Chapitre 2
Les Protéines

Les protéines représentent plus de la moitié du poids sec des cellules. Leur importance biologique vient de leur aptitude à reconnaître d'autres molécules avec lesquelles elles s'associent de manière transitoire d'après leur configuration spatiale. Ce premier chapitre récapitulera les principales bases biochimiques des protéines, et s'attachera à différentes méthodes d'analyse et de prédiction de leur structure tridimensionnelle. La structure d'une protéine est essentiellement déterminée par sa séquence primaire en acides aminés. Dans une cellule, une séquence protéique donne lieu à un seul type de structure tridimensionnelle. Toutefois, il est particulièrement difficile encore à l'heure actuelle de prévoir la structure d'une protéine en se basant uniquement sur sa séquence [119].

2.1 La structure protéique

2.1.1 Les acides aminés

Les protéines sont des polymères, des macromolécules biologiques primordiales dans l'ensemble du règne animal et végétal. Depuis la découverte de la structure de l'ADN (Acide Désoxyribo Nucléique), le dogme de la biologie moléculaire peut se résumer en trois points:

(1) l'ADN porte l'information génétique,

(2) cette information est transcrite en ARN mono-brin qui possède une information souvent codante,

(3) avec l'aide du complexe ribosomique, l'ARN est traduit en protéine.

Les protéines sont des successions d'acides aminés simples. Il en existe 20 principaux. Un

acide aminé est composé d'un carbone asymétrique dit carbone α (C_α). Ce carbone tétravalent est lié (cf. figure 2.1) à une fonction amine NH_2 (*N: azote et H: hydrogène*), à un atome d'hydrogène H, à une fonction acide COOH et à un radical, un groupement de taille plus ou moins important appelé chaîne latérale (R).

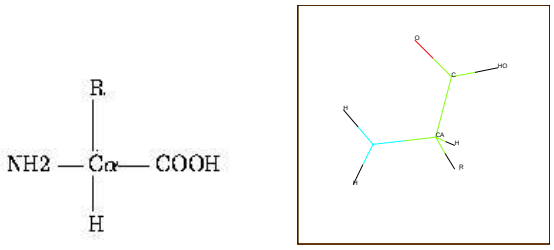

FIG. 2.1 – *Structure schématique d'un acide aminé à gauche plane et à droite visualisé en 3D par le logiciel XmMol [194]*

Les acides aminés libres sont solubles dans l'eau du fait des groupements polaires qu'ils portent. Il existe plusieurs types de chaînes latérales (R) qui se différencient selon leurs propriétés physico-chimiques :

- R hydrophobes : Alanine, Valine, Leucine, Isoleucine, Phénylalanine, Proline et Méthionine.
- R chargés : Aspartate, Glutamate, Lysine et Arginine.
- R polaires : Sérine, Thréonine, Cystéine, Acide Aspartique, Acide Glutamique, Histidine, Tyrosine et Tryptophane.
- sans R : Glycine.

Ils se différencient aussi suivant leur encombrement stérique (cf. Tableau 2.1). Certains possèdent des cycles aromatiques comme la Phénylalanine, la Tyrosine et le Tryptophane. La Proline est le seul acide aminé dont la chaîne latérale est liée aussi à l'azote du squelette polypeptidique et donc dépourvu d'hydrogène. La Cystéine possède un groupement souffré qui peut s'oxyder, ce qui permet la création de pont disulfure qui peut stabiliser de manière primordiale la structure protéique. De nombreuses études portant sur les relations et les équivalences au niveau physico-chimique des acides aminés ont été menées. On peut noter le travail de William

Taylor qui a mis au point une méthode de représentation simple prenant en compte à la fois les propriétés physico-chimiques, mais aussi le volume des acides aminés (cf. figure 2.2 [190]).

