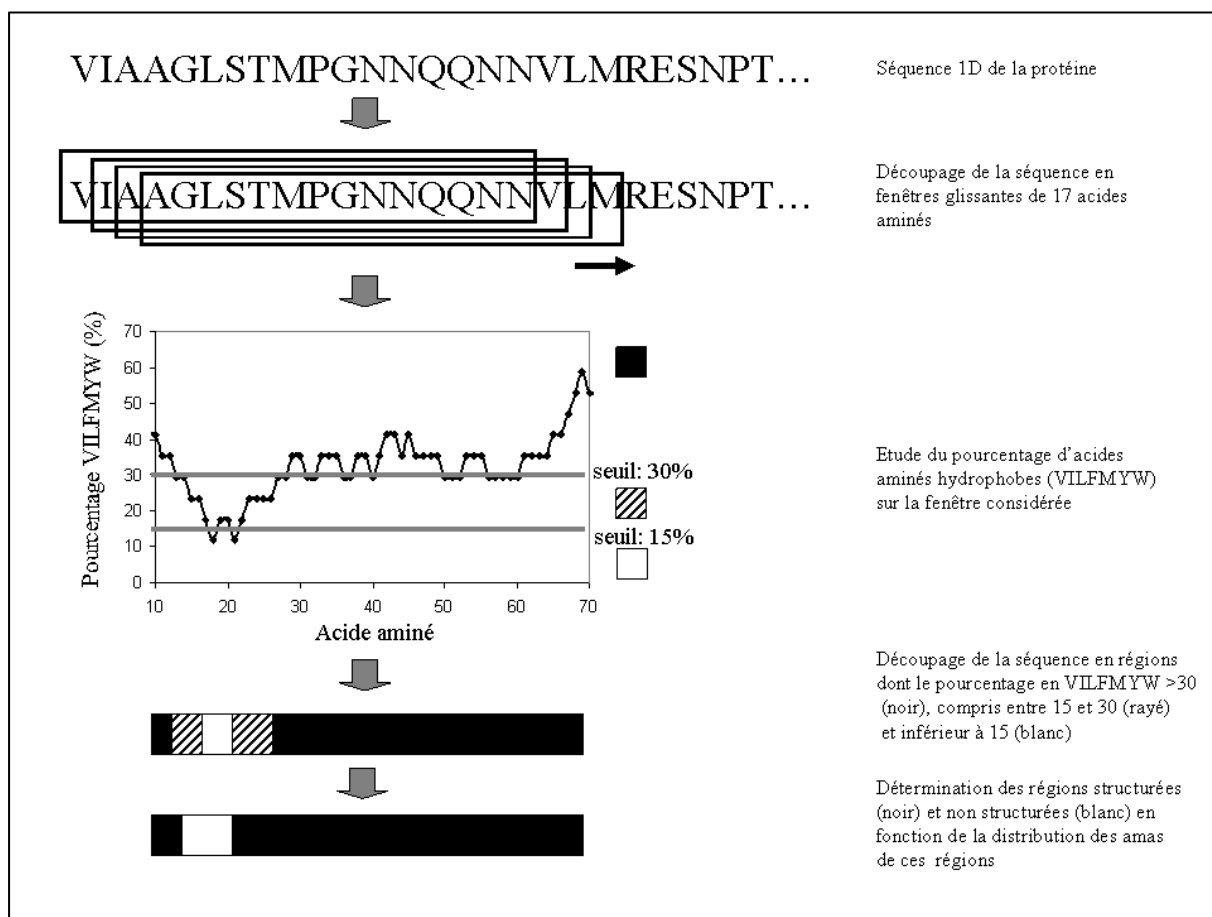


Figure 43 : Schématisation de la procédure DomHCA pour la détection des régions structurées dans les séquences de protéines.

4 Ajustement des bornes des régions structurées

Les séquences protéiques du génome de *Plasmodium falciparum*, récemment séquencé, contiennent de nombreuses régions de faible complexité et un taux beaucoup plus élevé de résidus asparagine et lysine que celles issues de génomes classiques [PIZZI, E. et al., 2001]. La représentation HCA des séquences de *Plasmodium falciparum* permet d’isoler visuellement les régions de faible complexité des régions structurées. Ainsi, une banque

« Témoin » de 20 protéines issues principalement de *Plasmodium falciparum* a été constituée (Tableau 7).

Tableau 7 : Banque « Témoin » de *Plasmodium falciparum*.

Nom de la protéine	Longueur (acides aminés)	Nom de la protéine	Longueur (acides aminés)
PF11_0204	352	PF10_0265	1811
PFE1110w	116	PF11_0449	297
PF11_0458	317	MAL13P1.32	4273
PFI0510c	1697	MAL13P1.163	284
MAL8P1.111	1259	PF14_0600	308
PFA0040w	344	PF00_0003	2129
PF10_0161	928	PF11_0241	2467
MAL6P1.219	294	PFE1590w	181
MAL7P1.37	728	PFC0425w	4550
PF10_0093	448	PFB0375w	1802

Une analyse minutieuse des tracés bidimensionnels HCA de ces 20 protéines a permis d'établir des règles issues de l'expérience pour ajuster au mieux les bornes des régions structurées de ces protéines. Les régions structurées identifiées et leurs bornes sont affinées en ajoutant certains critères de sélection ainsi qu'explicité ci-dessous.

4.1 Ajustement global

- si la taille de la région structurée prédite est inférieure ou égale à 10 acides aminés, le segment identifié n'est pas considéré comme structuré,
- si la taille de la région structurée prédite est comprise entre 11 et 25 acides aminés, cette région structurée est qualifiée de « région pseudo-structurée »,
- si la région structurée prédite a une taille supérieure à 25 acides aminés, la région détectée est dite structurée,
- si deux régions structurées sont séparées par moins de 15 acides aminés avec au moins un acide aminé hydrophobe (V, I, L, F, M, Y, W) dans la zone inter-région, les deux régions sont regroupées en une seule région structurée.

4.2 Ajustement des bornes de début et de fin de la région structurée

- si dans les 5 acides aminés (longueur classique d'une boucle additionnée de 1, [LESZCZYNSKI, J. F. et al., 1986]) précédant (ou suivant) le segment structuré prédit, il y en a au moins un inclus dans un amas hydrophobe (sauf amas 1 et 11: un acide aminé hydrophobe seul ou deux acides aminés hydrophobes consécutifs, respectivement séparés des autres acides aminés hydrophobes par au moins 4 acides aminés non hydrophobes ou une proline), la région structurée est ajustée à la borne de début (ou de fin) de l'amas,

- si la borne de début (ou de fin) de la région structurée prédite est incluse dans un amas, celle-ci est allongée en amont (ou en aval) à la première position de l'amas,

4.3 Correction apportée à la prédiction de régions structurées de petite taille entourées par de grandes régions charnières

Les « charnières » ou régions séparant deux régions structurées, ont été étudiées en examinant les amas hydrophobes qu'éventuellement elles contiennent. Il arrive que des régions « charnières » (régions séparant deux régions structurées) soient prédites autour de régions structurées de petite taille. Un exemple d'une telle situation est illustré Figure 44, où DomHCA prédit initialement trois régions structurées (résidu 1 à 45, 63 à 68 et 110 à 185).

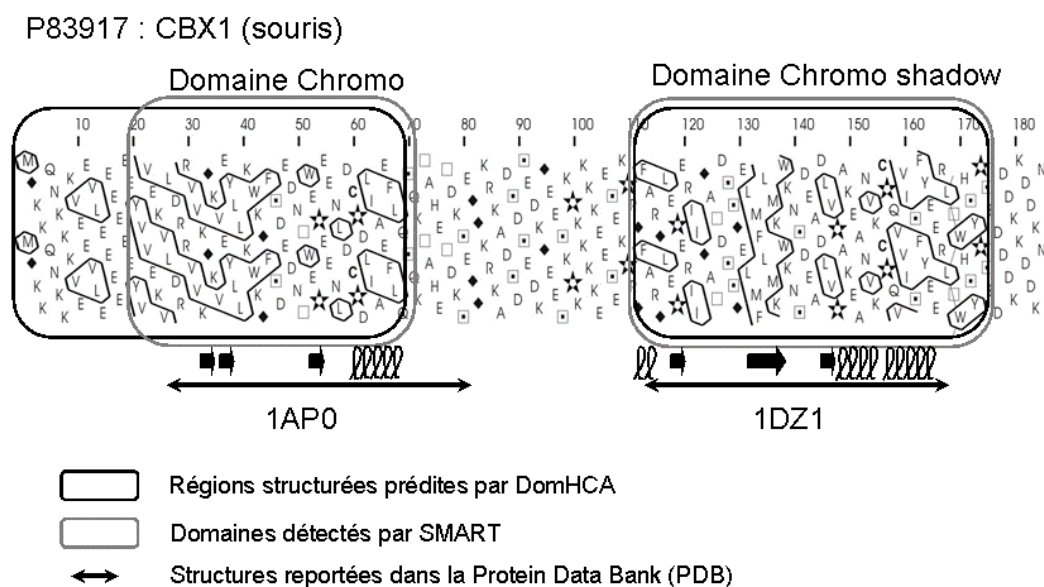


Figure 44 : Exemple de protéine « multidomaine », avec une région charnière séparant les deux domaines. La séquence de la protéine HP1 de souris est composée de deux domaines globulaires: un domaine CHROMO et un domaine CHOMO SHADOW [AASLAND, R. et al., 1995; YE, Q. et al., 1997]. Les régions prédites comme structurées d'après notre procédure sont encadrés en noir. Les bornes des domaines, telles que répertoriées dans la banque SMART, sont encadrées en gris.

A l'opposé, des valeurs largement supérieures à 33%, pourraient laisser présager des régions membranaires multiples (cf. paragraphe suivant). Dans le cas d'un passage membranaire unique, une forte concentration locale en acides aminés hydrophobes est observée (cf. paragraphe suivant).

6 Information déduite du score d'hydrophobie sur la présence éventuelle de passages membranaires dans les régions structurées

6.1 Introduction

Dans un contexte de découpage en régions structurées des protéines, il s'est avéré utile d'ajouter une information sur la présence éventuelle de passages membranaires dans ces segments. Les protéines membranaires se classent en deux groupes : les protéines transmembranaires (ou protéines membranaires intrinsèques) et les protéines membranaires périphériques (intracellulaire ou extracellulaire). Les protéines membranaires périphériques se fixent à la double couche lipidique externe ou interne des membranes par des liaisons covalentes, des forces électrostatiques ou des liaisons hydrophobes. Ces protéines membranaires sont soit extra ou intracellulaires. Les protéines transmembranaires traversent la membrane, soit une seule fois (protéine à traversée unique), soit plusieurs fois (protéines à traversées multiples) (Figure 46). Ces protéines asymétriques se caractérisent par leur amphipathie : elles possèdent généralement deux pôles hydrophiles, en contact l'un avec la phase aqueuse extracellulaire, l'autre avec la phase aqueuse cytoplasmique, et une partie hydrophobe plongée dans la couche lipidique. La presque totalité des protéines transmembranaires portent des chaînes polysaccharidiques (protéines glycosylées) plus ou moins ramifiées et longues, qui occupent la région externe à la membrane [MAILLET, M 1995].

Les protéines transmembranaires à traversées multiples (encore appelées « polytopiques ») traversent plusieurs fois la bicouche lipidique, en constituant des hélices α régulières (Figure 46), perpendiculaires ou obliques par rapport à la membrane. L'oblicité observée pour certaines d'entre elles entraîne des longueurs variables (de 8 à 38 résidus [CUTHBERTSON, J. M. et al., 2005]). Plusieurs programmes de prédiction automatique des passages membranaires existent ([ROST, B. et al., 1996; MOLLER, S. et al., 2001; TUSNADY, G. E. et al., 2001]). Les protéines de type porine forment une classe particulière

dans les protéines membranaires puisqu'elles traversent les membranes via des brins β , dont l'agencement en feuillet circulaire forme un pore. Plusieurs outils ont également été développés pour les prédire ([JACOBONI, I. et al., 2001; GROMIHA, M. M. et al., 2004; NATT, N. K. et al., 2004]).

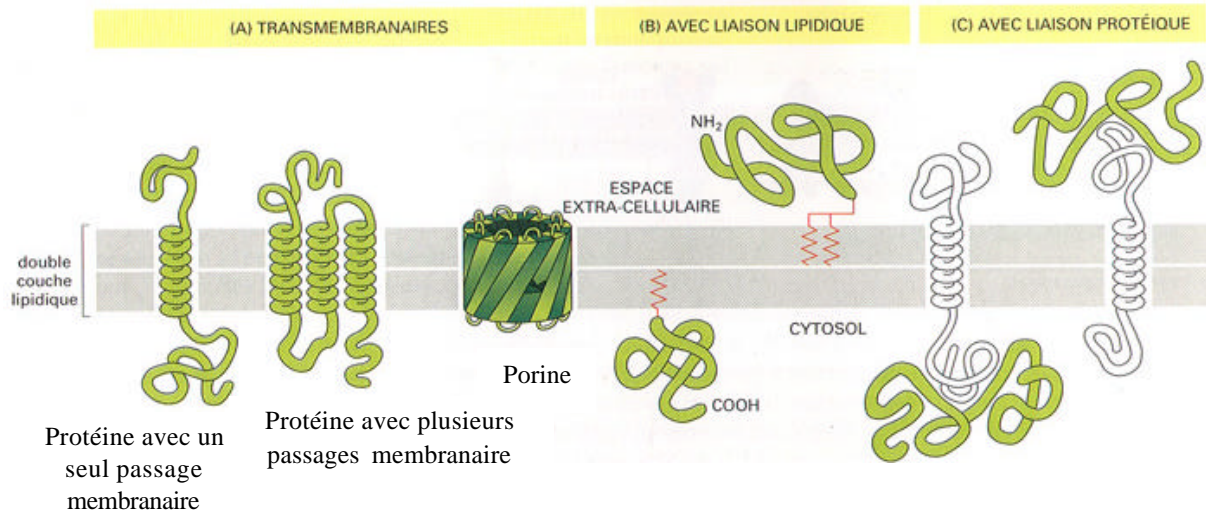


Figure 46 : Différentes configurations de protéines membranaires.

Image extraite du site internet (<http://www.humans.be/bio%20cell%20mb%20plasmique.html>).

Les passages membranaires en hélices des protéines à traversée simple (encore appelées « monotopiques ») sont caractérisés par un pourcentage local en résidus VILFMYW très élevé (supérieur à 80%) [KYTE, J. et al., 1982] et mesurent en général 21-22 acides aminés de long. Un grand nombre d'algorithmes a été développé pour prédire les hélices transmembranaires. Ces méthodes identifient dans 90 à 95% des cas les hélices transmembranaires, avec un excès de prédiction de quelques pourcents [VON HEIJNE, G. 1992; JONES, D. T. et al., 1994; ROST, B. et al., 1995; PERSSON, B. et al., 1996; ROST, B. et al., 1996; CUTHBERTSON, J. M. et al., 2005]. Ces méthodes sont souvent basées sur des alignements multiples ou des réseaux de neurones et peuvent être combinées pour de meilleures prédictions [CSERZO, M. et al., 1997].

6.2 Indications quant à la présence éventuelle de passages membranaires dans le cadre de la prédiction DomHCA

Lors de notre prédiction présentée ci-dessus de régions structurées, nous examinons le pourcentage d'acides aminés hydrophobes (VILFMYW) de chaque fenêtre glissante le long de la séquence. Les régions structurées sont définies à partir du moment où est observé un pourcentage d'acides aminés hydrophobes supérieur ou égal à 30%. Le score d'hydrophobie

moyen des régions structurées standard est proche de 33%. Cependant, certaines régions structurées présentent des taux d'hydrophobes bien supérieurs (au-delà de 38%) et correspondent à des segments présentant des passages membranaires. Lorsque l'on étudie le profil d'hydrophobie de ces segments, de nombreux pics d'hydrophobie peuvent être observés (entre 50% et 90% d'hydrophobie) (cf. Figure 47). Ces deux observations reposant sur l'hydrophobie de la région structurée nous ont permis d'établir un indicateur d'hélices transmembranaires potentielles et ainsi de proposer une information supplémentaire sur la séquence protéique étudiée, sans mettre en œuvre des méthodes calculatoires spécifiques de prédictions de passages transmembranaires.

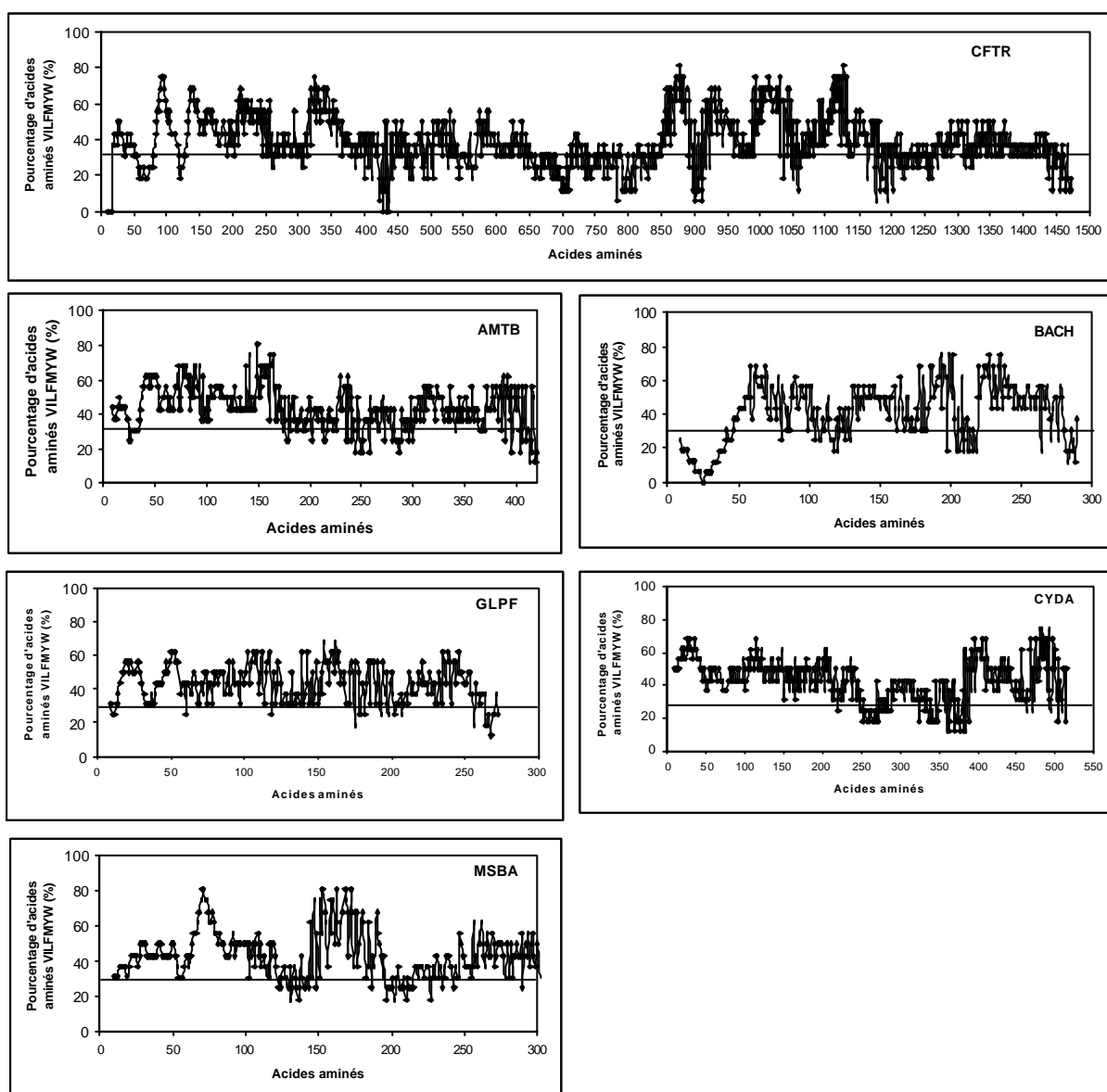


Figure 47 : Profils hydrophobes de protéines possédant des passages membranaires multiples. Le pourcentage moyen d'acides aminés hydrophobes typiquement présents dans les domaines globulaires (~33%) est indiqué par un trait horizontal noir.

6.3 Passages membranaires hélicoïdaux multiples

Dans le cas de segments structurés possédant plusieurs passages membranaires, nous avons développé un indicateur qui permet d'informer sur la position du réseau d'hélices transmembranaires. Il convient ensuite d'approfondir cette information par d'autres méthodes développées spécifiquement et/ou par expertise HCA.

Nous travaillons dans ce cadre uniquement sur les régions structurées ayant un pourcentage d'hydrophobie supérieur à 38%. Notre procédure, basée sur le déplacement d'une fenêtre de 21 acides aminés (taille moyenne des passages membranaires [DEBER, C. M. et al., 1986]), recherche les régions de plus forte hydrophobie (taux d'hydrophobes supérieur à 60%). Si ces segments de séquence sont à une distance inférieure de 100 acides aminés les uns des autres, ils sont assemblés pour former au sein de la région structurée une sous-région susceptible de contenir plusieurs hélices transmembranaires. Nous avons calibré notre détection sur cinq protéines de la Swiss Prot possédant de multiples passages membranaires, dont nous connaissons les différents domaines constitutifs (les structures de AMTB, GPLF et BACH sont connues, de nombreuses études expérimentales ont été réalisées sur CYDA, une prédiction fine de CFTR a été faite au laboratoire [CALLEBAUT, I. et al., 2004]). Leurs positions et les profils hydrophobes correspondants sont présentés Figures 47 et 48 :

- Protéine CFTR de l'Homme (« Cystic Fibrosis Transmembrane conductance Regulator »), transporteur d'anions (chlore) (code P13569),
- Transporteur de NH₃ (AMTB, code P69680) d'*Escherichia coli*,
- Facilitateur du glycérol (« Glycerol uptake facilitator protein », GLPF, code P11244) d'*Escherichia coli*,
- Halorhodopsine (BACH, code O93741) de *Halobacterium sp.*,
- Cytochrome D ubiquinol oxidase, sous-unité I (CYDA, code P11026) d'*Escherichia coli*.

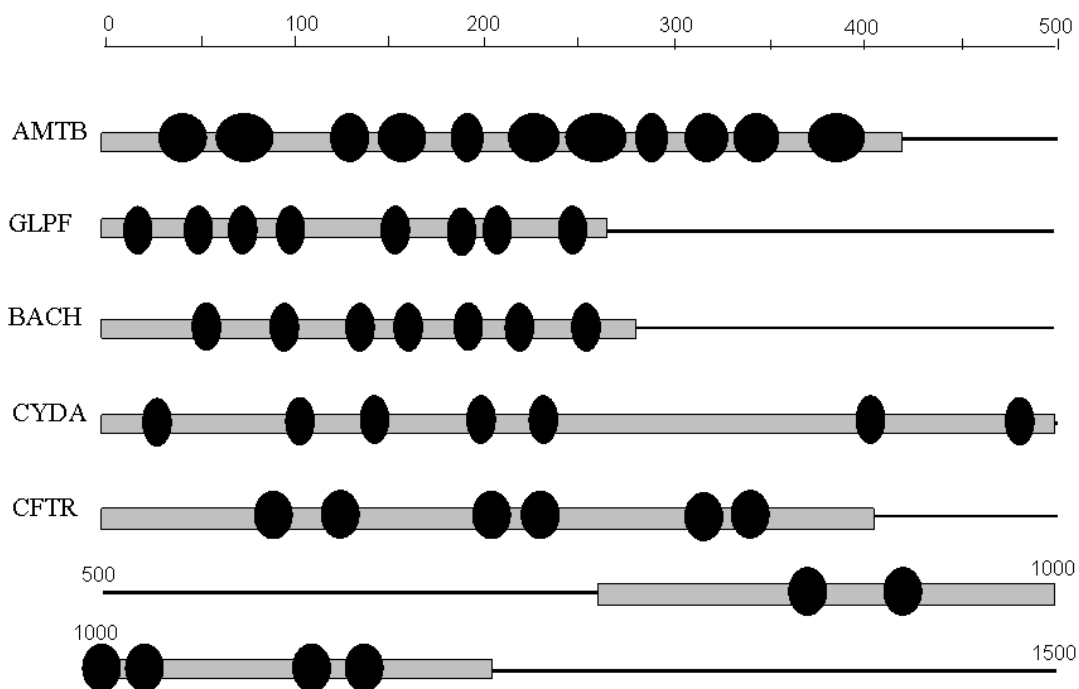


Figure 48 : Passages membranaires de cinq protéines polytopiques.

Les hélices transmembranaires sont indiquées en noir et les régions prédites comme possédant des passages membranaires multiples en gris.

Nous présentons ici l'exemple de CFTR, protéine impliquée dans le transport de chlore et dont les mutations sont à l'origine de la mucoviscidose (Figure 49). Cette protéine, étudiée dans l'équipe, contient des domaines globulaires (NBD1 et NBD2) impliqués dans l'hydrolyse de l'ATP, deux régions membranaires (MSD1 et MSD2) et un domaine régulateur R, peu structuré (Figure 49). Trois régions ont été prédites structurées par DomHCA (11-409, 428-679 et 705-1450) ; elles ont respectivement des scores d'hydrophobie de 44,86%, 36,41% et 39,81%. Les régions potentielles incluant des hélices transmembranaires multiples sont situées de 1 à 409 et de 760 à 1205, ce qui recouvre bien les deux régions membranaires MSD1 et MSD2 (Figure 49, page suivante).

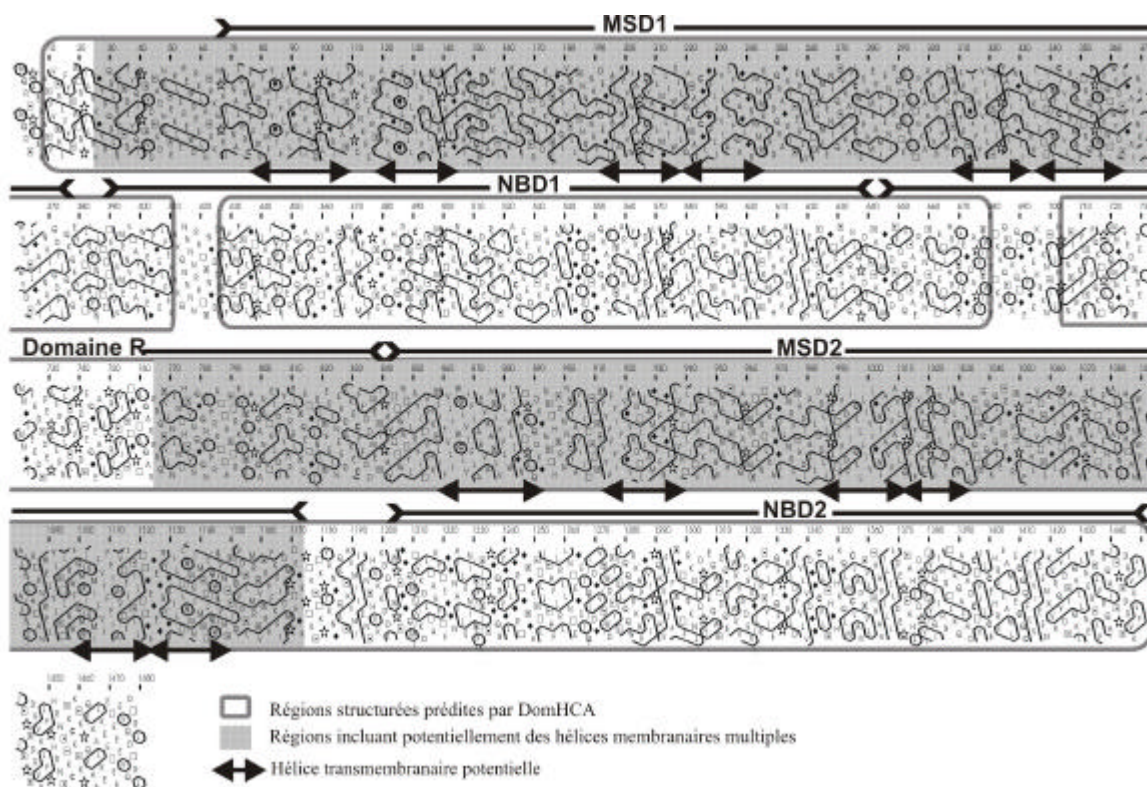


Figure 49 : Tracé HCA de la protéine CFTR renfermant deux régions membranaires (MSD1 et MSD2) possédant chacune six hélices membranaires.