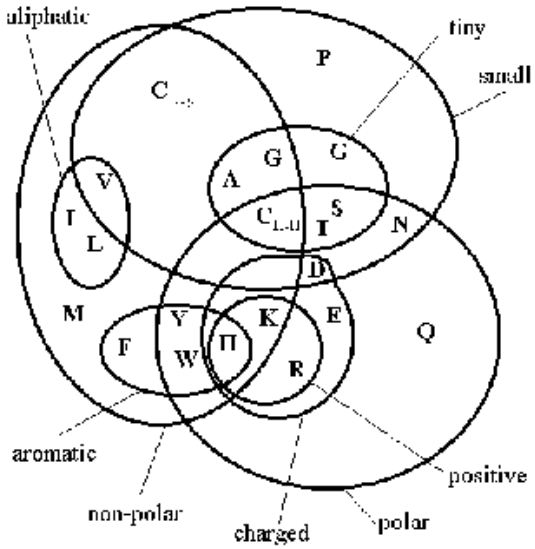

FIG. 2.2 – *Diagramme de Venn portant sur l'équivalence entre les différents acides aminés par rapport à leurs propriétés physico-chimiques et leurs volumes*

Deux acides aminés vont se lier en créant une liaison polypeptidique (cf. Figure 2.3), par condensation entre le groupement carboxyle du premier et le groupement amine du second. C'est une polymérisation qui s'effectue par perte d'une molécule d'eau. Cette nouvelle liaison implique une rigidité importante. Quand le nombre d'acides aminés est faible, la macromolécule est souvent nommée oligopeptide. Un nombre important de résidus est nécessaire pour parler de polypeptide ou protéine (les chiffres varient dans la littérature entre 20 et 50 résidus).

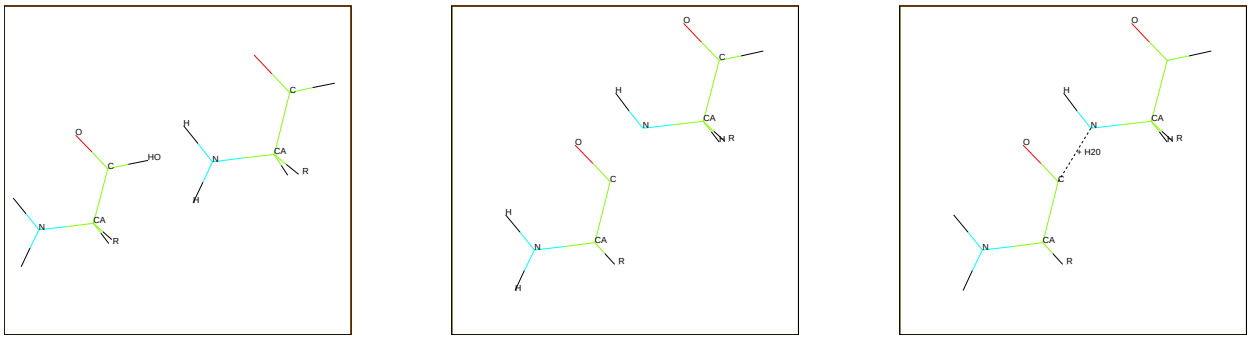

FIG. 2.3 – *Polymérisation de deux acides aminés par condensation et création d'une liaison peptidique.*

Nom	Name	3	1	masse	surface	volume	formule	représentation
Alanine	Alanine	ALA	A	71.09	115	88.6	CH_3-CH(NH_2)-COOH	
Arginine	Arginine	ARG	R	156.19	225	173.4	HN=C(NH_2)-NH-(CH_2)3-CH(NH_2)-COOH	
Acide Aspartique	Aspartic Acid	ASP	D	114.11	150	111.1	HOOC-CH_2-CH(NH_2)-COOH	
Asparagine	Asparagine	ASN	N	115.09	160	114.1	H_2N-CO-CH_2-CH(NH_2)-COOH	
Cystéine	Cysteine	CYS	C	103.15	135	108.5	HS-CH_2-CH(NH_2)-COOH	
Acide Glutamique	Glutamic Acid	GLU	E	129.12	190	138.4	HOOC-(CH_2)$_2$-CH(NH_2)-COOH	
Glutamine	Glutamine	GLN	Q	128.14	180	143.8	H_2N-CO-(CH_2)$_2$-CH(NH_2)-COOH	
Glycine	Glycine	GLY	G	57.05	75	60.1	H_2N-CH_2-COOH	
Histidine	Histidine	HIS	H	137.14	195	153.2	N*H-CH=N-CH=C*-CH_2-CH(NH_2)-COOH	