6.4 Passages membranaires hélicoïdaux isolés

Les passages membranaires isolés en hélices sont caractérisés par un pourcentage local en résidus VILFMYW souvent très élevé (supérieur à 80%) et comportent en général 20-22 acides aminés. La présence d'un pic d'hydrophobie correspondant à ce passage n'augmente pas significativement le taux d'hydrophobie moyen de la région structurée détectée par DomHCA, lorsque le passage est entouré de part et d'autre par des domaines globulaires. Ce taux moyen est en général proche de 33%. Nous détectons donc spécifiquement ces pics d'hydrophobie et nous étudions la composition en acides aminés de la région entourant le pic afin de délimiter plus précisément le passage membranaire. A partir de l'expertise HCA appliquée sur de nombreuses protéines, nous savons que dans les passages membranaires sont retrouvés des acides aminés hydrophobes forts (V, I, L, F, M, Y et W) mais aussi des acides aminés mimétiques comme A, S, T, G et également les acides aminés C et P; et le plus souvent au plus un acide aminé différent de ceux cités ci-dessus. Cette information de composition permet d'évaluer si la région structurée étudiée contient bien un passage membranaire isolé. Un exemple de tracé HCA d'une protéine renfermant une hélice transmembranaire isolée est présenté Figure 50.

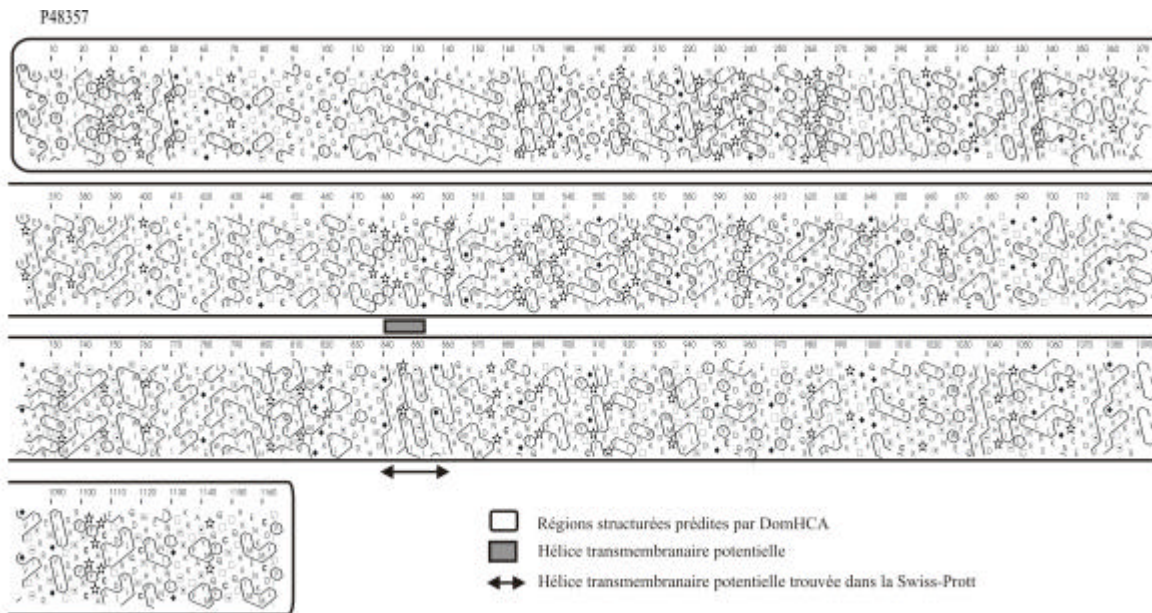


Figure 50 : Hélice transmembranaire détectée par DomHCA.

Le récepteur humain de la leptine (P48357) contient un passage membranaire potentiel de 840 à 862. DomHCA prédit deux passages membranaires (position 4 à 23 et position 839 à 858). Le premier passage membranaire détecté correspond au peptide signal de la protéine.

6.5 Cas particuliers des porines

Les porines sont des protéines situées dans la membrane plasmique des cellules. On en trouve principalement dans les bactéries, mais on peut aussi en isoler dans les cellules eucaryotes. Elles forment dans la membrane des tonneaux, composés de brins β , qui permettent la diffusion et le transport de sels et de petites molécules organiques. L'analyse des structures de porines montre une succession de brins β connectés par des boucles plus ou moins longues (Figure 51).

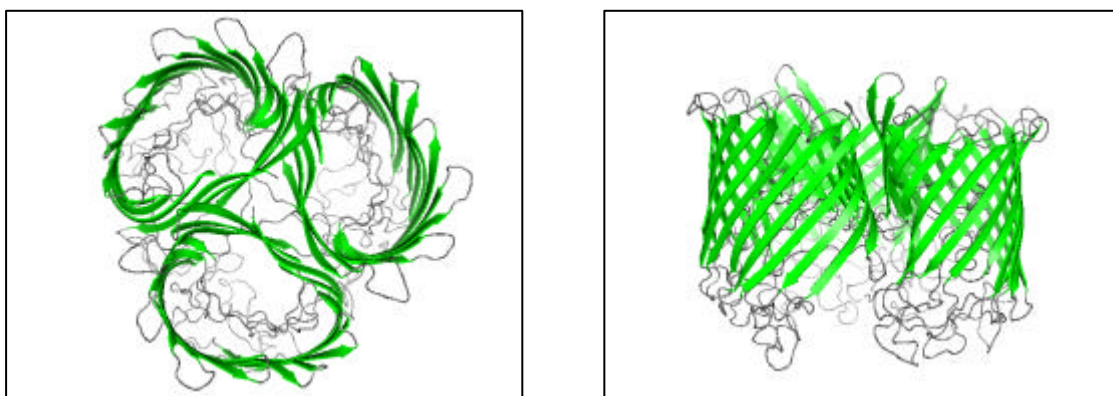


Figure 51 : Structure tridimensionnelle de la maltoporine (code PDB : 1AF6).

A gauche, le canal vu de dessus, à droite, vue de profil du tonneau b.

Les résidus polaires exposés sont retrouvés au cœur du pore et les résidus hydrophobes sont en contact avec la membrane dans laquelle le pore est inséré. Le profil d'hydrophobie est voisin de celui des régions structurées de type globulaire (score moyen d'hydrophobie voisin de 33%). De plus, celui-ci ne semble pas présenter de maxima locaux d'hydrophobie. Il s'avère difficile d'indiquer si une région structurée pourrait correspondre à une porine à partir de son seul profil d'hydrophobie. Dans ce contexte, nous ne fournirons pas d'information sur la possibilité d'avoir affaire à ce type de protéine (Figure 52).

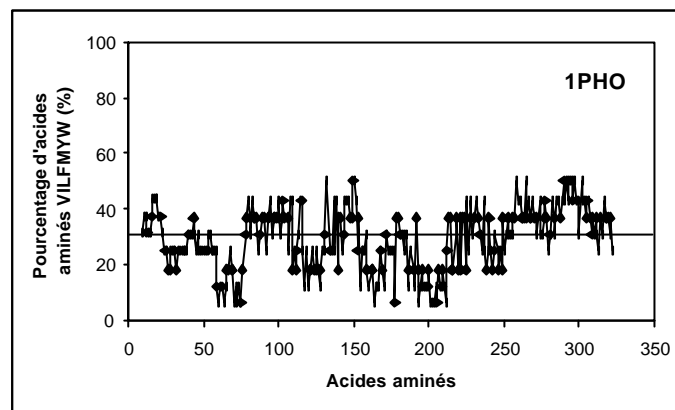


Figure 52 : Profil d'hydrophobie de la phosphoporine (code PDB : 1PHO).

Néanmoins, nous avons observé dans ces protéines une texture particulière avec de fortes concentrations locales en glycines et la présence d'amas hydrophobes typiques de brins β amphiphiles mais qui sont généralement longs (comme par exemple l'amas hydrophobe compris entre les acides aminés 76 et 108 de la Figure 53 page suivante). Ces informations suggèrent la présence de brins β constituant le pore d'une protéine membranaire. Cette hypothèse pourrait être confirmée par des méthodes de détection spécifiques des porines [BAGOS, P. G. et al., 2005].

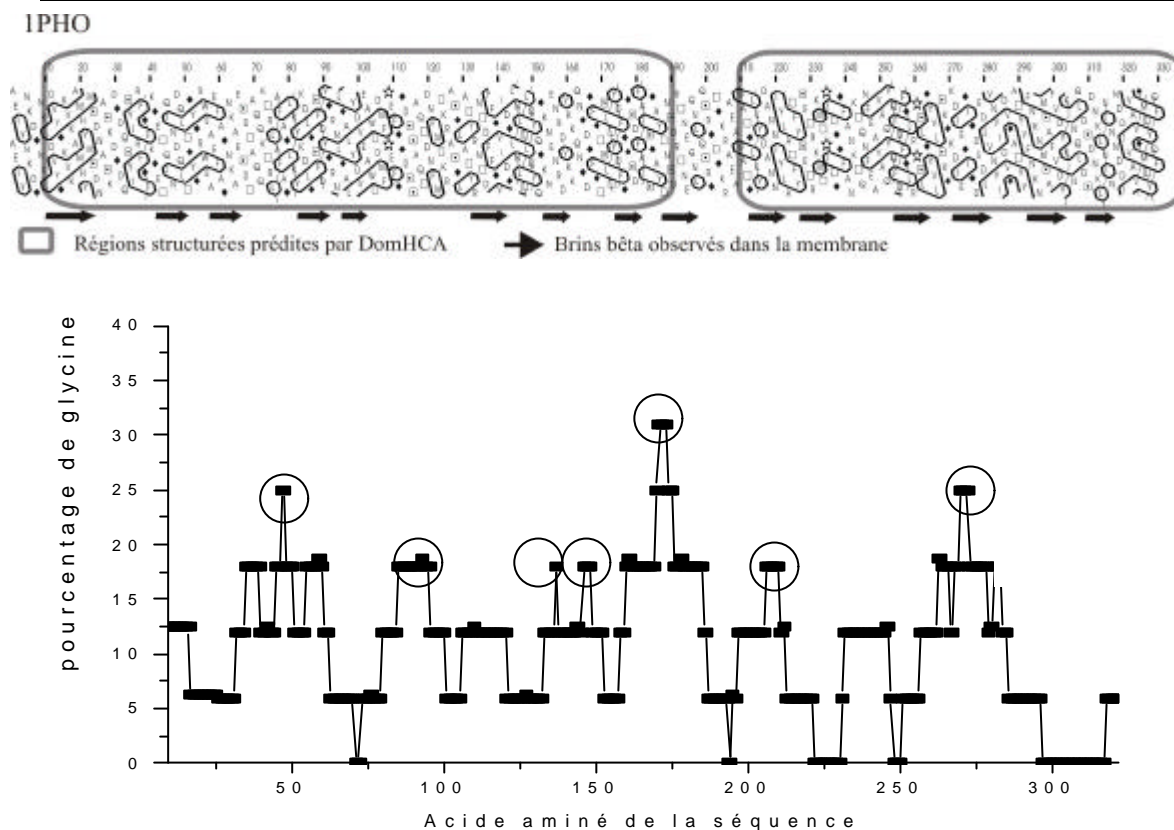


Figure 53 : Exemple de passage membranaire bêta : la protéine porine 1PHO.
 Sur le tracé HCA, les brins **b** sont représentés par des flèches. Sur la courbe du pourcentage de glycine, les pics de glycine sont indiqués par des cercles.

7 *Evaluation de la prédiction DomHCA*

Le taux de réussite des méthodes de prédictions de limites de domaines peut être estimé de différentes manières. L'efficacité de ces méthodes est généralement évaluée par une approche statistique, dans laquelle toutes les séquences ou groupes de séquences sont utilisés pour, d'une part, déterminer et, d'autre part, tester les paramètres (voir par exemple [TANAKA, T. et al., 2003]). Cependant, ce type d'approche ne peut s'appliquer à notre étude puisque la mise au point de DomHCA ne repose pas sur un jeu de données d'apprentissage, mais exploite les caractéristiques physicochimiques des **coeurs hydrophobes** des protéines.

Pour valider l'efficacité de DomHCA, nous avons choisi d'estimer, d'une part, la capacité de notre procédure à identifier pour une séquence protéique les segments structurés de cette séquence, tels que reportés dans la PDB et, d'autre part, dans le cas de protéines multidomaines, la capacité à identifier les linkers, régions séparant les domaines globulaires.

7.1 Méthodologie

7.1.1 Constitution des échantillons tests

Nous avons utilisé les 1403 chaînes protéiques structurées collectées précédemment (dont 1113 fichiers PDB correspondent à des « monodomaines » et 290 à des multidomaines, cf. paragraphe 3.3). Les séquences complètes dont sont issues les chaînes monodomaines et pluridomaines sont récupérées en comparant les séquences PDB aux séquences de protéines complètes répertoriées dans la Swiss Prot [BAIROCH, A. et al., 2000]. Ne sont conservées que les séquences pour lesquelles une entrée Swiss Prot est disponible.

Après ce traitement, les jeux de données de séquences monodomaines et pluridomaines comptent respectivement 927 et 255 représentants. Ceux-ci sont traités différemment.

Les séquences « monodomaines » sont triées en deux groupes : 434 séquences dont la taille de la séquence Swiss Prot est largement supérieure à celle issue de la PDB (échantillon SLP, pour SW Larger than PDB) et 493 séquences dont les tailles respectives sont similaires (échantillon SEP, pour SW Equal than PDB). Les attributions de structures secondaires sur les chaînes « pluridomaines », utiles pour déterminer par la suite la nature des linkers interdomaines, sont réalisées par la méthode DSSP [KABSCH, W. et al., 1983]. La codification du serveur EVA [KOH, I. Y. et al., 2003] est appliquée pour convertir les 8 états attribués de DSSP en trois états classiques (hélice, brin, boucle) (cf. Tableau 8).

Tableau 8 : Conversion des états DSSP en 3 états.

Les états DSSP «H», «G» et «I» sont définis comme hélices (H) ; les états «E» et «B» comme brins (E) et les états «T», «S» et «'» comme des boucles (L) (site internet : http://cubic.bioc.columbia.edu/eva/doc/measure_sec.html).

Méthode d'attribution	Motif reconnu							
	Hélice alpha	hélice 3_{10}	hélice π	Brin étendu	Résidu dans un pont bêta isolé	“turn” avec liaison hydrogène	“turn”	Non assigné
DSSP	H	G	I	E	B	T	S	'
EVA	H	H	H	E	E	L	L	L

Les différentes données caractérisant les chaînes sont codées comme présenté en Figure 54.

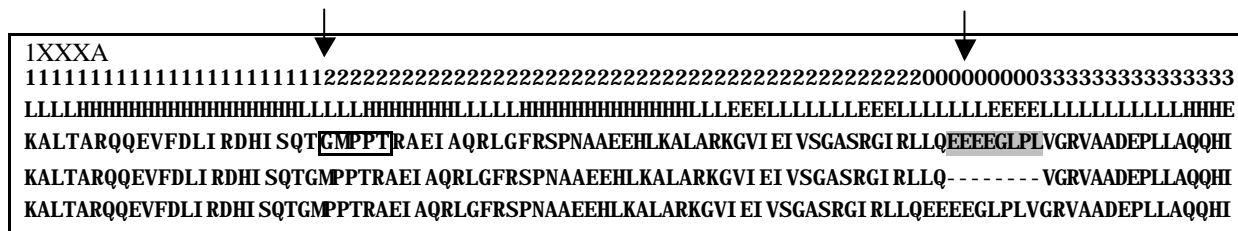


Figure 54 : Codification des différentes séquences protéiques étudiées.

La première ligne correspond au code PDB de la séquence suivi du nom de la chaîne. La seconde ligne reprend le découpage en domaines, tel que reporté dans la banque CATH. Chaque domaine est identifié par un chiffre de 1 à 9. La flèche indique un changement de domaine et donc un linker. La troisième ligne correspond aux structures secondaires observées (attribution DSSP simplifiée en un code à 3 lettres [H, E, L]). Les lignes suivantes (4 et 5) correspondent respectivement aux séquences observées dans la PDB et dans la banque CATH (les tirets indiquent des régions inter-domaines). La dernière ligne correspond à la séquence traitée par DomHCA (les régions prédites comme structurées présentent une séquence identique aux deux lignes supérieures).

7.1.2 Cas des protéines monodomaines

Plusieurs moyens peuvent être envisagés afin de mesurer l'accord entre prédiction et observation dans le cas de chaînes « monodomaines ».

a) Recouvrement de domaines ou “Domain Overlap” (DO)

L'accord entre l'état observé et la prédiction est estimé par le calcul d'un score similaire à celui calculé pour mesurer la correspondance entre prédiction et attribution des structures secondaires [ZEMLA, A. et al., 1999]. Ce score « Domain Overlap », exprimé en pourcentage, est calculé de la manière suivante :

Score (DO) = (nombre d'acides aminés correctement prédits comme appartenant à une région structurée/ nombre d'acides aminés réellement observés dans la région structurée (séquence PDB))*100.

D'après la Figure 55, le score DO est le rapport de N sur PDB. Ce score est déterminé par rapport à la séquence PDB et ne tient pas compte des positions de la prédiction DomHCA qui sont en dehors des limites de la séquence PDB (celles-ci seront évaluées en prenant en compte la valeur n').

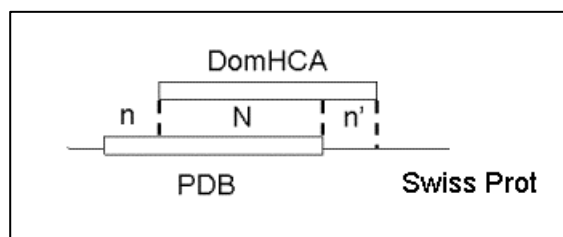


Figure 55 : Paramètres pris en compte pour le calcul de score de validation.

b) Extended Domain Overlap (EDO)

Afin de mettre en évidence le recouvrement de la prédiction faite avec DomHCA avec la séquence PDB, nous avons calculé une valeur appelée « EDO » pour Extended Domain Overlap. Cette valeur a été estimée uniquement sur le jeu de données de séquences SLP (lorsque la prédiction est calculée à partir des séquences beaucoup plus grandes que les domaines PDB correspondants). Elle se calcule selon la formule suivante (Figure 55):

$$\text{EDO} = (\text{Nx2})/(\text{PDB}+\text{DomHCA})$$

La valeur EDO permet de mettre en évidence l'adéquation entre la prédiction DomHCA et l'observation PDB.

c) Calcul du débordement en N-terminal/C-terminal

Une estimation du débordement de la prédiction DomHCA par rapport aux limites de la séquence PDB peut être réalisée en considérant le jeu de données SLP, dans lequel la prédiction est faite sur une séquence beaucoup plus grande que la séquence PDB, et en calculant la proportion d'acides aminés prédits dans la région structurée par DomHCA et qui sont en dehors des limites de la séquence PDB correspondante. Ce calcul est effectué en N-terminal et en C-terminal de la prédiction (valeurs n' de la Figure 55).

7.1.3 Cas des protéines « pluridomaines »

Pour le jeu de données pluridomaines, les prédictions de régions structurées sont évaluées en déterminant la capacité de DomHCA à mettre en évidence les «linkers» séparant les régions structurées. Ainsi, les limites des régions structurées prédites sont comparées aux limites définies dans la classification CATH.

Tanaka et collaborateurs définissent un **linker** comme une « boucle » séparant deux domaines continus, tels que répertoriés dans CATH [TANAKA, T. et al., 2003]. Une **boucle** est définie comme un court segment de plus de quatre acides aminés dépourvu d'éléments de structures secondaires régulières. Intuitivement, et d'après la définition de Tanaka et collaborateurs, on pourrait s'attendre à ce que les linkers soient uniquement des boucles. Or il n'est pas rare que les linkers soient structurés, comprenant des hélices alpha ou des brins bêta. Nous définirons donc comme linker tout peptide séparant deux domaines continus, tels que définis dans la classification CATH, quel que soit son état, structuré ou non. Ainsi, selon notre

définition, un linker peut contenir des éléments de structures secondaires. Il peut être une boucle, une hélice ou un brin.

Deux types de linkers peuvent être identifiés:

- les fragments de séquence, relativement longs, correspondant à de véritables régions hors domaines et identifiables d'après le code CATH par une suite de « 0 » (par exemple, la séquence EEEGLPL surlignée en gris dans la Figure 54) que l'on désigne comme « linkers 0 »,
- les fragments de séquence, généralement courts, pour lesquels est observée une même structure secondaire chevauchant deux domaines différents (ayant d'après le code CATH un numéro différent). Ceux-ci sont désignés comme « linkers sans rupture ». Un exemple de ce type de linker est la séquence encadrée GMPPT dans la Figure 54.

Les **vrais positifs** (VP) sont des linkers réellement observés, ainsi que prédits par DomHCA. Les **faux positifs** (FP) sont des linkers prédits par notre procédure, mais qui ne le sont pas d'après la classification CATH. Les **faux négatifs** (FN) sont des linkers réellement observés, mais non prédits par DomHCA.

7.2 Résultats

7.2.1 Hydrophobie et taille des régions prédites

La Figure 56 présente la comparaison entre les profils hydrophobes des régions structurées prédites par DomHCA (pour les jeux de séquences SLP et SEP) et ceux des domaines globulaires issus de SCOP (cf. précédemment Figure 39). Nous observons que les trois profils se superposent bien et sont centrés sur la valeur 33%. Cette juxtaposition permet de valider la prédiction sur ce critère d'hydrophobie.

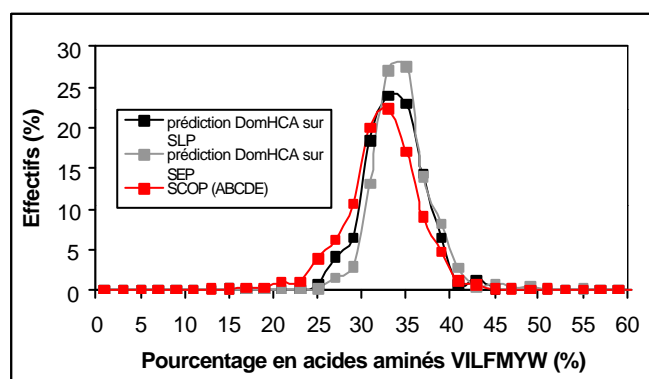


Figure 56 : Profil hydrophobe des régions structurées prédites avec DomHCA.
Les prédictions sur les deux jeux de données SLP et SEP sont comparées aux domaines globulaires de SCOP (A, B, C, D et E, 7530 séquences).

Nous avons également calculé la distribution de la taille des segments structurés prédits par DomHCA sur nos deux échantillons de données (SEP et SLP) et comparé ces distributions à celle calculée initialement sur le jeu de données « monodomaines » (Figure 57). Nous constatons que les domaines prédits par DomHCA ont des longueurs similaires à celles observées dans le jeu de « monodomaines », ce qui montre la capacité de DomHCA à extraire les domaines structurés à partir des séquences complètes dont ils sont issus. Néanmoins, nous observons aussi des longueurs de séquences prédites plus grandes que celles des monodomaines, notamment sur l'échantillon SLP, ce qui comme précédemment illustre sans doute la capacité de DomHCA à détecter des régions structurées contenant plusieurs domaines contigus. Nous notons également les fréquences supérieures des régions prédites de faible taille. Ceci met vraisemblablement en évidence la particularité de DomHCA à prédire les limites minimales des domaines.

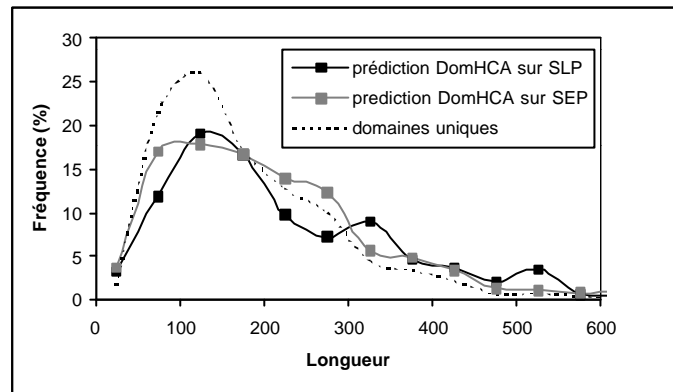


Figure 57 : Distribution de la longueur des segments structurés prédits par DomHCA et comparaison avec les tailles des segments monodomaines définis à partir de notre échantillon « monodomaines » d'après CATH (cf. Figure 41).

Les courbes sont tracées avec des intervalles de 50 acides aminés.

7.2.2 Prédiction de régions structurées à partir de chaînes « monodomaines »

Les prédictions sont évaluées par le calcul du score « Domain Overlap ». Pour chaque jeu de données (SEP et SLP), la distribution des scores DO est représentée en Figure 58. Ces scores DO correspondent au nombre de positions de la séquence PDB prédites comme structurées. Nous n'examinons ici que le résultat de la prédiction DomHCA sur la longueur de la séquence PDB. D'après la Figure 58, près de 80 % des protéines issues respectivement des échantillons SEP et SLP ont des scores DO compris entre 95 et 100%. Pour les SEP (séquences de même taille), le score DO moyen est de 97,5 % avec un écart type de 5,7. Pour les SLP (séquences de tailles différentes), le score DO moyen est de 95,3 % avec un écart type de 10,9. Les scores de prédiction sont bons, quelle que soit la taille de la séquence Swiss Prot de départ.

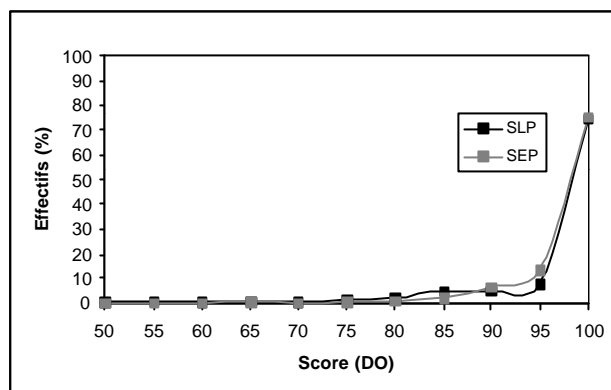


Figure 58 : Distribution des scores DO obtenus avec DomHCA pour les jeux de données SLP et SEP.

Ensuite, nous avons calculé le nombre d'acides aminés des séquences PDB non prédits comme faisant partie de la région structurée par DomHCA (n dans la Figure 55). Nous appelons cette valeur « imprécision », en partant du principe que 100% des acides aminés de la séquence PDB prédits comme « structurés » correspondent à une imprécision nulle. Cette imprécision correspond donc à l'erreur dans la position de la limite du domaine, exprimée en acides aminés. Nous avons testé différentes imprécisions (de 0 à 50 acides aminés) sur nos deux échantillons SEP et SLP (Figure 59).

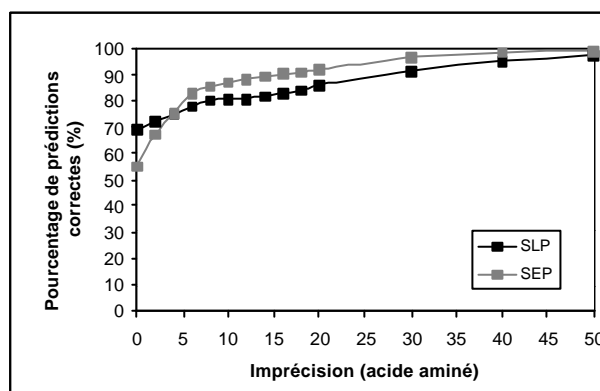


Figure 59 : Distribution du nombre de prédictions correctes en fonction de l'imprécision autorisée.

La prédiction DomHCA obtient pour le jeu de données SLP 69%, 81% et 92% de prédictions correctes pour des imprécisions respectives de 0, 10 et 30 résidus par rapport aux limites de la séquence PDB. Pour le jeu de données SEP, les valeurs obtenues sont 55%, 87% et 97% (Figure 59).

La Figure 60 présente deux exemples de prédictions correctes observées dans le jeu de données SLP (1CXZB et 1ECMA). Les prédictions DomHCA ont été effectuées à partir des séquences Swiss Prot, puis celles-ci ont été alignées avec la séquence PDB

correspondante et la séquence CATH. Dans le cas où plusieurs segments structurés DomHCA sont prédits sur la séquence Swiss Prot, seul le segment recouvrant la séquence PDB est aligné. Pour la chaîne recensée dans la PDB sous le code 1CXZ (chaîne B), DomHCA a prédit sept régions structurées (13-100, 125-188, 208-239, 264-331, 357-523, 610-850, 873-908). La région 13-100 correspond exactement (à deux acides aminés près) à la séquence CATH. Les autres régions structurées ne peuvent pas être validées en raison de l'absence actuelle de données expérimentales pour celles-ci. D'après la classification CATH, cette protéine de 942 acides aminés ne contient qu'un seul domaine (88 acides aminés), mais il semble évident que d'autres domaines globulaires composent cette protéine. Ils ne sont cependant pas recensés par CATH.

Pour la chaîne recensée dans la PDB sous le code 1ECM (chaîne A), le segment structuré prédit par DomHCA comprend 369 acides aminés alors que le domaine CATH ne comprend que 91 acides aminés. La région prédite englobe le domaine CATH dans sa totalité, mais se poursuit par un fragment de 280 acides aminés. Il est fort probable que le segment DomHCA contienne au moins deux domaines structurés, un premier recensé dans CATH et un second non recensé. Ces deux domaines seraient séparés par une région linker non détectable par DomHCA.

```
86_1CXZB (PDB)
WSLLEQLGLAGADLAAPGVQQQLLELERERLRREIRKELKKEGAENURRATTDLGRSLGPVELLLRGSSRRLDLHQQLQELH#HV
86_1CXZB (CATH)
WSLLEQLGLAGADLAAPGVQQQLLELERERLRREIRKELKKEGAENURRATTDLGRSLGPVELLLRGSSRRLDLHQQLQELH#HV
88_1CXZB#13/100 (DomHCA)
WSLLEQLGLAGADLAAPGVQQQLLELERERLRREIRKELKKEGAENURRATTDLGRSLGPVELLLRGSSRRLDLHQQLQELH#VWL

109_1ECMA (PDB)
MTSENPLLALREKISALDEKLLALLAERRELAWEVGGKAKLLSHRPVVRDIDRERDILLERLITLGAHHLDAHYITRFLQIIEDSVLTQQALLQQHUNKINPHSARIAFL
91_1ECMA (CATH)
NPLLALREKISALDEKLLALLAERRELAWEVGGKAKLLSHRPVVRDIDRERDILLERLITLGAHHLDAHYITRFLQIIEDSVLTQQALLQQH
369_1ECMA#7/375 (DomHCA)
LLALREKISALDEKLLALLAERRELAWEVGGKAKLLSHRPVVRDIDRERDILLERLITLGAHHLDAHYITRFLQIIEDSVLTQQALLQQHNLKINPHSARIAFL—(266 aa)
```

Figure 60 : Alignement de la séquence PDB, de la séquence du domaine CATH et de la séquence de la région structurée prédite par DomHCA.

La première ligne correspond à la taille de la séquence PDB et au code PDB suivi de l'identificateur de la chaîne. La seconde correspond à la séquence de la chaîne PDB. La troisième ligne correspond à la taille du domaine CATH et à son code. La ligne suivante correspond à la séquence de la chaîne du domaine CATH. La cinquième ligne correspond à la taille et aux bornes du domaine prédit par DomHCA. La dernière ligne correspond à la séquence du domaine DomHCA.

Les scores précédemment calculés ne rendent pas compte du recouvrement de la prédiction en dehors des limites des séquences PDB. Pour estimer ce recouvrement, nous calculons la valeur EDO (voir définition page 105 et Figure 55). La Figure 61 présente la distribution des valeurs EDO pour le jeu de donnée SLP. Ce calcul n'a de sens que sur le jeu de données SLP, dont les séquences qui ont servies à établir la prédiction DomHCA ont des tailles beaucoup plus élevées que les séquences PDB.

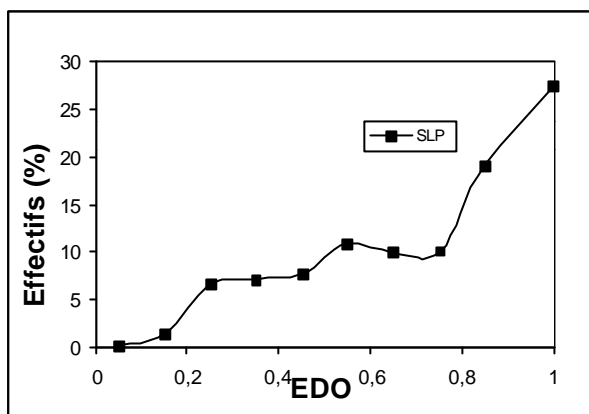


Figure 61 : Distribution de la valeur EDO.

Nous remarquons que 2% des protéines ont un recouvrement inférieur à 20%. Il n'existe pas de séquence PDB pour laquelle DomHCA n'a prédit aucun résidu dans le segment structuré. Enfin, nous constatons que 56% des séquences considérées ont une valeur EDO supérieure ou égale à 75%.

Dans un second temps, nous avons calculé la valeur n' qui correspond au nombre d'acides aminés, prédits à partir du jeu de séquences SLP, comme faisant partie de la région structurée, mais situés en dehors des séquences PDB. Ce nombre peut être calculé en N-terminal et en C-terminal des limites de la séquence PDB (Figure 62). Ce calcul n'a également de sens que sur le jeu de données SLP, dont les séquences qui ont servies à établir la prédiction DomHCA ont des tailles beaucoup plus élevées que les séquences PDB.

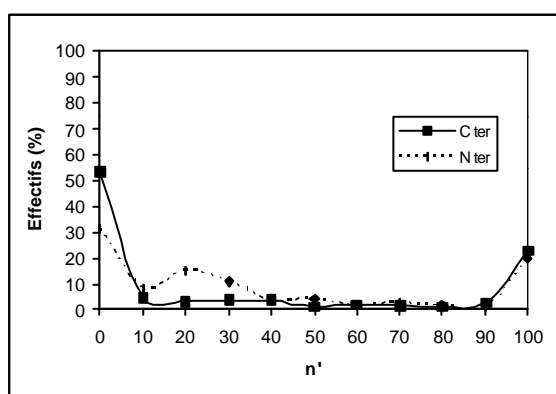


Figure 62 : Distribution des valeurs n' .

Nous constatons qu'une bonne partie des limites des prédictions DomHCA chevauchent à au plus 10 acides aminés près les limites de la séquence PDB considérée, que ce soit en N-terminal (39% des séquences) et C-terminal (59% des séquences). Il faut noter

qu'en C-terminal, la prédiction de limites semble meilleure. Il semble ainsi que DomHCA prédise moins facilement le début de la région structurée. En effet, l'extension en N-terminal peut parfois mesurer jusqu'à 40 acides aminés. La détection précoce résulte sans doute de la présence de quelques acides aminés hydrophobes parsemés en amont des séquences PDB. D'autre part, 21% et 25% des séquences prédites par DomHCA sont étendues respectivement en N-terminal et C-terminal de plus de 90 acides aminés. Ce résultat laisse supposer qu'il existe d'autres domaines structurés, situés en amont et en aval des séquences PDB testées, et contigus aux domaines étudiés. Un exemple de tel cas est celui de la glucuronidase citée précédemment (Figure 42, page 89). Cette protéine contient trois domaines distincts mais elle est prédite comme une seule région structurée d'après DomHCA, alors qu'elle est classée dans les protéines monodomaines d'après CATH.

7.2.3 Prédiction des régions structurées à partir de chaînes pluri-domaines

Pour évaluer la qualité de notre méthode de prédiction des régions structurées dans ce cas, nous avons mesuré l'aptitude de DomHCA à détecter les « linkers » séparant les domaines dans notre jeu de séquences pluridomaines (255 séquences). Les limites des segments structurés sont comparées aux limites de domaines définies par CATH. Nous avons identifié et comptabilisé 289 linkers. Ces « linkers » peuvent être séparés en deux classes.

La première contient 98 linkers, que nous appellerons linkers « 0 ». Ces linkers, de taille variable (taille minimale 2 résidus et taille maximale 302), correspondent à des fragments de séquence qui ne sont pas inclus dans un domaine d'après la classification CATH version 2.5.1 (Figure 63). 39 de ces « linkers » « 0 » ont une taille supérieure à 40 acides aminés et correspondent à des domaines ou fragments de séquence non assignés dans la base CATH. Récemment, la base CATH (version 2.6.0) a été mise à jour et CATH assigne maintenant une partie de ces linkers comme des domaines globulaires. Nous avons donc ôté les linkers ayant une taille supérieure à 40 acides aminés pour la suite de notre étude. La taille moyenne des linkers « 0 » est alors de 14 acides aminés. Nous comptabilisons respectivement 40 boucles (L), 14 hélices (H) et 5 brins (E) associées à ce type de linker « 0 ».

La seconde classe de linkers compte 192 séquences (66% de l'ensemble des linkers). Ces « linkers » sont qualifiés de « linkers sans rupture » car la séparation entre deux domaines consécutifs n'est marquée par aucune séquence peptidique particulièrement détectable (Figure 63). Ce sont des séquences correspondant à une structure secondaire, régulière (hélice α ou brin β) ou non (une boucle) et chevauchant les deux domaines consécutifs. Dans nos séquences pluridomaines, nous comptabilisons respectivement 117

boucles (L), 47 hélices (H) et 28 brins (E) associées à ce type de linker sans rupture (Figure 64).

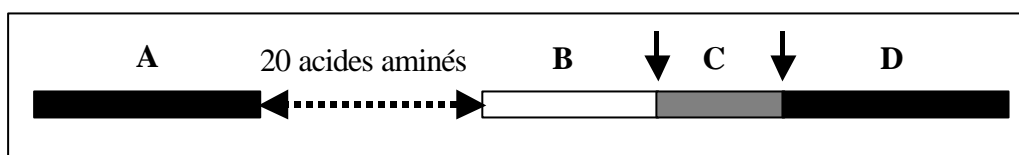


Figure 63 : Linkers « 0 » et linkers « sans rupture ».

20 acides aminés séparent les domaines A et B, c'est un linker « 0 », indiqué par une flèche horizontale en pointillé noir. Les domaines B et C et C et D ne sont séparés par aucun fragment de séquence, la séquence correspondant à la structure secondaire régulière (hélice α , brin b ou boucle) chevauchant les deux domaines est qualifiée de linker « sans rupture ». Ceux-ci sont indiqués par des flèches verticales noires.

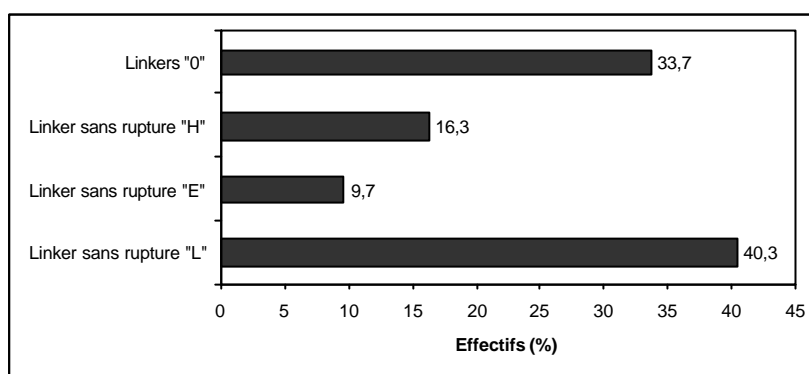


Figure 64 : Répartition des linkers totaux identifiés dans les chaînes pluridomains.

Beaucoup de « linkers » appartenant aux deux classes citées ci-dessus contiennent des structures secondaires régulières (94 linkers sont structurés, soit 37 % de l'ensemble de nos 251 linkers de la banque de chaînes multidomains). Ce sont essentiellement des structures en hélice α , qui joueraient le rôle d'« espaceurs » rigides entre les domaines [GEORGE et al., 2003]. DomHCA ne peut pas détecter efficacement ces linkers structurés car ils renferment souvent des amas hydrophobes, associés à des structures secondaires régulières. Ce type de « linker », non prédit, est qualifié de « faux négatif ». Un exemple de linker structuré « hélice α » est présenté Figure 65A à gauche. Ce linker est constitué de deux hélices α reliées par une courte boucle. Un autre exemple de faux négatif est présenté en Figure 65A, à droite. Ce linker correspond à une longue boucle qui adopte une structure étendue, avec un amas hydrophobe typique de ce type de structure secondaire. DomHCA se révèle efficace pour la prédiction de linkers ne contenant pas de structures secondaires régulières (« linkers » boucles) et de taille suffisante (vrai positif illustré en Figure 65B).

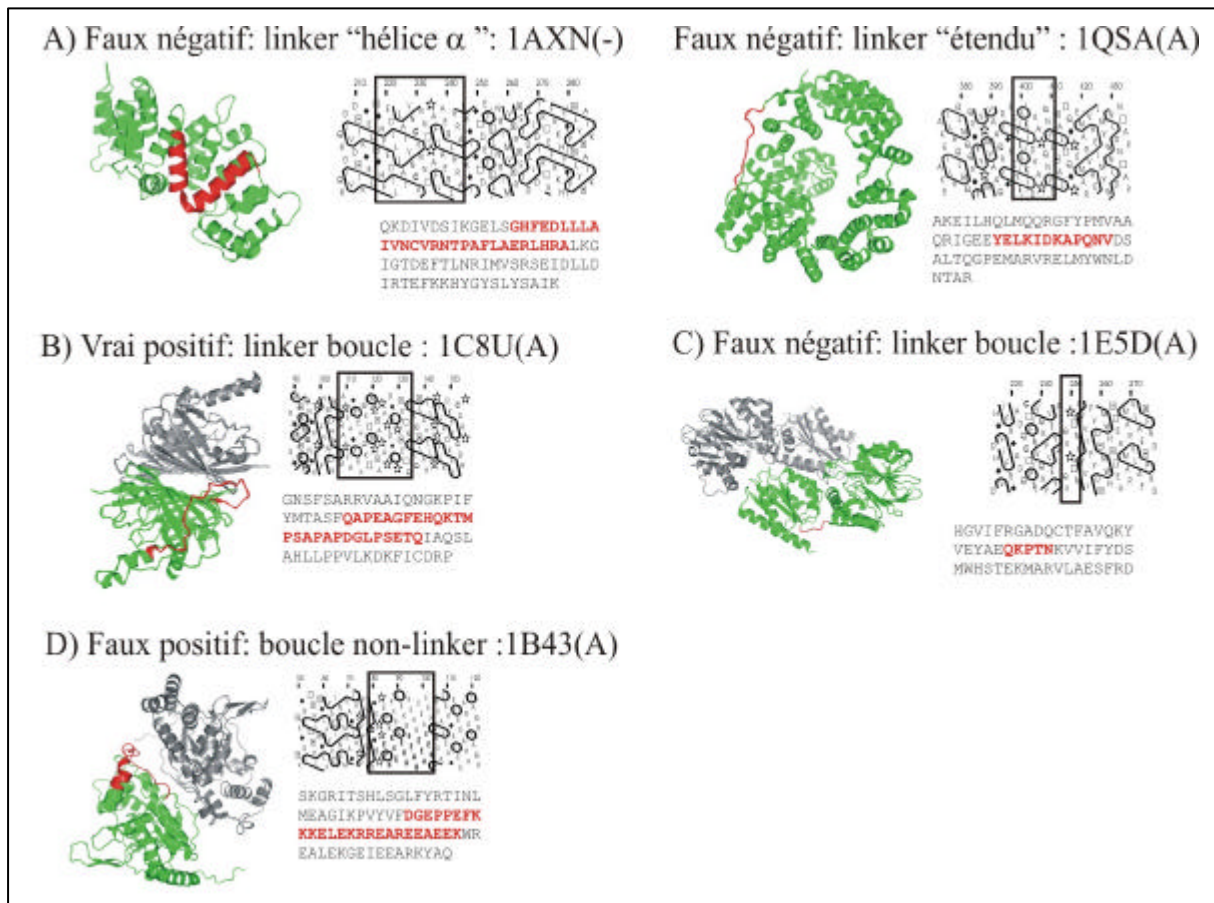


Figure 65 : Plusieurs exemples de résultats positifs et négatifs obtenus avec DomHCA. Le tracé HCA et la séquence 1D sont représentés avec la structure tridimensionnelle correspondante.

A. Exemples de faux négatif (FN). A gauche, le linker inclut une structure secondaire régulière hélice **a** et à droite, le linker s'il ne renferme aucune structure secondaire régulière, adopte une forme étendue, proche de celle d'un brin **b**.

B. Exemple de vrai positif (VP) dans lequel le linker est assez grand pour être détecté par DomHCA.

C. Exemple de faux négatif (FN) dans lequel le linker est trop petit pour être détecté par DomHCA.

D. Exemple de faux positif (FP) dans lequel le linker prédit correspond à une structure secondaire exposée ou à une grande boucle.

A l'inverse, DomHCA ne peut détecter les linkers boucles de petite taille (Figure 65C), qui constituent une grande partie de notre base de données. Ces linkers boucles « faux négatif » constituent 48% de l'ensemble des faux négatifs.

Comme précisé précédemment, les linkers non structurés que nous appellerons « boucles linkers » peuvent être détectés par notre méthode s'ils sont assez longs (Figure 65B). Ainsi, les six linkers identifiés par DomHCA sont des linkers « 0 ». Parmi ces six linkers, quatre sont des linkers non structurés (Figure 66). La procédure DomHCA ne peut détecter les linkers ruptures, sauf si ceux-ci correspondent à de grandes boucles.

Idéalement, une grande partie des linkers «0» devrait être identifiée par DomHCA. Cependant, il semble que ces linkers, correspondant essentiellement à des hélices et des boucles, contiennent un certain nombre d'acides aminés hydrophobes groupés en amas, de telle manière à ce que la distribution en amas hydrophobes qui leur est propre ne permette pas la prédiction d'une région non structurée (Faux négatifs).

Les « linkers » identifiés dans les protéines 1C8U, 1C96, 1F60 et 1KSI sont tous situés dans des régions de séquence où il n'y a pas ou peu d'amas hydrophobes. Ils correspondent à de longues boucles (Figure 66A, B, C et D). Deux autres « linkers » «0» ont été correctement prédits avec DomHCA (Figure 66E et G). Ils sont situés dans des régions contenant de petits amas hydrophobes et contiennent des hélices α . Ces hélices contiennent une forte proportion d'acides aminés A, E et G et peu d'acides aminés hydrophobes, ce qui rend difficile la prédiction par DomHCA. Les positions des linkers prédits par DomHCA et définis par CATH sont données Tableau 9.

Tableau 9 : Résultats des prédictions de DomHCA.

Protéine	Positions des domaines CATH	Position du linker prédit par DomHCA	Position du linker selon la définition CATH
1C8U(A)	1: 2-116 2: 131-286	108-134	117-130
1C96(A)	1: 2-202 2: 203-315 3: 316-490 4: 534-754	485-537	491-533
1F60(A)	1: 2-236 2: 239-327 3: 333-440	320-336	328-332
1KSI(A)	1: 6-101 2: 102-197 3: 231-646	198-230	198-230
1IAT(A)	Non assigné: 2-98 1: 99-293 Non assigné: 294-514 2 : 515-555	430-457	294-514
1IKP(A)	1: 1-223 Non assigné: 224-398 2: 399-605	330-366	224-398

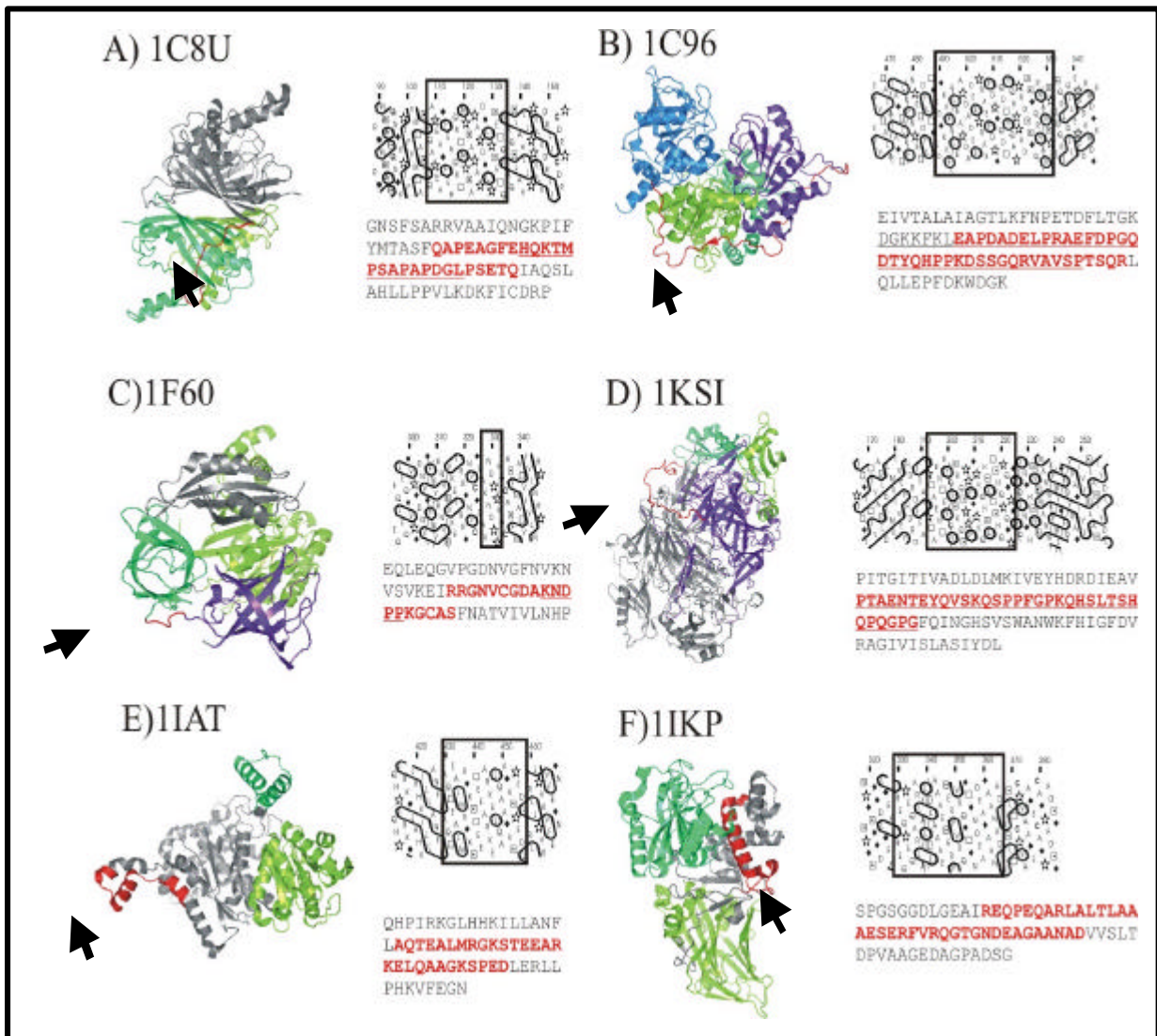


Figure 66 : Linkers correctement prédits par DomHCA.

Les linkers sont issus des séquences PDB : 1C8U, 1C96, 1F60, 1KSI, 1IAT et 1IKP. Les linkers sont colorés en rouge et sont indiqués par une flèche. Les domaines, composant la chaîne protéique dont est extrait le linker, sont respectivement colorés en vert clair, vert foncé, bleu clair et bleu foncé. Les autres chaînes protéiques éventuellement présentes dans la structure sont en gris ainsi que les domaines non assignés par CATH. Pour chaque linker sont indiquées sa séquence et la représentation HCA correspondante. La séquence du linker prédit par DomHCA est en rouge et la séquence du linker défini par CATH est soulignée. Pour 1IAT et 1IKP, les linkers définis par CATH sont de grande taille et ne sont pas représentés sur la séquence 1D.

DomHCA se révèle donc très efficace pour détecter les linkers non structurés, pourvu qu'ils soient de taille suffisante. Néanmoins, nous avons observé plusieurs « faux positifs ». La plupart de ces faux positifs correspondent à de grandes boucles au sein des domaines globulaires. Un exemple de ce type de linker « faux positif » est montré Figure 65D. D'autres faux positifs correspondent à des régions dans lesquelles sont présentes des structures secondaires, très exposées au solvant ou composées d'acides aminés mimétiques

(A, C, S, T). Ces propriétés rendent la détection de ces régions structurées délicates, en raison de l'absence d'amas hydrophobes.

Pour tenter de distinguer les «boucles linkers» séparant deux domaines des boucles classiques constitutives des domaines globulaires des protéines, nous avons comparé leurs longueurs et compositions dans notre jeu de données multidomaines. Les définitions utilisées sont similaires à celles établies par Tanaka et collaborateur [TANAKA, T. et al., 2003]. Une boucle correspond à une séquence de plus de 4 résidus sans structure secondaire (attribution par DSSP [KABSCH, W. et al., 1983], alors qu'une « boucle linker » renferme en plus une limite de domaine d'après CATH. Sont donc exclues de cette définition, les séquences « linkers » présentant des structures secondaires régulières. Comme illustré sur la Figure 67 et en accord avec les résultats de Tanaka [TANAKA, T. et al., 2003], la longueur moyenne des boucles « linkers » est plus élevée que celles des boucles « non-linkers ».

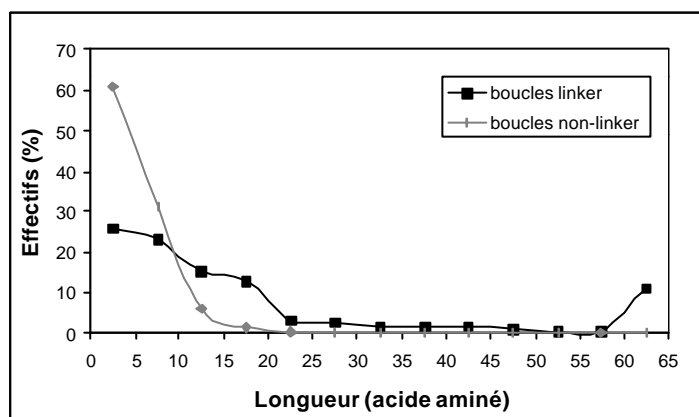


Figure 67 : Distribution des longueurs des boucles « linkers » et des boucles « non-linkers » (boucles simples).

Néanmoins, les fréquences des acides aminés ne diffèrent pas significativement (Figure 68), exceptés pour les acides aminés G, D et N qui sont plus fréquents dans les boucles « non-linkers ». Ces acides aminés sont des marqueurs de boucles, comme le montre leurs propensions pour les différents états de structure secondaire calculées à partir d'une banque PDB non redondante [CALLEBAUT et al., 1997] (cf. Figure 69). A l'opposé, certains acides aminés hydrophobes (phénylalanine, valine et leucine) sont plus fréquents dans les boucles linkers.

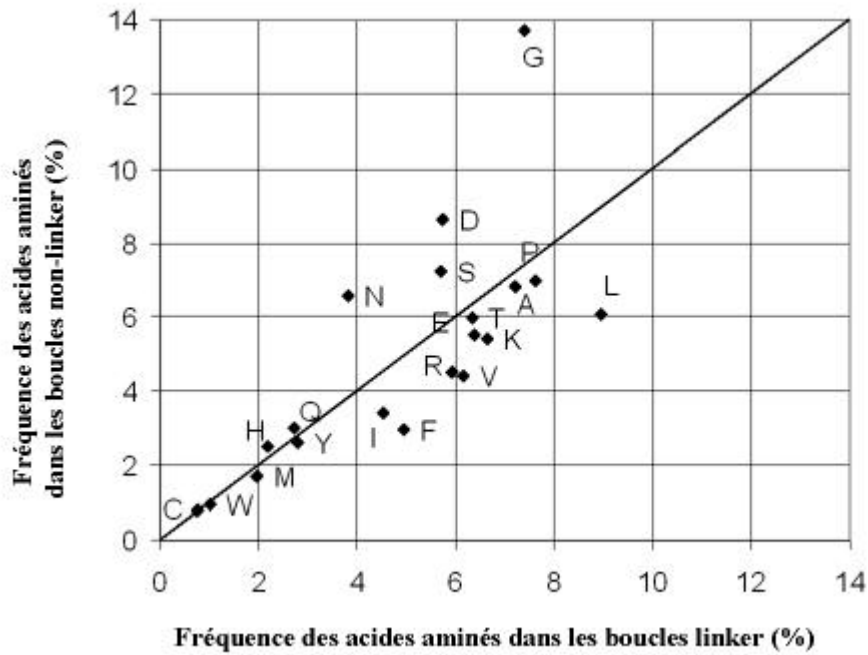


Figure 68 : Corrélation entre la composition en acides aminés des boucles « linkers » et des boucles « non-linkers ».