Nom	Name	3	1	masse	surface	volume	formule	représentation
Isoleucine	Isoleucine	ILE	I	113.16	175	166.7	CH_3-CH_2-CH(CH_3)-CH(NH_2)-COOH	
Leucine	Leucine	LEU	L	113.16	170	166.7	$(CH_3)_2$-CH-CH_2-CH(NH_2)-COOH	
Lysine	Lysine	LYS	K	128.17	200	168.6	H_2N-$(CH_2)_4$-CH(NH_2)-COOH	
Méthionine	Methionine	MET	M	131.19	185	162.9	CH_3-S-$(CH_2)_2$-CH(NH_2)-COOH	
Phénylalanine	Phenylalanine	PHE	F	147.18	210	189.9	Aro-CH_2-CH(NH_2)-COOH	
Proline	Proline	PRO	P	97.12	145	112.7	N*H-(CH_2)3-C*H-COOH	
Sérine	Serine	SER	S	87.08	115	89.0	HO-CH_2-CH(NH_2)-COOH	
Thréonine	Threonine	THR	T	101.11	140	116.1	CH_3-CH(OH)-CH(NH_2)-COOH	
Tryptophane	Tryptophane	TRP	W	186.12	255	227.8	Aro*-NH-CH=C*-CH_2-CH(NH_2)-COOH	
Tyrosine	Tyrosine	TYR	Y	163.18	230	193.6	HO-Aro-CH_2-CH(NH_2)-COOH	
Valine	Valine	VAL	V	99.14	155	140.0	$(CH_3)_2$-CH-CH(NH_2)-COOH	

TAB. 2.1 – *les 20 acides aminés classiques avec leur nom, en français (Nom), en anglais (Name), le code à trois lettres pour abréger le nom des acides aminés (3),le code à une lettre (1), leur masse, leur surface, une formule linéaire de leur chaîne latérale (avec * pour la fermeture du cycle et Aro pour un groupement aromatique) et enfin une représentation 3D avec le logiciel rasmol.*

On distingue alors couramment la région axiale, monotone et de composition constante (carbones et azote), appelée squelette de la chaîne polypeptidique et la succession des chaînes latérales variables selon le type d'acides aminés.

L'extrémité N-terminale correspondant au premier acide aminé de la protéine porte un groupement amine encore libre et l'extrémité C-terminale le dernier acide aminé de la protéine avec un groupement carboxyle libre aussi. Ces deux groupements à pH physiologique sont le plus souvent ionisés. La synthèse s'effectue de l'extrémité N-terminale vers l'extrémité C-terminale.

2.1.2 Les différents niveaux de la structure protéique

Les protéines peuvent-être subdivisées en trois catégories principales: (i) des protéines solubles en général globulaires, le plus souvent compartimentales (cytoplasme, noyau, mitochondrie et chloroplaste), (ii) des protéines membranaires qui sont incluses dans les membranes lipidiques, et (iii) des protéines fibreuses nécessaires pour des actions biologiques particulières telles la contraction musculaire ou le maintien du cytosquelette. Parmi les protéines membranaires, certaines possèdent à la fois des segments totalement enfouis et des régions qui baignent dans le cytoplasme.

Le calcul systématique des angles et des distances de tous les atomes des protéines cristallisées montrent qu'un certain nombre d'angles et de distances restent assez constant (cf. figure 2.4a), alors que d'autres sont beaucoup plus variables. Il s'agit par exemple des angles de valence et des longueurs de liaison tandis que les angles dièdres (ou de torsion) assurent la diversité du repliement 3D et sont donc nettement plus variables.

Différents angles sont utilisés pour décrire les protéines :

- les angles dièdres : ϕ, ψ et ω, qui décrivent le squelette; en prenant en compte pour l'angle ϕ_x (ϕ en position x dans la séquence), les atomes C'$_{x-1}$, N$_x$, Cα_x et C'$_x$, pour ψ_x,les atomes N$_x$, Cα_x et C'$_x$ et N$_{x+1}$ et pour ω, Cα_x, C'$_x$, N$_{x+1}$ et Cα_{x+1}. La figure 2.4c montre leur positionnement sur la chaîne polypeptidique.

- l'angle α : angle formé par 4 carbones α successifs.

- l'angle τ : angle formé par 3 carbones α successifs.

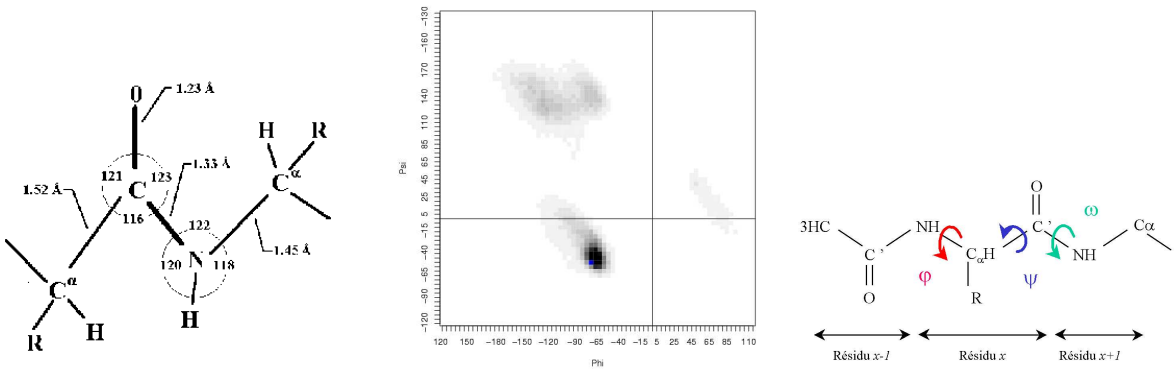

FIG. 2.4 – *(a) Valeurs classiques de certains angles et distances restant constantes dans les protéines. (b) Diagramme de Ramachandran, avec l'angle ϕ en abcisse et l'angle ψ en ordonnée. (c) Schéma des angles ϕ, ψ et ω.*

Les angles ϕ et ψ sont classiquement représentés l'un par rapport à l'autre dans un diagramme de Ramachandran (cf. figure 2.4b), qui fut utilisé pour la première fois en 1968 par G.N. Ramachandran se servant à l'époque de modèles énergétiques pour caractériser les zones préférentielles de ces angles de torsions. La succession des acides aminés correspond à la séquence primaire (1D). La structure tridimensionnelle résultant du repliement de la protéine est appelée structure tertiaire (3D). Nous nous attarderons sur une classification intermédiaire notée structure secondaire (2D). La base de données qui contient la totalité des structures 3D protéiques se nomme la Protein Data Bank ou PDB [6].

2.1.2.1 La structure secondaire

La structure secondaire se caractérise classiquement par les hélices α et les feuillets β, structures possédant des caractéristiques répétitives d'intérêt.

hélices et feuillets Hélice α et feuillet β sont les deux formes répétitives les plus importantes des protéines représentant chacune respectivement 30 et 20 % des résidus. Cette importance vient de leur stabilité énergétique particulière [184].

Les hélices α tournent dans le sens dit "main droite". Les chaînes latérales sont situées à l'extérieur de l'hélice. L'ensemble s'inscrit dans un cylindre de 10,5 Å de diamètre, le tour de spire ou pas de l'hélice fait 5,4 Å soit 3.6 résidus. L'hélice α est une structure thermodynamiquement stable du fait de nombreuses liaisons hydrogènes entre groupements amines (NH_2) et

carboxyles($C = O$). La figure 2.5a montre la structure de l'hémoglobine (code PDB : 1bbb) qui est une protéine pratiquement tout-α, ces hélices assurent intégralement la fonction biologique de la protéine. La figure 2.5b montre le domaine 3 de l'hélicase de *Escherichia coli* (code PDB: 1cuk) qui est un domaine comportant trois hélices α. La figure 2.5c montre le squelette de la chaîne polypeptidique. La répétitivité du motif est nette, les groupements carboxyles sont presque parallèles. Les liaisons hydrogènes existantes entre les résidus sont indiquées en pointillés. Ce sont des liaisons entre l'oxygène d'un résidu en position i et le groupement azoté d'un résidu en position $i+4$; cette liaison est notée [i:i+4].