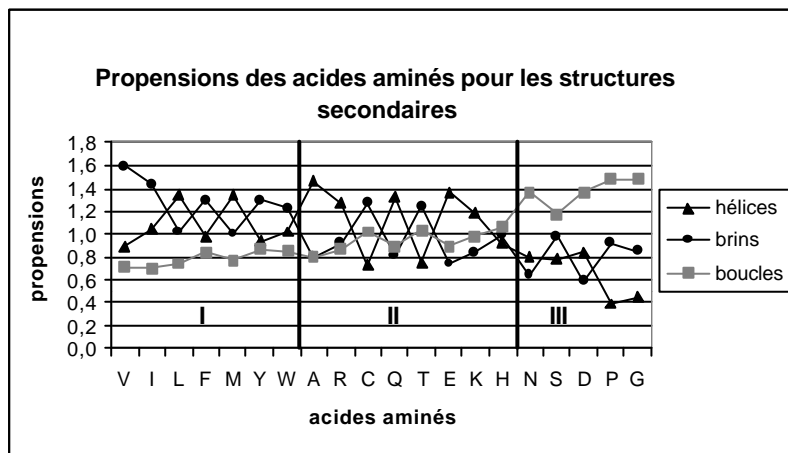


Figure 69 : Propensions des acides aminés pour les structures secondaires [CALLEBAUT, I. et al., 1997a].

Les acides aminés sont classés par ordre décroissant de la somme de leurs trois propensions, hélice + brin + boucle, de la valine (V) à la glycine (G). On peut distinguer trois classes d'acides aminés. La classe I correspond aux résidus participant majoritairement aux structures secondaires régulières; ils sont les plus hydrophobes et définissent l'alphabet standard HCA. La classe II rassemble les résidus présentant une préférence marquée pour un seul type de structure secondaire régulière. Ils forment une classe intermédiaire. La classe III présente les acides aminés à forte propension pour les boucles. Ils délaissent les structures secondaires régulières et sont considérés comme des constructeurs de boucles.

D'une part, comme en témoigne la Figure 70, les faux positifs (séquences prédites comme « linkers » mais qui en réalité ne le sont pas) contiennent des fréquences relativement

plus élevées en acides aminés qui ne sont pas des hydrophobes forts, mais qui sont souvent présents dans les structures en hélice α (alanine) ou brins β (thréonine, cystéine).

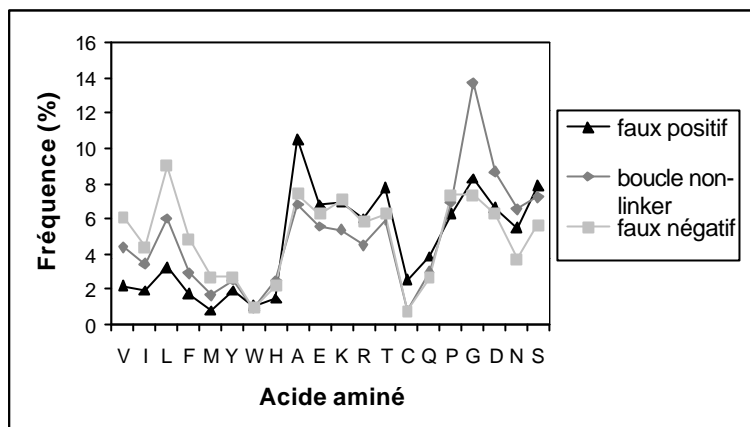


Figure 70 : Composition en acides aminés des boucles non-linkers, des faux positifs et des faux négatifs de notre banque de séquences pluridomaines.

Ces mêmes faux positifs contiennent moins d'acides aminés de la classe III (les marqueurs de boucle : P, G, D, N et S). D'autre part, les faux négatifs (linkers non prédits par DomHCA) contiennent des fréquences plus élevées en acides aminés hydrophobes forts (Figure 70). L'étude de ces caractéristiques de séquence pourrait permettre d'affiner la prédiction brute réalisée par DomHCA.

7.2.4 Comparaison avec d'autres méthodes

Il est difficile de comparer les résultats obtenus par les différentes méthodes de prédiction de limites de domaines ou de linkers. En effet, ces méthodes sont basées sur des caractéristiques différentes et différents critères doivent être utilisés pour les comparer. Néanmoins, nous avons effectué une comparaison grossière de plusieurs méthodes de prédiction avec DomHCA, à partir des jeux de séquence de nos banques test.

Un exemple de comparaison des résultats de prédiction appliquée à une séquence de notre jeu de données est montré en Figure 71. Les résultats des méthodes DomHCA, DomPred [MARSDEN, R. L. et al., 2002], DomCut [SUYAMA, M. et al., 2003] et du serveur Prelink [COEYTAUX, K. et al., 2005] sont présentés. Nous remarquons que DomHCA prédit exactement les limites des régions structurées, telles que répertoriées dans la banque CATH. Cependant, à l'instar des autres méthodes de prédiction, DomHCA ne peut subdiviser la première région structurée en ses trois domaines constitutifs. Etant donné qu'aucune des méthodes ne prédit de linkers entre les trois premiers domaines assignés par CATH et que, par contre, toutes les méthodes semblent prédire un linker entre le troisième et

le quatrième domaine, nous pouvons penser que les trois premiers domaines CATH de l'aconitase forment une structure compacte qui ne permet pas de les distinguer séparément.

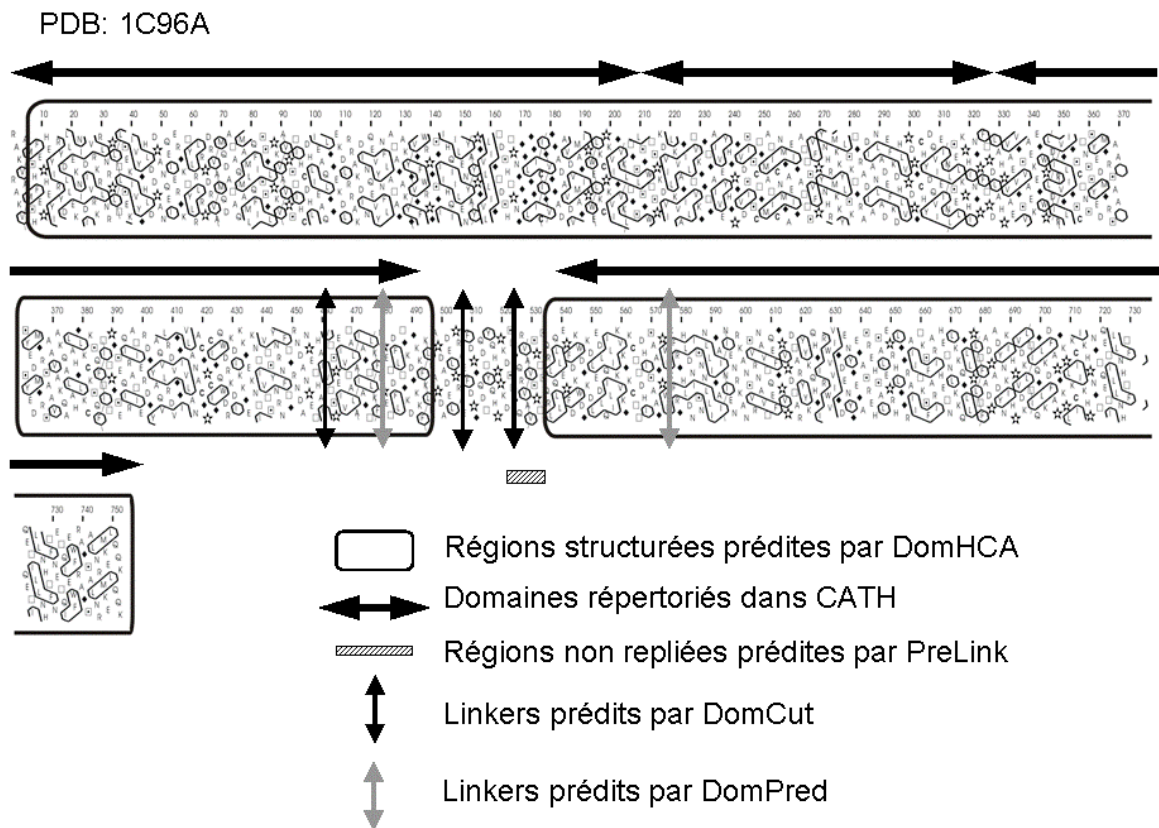


Figure 71 : Comparaison de méthodes de prédiction sur la protéine aconitase (1C96), chaîne A. Les prédictions DomHCA sont en accord avec celles de PreLink et celles de DomCut. DomCut prédit cependant un linker au sein d'un domaine structuré, ainsi que le fait DomPred.

Finalement, nous avons évalué DomHCA par rapport aux résultats de prédiction de cibles de CAFASP4 (<http://cafasp4.bioinformatics.buffalo.edu/dp/update.html>). Le but de CAFASP (Critical Assessment of Fully Automated Structure Prediction) est d'évaluer la performance des serveurs de prédiction automatique accessibles à la communauté scientifique [SAINI, H. K. et al., 2005]. 58 cibles (dont 41 présentent un domaine unique et 17 deux domaines) ont été soumises à 12 méthodes de prédiction. Nous avons comparé les prédictions faites par DomHCA sur ces 58 cibles. Une prédiction est considérée comme correcte si le nombre de domaines prédits est correct. Dans le cas des cibles ayant deux domaines, les domaines prédits peuvent être continus avec une absence de séparation entre les deux domaines. Pour chaque méthode, la sensibilité, la spécificité et le score de recouvrement (overlap score) ont été calculés et représentés. La méthode DomHCA a détecté correctement

36 cibles ayant un seul domaine sur 41 (87.8 % de bonnes prédictions). Cinq de ces cibles sont détectées comme ayant deux domaines (Figure 72). DomHCA semble obtenir de bons résultats comparativement aux autres méthodes.

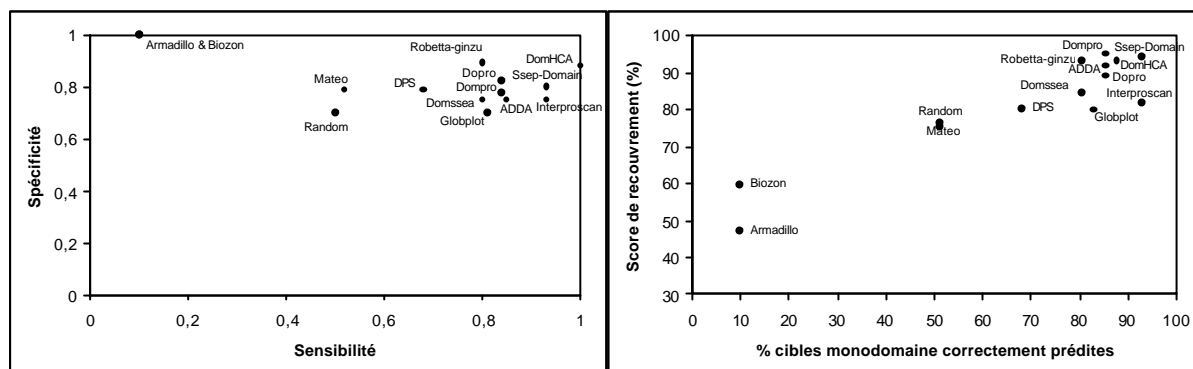


Figure 72 : Résultats des méthodes de prédiction testées dans CAFASP.

Le contrôle (Random) correspond à une valeur au hasard entre la valeur du contrôle 1 et la valeur du contrôle 2. Le contrôle 1 correspond à la prédiction de toutes les cibles comme des protéines ayant un seul domaine et le contrôle 2 correspond à la prédiction de toutes les cibles comme des protéines ayant deux domaines.

7.3 Application aux séquences de *Plasmodium falciparum*

La particularité remarquable des génomes de *Plasmodium falciparum* et *Dictyostelium discoideum* réside en leur richesse en nucléotides A+T [GARDNER, M. J. et al., 2002; EICHINGER, L. et al., 2005]. Il en résulte une abondance d'un certain nombre d'acides aminés (asparagine, lysine, isoleucine, tyrosine) (voir Figure 73 pour le cas de *Plasmodium falciparum*). Au début de cette étude, seul le génome de *Plasmodium falciparum* était disponible, nous l'avons donc choisi comme organisme d'étude.

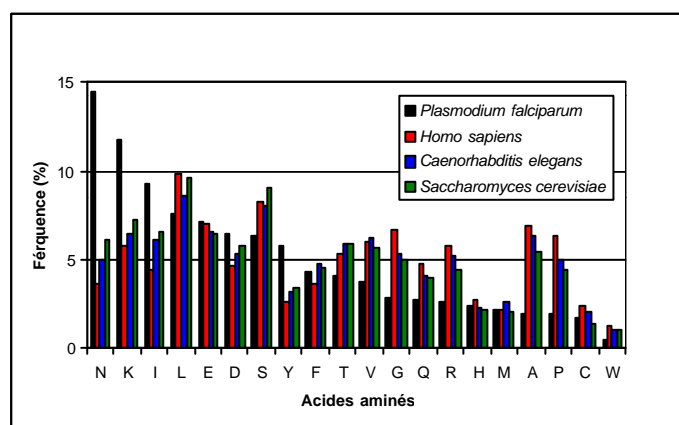


Figure 73 : Comparaison de la distribution des acides aminés dans les protéomes (protéines prédites) de *Plasmodium falciparum*, *Homo sapiens*, *Caenorhabditis elegans* et *Saccharomyces cerevisiae* (5334, 32035, 21629 et 6699 séquences, respectivement) [CALLEBAUT, I. et al., 2005]. Les acides aminés sont classés par ordre de fréquence décroissante chez *Plasmodium falciparum*.

La richesse en nucléotides A+T entraîne des biais dans les séquences qui rendent difficile l'utilisation des procédures automatiques de comparaison entre séquences. Ainsi, plus de 60% des protéines restent orphelines d'informations structurales et/ou fonctionnelles déduites de la comparaison des séquences. Il est probable qu'un nombre important de ces protéines orphelines correspond en réalité à des orthologues "cachés". Une inspection des tracés HCA de quelques séquences représentatives de *Plasmodium falciparum* montre cependant que les caractéristiques hydrophobes des domaines globulaires sont maintenues (pourcentage global d'acides aminés hydrophobes conservé, Figure 73) et que nombre des acides aminés surreprésentés sont localisés dans des régions de faible complexité, en dehors des domaines globulaires. De telles observations ont été également réalisées par d'autres groupes [ARAVIND, L. et al., 2003]. Ce sont ces régions de faible complexité qui affectent particulièrement l'efficacité des méthodes automatiques de comparaison des séquences. L'utilisation de DomHCA permet de s'abstraire de ces régions de faible complexité puisqu'elles ne sont généralement pas incluses dans les régions structurées prédites par le programme. Il permet de plus de traiter rapidement l'ensemble des séquences d'un génome.

Ainsi, l'ensemble du génome de *Plasmodium falciparum* a été soumis à ce découpage en régions structurées. 18 808 segments structurés et 866 segments «pseudo-structurés» ont été recensés. Nous présentons cette étude au chapitre VII.

8 Discussion et Conclusion

Du fait de l'organisation modulaire des protéines, il est intéressant de prédire les limites de domaines biologiquement actifs. Ce concept peut être particulièrement intéressant dans le cadre de la détermination de la structure 3D de protéines par RMN ou par cristallographie, afin de diminuer la taille de la molécule étudiée (RMN) et d'augmenter les chances d'obtention de cristaux (cristallographie).

Sans identifier précisément les domaines d'une séquence protéique, il peut être important dans un premier temps de prédire les limites des régions structurées et non-structurées. Nous avons ainsi développé un outil automatique, DomHCA, basé sur la distribution des acides aminés hydrophobes forts (V, I, L, F, M, Y et W) et des amas hydrophobes HCA de la séquence, qui prédit les régions structurées dans une protéine. La méthode HCA permet en effet d'identifier les limites de nombreux domaines, mais nécessite souvent les compétences d'un expert. DomHCA prédit de façon automatique les régions potentiellement structurées, domaines globulaires et/ou passages membranaires. Ces dernières

régions peuvent en effet être distinguées, une forte fréquence moyenne ou locale en acides aminés hydrophobes est généralement observée, respectivement, dans les passages membranaires multiples et les passages membranaires uniques. D'autres hypothèses peuvent être ajoutées aux prédictions DomHCA, comme la présence de brins β caractéristiques des porines, qui présentent des segments locaux riches en acide aminé glycine, celle de peptides de fusion qui présentent des caractéristiques de séquence particulières [DEL ANGEL, V. D. et al., 2002] ou de régions de répétitions (cf. chapitre VI). Ces hypothèses peuvent être affinées par des outils plus spécialisés (par exemple pour les hélices transmembranaires (PHDhtm [ROST, B. et al., 1996], TMHMM [MOLLER, S. et al., 2001], HMMTOP [TUSNADY, G. E. et al., 2001]; pour les protéines membranaires en tonneaux bêta [JACOBONI, I. et al., 2001; GROMIHA, M. M. et al., 2004; NATT, N. K. et al., 2004] et pour les peptides signaux [NIELSEN, H. et al., 1999]).

Le calcul du pourcentage moyen d'acides aminés hydrophobes a permis de mettre en évidence que les domaines globulaires ont des valeurs voisines de 33% alors que des plus faibles valeurs peuvent être éventuellement associées à des régions intrinsèquement désordonnées (voir par exemple le domaine pKID de la protéine CREB page 94). Il semble possible de détecter ces régions désordonnées des régions structurées en identifiant localement de plus faibles fréquences en acides aminés hydrophobes (en utilisant une fenêtre glissante de taille similaire à la taille des régions intrinsèquement désordonnées, environ 50 acides aminés [UVERSKY, V. N. et al., 2000]).

DomHCA est efficace dans le cas de protéines « monodomaines » pour lesquelles nous avons montré un bon recouvrement entre les limites des domaines prédites par DomHCA et les limites des chaînes observées dans la Protéine Data Bank. DomHCA prédit les linkers dans les protéines multidomaines dans certains cas. En effet, de nombreux linkers contiennent soit des structures secondaires, qui ne peuvent pas être distinguées de celles des domaines globulaires classiques, soit correspondent à de petites séquences non distinguables des boucles « non-linkers » des protéines. Néanmoins, ces chaînes multidomaines ne doivent souvent pas être découpées pour préserver leur intégrité fonctionnelle, spécialement quand elles forment une structure compacte comme celle de la glucuronidase (Figure 42, page 89). Dans ce contexte, il ne serait pas judicieux de suggérer un découpage possible de cette chaîne polypeptidique pour une caractérisation fonctionnelle et structurale. En combinant plusieurs méthodes de prédictions de limites de domaines, il serait possible d'améliorer la fiabilité des prédictions de domaines [SAINI, H. K. et al., 2005].

Finalement, DomHCA est un outil particulièrement intéressant pour extraire les domaines structurés des séquences orphelines, dont beaucoup sont présentes dans des génomes ayant des biais de composition en acides aminés, comme par exemple *Plasmodium falciparum* ou *Dictyostelium discoideum*. La sélection des régions structurées prédites par DomHCA, comme séquences sondes pour des recherches de similarité ou de domaines, peut alors permettre d'identifier dans le bruit de fond des séquences alors que la requête n'aurait pas été concluante si elle avait été faite sur la séquence entière. Un exemple d'identification de domaine dans la protéine MAL8P1.111 de *Plasmodium falciparum* est présenté en Figure 74.

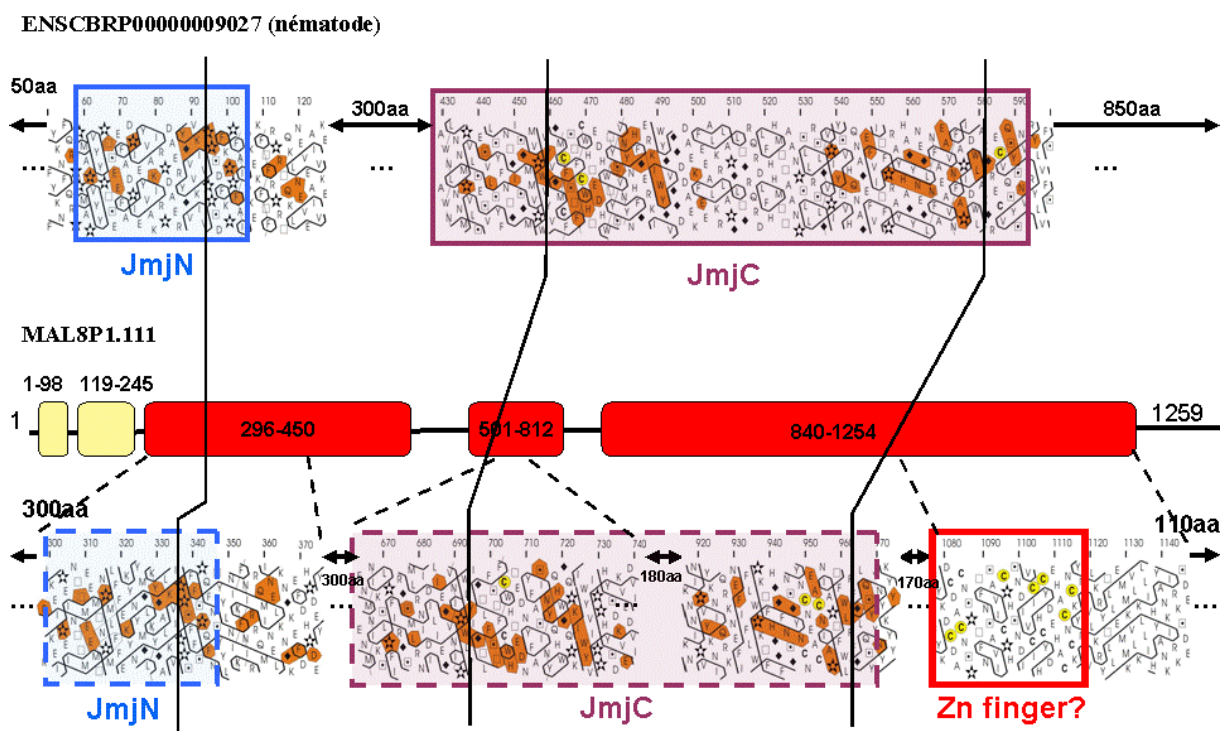


Figure 74 : Exemple de découpage permettant une meilleure recherche automatique de domaines. La protéine MAL8P1.111 de *Plasmodium falciparum* compte 1259 acides aminés. Après traitement avec DomHCA (Tableau 10), la protéine MAL8P1.111 est découpée en cinq segments structurés (1-90, 119-245, 296-450, 501-812, 840-1259). La recherche de domaines sur la protéine entière dans la banque SMART détecte le seul domaine alors qu'après le découpage, trois domaines sont identifiés (JmjN, JmjC et Zf-C5HC2) avec des scores significatifs.

Le traitement de la séquence MAL8P1.111 par DomHCA permet d'isoler cinq régions structurées potentielles. Soumises séparément à des recherches automatiques de domaines dans les banques de domaines comme SMART ou PFAM [LETUNIC, I. et al.,

2002], elles permettent d'identifier des domaines non détectés lors de la recherche sur la protéine complète, comme par exemple ici le domaine JmjN (Tableau 10).

Tableau 10 : Domaines identifiés dans MAL8P1.111 à partir d'une recherche de type HMMER dans les banques de domaines SMART et PFAM.

Fragment testé	SMART + PFAM	Score
1-90	/	/
119-245	/	/
296-450	JmjN (302-343)	$8,17.10^{-3}$
501-812	JmjC (660-791)	$9,24.10^{-7}$
840-1259	JmjC (840-974)	$5,17.10^{-2}$
	Zf-C5HC2 (1082-1133)	$1,20.10^{-2}$
Protéine entière	JmjC (660-974)	$1,40.10^{-33}$

Enfin, la rapidité de DomHCA est aussi un avantage pour le traitement à grande échelle des génomes et pour une analyse globale de l'architecture en domaines des séquences. DomHCA sera prochainement accessible sur le serveur du laboratoire.

Chapitre VI

Autres développements et caractérisation de régions spécifiques dans les protéines

1 *Introduction*

L'analyse de la texture peut se traduire par la recherche de motifs ou de domaines particuliers. L'extraction de telles régions pourrait permettre des comparaisons de motifs ou de domaines entre protéines, l'identification de nouvelles fonctions dans des séquences ne présentant aucune similarité avec les protéines contenues dans les banques classiques. De plus, certains motifs, par leur composition, voire la répétition d'une même séquence, peuvent générer des biais lors des requêtes automatiques de recherche de similarités. Ainsi, détecter ces régions de répétitions ou de composition particulière permettrait d'« épurer » ces séquences et ainsi d'identifier de nouvelles similitudes par les méthodes classiques de recherche de similitudes, jusqu'alors non détectées en raison de la focalisation du moteur de recherche sur les premières régions. Nous présentons ci-après les développements poursuivis en vue d'identifier des régions de texture particulière.

2 *Identification de régions de répétition*

2.1 Méthode d'identification

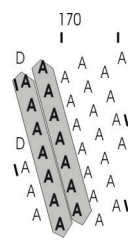
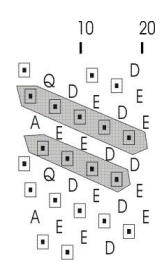
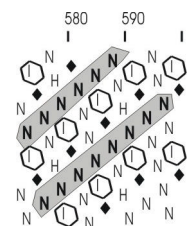
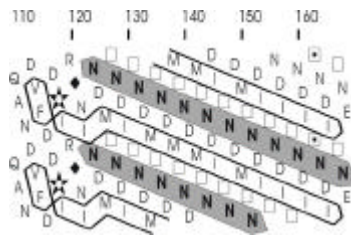
2.1.1 Harmonique standard

Des régions de répétitions régulières peuvent être aisément identifiées visuellement sur le tracé HCA bidimensionnel. Nous les appellerons « harmoniques ». Elles sont formées sur la répétition d'un motif de base. Quatre motifs de base ont été identifiés: « λ », « $\lambda x \lambda$ », « $\lambda x x \lambda$ » et « $\lambda x x x \lambda$ » (λ étant un acide aminé quelconque toujours retrouvé en une ou plusieurs positions précises et x n'importe quel acide aminé y compris celui symbolisé par λ).

Les harmoniques de niveau 1 sont celles constituées à partir du motif de base « λ »; de niveau 2, celles constituées à partir du motif de base « $\lambda x \lambda$ »; de niveau 3, celles

constituées à partir du motif de base « λxxλ » et de niveau 4, celles constituées à partir du motif de base « λxxxλ ». Un exemple de chaque harmonique est présenté dans le Tableau 11.

Tableau 11 : Présentation des quatre types d'harmoniques.

Type	Tracé HCA correspondant	Base de l'harmonique
'λ'		O00358 (<i>Homo sapiens</i>) AAAAAAAAAAAAAAAA
'λxλ'		S0000023 (<i>Saccharomyces cerevisiae</i>) SASQSESDSESESDSDS
'λxxλ'		PFI0510c (<i>Plasmodium falciparum</i>) NIGNNHNIGNNNNIGNNHNIGNNNN
'λxxxλ'		MAL8P1.111 (<i>Plasmodium falciparum</i>) NDMTNDITNDMTNDMTNDMTNDITNDMT NDITNDITNDISNDITN

Une harmonique de niveau supérieur peut se superposer à une harmonique de niveau inférieur. Cependant, nous retiendrons comme harmonique celle correspondant à l'harmonique de plus grande taille (Figure 75), et à taille identique, celle de plus haut niveau (Figure 76).

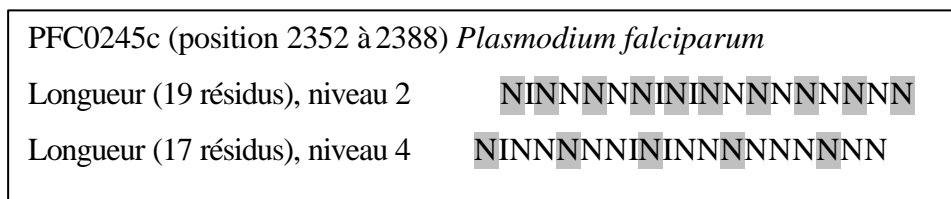


Figure 75 : Comparaison de deux harmoniques de longueurs et niveaux différents. L'harmonique retenue est celle de longueur 19 et de niveau 2.

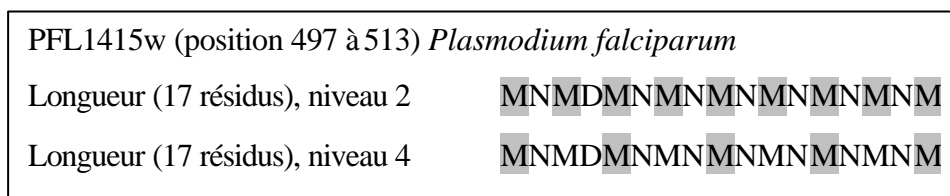


Figure 76 : Comparaison de deux harmoniques de niveaux différents. L'harmonique retenue est celle de longueur 17 et de niveau 4.

Dans certains cas, deux harmoniques peuvent se chevaucher. L'harmonique considérée est l'harmonique totale englobant les deux harmoniques quels que soient leurs longueurs et leurs niveaux (Figure 77).

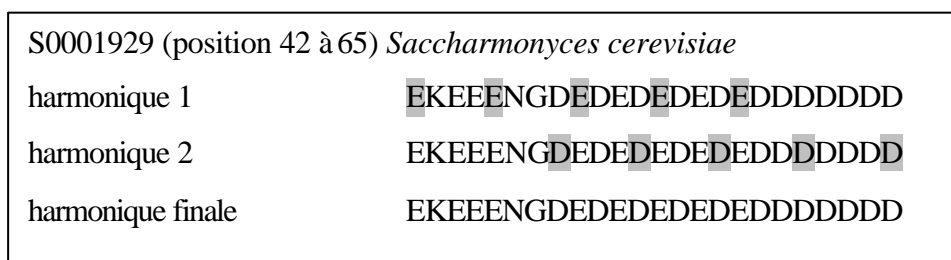


Figure 77 : Exemple de deux harmoniques se chevauchant, celles-ci sont assemblées dans notre traitement. L'harmonique retenue est l'harmonique finale.

Une procédure automatique a été mise au point pour détecter les harmoniques. Elle est basée sur les principes suivants :

- une harmonique doit renfermer au moins 4 fois successivement le motif de base qui la compose (pour les motifs de base «λxλ», «λxxλ» et «λxxxλ») et 10 fois pour le motif de base «λ»,
- si plusieurs harmoniques de différents niveaux se superposent, l'harmonique détectée est toujours celle de plus grande taille et de plus haut niveau,
- si deux harmoniques se chevauchent, l'harmonique détectée englobe la totalité des deux harmoniques,

- une harmonique commence et finit toujours par un acide aminé « λ ».

2.1.2 Harmonique hydrophobe

Après avoir rencontré un autre type de répétition dans les séquences protéiques de *Plasmodium falciparum*, nous avons ajouté une flexibilité dans notre algorithme de détection. En effet, nous avons défini un nouveau type d'harmonique : l'harmonique hydrophobe. C'est une harmonique standard de n'importe quel niveau dont l'acide aminé « λ » n'est pas unique mais représente n'importe quel acide aminé hydrophobe (V, I, L, F, M, Y ou W). Les acides aminés « x » peuvent correspondre à n'importe quel acide aminé y compris des acides aminés hydrophobes.

Prenons l'exemple de l'harmonique λxxλxxλxxλxxλ. Le premier résidu « λ » peut correspondre à une valine alors que le second résidu « λ » (et les suivants) peuvent correspondre à une leucine ou à tout autre acide aminé hydrophobe (V, I, L, F, M, Y ou W). Un exemple d'harmonique hydrophobe, détectée dans la protéine hypothétique PF11_0204 de *Plasmodium falciparum*, est présenté en Figure 78.

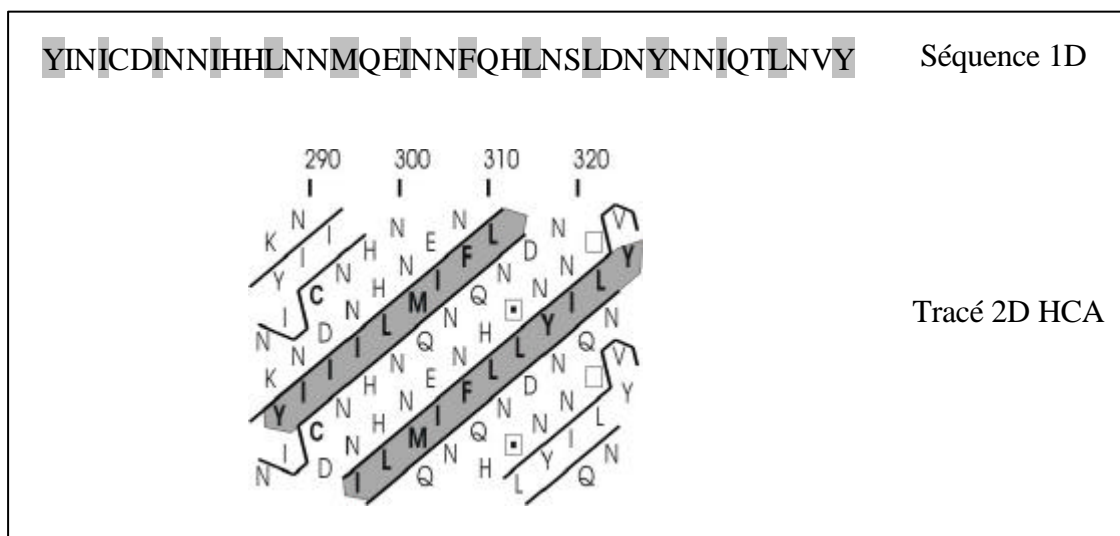


Figure 78 : Exemple d'harmonique hydrophobe (PF11_0204).

Les résidus en gris sont les résidus « l » de l'harmonique, les résidus entre les deux régions grisées sont les résidus « x » de l'harmonique.

2.1.3 Harmonique dégénérée

D'après les définitions établies, une harmonique commence et finit toujours par un acide aminé « λ ». Chaque harmonique est formée en général de quelques acides aminés. Dans l'exemple ci-dessous (Figure 79), les acides aminés constitutifs de l'harmonique sont la

sérine, la thréonine, l'alanine et la valine. L'acide aminé qui suit l'harmonique dans la séquence est une thréonine. Nous avons admis que l'harmonique pouvait être prolongée tant que les acides aminés qui la suivent sont des constituants majeurs de l'harmonique. L'harmonique initiale **SESSSTSASTSASASTSASTSASTSVSTSVSTSVSTAS** devient l'harmonique dégénérée **SESSSTSASTSASASTSASTSASTSVSTSVSTSVSTASTT**.

PFB0580w (position 263 à 303) <i>Plasmodium falciparum</i>
SESSSTSASTSASASTSASTSASTSVSTSVSTSVSTASTT ...
SESSSTSASTSASASTSASTSASTSVSTSVSTSVSTASTT ..

**Figure 79 : Exemple d'harmonique dégénérée (PFB0580w).
Les harmoniques initiales et après élongation sont encadrées.**

Toutes les harmoniques (classiques, hydrophobes et dégénérés) sont recherchées dans les séquences des protéines et peuvent être caractérisées par plusieurs critères.

2.2 Caractérisation des harmoniques

2.2.1 Répétitions structurées, pseudo-structurées

Nous classerons les répétitions en trois catégories principales: une région de répétition incluse dans un segment structuré est dite « répétition structurée (RS) », une région de répétition chevauchant un segment structuré est dite « répétition pseudo-structurée (RPS) » et une région de répétition hors segment structuré est nommée de « répétition (R) » (Tableau 12).

Tableau 12 : Exemples de chacun des trois types de régions répétitives.

Type	Code de la protéine	Bornes	Séquence de la région de répétition
RS	ENSP00000167586	81/109	GSGFGGGYGGGLGAGLGGGFGGGFAGGDG
RPS	ENSP00000223269	180/196	QLQLQLQQQQQQQQ
R	ENSP00000005545	226/255	DEEELLEDEEDEDEEEELLEDEEELLEDD

2.2.2 Répétitions simples ou mixtes

Les régions de répétition basées sur l'harmonique de niveau 1 ou constituées majoritairement par un acide aminé (fréquence supérieure ou égale à 60%) sont dites « simples » ; les autres répétitions sont dites « mixtes » (Tableau 13).

Tableau 13 : Illustration d'harmoniques structurées simples et mixtes.

La région de répétition de la protéine ENSP00000167586 (*Homo sapiens*) renferme 67% de résidus glycine. La région de répétition de la protéine ENSP00000007740 (*Homo sapiens*) renferme 38 % de résidus valine, 15% de résidus glycine, 8% de résidus acide glutamique et 38% de résidus histidine.

Type	Complexité	Code de la protéine	Bornes	Séquence de la région répétitive
S	Simple	ENSP00000167586	81/109	GSGFGGGYGGGLGAGLGGGF GGGFAGGDG
S	mixte	ENSP00000007740	144/158	VGHVGHVGHVEHVH

3 Identification de peptides de fusion

3.1 Introduction

La fusion entre membranes est un évènement très important qui se produit en continu dans toutes les cellules. Ainsi, certaines protéines ont la capacité d'induire la fusion entre deux membranes phospholipidiques, comme par exemple les glycoprotéines des virus enveloppés. Ces glycoprotéines permettent aux virus d'injecter leur génome dans le cytoplasme des cellules hôtes. En général, les glycoprotéines virales fusogènes subissent un changement de conformation qui leur permet d'exposer à leur surface un domaine hydrophobe (le peptide de fusion) qui a la propriété de pénétrer dans la bicouche des membranes et de provoquer ultérieurement le rapprochement des deux membranes.

Les peptides de fusion sont de petites séquences, entre 20 et 30 acides aminés relativement hydrophobes et qui sont conservées à l'intérieur d'une famille de virus, mais pas entre les familles [PECHEUR, E. I. et al., 2000]. Ils s'insèrent dans la bicouche lipidique et provoquent des changements de disposition des lipides entraînant ultérieurement la formation du pore de fusion. Les peptides de fusion peuvent être classés selon leur position dans la glycoprotéine (amino-terminale ou interne), leur pH de fusion (neutre ou acide), et de la présence ou de l'absence de motifs « coiled-coil » en aval dans la protéine fusogène [DOMINGUEZ DEL ANGEL, V. 2003]. Les glycoprotéines virales qui font l'objet d'un clivage (comme celle du virus de l'immunodéficience humaine (HIV), l'Hémagglutinine de l'Influenza (HA) ou encore celle du virus de la leucémie humaine (HTLV)) disposent ainsi de leur peptide de fusion en partie amino-terminale de la protéine transmembranaire, suivi d'une structure « coiled-coil » (tresse d'hélice). A l'inverse, les protéines non clivées ont un peptide de fusion interne, non suivi d'un motif de type « coiled-coil »

Une étude précédente menée au laboratoire par V. Dominguez Del Angel sur un ensemble de référence de 39 peptides de fusion bien caractérisés et non redondants [DEL ANGEL, V. D. et al., 2002] a permis de décrire les peptides de fusion au moyen des caractéristiques suivantes :

- une fréquence des résidus A et G supérieure à 23%,
- une fréquence des résidus V, I, L, F, M, Y et W supérieure à 33%,
- une fréquence des résidus Y, W, C, P, D, E, K, R et H inférieure à 19%,
- une longueur entre 22 et 33 acides aminés,
- une forte fréquence en résidus V, I, L, F, M, Y, W, A, G, S et T, avec une prédominance de résidus hydrophobes fort en N-terminal du peptide et d'alanine et glycine en C-terminal.

Un premier criblage avait été développé sur ces bases [DEL ANGEL, V. D. et al., 2002]. Dans notre découpage en régions structurées des séquences, il nous est apparu intéressant de donner une indication sur la présence éventuelle d'un peptide de fusion au sein de la région structurée prédite. Dans ce contexte, nous avons établi une procédure automatique, reprenant les caractéristiques mises en évidence par V. Dominguez Del Angel.

3.2 Méthode de détection

Un algorithme basé sur les caractéristiques précédemment citées a été mis au point. Il découpe les séquences en fenêtres glissantes de 17 acides aminés (longueur utilisée pour la recherche de segments structurés) et calcule les fréquences en acides aminés A et G, en acides aminés V, I, L, F, M, Y et W, et en acides aminés Y, W, C, P, D, E, K, R et H pour chaque fenêtre. Les régions susceptibles de correspondre à un peptide fusion sont des régions de plus de 10 acides aminés de long pour lesquelles la fréquence en acides aminés cités ci-dessus est supérieure (A, G / VILFMYW) et inférieure (Y, W, C, P, D, E, K, R et H) aux valeurs seuils observées. Le peptide potentiel est étendu en amont et en aval de 6 acides aminés. Ces bornes de début et de fin sont ajustées en fonction de la présence de résidus V, I, L, F, M, Y, W, A, G, S ou T dans les 10 acides aminés de début et de fin. Ensuite, les fréquences cumulées en acides aminés A et G, et en acides aminés V, I, L, F, M, Y, W sont calculées. L'hypothèse d'un peptide de fusion se trouve renforcée si les fréquences cumulées pour les acides aminés VILFMYW sont décroissantes et si les fréquences cumulées pour les acides aminés AG sont croissantes. Un exemple de détection d'un peptide de fusion potentiel est présenté en Figure 80.

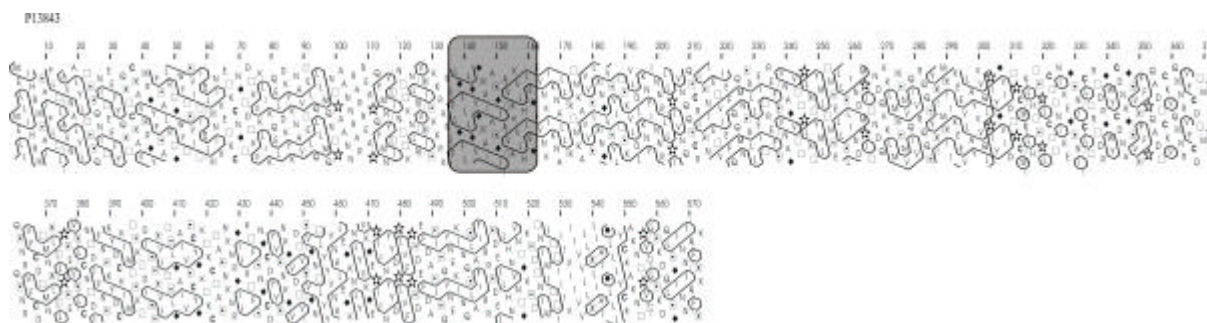


Figure 80 : Prédiction d'un peptide de fusion.

La glycoprotéine de virus respiratoire syncytial humain (VGLF_HRSV1, code P13843) contient un peptide de fusion entre les acides aminés 137 et 160. A l'aide de notre procédure, nous prédisons un peptide de fusion potentiel entre les résidus 137 et 162 (région grise encadrée).

4 Prédiction de la classe de repliement (A, B, C, D)

DomHCA permet de prédire les régions structurées d'une protéine et de donner des indications quant à la présence éventuelle de passages membranaires, de peptides de fusion ou de régions de répétitions (cf. ci-dessus). Dans cette même optique, il serait intéressant de pouvoir indiquer le type de repliement (alpha, bêta, alpha+bêta, alpha/bêta) d'une région prédite comme structurée. Nous avons donc évalué plusieurs approches tendant à donner une information sur le type de repliement de la région structurée, reposant sur l'analyse des propensions des différents acides aminés pour les différents types de structures secondaires, de la distribution de ces mêmes acides aminés dans les différents types de repliements, des amas hydrophobes HCA (corrélation avec structures secondaires, distribution par rapport à l'aléatoire, longueur et nombre des amas). Nous avons pris en référence un ensemble de protéines présentant un repliement tout α , tout β , α/β et $\alpha+\beta$ (classes A, B, C et D de SCOP constituées précédemment).

4.1 Propension des acides pour un type de structure secondaire

Beaucoup d'acides aminés présentent des préférences pour un type de structures secondaires (hélice, brin ou boucle). La propension d'un acide aminé pour une structure correspond au rapport de sa fréquence dans cette structure et de la fréquence de tous les acides aminés pour cette même structure. Prenons l'exemple de la valine (V), qui apparaît nVa dans les hélices, nVb dans les brins et nVc dans les boucles. Soit $nV = nVa + nVb + nVc$, le nombre total de valines observées et $N = Na + Nb + Nc$, le nombre total de résidus considérés, la propension de la valine pour l'hélice se calcule ainsi:

$$\text{propension} = \frac{\frac{nVa}{nV}}{\frac{Na}{N}}$$

Nous avons présenté ces propensions sur la Figure 69 (page 118) et pouvons constater que de nombreux acides aminés ont des préférences pour un type de structure secondaire. Nous avons donc souhaité poursuivre notre analyse afin d'appréhender dans quelle mesure les distributions des acides aminés pouvaient rendre compte des différents types de repliements α , β , $\alpha+\beta$ et α/β .

4.2 Distribution des acides aminés au sein des différents types de repliements

Etant donné la préférence de certains acides aminés pour un type de structures secondaires, nous avons calculé la distribution des acides aminés pour chaque banque A, B, C et D de SCOP (Figure 81).

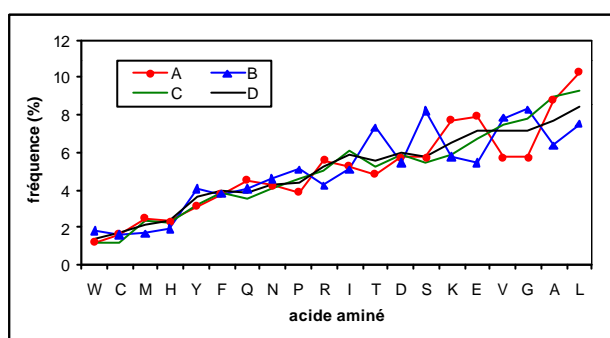


Figure 81 : Distribution des acides aminés dans les banques SCOP A, B, C et D. Les acides aminés sont classés par ordre croissant de leurs fréquences dans les quatre banques A, B, C et D.

Les distributions des banques A, B, C et D sont relativement similaires mis à part pour les acides aminés A, E, K, R et T. Cependant, ces différences sont faibles et semblent peu exploitables pour attribuer une tendance de repliement à un domaine à partir de sa distribution en acides aminés, sauf peut-être pour discriminer les banques A et B. Aucune discrimination n'est possible entre les banques C et D.

La répartition des acides aminés au sein des différents types de structure secondaires ne permettant pas de caractériser ni les structures secondaires ni leur enchaînement dans la

protéine, nous avons étudié les répartitions des amas hydrophobes dans nos 4 banques A, B, C et D.

4.3 Dictionnaire d'amas hydrophobes

Dans le cadre de la thèse de K.Le Tuan [LE TUAN, 2003], la fréquence d'apparition de chacun des amas hydrophobes (allant jusqu'à l'amas 110110011), dans les structures secondaires alpha, bêta et boucle a été calculée à partir d'une banque PDB à 90% de redondance (10 631 chaînes protéiques différentes, 4704 protéines [LE TUAN, 2003]). Les résultats obtenus sont comparables à ceux issus de la compilation d'une banque non redondante (25% d'identité). La redondance permet cependant d'augmenter le nombre de représentants de chaque amas, et ainsi d'avoir accès à une information statistiquement pertinente pour un plus grand nombre d'amas. Le dictionnaire d'amas avec les fréquences correspondantes des structures secondaires observées est présenté en Annexe.

4.3.1 Règles d'attribution des états A, B et ?

Pour un amas hydrophobe donné, si la fréquence est supérieure à 50% pour un des deux états alpha ou bêta, nous considérons que l'amas peut être généralement en l'état «A» ou «B» («A» correspond à l'état alpha, «B» à l'état bêta). Si la fréquence est comprise entre 40% et 50% pour un des deux états, mais qu'elle est supérieure de 10% à la fréquence de l'autre état, on attribue également à l'amas le premier état «a» ou «b» («a» correspond à l'état alpha, «b» à l'état bêta; les lettres minuscules indiquant une probabilité moindre). L'état indéterminé, noté «?», correspond à une fréquence inférieure à 40% dans chacun des états alpha ou bêta. Les deux plus petits amas hydrophobes, de taille 1 et 2 (1 et 11), sont majoritairement associés à des boucles.

4.3.2 Prédiction de la tendance de repliement de régions structurées

Une région structurée est représentée comme une succession d'amas hydrophobes, et donc d'états attribués à chacun des amas (A, B, a, b ou ?). La succession des différents états correspondant au segment structuré est interprétée de la manière suivante :

- S'il n'y a que des «?», la tendance de repliement du domaine est indéterminée,
- S'il y a un ou plusieurs «?» et au moins un ou plusieurs «A,a» et aucun «B,b», la tendance est alpha,
- S'il y a un ou plusieurs «?» et au moins un ou plusieurs «B,b» et aucun «A,a», la tendance est bêta,
- S'il y a des «A,a» et des «B,b» et aucun ou plusieurs «?», plusieurs cas sont considérés :

- Si la fréquence de A est inférieure à 10% et que la fréquence de B est au moins supérieure à 10%, la tendance est bêta,
- Si la fréquence de B est inférieure à 10% et que la fréquence de A est au moins supérieure à 10%, la tendance est alpha,
- Si les fréquences de A et B sont supérieures à 10% chacune, la tendance est alpha/bêta.

La fréquence des états A et B dans la région structurée est déterminée ainsi :

Par exemple pour l'état A, $\text{freqA} = \text{nbAtotal} / (\text{nbAtotal} + \text{nbBtotal}) * 100$

avec nbAtotal = nombre d'état « A et a » du domaine et nbBtotal = nombre d'état « B et b » du domaine

Un score de confiance de la tendance peut être calculé selon la formule suivante :

$\text{score} = (\text{nbB} + \text{nbA} + (\text{nba}/2) + (\text{nbb}/2)) / (\text{nbA} + \text{nbB} + \text{nbC})$;

avec nbA = nombre d'états « A » du domaine,

nbB = nombre d'états « B » du domaine,

nbC = nombre d'états « ? » du domaine,

nba = nombre d'états « a » du domaine,

et nbb = nombre d'états « b » du domaine.

Le score de confiance est compris entre 0 et 1. Un score proche de 1 signifie une fiabilité élevée de la tendance attribuée au domaine globulaire. Le score de confiance n'est déterminé que pour les domaines présentant au moins un état A, a, B ou b dans leur séquence. Dans le cas où il n'y a que des états « ? » dans le domaine globulaire, le score de confiance vaut 0.

Nous avons tenté de discriminer les domaines tout α de ceux tout β . Les résultats obtenus à partir des règles ci-dessus ne sont pas très bons. En effet, seuls un tiers des domaines protéiques de la banque A sont prédits comme tout alpha. Les autres n'ont pas de repliements définis. Dans le cas de la banque B, moins d'un tiers des domaines protéiques sont prédits comme tout bêta. Dans les deux banques cependant, moins de 10 % des protéines sont prédites dans le repliement opposé à laquelle elles appartiennent. L'information déduite des amas hydrophobes est intéressante pour les amas ayant une forte présence dans une des deux structures secondaires hélice ou brin. Cependant, dans l'enchaînement de plusieurs structures secondaires composant le repliement de la protéine, cette information est difficilement exploitable. De nombreux amas hydrophobes sont représentés aussi bien dans les structures en hélice que dans les structures en brin sans avoir une préférence particulière,

et il conviendrait d'ajouter d'autres informations, tels les profils de séquence associés à chacun des amas et pour chacun des états de structure secondaire, pour affiner la prédiction.

4.4 Fréquence d'apparition des amas hydrophobes par rapport à l'aléatoire (Z-score) et corrélation aux structures secondaires

Pour mesurer la fréquence d'apparition des amas hydrophobes par rapport à l'aléatoire, des Z-scores ont été calculés dans un travail précédent [HENNETIN, 2003]. A partir de séquences protéiques, des séquences « aléatoires » ont été produites en permutant aléatoirement la position des acides aminés. Un grand nombre de banques de séquences aléatoires ont ainsi été produites, à partir de la banque de séquences réelles. Les Z-scores sont calculés pour chacun des amas. Ils permettent de comparer un évènement par rapport à l'aléatoire. Le Z-score se calcule de la manière suivante:

$$Z = \frac{\text{occ} - M}{\sigma}$$

où Z est le Zscore ; occ est l'occurrence d'un amas dans une banque protéique ; M est la moyenne des occurrences de ce même amas dans les banques aléatoires et σ en est l'écart-type de cette distribution.

Il a été montré que pour les hélices α , les Z-scores sont généralement positifs et élevés [HENNETIN, J. et al., 2003]. Par contre, les brins, tout aussi fréquents que les hélices, ne conduisent pas à des Z-scores élevés. En tant que structure tendant à l'agrégation (formation de feuillets de brins), on peut considérer que les amas de type «brins » sont sous contrôle, afin d'éviter une agrégation excessive donc pathogène [RICHARDSON, J. S. et al., 2002; WANG, W. et al., 2002]. Leurs Z-scores sont ainsi neutres, voir négatifs [RICHARDSON, J. S. et al., 2002].

En utilisant la table répertoriant le Z-score déterminé pour chaque amas de taille inférieure à 15 acides aminés (réalisée par J.Hennetin [HENNETIN, 2003]), nous avons calculé dans chaque banque A, B, C et D, la somme des Z-scores de chaque séquence pondérée par le nombre d'amas contenus dans cette séquence. La Figure 82 présente ces distributions pour les différentes banques considérées.

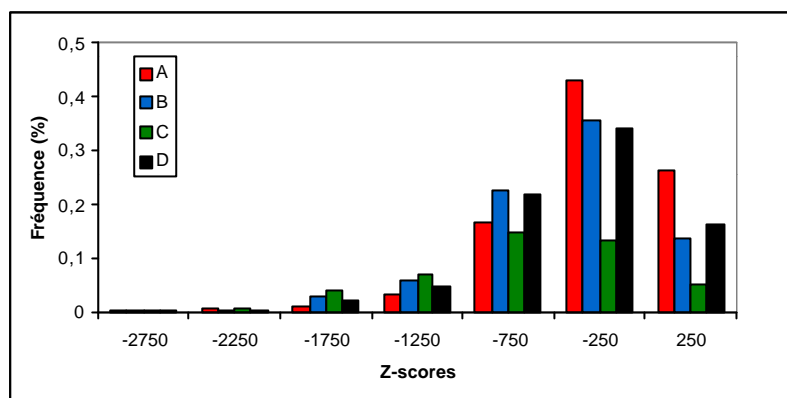


Figure 82 : Distribution des Z-scores associés aux amas présents dans les banques A, B, C et D de SCOP (normalisation par le nombre d'amas de chaque banque).

Nous avons constaté qu'il y a des différences notables de distribution des Zscores dans les banques A, B, C et D. Cependant, il est difficile d'interpréter ces différences et d'en déduire des règles pour fournir une information sur le repliement. Une valeur de Z-score pondérée globale ne permet pas d'identifier à quelle banque appartient une protéine.

4.5 Distribution des amas en fonction de leur longueur et de leur nombre d'acides aminés hydrophobes

Comme nous l'avons vu précédemment, les amas hydrophobes peuvent servir de marqueurs assez spécifiques pour les hélices α ou les brins β . Ils sont de bons marqueurs de structures secondaires. Chaque amas présente une préférence relative, pour l'une ou l'autre de ces structures secondaires, souvent autour de 60% (cf. paragraphe 4.3 et Annexe). On notera également que les amas courts sont souvent plus généralement associés à des brins β , les amas longs à des hélices α . Il est alors envisageable de visualiser la distribution de la taille des amas hydrophobes (de 2 à 20 acides aminés) dans chacun des repliements (banques A, B, C, D, F et G de SCOP) (Figure 83).

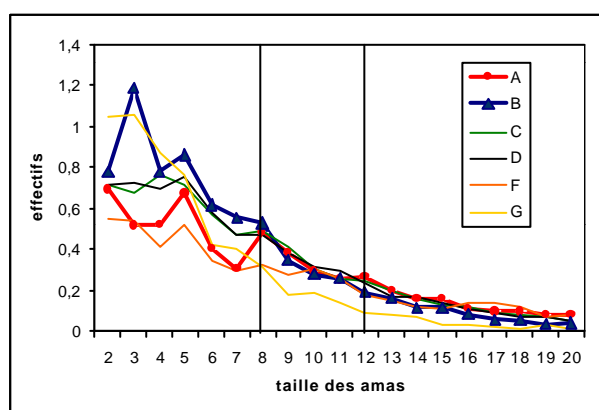


Figure 83 : Distribution des tailles des amas dans les banques A, B, C, D, F et G de SCOP.