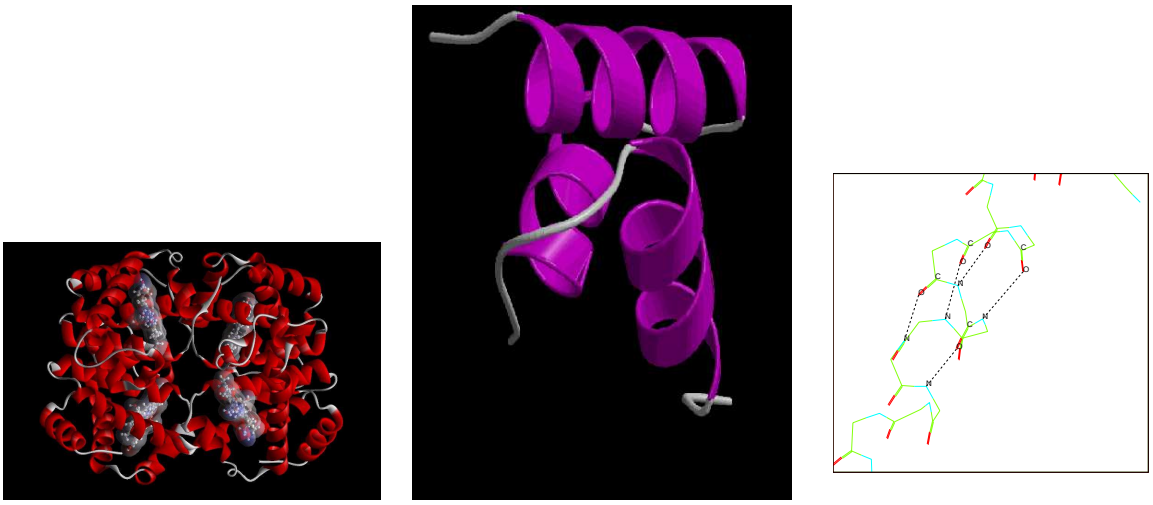

FIG. 2.5 – *a. Une proteine composée en majorité d'hélice α, la 1bbb (les hélices sont en rouges et maintiennent les hèmes). b. Squelette de la première hélice d'une hélicase de Escherichia coli (code PDB: 1cuk). c. Les liaisons entre les résidus 154 à 163 de cette protéine sont indiquées en pointillés.*

Les hélices, malgré leur définition, ne sont pas des tubes rigides. Plus d'un quart des hélices α sont fortement non-régulières [3]. Un lien direct entre la longueur de l'hélice et son degré de courbure a d'ailleurs été déterminé [113].

Les hélices possèdent une composition en acides aminés particulière. De nombreuses études ont montré aussi l'existence d'acides aminés sur-représentés aux extrémités C- et N-terminales des hélices. Une dizaine de successions spécifiques qui forment des terminaisons stables et qui induisent ces fins d'hélices ont été caractérisées [154, 2]. En plus des acides aminés classiques du type Proline ou Glycine, certains acides aminés ont un rôle structural dans ces motifs. Les Glycines se trouvent dans la zone des Ψ positifs (cf. figure 2.4b). Les hélices sont présentes

dans la zone des Φ et Ψ négatifs. Certaines classes d'hélices α ont été étudiées du fait de leur importance fonctionnelle comme les hélices amphipatiques (ayant une face polaire et une autre non-polaire) [176].

Les feuillets sont moins stables thermodynamiquement que les hélices α [184]. Ils sont dus à un aller-retour de la chaîne polypeptidique qui fait que les segments deviennent adjacents. Ils se retrouvent dans la zone des Φ négatifs et Ψ positifs (cf. figure 2.4b).

La figure 2.6a montre une porine, protéine tout-β transmembranaire, son ouverture permet le passage de divers solutés. La figure 2.6b représente le premier domaine de la protéine ldu TNF (code PDB : 1ext) d'*Escherichia coli*, où se trouvent plusieurs feuillets β. La figure 2.6c montre un agrandissement du feuillet comprenant les résidus 126-130 et 150-154. Les liaisons stabilisantes existant entre les groupements CO et NH ont été notées en pointillés. Quand les segments sont orientés dans des directions opposées, les feuillets sont dits anti-parallèles. S'ils sont orientés dans la même direction, ils sont dits parallèles.