D'après la Figure 83, nous observons que les petits amas (taille 2 à 8) sont plus abondants dans la banque B. Ainsi, dans les repliements construits principalement à partir de brins bêta (banque B), les petits amas sont logiquement dominants. Pour des amas de taille 8 et 11 acides aminés, toutes les courbes sont superposées. Les amas de taille 8 à 11 ne permettent pas de différencier les quatre premières classes (A, B, C et D). Quant aux amas de grande taille (supérieur à 12 acides aminés), ils sont retrouvés préférentiellement dans les hélices alpha (banque A). Nous notons que les courbes C et D sont situées entre les deux courbes A et B. Elles correspondent respectivement aux banques alpha/bêta et alpha+bêta. Il est normal de les retrouver entre les deux courbes A et B puisqu'elles renferment à la fois des petits amas (plutôt retrouvés dans les brins bêta) et des grands amas (plutôt retrouvés dans les hélices alpha). La courbe G correspond aux petites protéines et présente relativement moins d'amas de tailles supérieures à 8. Cette banque contient beaucoup moins d'acides aminés hydrophobes que les autres banques. Elle est constituée de protéines de petites tailles renfermant peu d'acides aminés hydrophobes et peu de structures secondaires. Elles contiennent donc moins d'amas de moyennes et grandes tailles. La courbe F correspond aux protéines membranaires et de surface. Ces protéines contiennent à l'inverse de nombreux amas hydrophobes de grande taille et peu d'amas hydrophobes de petites tailles. Enfin, nous pouvons remarquer des pics centrés sur les valeurs 5 et 8, correspondant à des périodicités d'hélice alpha. Il a été montré que les tailles d'amas correspondant aux multiples de tours d'hélices alpha sont plus représentées dans les banques [HENNETIN, J. 2003].

D'après les observations précédentes, nous pouvons trier les amas hydrophobes en deux classes : les amas de petites tailles (inférieures à 8 acides aminés) que nous appelons « amas S » pour Short et les amas de grandes tailles (supérieures à 12 acides aminés) que nous appelons « amas L » pour Long. Pour les quatre banques A, B, C et D, nous avons calculé le rapport S/L (Figure 84).

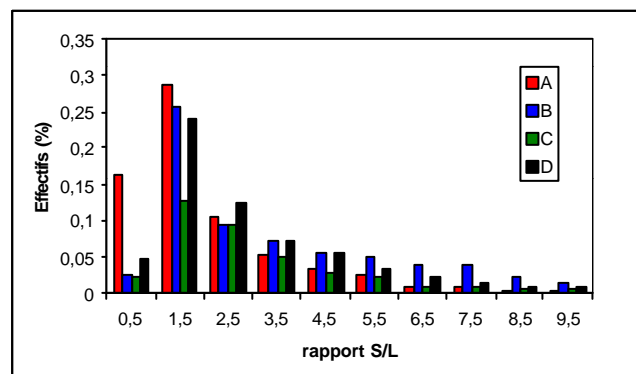


Figure 84 : Distribution du rapport S/L dans les banques A, B, C et D de SCOP.

Dans la banque A, le rapport S/L est généralement très faible (inférieur à 2) pour l'ensemble des protéines. A l'inverse, la banque B présente très peu de valeurs inférieures à 1 et un grand nombre de valeurs plus élevées (allant jusqu'à 10). Les banques C et D ont des distributions similaires. Nous pouvons conclure que généralement les protéines ayant un repliement tout alpha (banque A) ont des rapports S/L inférieurs à 0,5 et que les protéines ayant un repliement tout bêta (banque B) ont des rapports S/L supérieurs à 6. Sur chaque segment structuré prédit par DomHCA, nous pouvons ainsi fournir une information quant au type de repliement potentiellement adopté par le segment.

5 Conclusion

Nous avons développé plusieurs procédures pour identifier des régions particulières dans les séquences de protéines (régions structurées/non structurées, passages membranaires, régions de répétition, peptides de fusion). Dans le même contexte, nous pouvons fournir une indication sur le type de repliement adopté par les régions structurées prédites par DomHCA. Il serait donc envisageable de rassembler toutes ces informations et de caractériser chaque domaine structuré prédit par DomHCA par son repliement potentiel, sa taille, son score d'hydrophobie, etc...

Dans la perspective d'un traitement à grande échelle des génomes et de comparaison des protéomes, les séquences de protéines pourraient être découpées en régions caractéristiques d'une texture bien particulière. Nous avons choisi dans une première approche de comparer le génome de *Plasmodium falciparum* à trois génomes bien connus : *Homo sapiens*, *Saccharomyces cerevisiae* et *Caenorhabditis elegans*.

Chapitre VII

Application à *Plasmodium falciparum* : Etude d'un génome particulier

1 Introduction

Plasmodium falciparum, agent responsable du paludisme, est un protozoaire parasite. Le paludisme est transmis à l'homme via la piqûre d'un moustique femelle Anopheles, elle-même infectée après avoir piqué un homme impaludé. Le génome de *Plasmodium falciparum*, récemment séquencé [GARDNER, M. J. et al., 2002], présente d'une part, des portions répétées et d'autre part, une grande richesse en bases A+T (82%) [BASTIEN, O. et al., 2004]. Nous avons réalisé dans un premier temps une étude comparative des codons et acides aminés de ce génome par rapport à des génomes standard connus (*Homo sapiens*, *Saccharomyces cerevisiae* et *Caenorhabditis elegans*). Puis, dans un second temps, nous avons soumis les protéines issues de ces génomes au découpage DomHCA afin d'éventuellement mettre en évidence des différences au niveau des domaines structurés les composant. Enfin, nous avons mis au point une méthode simple ayant pour but d'ôter, au moins partiellement, les acides aminés surabondants présents dans les séquences issues de ce génome, afin in fine de pouvoir ôter le biais introduit par ces acides aminés dans les procédures de recherche.

2 Comparaison des génomes

2.1 Distribution des bases et des codons dans les génomes

2.1.1 Introduction

Les séquences d'ADN sont riches en redondances, résultat des processus d'évolution et de sélection naturelle auxquels elles sont soumises. L'étude de ces biais permet de mieux comprendre comment la cellule exploite son information génétique, d'effectuer la recherche systématique dans de nouvelles séquences de caractéristiques inconnues et de fournir des outils de prédiction pour l'analyse de leurs propriétés.

2.1.2 Composition des génomes

Les quatre bases constituant les génomes des espèces vivantes sont l'adénine (A), la thymine (T), la guanine (G) et la cytosine (C). La complémentarité des bases dans la structure en double hélice de l'ADN impose que globalement le nombre de A soit égal au nombre de T et que le nombre de G soit égal au nombre de C. Il est donc intéressant de déterminer les rapport $(G+C)/(A+T)$, ou plus simplement le pourcentage de GC ou AT d'un génome.

Chez le colibacille, la distribution des bases est à peu près équilibrée, puisqu'il y a environ 51% de GC contre 49% de AT dans les phases codantes (CDS) du génome. L'équidistribution n'est cependant pas la règle générale et ce pourcentage peut varier dans une fourchette comprise approximativement entre 15% et 70% de GC. On trouve ainsi 24% de GC calculé dans les phases codantes (CDS) du génome de *Plasmodium falciparum* (Figure 85), contre près de 68% chez *Thermus thermophilus*, une bactérie peuplant des sources chaudes et capable de se développer à des températures dépassant 80°C. Cette forte proportion en GC permet probablement à l'ADN de mieux résister à la dénaturation thermique (l'appariement G-C étant plus stable que l'appariement A-T).

```
TGGAAATAAAATAAATAATCAGATGAAAAGGATTGGTGCTTCTGTTGATTGGTCAAGAGAATATTT
TACCATGAATGAAAATTTATCAAATGCGGTAAGAAGCTTTTATTAAATTTTATGAAAGTGGTTT
ATATATAGAGATAATAGATTAGTTGCTTGGTGTCTCATTAAAAACTGCCTTATCAGATATTGAA
GTAAATCTAGAAGAAATTAACCAACCAAAATCAAATACCATCCTTTGATCATTTAGTTGAA
GTAGGTGTTCTATATAAATTTTTTATCAAATAAAGATAGTGAAGAAAAAATAGAAATAGCAACA
ACTCGTATTGAAACCATGCTAGGAGATGTTGCTGTTGCTGTCCATCCAAAAGATAAAAGATATGCA
CATTTAATTGGTAAAGAAATTGTACATCCATTTATTCCTAATAGGAAAATTATTATTATTGCTGATG
ATTTTGTGATATGCAATATGGTACTGGTGCTGTGAAAATTACTCCAGCTCATGATAAAAATGATTA
```

Figure 85 : Séquence d'un fragment de gène issu de *Plasmodium falciparum* (24% GC).

En général, les génomes de vertébrés se situent souvent au centre de la fourchette avec des taux de GC pour les CDS compris entre 40% et 45% et avec une segmentation en grandes régions ou « isochores » où le taux est localement constant mais diffère de la région voisine. Notre étude a porté sur la comparaison du génome de *Plasmodium falciparum* avec quatre autres génomes : *Homo sapiens*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae* et *Caenorhabditis elegans*. Le Tableau 14 présente les valeurs observées pour ces cinq génomes.

Tableau 14: Distribution des fréquences GC dans les séquences codantes (CDS) de plusieurs génomes. Le calcul a été effectué à l'aide du programme Countcodon (site internet : <http://www.kazusa.or.jp/codon/countcodon.html>) sur l'ensemble des CDS de chacun de génomes.

Génomes	<i>Plasmodium falciparum</i>	<i>Homo sapiens</i>	<i>Arabidopsis thaliana</i>	<i>Saccharomyces cerevisiae</i>	<i>Caenorhabditis elegans</i>
GC (%)	23.83%	52.59%	44.60%	39.76%	42.94%

Les séquences codantes du génome de *Dictyostelium discoideum* [GLOCKNER, G. et al., 2002], amibe unicellulaire, présentent un taux de GC de 29%, proche de celui de *Plasmodium falciparum*.

Le biais en A-T (presque 80%) dans la composition des nucléotides des séquences codantes de *P. falciparum* peut introduire de grandes modifications dans les séquences protéiques et ainsi entraîner un biais dans les procédures de recherches.

2.2 Analyse de la composition en codons et acides aminés codés par les codons

A l'inverse de l'ADN pour lequel il existe des variations importantes de composition d'une espèce à une autre, la composition moyenne des protéines est relativement constante dans le monde du vivant. Il existe de petites variations en fonction de la nature des protéines considérées : ainsi des protéines membranaires seront plus riches en résidus hydrophobes, des protéines interagissant avec les acides nucléiques seront plutôt basiques. Par rapport à la composition «moyenne» des protéines (Tableau 15), on constate que les 20 acides aminés ne sont pas équivalents et certains sont relativement plus abondants que d'autres.

Tableau 15 : Composition moyenne des protéines (site internet : (<http://www.esil.inuvmrs.fr/~dgaut/Cours/biais.html>)).

Alanine	A : 83 ‰	Méthionine	M : 24 ‰
Cystéine	C : 17 ‰	Asparagine	N : 44 ‰
Aspartate	D : 53 ‰	Proline	P : 51 ‰
Glutamate	E : 62 ‰	Glutamine	Q : 40 ‰
Phénylalanine	F : 39 ‰	Arginine	R : 57 ‰
Glycine	G : 72 ‰	Sérine	S : 69 ‰
Histidine	H : 22‰	Thréonine	T : 58 ‰
Isoleucine	I : 52 ‰	Valine	V : 66 ‰
Lysine	K : 57 ‰	Tryptophane	W : 13 ‰
Leucine	L : 90 ‰	Tyrosine	Y : 32 ‰

Le fait que l'ADN soit composé de gènes codant pour des protéines impose des contraintes supplémentaires sur la séquence des nucléotides qui le composent. La traduction en protéine faisant appel à des mots constitués de 3 bases, les codons. Il est intéressant de déterminer les fréquences des 64 triplets possibles et de construire la statistique d'apparition des différents codons pour une espèce donnée (Figure 86). La distribution des codons doit nécessairement suivre celle des acides aminés qui leur correspondent dans le code génétique. Dans les protéines contenant par exemple 1,3 % de résidus tryptophane, on s'attend à trouver 1,3 % de TGG, l'unique codon correspondant à cet acide aminé, à l'intérieur des gènes. Pour le cas des acides aminés codés par plusieurs codons synonymes, il est intéressant de voir si les deux codons sont représentés de manière équilibrée ou non. Lorsque les codons synonymes ne sont pas employés avec la même fréquence, on parle de biais dans l'usage des codons ou « biais de codon ». Le biais de codon diffère d'une espèce à l'autre, selon les contraintes propres à chaque espèce (richesse en GC, taux de mutation, réparations) (<http://www.esil.inuv-mrs.fr/~dgaut/Cours/biais.html>).

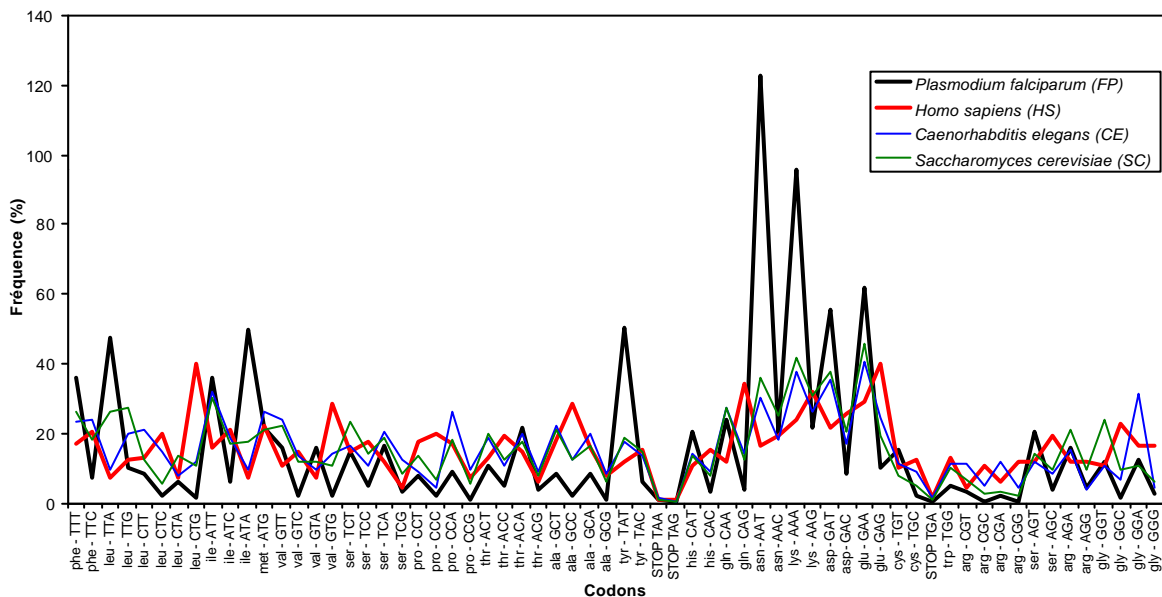


Figure 86 : Usage du code génétique.

Les fréquences des différents codons dans les gènes issus des génomes de *Plasmodium falciparum*, *Homo sapiens*, *Caenorhabditis elegans* et *Saccharomyces cerevisiae* ont été calculées à l'aide du programme Countcodon (<http://www.kazusa.or.jp/codon/countcodon.html>). Elles sont exprimées en %.

On notera que chez *Plasmodium falciparum*, les codons STOP (TAA, TAG et TGA) ont une fréquence de 0,13 % contre 0,25 % chez le génome d'*Homo sapiens*.

Les différences de fréquences observées sont le résultat de deux effets superposés : la composition en acide aminé des protéines qui n'est pas uniforme et la préférence systématique pour certains codons parmi les différents possibles. L'usage préférentiel de

certain codons synonymes est probablement évolutivement conservé. Deux espèces voisines auront généralement des tables d'utilisation de codons très similaires. Individuellement, chaque gène se conforme à ces règles de préférence qui sont en quelque sorte la «signature» du génome. Chez *P. falciparum*, qui contient un taux très faible de GC (24%), cela se répercute de manière significative sur le choix des codons. Dans le cas d'un génome riche en AT, les codons se terminant par A ou T sont très fortement préférés (l'inverse sera vrai pour les génomes riches en GC).

Les codons codant pour les acides aminés asparagine et lysine (AAT, AAC, AAA, AAG) représentent 26% des codons chez *P.falciparum* (9% chez *Homo sapiens*). Les deux codons surreprésentés sont AAT et AAA.

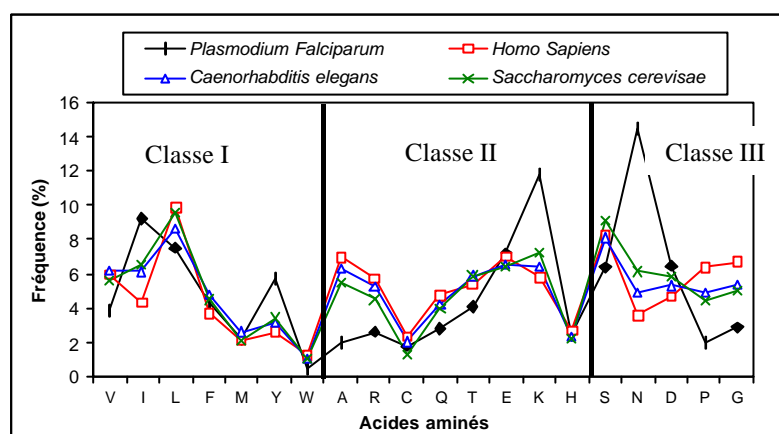


Figure 87 : Comparaison de la distribution des acides aminés dans les protéomes (protéines prédites) de *Plasmodium falciparum*, *Homo sapiens*, *Caenorhabditis elegans* et *Saccharomyces cerevisiae* (5334, 32035, 21629 et 6699 séquences, respectivement) [CALLEBAUT, I. et al., 2005].

Les acides aminés sont ordonnés suivant la classification établie par Callebaut et al. [CALLEBAUT, I. et al., 1997a]. La classe I regroupe les acides aminés hydrophobes (V, I, L, F, M, Y et W), pour lesquels les propensions pour les structures secondaires régulières (hélice α et brin β) sont plus élevées que celles pour les boucles. La classe III rassemble les acides aminés formateurs de boucles (P, G, D, N et S). La classe II est la classe intermédiaire et regroupe les acides aminés (A, R, C, Q, T, E et K).

Les acides aminés hydrophobes V, I, L, F, M et Y sont les acides aminés retrouvés dans les cœurs des domaines globulaires et constituent la classe I établie par Callebaut et al. [CALLEBAUT, I. et al., 1997a]. Les acides aminés de la classe I pris dans leur ensemble sont similaires chez *Plasmodium falciparum* et les autres génomes (Figure 87). La fréquence de la cystéine chez *P. falciparum*, acide aminé fréquent des cœurs hydrophobes, ne diffère pas non plus des autres génomes. De plus, la fréquence de l'histidine, acide aminé très neutre vis-à-vis des structures secondaires (acide aminé qui peut aussi bien être inclus dans une boucle, une hélice α ou un brin β) est également stable. Deux acides aminés sont très fréquents chez *P.*

falciparum: l'asparagine (N) qui est trois fois plus élevée (14.5% contre 5%) et la lysine (K) qui est deux fois plus élevée que dans les autres génomes (12% contre 6%).

Le biais en A-T, observé dans le génome de *P.falciparum*, est responsable de cette forte proportion en asparagine et en lysine et dans une moindre mesure l'isoleucine (I) et la tyrosine (Y). A l'inverse les proportions en arginine, alanine, proline et glycine, codés par des codons riches en GC, sont beaucoup plus faibles chez *P. falciparum* [BASTIEN, O. et al., 2004]. Comme expliqué précédemment (chapitre V), une part importante des acides aminés, asparagine et lysine, est retrouvée dans les régions de faible complexité, localisées dans les régions charnières entre domaines fonctionnels. En effet, plus de 14% des résidus asparagine et 4% des résidus lysine du génome sont retrouvés dans des régions dites de répétitions (cf. chapitre VI). Ces deux acides aminés constituent plus de la moitié des régions de répétitions (55%). Les acides aminés hydrophobes, leucine, phénylalanine, méthionine et tryptophane, ont des fréquences identiques chez *P. falciparum* et les autres génomes (classe I, Figure 87). On peut noter qu'il existe une balance presque parfaite entre la valine et l'isoleucine (codons GTT, GTC, GTA et GTG pour la valine et codons ATT, ATC et ATA pour l'isoleucine), deux acides aminés qui sont très proches chimiquement et souvent interchangeable au niveau structural. Le dernier acide aminé de la classe hydrophobe, la tyrosine, fait partie des résidus formateurs de boucles [CALLEBAUT, I. et al., 1997a] et l'augmentation de sa fréquence chez *P. falciparum* peut ne pas affecter l'équilibre général des coeurs hydrophobes des domaines. La haute fréquence en lysine, qui est en moyenne l'acide aminé le plus exposé dans les domaines globulaires [SOYER, A. et al., 2000] [PINTAR, A. et al., 2003a; PINTAR, A. et al., 2003b], peut être compensée par les faibles fréquences en arginine et alanine. Dans la classe III, les faibles fréquences en proline et glycine, très présents dans les boucles, peuvent ensemble compenser la très forte fréquence en asparagine, qui juste après la glycine, généralement représentée dans les régions de boucles, partage avec elle sa capacité à adopter des conformations en hélice gauche [CALLEBAUT, I. et al., 2005].

La conservation de la proportion totale des acides aminés hydrophobes chez *P. falciparum* et le comportement compensatoire des autres couples d'acides aminés permettent de supposer que les coeurs hydrophobes des domaines fonctionnels sont conservés et qu'ils peuvent être identifiés en utilisant notamment la méthode HCA.

Nous avons également examiné les fréquences des acides aminés dans le génome *Dictyostelium discoideum*, autre génome riche en AT (Figure 88).

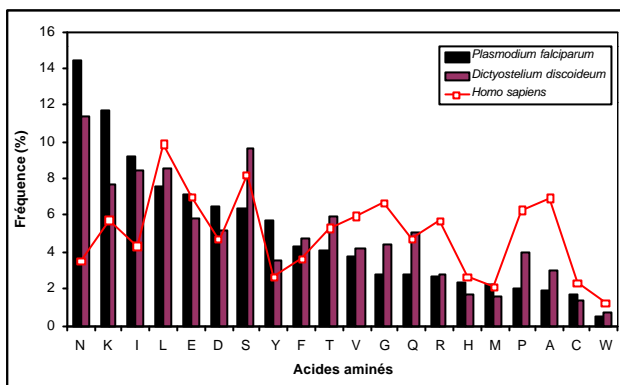


Figure 88 : Comparaison de la distribution des acides aminés dans les protéomes de *Plasmodium falciparum* et *Dictyostelium discoideum* (5334 et 13574 séquences).

Les acides aminés sont classés par ordre décroissant de leurs fréquences chez *P.falciparum*.

Les séquences codantes de celui-ci contiennent une forte proportion en acides aminés N et K, (comme celui de *Plasmodium falciparum*). Elles s'en distinguent notamment par une plus forte proportion de sérine (Figure 88).

2.3 Analyse des protéines

Nous avons comparé les effectifs et les tailles des protéines dans nos quatre génomes (Figure 89 et Tableau 16).

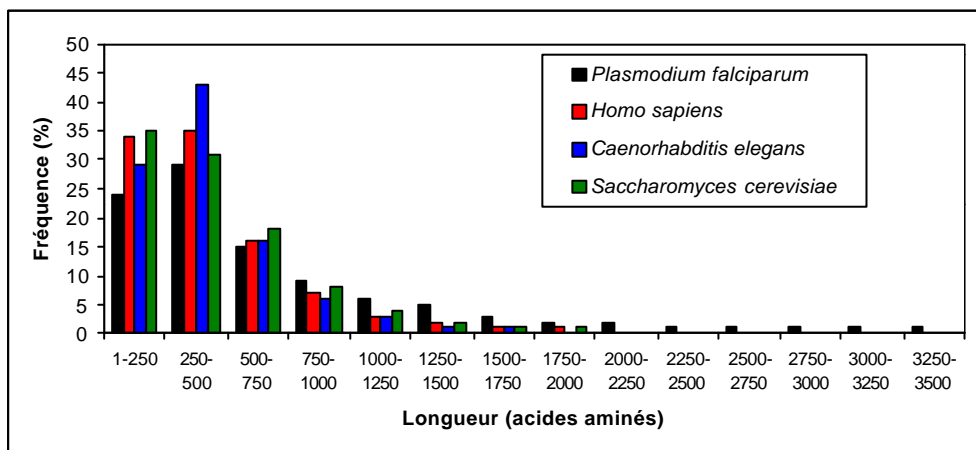


Figure 89 : Comparaison des tailles des protéines des protéomes de *Plasmodium falciparum*, *Homo sapiens*, *Caenorhabditis elegans* et *Saccharomyces cerevisiae*.

Tableau 16 : Bilan des caractéristiques des séquences incluses dans les protéomes de *Plasmodium falciparum*, *Homo sapiens*, *Caenorhabditis elegans* et *Saccharomyces cerevisiae*.

	PF	HS	CE	SC
Nb de protéines	5 291	32 034	21 629	6 699
Nb de résidus total	4 013 502	14 773 391	9 535 848	3 018 115
Taille maximale des protéines	10 589	10 425	13 100	4 910
Taille minimale des peptides/protéines	17	19	6	16
Taille moyenne des protéines	759	461	441	450

La taille moyenne des protéines dans les génomes *Homo sapiens*, *Caenorhabditis elegans* et *Saccharomyces cerevisiae* est de 450 acides aminés et de 759 chez *Plasmodium falciparum*. Chez ce dernier, les protéines ont des tailles beaucoup plus élevées.

2.4 Analyse des segments prédits structurés par DomHCA

Nous avons extrait les segments structurés des protéomes de *Plasmodium falciparum*, *Homo sapiens*, *Caenorhabditis elegans* et *Saccharomyces cerevisiae* avec DomHCA (Tableau 17) et nous en avons tracé la distribution des tailles (Figure 90). Nous avons remarqué que même si les protéines de *Plasmodium falciparum* sont beaucoup plus longues que celles des trois autres génomes, les régions structurées ont des tailles similaires (environ 150-160 acides, taille moyenne d'un domaine globulaire).

Tableau 17 : Bilan des régions structurées prédites par DomHCA dans les protéines des génomes de *Plasmodium falciparum*, *Homo sapiens*, *Caenorhabditis elegans* et *Saccharomyces cerevisiae*.

	PF	HS	CE	SC
Nb de régions structurées (RS)	17 779	68 353	42 661	13 219
Nb de régions pseudo-structurées (RPS)	7 670	41 361	21 669	6 257
Nb de résidus dans RS	3 012 721	9 367 973	6 905 680	2 243 856
Nb de résidus totaux dans le génome	4 013 502	14 773 391	9 535 848	3 018 115
% de résidus inclus dans les RS	75,06 %	63,41 %	72,42 %	74,35 %
Taille max des RS	9 635	4 483	2 201	3 038
Taille moyenne des RS	169	137	162	170

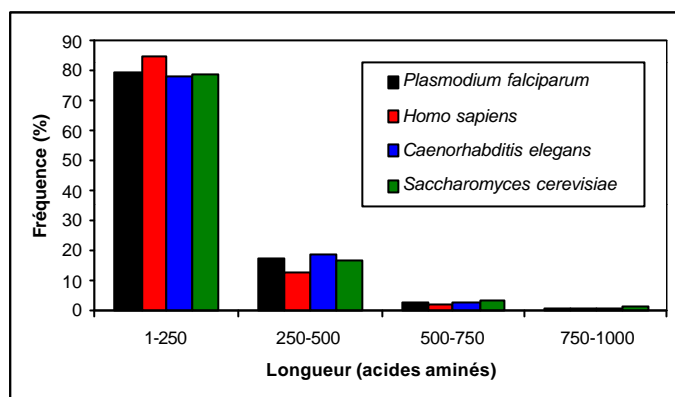


Figure 90 : Comparaison des tailles des régions structurées issues des protéomes de *Plasmodium falciparum*, *Homo sapiens*, *Caenorhabditis elegans* et *Saccharomyces cerevisiae*.

Les régions structurées de *Plasmodium falciparum*, *Caenorhabditis elegans* et *Saccharomyces cerevisiae* contiennent environ 75% des acides aminés des protéines. Il semble que le génome *Plasmodium falciparum* contient proportionnellement autant de résidus inclus dans les régions structurées que les autres génomes. Nous remarquons que le génome *Homo sapiens* présente un pourcentage en acides aminés présents dans les régions structurées plus faible que les autres génomes (64%). Il serait intéressant de traiter d'autres génomes plus proches évolutivement de l'homme afin de voir si cette caractéristique est retrouvée chez tous les vertébrés et à quel embranchement de l'évolution serait apparue cette modification.

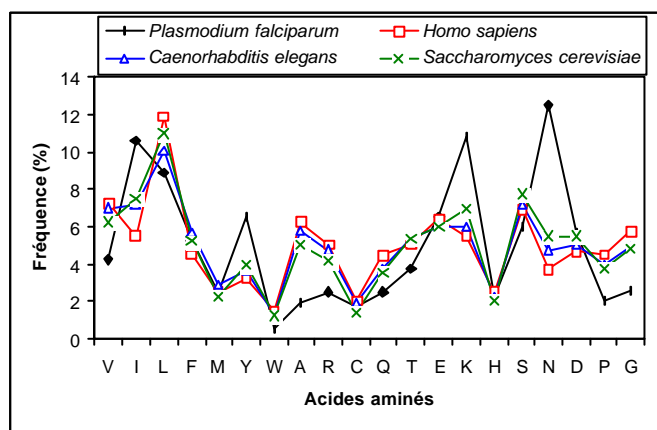


Figure 91 : Comparaison de la distribution des acides aminés dans les régions structurées prédites par DomHCA dans les protéomes (protéines prédites) de *Plasmodium falciparum*, *Homo sapiens*, *Caenorhabditis elegans* et *Saccharomyces cerevisiae*.

Nous avons tracé la distribution des acides aminés au sein des régions structurées prédites par DomHCA (Figure 91). Les distributions sont similaires à celles des protéines entières (voir Figure 87). Les acides aminés sont répartis uniformément dans les régions structurées et non structurées des protéines. La forte proportion en acides aminés N et K,

prises en évidence dans les protéines de *Plasmodium falciparum*, est également retrouvée au sein des régions structurées.

2.5 Analyse des régions de répétitions

Nous avons recherché les régions de répétitions (que nous avons définies au chapitre précédent), dans les protéomes de *Plasmodium falciparum*, *Homo sapiens*, *Caenorhabditis elegans* et *Saccharomyces cerevisiae* (Tableau 18).

Tableau 18 : Bilan des régions de répétitions identifiées dans les protéines des génomes de *Plasmodium falciparum*, *Homo sapiens*, *Caenorhabditis elegans* et *Saccharomyces cerevisiae*.

	PF	HS	CE	SC
Nb de répétitions	8 142	8 728	4 973	1 544
Nb de résidus dans les répétitions	194 041	234 622	100 146	27 846
Nb de résidus total dans le génome	4 013 502	14 773 391	9 535 848	3 018 115
% de résidus inclus dans les répétitions	4,83 %	1,59 %	1,05 %	0,92 %
Taille max des répétitions	277	1 013	463	113
Taille min	8	8	8	8
Taille moyenne	22	22	20	17

Près de 5% des acides aminés des séquences protéiques de *Plasmodium falciparum* sont inclus dans les régions de répétition. Dans nos trois autres génomes, seuls 1% des acides sont situés dans ces mêmes régions. La taille moyenne des régions de répétition est sensiblement la même d'un génome à l'autre. Les tailles maximales des régions de répétition sont très variables (allant de 113 acides aminés pour *Saccharomyces cerevisiae* à 1013 pour *Homo sapiens*). Ces tailles élevées peuvent s'expliquer par la présence de régions de répétition chevauchantes. La taille mesurée est la taille globale de la région de répétition. Par exemple, la région de répétition **RSRSRDRGRGGGGGGGGGGGG**, détectée dans la protéine ENSP00000291552 de *Homo sapiens* entre les résidus 203 et 223, se compose des deux répétitions suivantes :

- RRSRSRDRGR qui correspond à une harmonique 'λxλ',

- GRGGGGGGGGGGGG qui correspond à une harmonique 'λxxλ' (nous comptons toujours l'harmonique la plus grande possible).

Ces deux harmoniques se chevauchent sur les deux résidus (GR). Il est impossible de déterminer où finit la première région de répétition et où commence la seconde. La distribution des tailles des régions de répétition est présentée Figure 92.

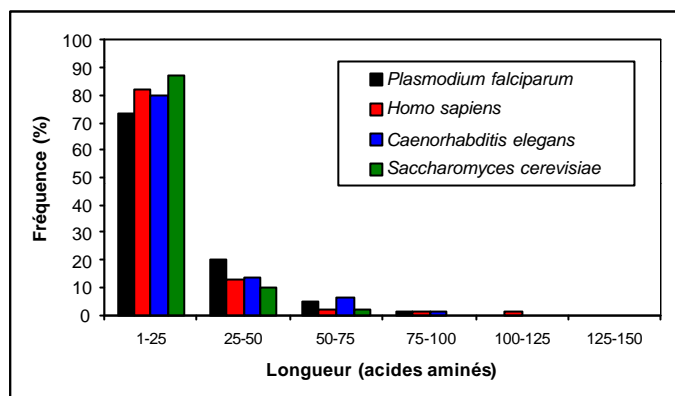


Figure 92 : Comparaison des tailles des régions de répétition issues des protéomes de *Plasmodium falciparum*, *Homo sapiens*, *Caenorhabditis elegans* et *Saccharomyces cerevisiae*.

Entre 70% et 85% des régions de répétition dans nos quatre génomes ont une taille comprise entre 8 et 25 acides aminés et 15-20% ont une taille comprise entre 25 et 50 acides aminés. Les distributions sont similaires pour ces quatre génomes. Nous avons également observé la distribution des acides aminés composant ces régions de répétition (Figure 93).

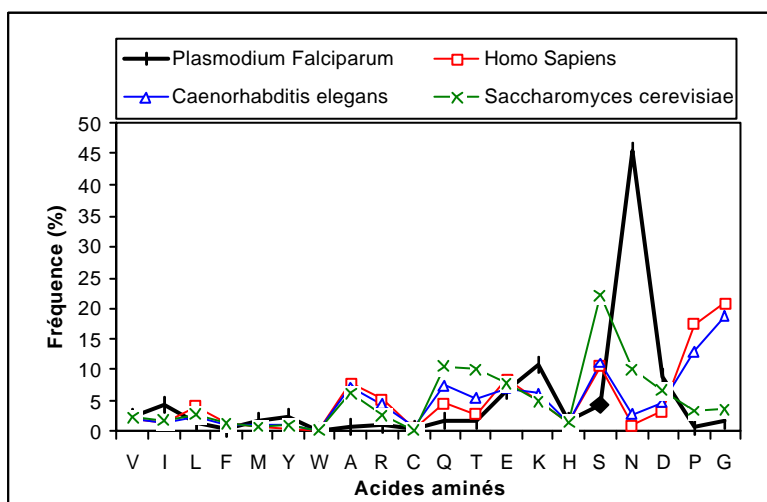


Figure 93 : Comparaison de la distribution des acides aminés dans les régions de répétitions issues des protéomes (protéines prédites) de *Plasmodium falciparum*, *Homo sapiens*, *Caenorhabditis elegans* et *Saccharomyces cerevisiae*.

Les acides aminés hydrophobes (V, I, L, F, M, Y et W) et la cystéine (C) sont très peu retrouvés dans les régions de répétition, quel que soit le génome (Figure 93). Les acides

aminés K et N sont par contre très présents dans les régions de répétition de *Plasmodium falciparum* (11% et 46%). Ainsi, 14% des résidus N et 4% des résidus K comptabilisés dans le génome entier sont situés dans des régions de répétition contre 0,4% et 1,2% chez l'homme.

Dans le génome de *Plasmodium falciparum*, nous avons également observé que près de 45 % des protéines présentent une répétition dans leur séquence. Nous avons tracé la distribution du nombre de régions de répétition par protéine chez *Plasmodium falciparum* (Figure 94).

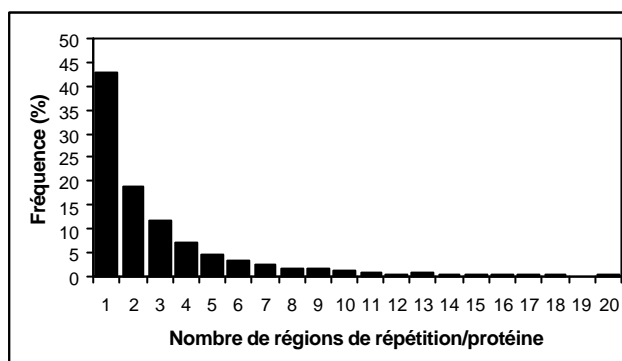


Figure 94 : Distribution du nombre de régions de répétition contenues dans les protéines de *Plasmodium falciparum*.

2668 séquences contiennent des régions de répétition sur 5329, soit la moitié du protéome complet. 2661 séquences ne présentent pas de régions de répétition dans leur séquence. Le nombre de régions de répétition identifiées dans une protéine varie de 0 à 20. Plus de 60% des protéines ont entre une à deux régions de répétition dans leur séquence (Figure 94).

3 Conclusion

Les protéines de *Plasmodium falciparum* sont plus longues que celles issues des génomes d'*Homo sapiens*, de *Saccharomyces cerevisiae* et de *Caenorhabditis elegans*. Elles sont aussi particulièrement riches en acides aminés N et K. Même si ces acides aminés sont localisés tout au long des séquences, une part importante de ces acides aminés est retrouvée dans les régions de faible complexité de séquence, qui sont souvent des régions charnières entre domaines fonctionnels. Les acides aminés hydrophobes (V, I, L, F, M, Y et W) sont globalement conservés dans les séquences et les autres acides aminés semblent se compenser. Cela nous permet de supposer que les coeurs hydrophobes des domaines fonctionnels des protéines sont conservés et peuvent être identifiés sur la base de leur dichotomie

hydrophobe/hydrophile. Nous avons également observé que les séquences de *Plasmodium falciparum* renferment de nombreuses régions de répétition, qui contiennent elles aussi des acides aminés N et K, mais en proportions égales à celles retrouvées sur l'ensemble de la séquence. Le découpage des séquences avec DomHCA nous a permis de montrer que les régions structurées de *Plasmodium falciparum* sont de même taille que celles des autres génomes et que la proportion d'acides aminés inclus dans ces régions est également similaire aux autres génomes (proche de 75%). Il semble que l'information nécessaire aux divers domaines fonctionnels soit présente mais plus ou moins masquée par la présence de nombreuses régions non structurées, de répétitions ou de régions de faible complexité dans les séquences. Il a été montré récemment lors d'une étude sur la longueur des protéines chez les eucaryotes, les archaées et les bactéries que les protéines eucaryotes ont des tailles plus grandes par rapport à celles des procaryotes. Cela suggère que l'évolution des protéines eucaryotes est influencée par des processus de fusion de protéines et l'acquisition de multiples fonctions pour une protéine ayant initialement une seule fonction [BROCCHIERI, L. et al., 2005]. Le fait que les protéines issues du génome humain présentent 63% des acides aminés dans les régions structurées (contre 75% chez *Plasmodium falciparum*, *Saccharomyces cerevisiae* et *Caenorhabditis elegans*) représente une particularité qu'il conviendrait d'explorer plus avant. Il serait intéressant d'étudier si cette particularité est retrouvée uniquement chez les vertébrés ou si elle remonte un peu plus loin dans l'évolution.

L'abondance en résidus N et K dans le génome de *Plasmodium falciparum* semble être un problème pour les recherches automatiques de similarité avec d'autres protéines. Nous envisageons dans ce contexte de mettre au point une procédure pour extraire les acides aminés N et K surabondants et pour les substituer par des acides aminés X, ceci dans l'optique de créer un « masque », perturbant moins les recherches automatiques de similitude. Cette procédure, associée à un découpage DomHCA des séquences, pourrait dans certains cas améliorer l'identification des fonctions de ces domaines.

Conclusion générale et perspectives

Dans le domaine industriel, et en particulier en contrôle qualité, de nombreuses applications nécessitent la recherche de défauts dans une surface qui n'est pas homogène (textile, bois, agro-alimentaire,...). L'information de texture, et la possibilité de détecter une rupture dans cette information, peut alors permettre la mise en évidence de ces défauts. Dans une séquence de protéine, cette recherche de « défaut » pourrait s'apparenter à l'identification d'une région de faible complexité, d'une région de répétition, d'une région désordonnée, d'un motif bien particulier, d'un domaine globulaire ou d'une région caractérisée par une texture spécifique.

La méthode Hydrophobic Cluster Analysis (HCA) permet de faire ressortir visuellement des régions présentant une texture particulière dans les séquences. Cependant, cette observation est qualitative et expert-dépendante. Dans ce contexte, nous avons souhaité caractériser la « texture » d'une protéine à partir de sa représentation 2D HCA en transposant certaines méthodes utilisées en analyse de texture d'images à notre système. L'utilisation d'une représentation en deux dimensions de la séquence, permet en effet beaucoup mieux qu'une séquence linéaire de souligner les changements de texture dans les séquences protéiques. La représentation bidimensionnelle HCA (2D) des séquences peut être considérée dans un certain sens comme une « image » dont les pixels seraient les acides aminés. Nous avons adapté plusieurs paramètres comme le contraste, l'entropie de la méthode des cooccurrences et les différents paramètres liés aux niveaux de gris (SRE, LRE, GLD, RLD et RLP) de la méthode des longueurs de plage. Nous avons utilisé le concept de fenêtre glissante le long de la séquence afin de prendre en compte l'environnement de chaque acide aminé et plusieurs alphabets basés sur l'hydrophobie et l'inclusion dans les amas hydrophobes des acides aminés. Cette première étude nous a permis de mieux caractériser les domaines globulaires issus des classes de repliement SCOP (A, B, C D, F et G) et leur texture. Les classes A, B, C et D, correspondant respectivement au repliement α , β , α/β et $\alpha+\beta$, se sont révélées avoir un comportement moyen similaire au regard des paramètres de texture alors que les deux classes F et G ont présenté des particularités. Ainsi, la classe F, correspondant aux protéines membranaires, est riche en amas de grande taille et inversement la classe G, renfermant les petites protéines, est pauvre en amas. Nous avons constaté que notre « image » n'est pas assez riche en pixels pour pouvoir isoler des textures à l'aide des paramètres calculés

mais que l'observation des profils d'hydrophobie au sens HCA peut être exploitée. L'absence ou la pauvreté en amas dans une région correspond à des valeurs extrêmes des paramètres de texture alors que des régions présentant une répartition uniforme des amas hydrophobes n'entraînent que peu de variations de ces paramètres. Nous avons réalisé que cette information est directement disponible au travers du profil d'hydrophobie et qu'elle peut être obtenue sans la nécessité de mettre en place des méthodes d'analyse de texture, parfois lourdes en calcul.

Nous avons donc mis au point une procédure automatique (DomHCA) qui détecte les régions de rupture ou charnières (séparant deux régions présentant une uniformité de la distribution des amas hydrophobes) uniquement à partir de l'information contenue dans la séquence. Cette procédure, écrite en langage C sous Visual.net et inspirée des résultats obtenus sur la texture et des propriétés d'HCA (alphabet hydrophobe basé sur les sept acides aminés (V, I, L, F, M, et W) et une fenêtre de 17 acides aminés se déplaçant le long de la séquence), permet de pré-découper les séquences en régions dites structurées ou pseudo-structurées. DomHCA attribue un score d'hydrophobie pour chaque prédiction, indiquant le pourcentage d'acides aminés hydrophobes contenus dans la région structurée prédite. DomHCA fournit aussi une information sur la présence de passages membranaires multiples ou isolés. Notre étude sur les régions de répétitions, les peptides de fusion et les porines pourra également être utilisée pour compléter les fonctionnalités à DomHCA. D'ici la fin 2005, DomHCA sera accessible via un serveur basé à l'Institut de Minéralogie et de Physique des milieux condensés (IMPMC). La particularité de cette méthode est d'être rapide et facile d'utilisation. Elle peut donc être utilisée pour des études à grande échelle des génomes, afin notamment d'effectuer un premier découpage des séquences en régions structurées potentielles.

Dans ce contexte, nous avons étudié le génome de *Plasmodium falciparum* qui a la particularité d'être riche en bases A+T. Cette composition biaisée rend peu efficace les recherches automatiques de similarité. Nous avons soumis ce génome et trois autres génomes à DomHCA (*Homo sapiens*, *Saccharomyces cerevisiae* et *Caenorhabditis elegans*). Bien que les protéines de *Plasmodium falciparum* soient beaucoup plus longues que celles des trois autres génomes (près du double), nous avons constaté que le nombre d'acides aminés contenus dans l'ensemble des régions structurées prédites par DomHCA à partir des séquences de chacun des génomes est sensiblement identique et avoisine 75%. Seul, le génome humain semble caractérisé par une valeur plus faible, de l'ordre de 64%. Il serait intéressant d'étudier d'autres génomes proches de l'homme afin de voir si cette valeur est

propre à celui-ci ou est constante parmi l'ensemble des vertébrés ou d'un clade supérieur. Chez *Plasmodium falciparum*, les régions de répétition contiennent près de 4% des acides aminés (contre moins de 1% chez les autres génomes), mais présentent des tailles moyennes similaires. Enfin, l'abondance des résidus N et K dans les protéines est retrouvée également dans les régions de répétitions et les régions structurées. Les résidus N et K sont en général uniformément répartis le long de la séquence et constituent parfois des régions de faible complexité encadrant des régions structurées. La conservation de la proportion totale des acides aminés hydrophobes chez *P. falciparum* et le comportement compensatoire des autres couples d'acides aminés permettent de supposer que les coeurs hydrophobes des domaines fonctionnels sont conservés et qu'ils peuvent être identifiés en utilisant notamment la méthode HCA.

Le découpage automatique des séquences peut s'avérer un outil utile pour la modélisation et l'étude expérimentale des structures protéiques. En effet, la délimitation des régions structurées contribue à focaliser les simulations ou les expériences sur les régions importantes dans le repliement. Ainsi, les régions fortement flexibles composées de boucles ou ayant peu de structures secondaires peuvent être mises de côté afin d'augmenter les chances de succès des investigations expérimentales et théoriques. Quelques études ont déjà été entreprises dans ce contexte, à l'aide de DomHCA (Friedrich Rippmann de Merck, communication personnelle). Avec le développement de la génomique et les nombreux séquençages, il est important de pouvoir traiter de nombreuses données rapidement. DomHCA permet de prédécouper rapidement l'ensemble des protéines issues de génomes complets et d'en extraire les régions structurées potentielles. Peut-être pourrions nous envisager de trier, d'indexer les domaines structurés des protéines à l'aide de paramètres de texture et d'identifier les inter-relations entre domaines. Nous pourrions ainsi comparer différentes protéines dans des génomes différents et mettre éventuellement en évidence des mécanismes de duplication ou de fusion de domaines. Il serait également intéressant de soumettre l'ensemble des protéines issues des génomes d'autres espèces de *Plasmodium*, du génome *Dictyostelium discoideum* et d'autres génomes de parasites à la procédure DomHCA pour vérifier si l'on retrouve les mêmes caractéristiques des régions structurées accompagnées d'un biais en A+T chez tous les parasites ou uniquement chez les espèces du genre *Plasmodium*. Dans le cadre d'étude des génomes, nous envisageons de stocker dans une base de données les résultats du découpage en domaines des protéines des quatre génomes

Plasmodium falciparum, *Homo sapiens*, *Saccharomyces cerevisiae* et *Caenorhabditis elegans*.

Mon travail de thèse principal sur la texture et la mise en place d'une procédure de découpage des protéines en régions structurées est associé à deux publications:

- DomHCA: automated prediction of protein structured region from a single sequence using secondary structure organization.

K.Prat-Albeau, JP.Mornon et I.Callebaut, soumis (article inséré à la fin de la thèse).

- Prediction of the general transcription factors associated with RNA polymerase II in *Plasmodium falciparum*: conserved features and differences relative to other eukaryotes, I.Callebaut, **K.Prat**, E.Meurice, JP.Mornon and S.Tomavo, BMC Genomics, 2005, sous presse.

Les six premiers mois de cette thèse, consacrés à la finition d'un travail sur la vérification de l'hypothèse de deux évènements de duplication entre la séparation des invertébrés et des vertébrés, ont donné lieu à deux publications :

- Phylogenetic analysis of *Ciona intestinalis* gene superfamilies supports the hypothesis of successive gene expansions.

M.Leveugle*, **K.Prat***, C.Popovici, D.Birnbaum and F.Coulier. J Mol Evol., 2004, **58** : 168-181 (article inséré à la fin de la thèse).

- Paradb : A tool for paralogy mapping in vertebrate genomes.

M.Leveugle, **K.Prat**, N.Perrier, D.Birnbaum and F.Coulier. Nucleic Acids Research., 2003, **31** (1) : 63-67.

Autre collaboration:

- Characterization and study of a kappa casein-like chymosin-sensitive linkage.

I.Callebaut, F.Schoentgen, **K.Prat**, JP.Mornon, P. Jollès. Biochimica et Biophysica Acta., 2005, **1749** (1):75-80.

* M.Leveugle et K.Prat ont contribué également à ce travail.

Bibliographie

- AASLAND, R. et Stewart, A. F. (1995). "The chromo shadow domain, a second chromo domain in heterochromatin-binding protein 1, HP1." Nucleic Acids Res **23**(16): 3168-73.
- ABI-RACHED, L., Gilles, A., Shiina, T., Pontarotti, P. et Inoko, H. (2002). "Evidence of en bloc duplication in vertebrate genomes." Nat Genet **31**(1): 100-5.
- ALBERTS, B., Bray, D., Lewis, J., Raff, M., Roberts, K. et D. Watson, J. (1997). Biologie moléculaire de la cellule. Flammarion Paris.
- ANDRADE, M. A., Ponting, C. P., Gibson, T. J. et Bork, P. (2000). "Homology-based method for identification of protein repeats using statistical significance estimates." J Mol Biol **298**(3): 521-37.
- ANFENSEN, C. B. (1973). "Principles that govern the folding of protein chains." Science **181**(96): 223-30.
- APOSTOLICO, A., Comin, M. et Parida, L. (2005). "Conservative extraction of over-represented extensible motifs." Bioinformatics **21 Suppl 1**: i9-i18.
- ARAVIND, L., Iyer, L. M., Wellems, T. E. et Miller, L. H. (2003). "Plasmodium biology: genomic gleanings." Cell **115**(7): 771-85.
- BAGOS, P. G., Liakopoulos, T. D. et Hamodrakas, S. J. (2005). "Evaluation of methods for predicting the topology of beta-barrel outer membrane proteins and a consensus prediction method." BMC Bioinformatics **6**(1): 7.
- BAIROCH, A. et Apweiler, R. (2000). "The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000." Nucleic Acids Res **28**(1): 45-8.
- BARBEAU, J., Vignes-Lebbe, R. et Stamon, G. (2002). A signature based on Delaunay graph and Co-occurrence matrix. International Conference on Computer Vision and Graphics.
- BARRETTE, I., Poisson, G., Gendron, P. et Major, F. (2001). "Pseudoknots in prion protein mRNAs confirmed by comparative sequence analysis and pattern searching." Nucleic Acids Res **29**(3): 753-8.
- BASTIEN, O., Lespinats, S., Roy, S., Metayer, K., Fertil, B., Codani, J. J. et Marechal, E. (2004). "Analysis of the compositional biases in Plasmodium falciparum genome and proteome using Arabidopsis thaliana as a reference." Gene **336**(2): 163-73.
- BATEMAN, A., Birney, E., Cerruti, L., Durbin, R., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M. et Sonnhammer, E. L. (2002). "The Pfam protein families database." Nucleic Acids Res **30**(1): 276-80.
- BERMAN, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. et Bourne, P. E. (2000). "The Protein Data Bank." Nucleic Acids Res **28**(1): 235-42.

- BEZY-WENDLING, J. (1997). Modélisation vasculaire et analyse de texture. Thèse de l'Université Rennes 1.
- BRACKEN, C., Iakoucheva, L. M., Romero, P. R. et Dunker, A. K. (2004). "Combining prediction, computation and experiment for the characterization of protein disorder." Curr Opin Struct Biol **14**(5): 570-6.
- BRANDEN, C et Tooze, J (1991). Introduction to protein structure. G. Publishing.
- BRENNER, S. E., Koehl, P. et Levitt, M. (2000). "The ASTRAL compendium for protein structure and sequence analysis." Nucleic Acids Res **28**(1): 254-6.
- BROCCHIERI, L. et Karlin, S. (2005). "Protein length in eukaryotic and prokaryotic proteomes." Nucleic Acids Res **33**(10): 3390-400.
- BRODATZ, P. (1966). Textures: A photographic album for artists and designers. Dover New York.
- BUSETTA, B. et Barrans, Y. (1984). "The prediction of protein domains." Biochim Biophys Acta **790**(2): 117-24.
- CALLEBAUT, I., Courvalin, J. C. et Mornon, J. P. (1999). "The BAH (bromo-adjacent homology) domain: a link between DNA methylation, replication and transcriptional regulation." FEBS Lett **446**(1): 189-93.
- CALLEBAUT, I., de Gunzburg, J., Goud, B. et Mornon, J. P. (2001). "RUN domains: a new family of domains involved in Ras-like GTPase signaling." Trends Biochem Sci **26**(2): 79-83.
- CALLEBAUT, I., Eudes, R., Mornon, J. P. et Lehn, P. (2004). Nucleotide-binding domains of human cystic fibrosis transmembrane conductance regulator: detailed sequence analysis and three-dimensional modeling of the heterodimer." Cell Mol Life Sci **61**(2): 230-42.
- CALLEBAUT, I., Labesse, G., Durand, P., Poupon, A., Canard, L., Chomilier, J., Henrissat, B. et Mornon, J. P. (1997a). "Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives." Cell Mol Life Sci **53**(8): 621-45.
- CALLEBAUT, I. et Mornon, J. P. (1997b). "From BRCA1 to RAP1: a widespread BRCT module closely associated with DNA repair." FEBS Lett **400**(1): 25-30.
- CALLEBAUT, I. et Mornon, J. P. (1997c). "The human EBNA-2 coactivator p100: multidomain organization and relationship to the staphylococcal nuclease fold and to the tudor protein involved in *Drosophila melanogaster* development." Biochem J **321** (Pt 1): 125-32.
- CALLEBAUT, I. et Mornon, J. P. (1998). "The V(D)J recombination activating protein RAG2 consists of a six-bladed propeller and a PHD fingerlike domain, as revealed by sequence analysis." Cell Mol Life Sci **54**(8): 880-91.
- CALLEBAUT, I., Prat, K., Meurice, E., Mornon, J-P. et Tomavo, S. (2005). "Prediction of the general transcription factors associated with RNA polymerase II in *Plasmodium falciparum*: conserved features and differences relative to other eukaryotes." BMC Genomics, sous presse.

CHAKRABARTI, S., Anand, A. P., Bhardwaj, N., Pugalenthi, G. et Sowdhamini, R. (2005). "SCANMOT: searching for similar sequences using a simultaneous scan of multiple sequence motifs." Nucleic Acids Res **33**(Web Server issue): W274-6.

CHOTHIA, C. et Gerstein, M. (1997). "Protein evolution. How far can sequences diverge?" Nature **385**(6617): 579, 581.

COEYTAUX, K. (2004). Inhibition de la ribonucléotide réductase par les oxydes d'azote, Détection des domaines non-repliés dans les séquences protéiques - Application aux protéines de la réparation de l'ADN. Thèse de l'Université Paris XI.

COEYTAUX, K. et Poupon, A. (2005). "Prediction of unfolded segments in a protein sequence based on amino acid composition." Bioinformatics **21**: 1891-900.

CORPET, F., Servant, F., Gouzy, J. et Kahn, D. (2000). "ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons." Nucleic Acids Res **28**(1): 267-9.

CSERZO, M., Wallin, E., Simon, I., von Heijne, G. et Elofsson, A. (1997). "Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method." Protein Eng **10**(6): 673-6.

CUTHBERTSON, J. M., Doyle, D. A. et Sansom, M. S. (2005). "Transmembrane helix prediction: a comparative evaluation and analysis." Protein Eng Des Sel **18**(6): 295-308.

DEBER, C. M., Brandl, C. J., Deber, R. B., Hsu, L. C. et Young, X. K. (1986). "Amino acid composition of the membrane and aqueous domains of integral membrane proteins." Arch Biochem Biophys **251**(1): 68-76.

DEL ANGEL, V. D., Dupuis, F., Mornon, J. P. et Callebaut, I. (2002). "Viral fusion peptides and identification of membrane-interacting segments." Biochem Biophys Res Commun **293**(4): 1153-60.

DOMINGUEZ DEL ANGEL, V. (2003). Prédiction de relations séquence-structure-fonction: modélisation de la botrocétine, une protéine du venin de B. jararaca - Prédiction de segments potentiellement fusogènes, Thèse de l'Université Paris VII.

DUNKER, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M. et Obradovic, Z. (2002). "Intrinsic disorder and protein function." Biochemistry **41**(21): 6573-82.

DUNKER, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. M., Hipps, K. W. et al. (2001a). "Intrinsically disordered protein." J Mol Graph Model **19**(1): 26-59.

DUNKER, A. K. et Obradovic, Z. (2001b). "The protein trinity--linking function and disorder." Nat Biotechnol **19**(9): 805-6.

DYSON, H. J. et Wright, P. E. (2005). "Intrinsically unstructured proteins and their functions." Nat Rev Mol Cell Biol **6**(3): 197-208.

EICHINGER, L., Pachebat, J. A., Glockner, G., Rajandream, M. A., Sugang, R., Berriman, M., Song, J., Olsen, R., Szafranski, K., Xu, Q. et al. (2005). "The genome of the social amoeba Dictyostelium discoideum." Nature **435**(7038): 43-57.