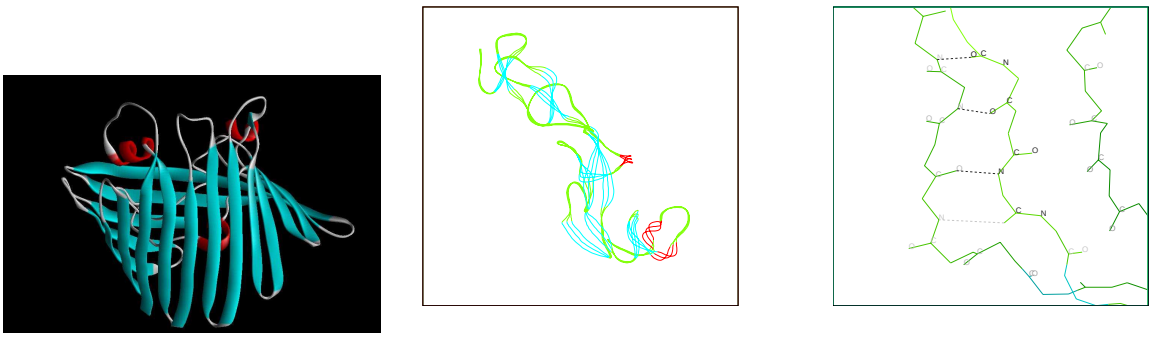

FIG. 2.6 – *(a) Représentation d'une porine, protéine fortement β, les feuillets sont en bleu. (b) Domaine 1 extracellulaire du récepteur au TNF (code PDB: 1ext) fortement β. (c) Liaisons stabilisant un feuillet β.*

Ces deux conformations ne font pas intervenir les chaînes latérales mais seulement le squelette polypetidique. Ceci distingue la structure secondaire de la structure tertiaire. Les hélices α représentent environ 30% des protéines, les feuillets β environ 20%. Le reste est souvent dénommé boucles et a fréquemment été considéré comme variable. Toutefois, on trouve d'autres formes régulières.

structures	nombre de résidus par tours	tours par résidu (Å)	rayon(Å)	Φ	Ψ
hélices α	3.6	1.5	2.3	-57	-47
feuillets β anti-parallèle	2.0	3.4	1.0	-139	135
feuillets βparallèle	2.0	3.2	1.0	-119	113

TAB. 2.2 – *valeurs classiques des angles des hélices α et des feuillets β*

L'attribution des structures secondaires est souvent un problème délicat. Cinq méthodes d'assignation sont actuellement utilisées :

(a) DSSP[100] (1983): Cette méthode est fondée sur une recherche des liaisons hydrogènes spécifiques des structures répétitives. Pour les hélices α, DSSP observe les liaisons en positions $(i, i+3)$ ou $(i, i+4)$, alors que pour les feuillets β des liaisons plus éloignées sont en jeu. DSSP ne prend en compte que des structures secondaires répétitives d'au moins 4 résidus de long.

(b) DEFINE[156] (1988): Elle se base sur les distances entre carbones α. Ces distances sont comparées à des distances de structures secondaires connues. Quand au moins 4 distances successives correspondent à un même type de structure secondaire, la structure secondaire est attribuée au fragment. Les auteurs décrivent des "super" structures secondaires, qui ne sont pas implémentées dans le logiciel disponible.

(c) P-CURVE[181] (1989): Les protéines sont décrites par un axe global, qui suit au mieux le squelette de la protéine. Cet axe est calculé avec une fonction de minimisation qui prend en compte la position des atomes de la chaîne polypeptidique et qui passe au mieux entre eux. Les structures secondaires sont alors définies en comparant les données obtenues avec des segments de structures idéales pour chaque structure secondaire.

(d) Consensus[31] (1993): Les trois méthodes précédentes se basent sur des critères distincts. Colloc'h et collaborateurs ont observé que les assignations qu'ils donnaient étaient parfois fort divergentes. Plus d'un tiers des résidus ne sont pas assignés de la même façon par les trois algorithmes. Ceci peut poser des problèmes, principalement pour la prédiction de ces structures. Les différences se situent principalement aux extrémités des structures répétitives qui sont traitées de manière différente selon les algorithmes. Les assignations incompatibles ("α" pour un algorithme "β" pour un autre) sont en nombre négligeable. Le consensus revient à choisir dans les cas litigieux la solution donnée par deux algorithmes.