- FINK, A. L. (2005). "Natively unfolded proteins." Curr Opin Struct Biol **15**(1): 35-41.
- GABORIAUD, C., Bissery, V., Benchetrit, T. et Mornon, J. P. (1987). "Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences." FEBS Lett **224**(1): 149-55.
- GALZITSKAYA, O. V. et Melnik, B. S. (2003). "Prediction of protein domain boundaries from sequence alone." Protein Sci **12**(4): 696-701.
- GARDNER, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., Carlton, J. M., Pain, A., Nelson, K. E., Bowman, S. et al. (2002). "Genome sequence of the human malaria parasite *Plasmodium falciparum*." Nature **419**(6906): 498-511.
- GEORGE, R. A. et Heringa, J. (2002a). "Protein domain identification and improved sequence similarity searching using PSI-BLAST." Proteins **48**(4): 672-81.
- GEORGE, R. A. et Heringa, J. (2002b). "SnapDRAGON: a method to delineate protein structural domains from sequence data." J Mol Biol **316**(3): 839-51.
- GEORGE, R. A. et Heringa, J. (2003). "An analysis of protein domain linkers: their classification and role in protein folding." Protein Eng **15**: 871-79.
- GIRAULT, J. A., Labesse, G., Mornon, J. P. et Callebaut, I. (1999). "The N-termini of FAK and JAKs contain divergent band 4.1 domains." Trends Biochem Sci **24**(2): 54-7.
- GLOCKNER, G., Eichinger, L., Szafranski, K., Pachebat, J. A., Bankier, A. T., Dear, P. H., Lehmann, R., Baumgart, C., Parra, G., Abril, J. F. et al. (2002). "Sequence and analysis of chromosome 2 of *Dictyostelium discoideum*." Nature **418**(6893): 79-85.
- GOVINDARAJAN, S., Recabarren, R. et Goldstein, R. A. (1999). "Estimating the total number of protein folds." Proteins **35**(4): 408-14.
- GRACY, J. et Argos, P. (1998). "Automated protein sequence database classification. II. Delineation Of domain boundaries from sequence similarities." Bioinformatics **14**(2): 174-87.
- GRAIS, B (1992). Méthodes statistiques. Dunod Paris.
- GROMIHA, M. M., Ahmad, S. et Suwa, M. (2004). "Neural network-based prediction of transmembrane beta-strand segments in outer membrane proteins." J Comput Chem **25**(5): 762-7.
- GUTMAN, R., Berezin, C., Wollman, R., Rosenberg, Y. et Ben-Tal, N. (2005). "QuasiMotifFinder: protein annotation by searching for evolutionarily conserved motif-like patterns." Nucleic Acids Res **33**(Web Server issue): W255-61.
- HADLEY, C. et Jones, D. T. (1999). "A systematic comparison of protein structure classifications: SCOP, CATH and FSSP." Structure Fold Des **7**(9): 1099-112.
- HENNETIN, J. (2003). Texture hydrophobe HCA des séquences protéiques. Regards sur les introns., Thèse de l'Université Paris VII.

HENNETIN, J., Le, T. K., Canard, L., Colloc'h, N., Mornon, J. P. et Callebaut, I. (2003). "Non-intertwined binary patterns of hydrophobic/nonhydrophobic amino acids are considerably better markers of regular secondary structures than nonconstrained patterns." Proteins **51**(2): 236-44.

HERLIDOU, S. (1999). Caractérisation tissulaire en imagerie par RMN par l'analyse de texture: Etude du tissu musculaire et de tumeurs intracrâniennes. Thèse de l'Université Rennes 1.

HOLM, L. et Sander, C. (1994). "Parser for protein folding units." Proteins **19**(3): 256-68.

HOLM, L. et Sander, C. (1995). "3-D lookup: fast protein structure database searches at 90% reliability." Proc Int Conf Intell Syst Mol Biol **3**: 179-87.

ISLAM, S.A., Luo, J. et Sternberg, M.J. (1995). "Identification and analysis of domains in proteins." Protein Eng **8**: 513-525.

JACOBONI, I., Martelli, P. L., Fariselli, P., De Pinto, V. et Casadio, R. (2001). "Prediction of the transmembrane regions of beta-barrel membrane proteins with a neural network-based predictor." Protein Sci **10**(4): 779-87.

JONES, D. T., Taylor, W. R. et Thornton, J. M. (1994). "A model recognition approach to the prediction of all-helical membrane protein structure and topology." Biochemistry **33**(10): 3038-49.

JONES, D. T. et Ward, J. J. (2003). "Prediction of disordered regions in proteins from position specific score matrices." Proteins **53 Suppl 6**: 573-8.

JONES, S., Stewart, M., Michie, A., Swindells, M. B., Orengo, C. et Thornton, J. M. (1998). "Domain assignment for protein structures using a consensus approach: characterization and analysis." Protein Sci **7**(2): 233-42.

KABSCH, W. et Sander, C. (1983). "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features." Biopolymers **22**(12): 2577-637.

KAM, L. et Blanc-Talon, J. (2000). Multifractal texture segmentation for off-road robot vision. Advances in Intelligent Systems: Theory and Applications. I. Press. Netherlands.

KENDREW, J.C. et Dickerson, R.E. (1960). "Structure of myoglobin. A three dimensional Fourier synthesis at Angstrom of resolution." Nature **185**: 422-27.

KIKUCHI, T., Nemethy, G. et Scheraga, H. A. (1988). "Prediction of the location of structural domains in globular proteins." J Protein Chem **7**(4): 427-71.

KOH, I. Y., Eyrich, V. A., Marti-Renom, M. A., Przybylski, D., Madhusudhan, M. S., Eswar, N., Grana, O., Pazos, F., Valencia, A., Sali, A. et al. (2003). "EVA: Evaluation of protein structure prediction servers." Nucleic Acids Res **31**(13): 3311-5.

KOLPAKOV, R., Bana, G. et Kucherov, G. (2003). "mreps: Efficient and flexible detection of tandem repeats in DNA." Nucleic Acids Res **31**(13): 3672-8.

- KOSHI, J. M. et Goldstein, R. A. (1997). "Mutation matrices and physical-chemical properties: correlations and implications." Proteins **27**(3): 336-44.
- KURTZ, S. et Schleiermacher, C. (1999). "REPuter: fast computation of maximal repeats in complete genomes." Bioinformatics **15**(5): 426-7.
- KYTE, J. et Doolittle, R. F. (1982). "A simple method for displaying the hydropathic character of a protein." J Mol Biol **157**(1): 105-32.
- LABESSE, G., Colloc'h, N., Pothier, J. et Mornon, J. P. (1997). "P-SEA: a new efficient assignment of secondary structure from C alpha trace of proteins." Comput Appl Biosci **13**(3): 291-5.
- LADUNGA, I. et Smith, R. F. (1997). "Amino acid substitutions preserve protein folding by conserving steric and hydrophobicity properties." Protein Eng **10**(3): 187-96.
- LEFEBVRE, A., Lecroq, T., Dauchel, H. et Alexandre, J. (2003). "FORRepeats: detects repeats on entire chromosomes and between genomes." Bioinformatics **19**(3): 319-26.
- LERCHER, M. J., Smith, N. G., Eyre-Walker, A. et Hurst, L. D. (2002). "The evolution of isochores: evidence from SNP frequency distributions." Genetics **162**(4): 1805-10.
- LESZCZYNSKI, J. F. et Rose, G. D. (1986). "Loops in globular proteins: a novel category of secondary structure." Science **234**(4778): 849-55.
- LETUNIC, I., Goodstadt, L., Dickens, N. J., Doerks, T., Schultz, J., Mott, R., Ciccarelli, F., Copley, R. R., Ponting, C. P. et Bork, P. (2002). "Recent improvements to the SMART domain-based sequence annotation resource." Nucleic Acids Res **30**(1): 242-4.
- LEVEUGLE, M. (2004). Evolution et duplications des génomes de vertébrés: Base de données et phylogénies. Thèse de l'Université d'Aix Marseille II.
- LEVEUGLE, M., Prat, K., Popovici, C., Birnbaum, D. et Coulier, F. (2004). "Phylogenetic analysis of Ciona intestinalis gene superfamilies supports the hypothesis of successive gene expansions." J Mol Evol **58**(2): 168-81.
- LEVITT, M. et Chothia, C. (1976). "Structural patterns in globular proteins." Nature **261**(5561): 552-8.
- LEVIVIER, E., Goud, B., Souchet, M., Calmels, T. P., Mornon, J. P. et Callebaut, I. (2001). "uDENN, DENN, and dDENN: indissociable domains in Rab and MAP kinases signaling pathways." Biochem Biophys Res Commun **287**(3): 688-95.
- LINDING, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J. et Russell, R. B. (2003a). "Protein disorder prediction: implications for structural proteomics." Structure (Camb) **11**(11): 1453-9.
- LINDING, R., Russell, R. B., Neduva, V. et Gibson, T. J. (2003b). "GlobPlot: Exploring protein sequences for globularity and disorder." Nucleic Acids Res **31**(13): 3701-8.
- LIU, J. et Rost, B. (2003). "NORSp: Predictions of long regions without regular secondary structure." Nucleic Acids Res **31**(13): 3833-5.

LIU, J. et Rost, B. (2004). "CHOP: parsing proteins into structural domains." Nucleic Acids Res **32**(Web Server issue): W569-71.

LORETTE, A. (1999). Analyse de texture par méthodes markovienne et par morphologie mathématique: application à l'analyse des zones urbaines sur des images satellitaires. INRIA.

MAILLET, M (1995). Biologie cellulaire. Masson.

MARSDEN, R. L., McGuffin, L. J. et Jones, D. T. (2002). "Rapid protein domain assignment from amino acid sequence using predicted secondary structure." Protein Sci **11**(12): 2814-24.

MICHIE, A. D., Orengo, C. A. et Thornton, J. M. (1996). "Analysis of domain structural class using an automated class assignment protocol." J Mol Biol **262**(2): 168-85.

MOLLER, S., Croning, M. D. et Apweiler, R. (2001). "Evaluation of methods for the prediction of membrane spanning regions." Bioinformatics **17**(7): 646-53.

MORNON, J. P., Prat, K., Dupuis, F., Boisset, N. et Callebaut, I. (2002a). "Structural features of prions explored by sequence analysis. II. A PrP(Sc) model." Cell Mol Life Sci **59**(12): 2144-54.

MORNON, J. P., Prat, K., Dupuis, F. et Callebaut, I. (2002b). "Structural features of prions explored by sequence analysis I. Sequence data." Cell Mol Life Sci **59**(8): 1366-76.

MURZIN, A. G., Brenner, S. E., Hubbard, T. et Chothia, C. (1995). "SCOP: a structural classification of proteins database for the investigation of sequences and structures." J Mol Biol **247**(4): 536-40.

NATT, N. K., Kaur, H. et Raghava, G. P. (2004). "Prediction of transmembrane regions of beta-barrel proteins using ANN- and SVM-based methods." Proteins **56**(1): 11-8.

NEVILL-MANNING, C. G., Wu, T. D. et Brutlag, D. L. (1998). "Highly specific protein sequence motifs for genome analysis." Proc Natl Acad Sci U S A **95**(11): 5865-71.

NIELSEN, H., Brunak, S. et von Heijne, G. (1999). "Machine learning approaches for the prediction of signal peptides and other protein sorting signals." Protein Eng **12**(1): 3-9.

NURIZZO, D., Nagy, T., Gilbert, H. J. et Davies, G. J. (2002). "The structural basis for catalysis and specificity of the *Pseudomonas cellulosa* alpha-glucuronidase, GlcA67A." Structure (Camb) **10**(4): 547-56.

OHNO, S. (1970). Evolution by gene duplication. B. Springer Verlag.

OHNO, S. (1993). "Patterns in genome evolution." Curr Opin Genet Dev **3**(6): 911-4.

ORENGO, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. et Thornton, J. M. (1997). "CATH--a hierarchic classification of protein domain structures." Structure **5**(8): 1093-108.

PARK, J. et Teichmann, S. A. (1998). "DIVCLUS: an automatic method in the GEANFAMMER package that finds homologous domains in single- and multi-domain proteins." Bioinformatics **14**(2): 144-50.

- PECHEUR, E. I., Martin, I., Bienvenue, A., Ruyschaert, J. M. et Hoekstra, D. (2000). "Protein-induced fusion can be modulated by target membrane lipids through a structural switch at the level of the fusion peptide." J Biol Chem **275**(6): 3936-42.
- PERSSON, B. et Argos, P. (1996). "Topology prediction of membrane proteins." Protein Sci **5**(2): 363-71.
- PINTAR, A., Carugo, O. et Pongor, S. (2003a). "Atom depth as a descriptor of the protein interior." Biophys J **84**(4): 2553-61.
- PINTAR, A., Carugo, O. et Pongor, S. (2003b). "Atom depth in protein structure and function." Trends Biochem Sci **28**(11): 593-7.
- PIZZI, E. et Frontali, C. (2001). "Low-complexity regions in Plasmodium falciparum proteins." Genome Res **11**(2): 218-29.
- PROMPONAS, V. J., Enright, A. J., Tsoka, S., Kreil, D. P., Leroy, C., Hamodrakas, S., Sander, C. et Ouzounis, C. A. (2000). "CAST: an iterative algorithm for the complexity analysis of sequence tracts. Complexity analysis of sequence tracts." Bioinformatics **16**(10): 915-22.
- RAMAKRISHNAN, C. et Ramachandran, G. N. (1965). "Stereochemical criteria for polypeptide and protein chain conformations. II. Allowed conformations for a pair of peptide units." Biophys J **5**(6): 909-33.
- RICHARDSON, J. S. et Richardson, D. C. (2002). "Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation." Proc Natl Acad Sci U S A **99**(5): 2754-9.
- RIGDEN, D. J. (2002). "Use of covariance analysis for the prediction of structural domain boundaries from multiple protein sequence alignments." Protein Eng **15**(2): 65-77.
- ROMERO, P., Obradovic, Z., Li, X., Garner, E. C., Brown, C. J. et Dunker, A. K. (2001). "Sequence complexity of disordered protein." Proteins **42**(1): 38-48.
- ROST, B., Casadio, R., Fariselli, P. et Sander, C. (1995). "Transmembrane helices predicted at 95% accuracy." Protein Sci **4**(3): 521-33.
- ROST, B., Fariselli, P. et Casadio, R. (1996). "Topology prediction for helical transmembrane proteins at 86% accuracy." Protein Sci **5**(8): 1704-18.
- SAINI, H. K. et Fischer, D. (2005). "Meta-DP: domain prediction meta server." Bioinformatics **21**: 2917-20.
- SANDER, C. et Schneider, R. (1991). "Database of homology-derived protein structures and the structural meaning of sequence alignment." Proteins **9**(1): 56-68.
- SERVANT, F., Bru, C., Carrere, S., Courcelle, E., Gouzy, J., Peyruc, D. et Kahn, D. (2002). "ProDom: automated clustering of homologous domains." Brief Bioinform **3**(3): 246-51.
- SIDDIQUI, A. S. et Barton, G. J. (1995). "Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions." Protein Sci **4**(5): 872-84.

- SIEW, N. et Fischer, D. (2004). "Structural biology sheds light on the puzzle of genomic ORFans." J Mol Biol **342**(2): 369-73.
- SIGRIST, C. J., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A. et Bucher, P. (2002). "PROSITE: a documented database using patterns and profiles as motif descriptors." Brief Bioinform **3**(3): 265-74.
- SOWDHAMINI, R., Rufino, S. D. et Blundell, T. L. (1996). "A database of globular protein structural domains: clustering of representative family members into similar folds." Fold Des **1**(3): 209-20.
- SOYER, A., Chomilier, J., Mornon, J. P., Jullien, R. et Sadoc, J. F. (2000). "Voronoi tessellation reveals the condensed matter character of folded proteins." Phys Rev Lett **85**(16): 3532-5.
- SUYAMA, M. et Ohara, O. (2003). "DomCut: prediction of inter-domain linker regions in amino acid sequences." Bioinformatics **19**(5): 673-4.
- SWINDELLS, M. B. (1995). "A procedure for detecting structural domains in proteins." Protein Sci **4**(1): 103-12.
- SZKLARCZYK, R. et Heringa, J. (2004). "Tracking repeats using significance and transitivity." Bioinformatics **20 Suppl 1**: I311-I317.
- TANAKA, T., Kuroda, Y. et Yokoyama, S. (2003). "Characteristics and prediction of domain linker sequences in multi-domain proteins." J Struct Funct Genomics **4**(2-3): 79-85.
- TAYLOR, W. R. et Orengo, C. A. (1989). "Protein structure alignment." J Mol Biol **208**(1): 1-22.
- TUSNADY, G. E. et Simon, I. (2001). "The HMMTOP transmembrane topology prediction server." Bioinformatics **17**(9): 849-50.
- UVERSKY, V. N., Gillespie, J. R. et Fink, A. L. (2000). "Why are "natively unfolded" proteins unstructured under physiologic conditions?" Proteins **41**(3): 415-27.
- VON HEIJNE, G. (1992). "Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule." J Mol Biol **225**(2): 487-94.
- VONDERVISZT, F. et Simon, I. (1986). "A possible way for prediction of domain boundaries in globular proteins from amino acid sequence." Biochem Biophys Res Commun **139**(1): 11-7.
- VUCETIC, S., Brown, C. J., Dunker, A. K. et Obradovic, Z. (2003). "Flavors of protein disorder." Proteins **52**(4): 573-84.
- WANG, W. et Hecht, M. H. (2002). "Rationally designed mutations convert de novo amyloid-like fibrils into monomeric beta-sheet proteins." Proc Natl Acad Sci U S A **99**(5): 2760-5.
- WANG, Z. X. (1998). "A re-estimation for the total numbers of protein folds and superfamilies." Protein Eng **11**(8): 621-6.

- WARD, J. J., McGuffin, L. J., Bryson, K., Buxton, B. F. et Jones, D. T. (2004a). "The DISOPRED server for the prediction of protein disorder." Bioinformatics **20**(13): 2138-9.
- WARD, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. et Jones, D. T. (2004b). "Prediction and functional analysis of native disorder in proteins from the three kingdoms of life." J Mol Biol **337**(3): 635-45.
- WHEELAN, S. J., Marchler-Bauer, A. et Bryant, S. H. (2000). "Domain size distributions can predict domain boundaries." Bioinformatics **16**(7): 613-8.
- WOLF, Y.I., Grishin, N.V. et Koonin, E. V. (2000). "Estimating the number of protein folds and families from complete genome data." J Mol Biol **299**: 897-905.
- WOODCOCK, S., Mornon, J. P. et Henrissat, B. (1992). "Detection of secondary structure elements in proteins by hydrophobic cluster analysis." Protein Eng **5**(7): 629-35.
- WOODSMALL, R. M. et Benson, D. A. (1993). "Information resources at the National Center for Biotechnology Information." Bull Med Libr Assoc **81**(3): 282-4.
- WOOTTON, J. C. (1994). "Non-globular domains in protein sequences: automated segmentation using complexity measures." Comput Chem **18**(3): 269-85.
- WRIGHT, P. E. et Dyson, H. J. (1999). "Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm." J Mol Biol **293**(2): 321-31.
- YE, Q., Callebaut, I., Pezhman, A., Courvalin, J. C. et Worman, H. J. (1997). "Domain-specific interactions of human HP1-type chromodomain proteins and inner nuclear membrane protein LBR." J Biol Chem **272**(23): 14983-9.
- ZEMLA, A., Venclovas, C., Fidelis, K. et Rost, B. (1999). "A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment." Proteins **34**(2): 220-3.
- ZHANG, C. et DeLisi, C. (1998). "Estimating the number of protein folds." J Mol Biol **284**(5): 1301-5.

Annexe

Répartition de chaque amas dans les hélices α , brins β et boucles et état attribué d'après nos règles.

Amas	Effectif	% hélice α	% brin β	% boucle	Etat attribué
1	16 195	15.1	22.0	62.9	/
11	8 394	21.0	32.6	46.3	/
101	5 702	11.5	53.8	34.0	B
111	3 242	11.7	64.8	23.5	B
1001	3 009	32.5	27.7	38.1	?
1011	2 092	17.0	57.5	24.9	B
1101	1 834	17.2	58.7	22.9	B
1111	1 095	6.3	78.6	14.6	B
10001	2 061	34.4	34.7	29.2	?
10011	1 306	50.4	25.4	23.0	A
10101	1 441	12.1	70.4	15.3	B
10111	780	20.6	65.1	13.5	B
11001	1 580	41.5	38.1	18.4	?
11011	751	45.8	34.5	18.6	A
11101	698	15.2	70.6	12.9	B
11111	294	16.0	73.1	8.8	B
100011	604	32.8	41.4	22.2	b
100101	611	19.0	48.6	28.5	b
100111	445	24.3	53.0	18.2	B
101001	698	21.6	51.0	23.4	B
101011	518	15.1	67.0	12.2	B
101101	397	14.6	68.0	15.1	B
101111	203	5.9	81.3	9.9	B
110001	731	35.8	32.6	25.0	?
110011	499	58.9	23.2	13.6	A
110101	435	15.6	67.8	12.9	B
110111	203	34.5	56.2	7.9	B
111001	397	23.9	53.1	17.4	B
111011	250	32.0	59.2	8.4	B
111101	184	6.5	78.8	10.3	B
111111	54	13.0	81.5	5.6	B
1000101	450	29.8	46.0	15.1	b
1000111	207	25.1	55.6	15.5	B
1001001	389	42.4	25.2	23.1	a
1001011	209	22.5	54.5	17.2	B
1001101	319	36.4	46.1	10.0	b
1001111	108	27.8	63.0	5.6	B
1010001	391	22.3	54.5	17.9	B
1010011	247	32.4	42.1	12.6	b
1010101	318	6.0	81.8	7.2	B
1010111	121	19.8	66.1	7.4	B
1011001	294	29.6	50.7	10.9	B
1011011	156	19.6	57.7	16.0	B
1011101	130	20.0	77.7	2.3	B
1011111	95	10.8	89.2	0.0	B

1100111	105	30.5	48.6	11.4	b
1101001	310	16.1	59.7	18.1	B
1101011	151	13.9	66.2	14.6	B
1101101	140	30.0	52.1	10.0	B
1110001	222	19.4	57.7	13.5	B
1110011	132	31.8	50.0	7.6	B
1110101	106	13.2	75.5	5.7	B
1111001	134	8.2	76.1	4.5	B
10001001	417	46.0	35.5	13.7	a
10001011	212	37.7	43.9	13.7	?
10001101	155	27.1	49.0	13.5	b
10001111	69	15.9	63.8	7.2	B
10010001	333	66.7	18.3	11.1	A
10010011	185	73.0	15.7	5.9	A
10010101	132	19.7	57.6	12.9	B
10010111	71	28.2	56.3	4.2	B
10011001	331	82.2	11.8	4.2	A
10011011	184	53.8	37.5	6.5	A
10011101	123	35.8	53.7	4.1	B
10100011	131	34.4	32.8	16.0	?
10100101	151	16.6	47.7	19.9	b
10100111	80	18.8	41.3	15.0	b
10101001	324	15.1	77.5	3.7	B
10101011	94	12.8	73.4	6.4	B
10101101	83	14.5	69.9	7.2	B
10110001	184	39.7	44.0	5.4	?
10110011	120	51.7	35.0	6.7	A
10110101	80	5.0	75.0	8.8	B
10111001	110	36.4	57.3	5.5	B
10111101	35	5.7	88.6	0.0	B
11000101	118	36.4	34.7	14.4	?
11000111	104	10.6	40.4	1.9	b
11001001	223	66.8	17.0	10.8	A
11001011	102	39.2	42.2	12.7	?
11001101	81	44.4	37.0	7.4	?
11010001	146	30.1	52.7	9.6	B
11010011	88	44.3	39.8	6.8	?
11010101	78	17.9	67.9	6.4	B
11011001	138	73.9	17.4	6.5	A
11100011	66	18.2	51.5	13.6	B
11100101	58	13.8	56.9	8.6	B
11101001	160	11.3	80.0	3.8	B
11110001	84	9.5	60.7	10.7	B
100010001	141	52.5	21.3	14.2	A
100010011	138	68.1	15.9	8.0	A
100010101	113	27.4	52.2	5.3	B
100010111	71	29.6	56.3	4.2	B
100011001	100	62.0	25.0	10.0	A

100011011	89	53.9	29.2	6.7	A
100100011	83	47.0	19.3	9.6	a
100100111	75	26.7	52.0	8.0	B
100101001	78	32.1	38.5	19.2	?
100101011	65	24.6	67.7	6.2	B
100110001	126	69.0	16.7	9.5	A
100110011	132	61.4	25.0	2.3	A
100110101	77	18.2	55.8	18.2	B
101000101	182	6.6	75.8	4.9	B
101001001	80	27.5	41.3	16.3	b
101010001	119	14.3	44.5	7.6	b
110001001	126	54.8	22.2	12.7	A
110010001	125	70.4	15.2	3.2	A
110010011	112	72.3	17.0	0.9	A
110011001	126	69.0	24.6	4.0	A
110011011	46	73.9	10.9	4.3	A
110100011	60	55.0	23.3	8.3	A
110110001	75	66.7	21.3	5.3	A
110110011	163	19.6	54.6	1.8	B

Résumé

Découper, a priori et de façon précise, les séquences en domaines est d'une grande importance dans le champ de la biologie, notamment pour optimiser les études de génomique structurale et de génomique fonctionnelle. Différentes approches basées sur la composition en acides aminés, la complexité de la séquence ou la construction de modèles 3D ab initio, ont été développées par le passé. Nous proposons, dans le cadre de ce travail, une approche nouvelle et originale pour le découpage automatique et sensible des séquences protéiques en domaines structurés distincts par exploitation de leur texture. Cette approche bénéficie de l'information de voisinage 2D apportée par la méthodologie «Hydrophobic Cluster Analysis » (HCA), qui permet d'intégrer à l'analyse une gestion directe de la structuration secondaire à partir de la connaissance d'une seule séquence. Ainsi, la distribution des différentes catégories d'amas hydrophobes, tels que définis par l'intermédiaire de HCA, ainsi que l'analyse de leurs caractéristiques en termes de structures secondaires, permettent d'appréhender de façon différenciée les textures des régions globulaires, non globulaires et/ou désordonnées, répétitives, passages membranaires isolés ou multiples.... Par contre, l'étude, au moyen d'outils classiques d'analyse de texture 2D (méthodes de cooccurrence et des longueurs de section classiquement utilisées en analyse d'images) ne s'est pas révélée prometteuse. L'approche développée, DomHCA, permet in fine de segmenter une séquence protéique en une série de régions et sous-régions caractérisées par des textures précises, segmentation qui, appliquée à l'échelle des génomes, autorise une comparaison rapide et originale de l'ensemble des séquences. Une des applications concerne les séquences du génome de *Plasmodium falciparum* qui, par leurs fortes proportions en acides aminés N et K, rendent les méthodes classiques de détection de similarité peu efficaces.

Abstract

The accurate delineation of domains within protein sequences is of count most importance in biology, especially for optimizing the structural and functional genomics investigations. Various approaches based on the composition in amino acids, the complexity of the sequence or the construction of ab initio 3D models, were developed in the past to predict functional domains. We propose a new and original approach to automatically delineate within proteins sequences distinct structured regions by exploring their texture. This approach is based on the Hydrophobic Cluster Analysis (HCA) method, which relies on the physico-chemical and topological principles underlying the fold of globular domains and allows a direct prediction of regular secondary structures from a single amino acid sequence. Hence, the distribution of the various categories of hydrophobic clusters, as defined by HCA, and the analysis of their secondary structures characteristics, discriminate different textures and is used as an efficient tool to distinguish structured regions, disordered regions, repeat regions and potential single or multiple transmembrane segments. The study performed by means of traditional tools for analysis of 2D texture (methods of cooccurrence and of section lengths classically used in image analysis) did not appear promising. The DomHCA approach, allows to delineate regions characterized by particular textures in a protein sequence. This segmentation, applied on the scale of genomes, authorizes a fast and original comparison of the whole sequences. As an example, we applied DomHCA on the *Plasmodium falciparum* genome sequences, for which standard tools of similarity searches are generally poorly efficient, due to their high content in N and K residues.

Discipline : Sciences de la Vie et de la Terre

Mots clés : domaines, Hydrophobic Cluster Analysis (HCA), *Plasmodium falciparum*, texture.

Laboratoire de génétique et biologie cellulaire – EPHE
Université Versailles Saint-Quentin-en-Yvelines
Bât Fermat, 45 avenue des Etats-Unis 78035 Versailles