	DSSP (%)	P-CURVE (%)	DEFINE (%)	STRIDE (%)	concensus (%)
hélice	93,8	92,4	83,6	93,2	94,2
feuillet	78,4	74,4	66,8	77,5	79,4
boucles	79,3	80,6	84,4	81,2	85,1
Total	83,4	82,4	79,0	84,1	86,5

TAB. 2.3 – *Assignement comparé entre DSSP, P-CURVE, DEFINE, STRIDE et la méthode consensus pour chacun des trois types de structures secondaires avec P-SEA (Table II, page 293 [116]).*

(e) STRIDE[58] (1995): STRIDE ajoute un critère angulaire défini dans DSSP [100] au notion des liaisons hydrogènes qui sont normalement présentes dans les structures répétitives. Pour cela, les mesures adéquates sont directement reliées à celles utilisées dans des travaux cristallographiques. En plus des hélices α et des feuillets β, ils utilisent leur approche pour définir d'autres types de structures caractéristiques. Celles-ci seront développées dans le paragraphe suivant.

(f) P-SEA[116] (1997): L'assignation des structures secondaires par P-SEA (pour *Protein Secondary Element Assignement*) s'effectue exclusivement sur les positions des carbones α. Trois distances et deux angles sont calculés avec d2i pour la distance Cα_{i-1} Cα_{i+1}, d3i pour la distance Cα_{i-1} Cα_{i+2}, d4i pour la distance Cα_{i-1} Cα_{i+3}, les angles α_i décrivant les carbones *i-1, i, i+1, i+2* et τ_i décrivant les carbones *i-1, i, i+1*. L'assignation se fait si les distances et/ou les angles correspondent à des valeurs types décrivant les hélices et les feuillets.

Comparaison entre les différents algorithmes :

L'ensemble de ces méthodes, à l'exception du consensus, dépend d'un certain nombre de définitions sur la valeur des angles ou des distances et des variations autorisées autour de ces valeurs. Les différentes méthodes d'assignation sont donc loin d'être équivalentes. La méthode consensus dérive d'ailleurs de cette constatation, seulement 64 % des résidus assignés par les trois méthodes DSSP, P-CURVE et DEFINE étant attribués au même type de structure secondaire, et propose de prendre en compte l'ensemble des 3 algorithmes précédents.

Le tableau 2.3 récapitule les attributions comparées suivant le type de structure assigné par P-SEA. De fortes différences sont visibles principalement pour les feuillets β qui sont les plus difficiles à caractériser. Dans le tableau 2.4, j'ai récapitulé les concordances d'assignation entre les 5 algorithmes actuellement disponibles (PSEA, DSSP, STRIDE, DEFINE, PCURVE)

	PSEA	DSSP	STRIDE	DEFINE	PCURVE
PSEA	—	74,81	83,03	68,73	76,81
DSSP		—	87,25	63,30	67,53
STRIDE			—	66,43	73,53
DEFINE				—	62,32

TAB. 2.4 – *Fréquence d'assignation commune (en %) entre différents algorithmes d'assignation de structures secondaires effectuée avec 906 chaînes issues de la base de données de R. Dunbrack possédant moins de 50% de similitudes de séquences. Les différentes versions des logiciels utilisés sont mises entre parenthèse avec P-SEA (version 2.0), P-CURVE (version 3.1), DSSP (version DsspCMBI-April-2000, modification du programme original en 1988 et 1994), STRIDE (version 1995) et DEFINE (version 2, 1994).*

type de structures	Φ	Ψ	sur les résidus	liaison CO-NH
hélices 3_{10} *classique*	-71	-18	*i+1* et *i+2*	*i* et *i+3*
hélices 3_{10} *"parfaite"*	-74	-4	*i+1* et *i+2*	*i* et *i+3*
hélices π	-57	-70	*i+1*, *i+2* et *i+3*	*i* et *i+5*

TAB. 2.5 – *valeurs nécessaires à l'assignation des hélices 3_{10} et π.*

pour une base de données de 906 chaînes protéiques possédant moins de 50% de similitudes de séquences. Les fichiers ne pouvant être analysés par un programme donné n'ont pas été pris en compte. Les résultats obtenus sont en forte concordance avec ceux de la littérature. Un maximum de similitude se trouve entre l'assignation effectuée par DSSP et STRIDE, ce qui est logique du fait de l'utilisation par STRIDE des valeurs angulaires définies dans DSSP. Le logiciel qui possède une assignation la plus divergente des autres méthodes est le logiciel DEFINE qui utilise uniquement des distances sur les C_α. Sa notice d'utilisation dit d'ailleurs qu'il est loin d'avoir les critères optimaux nécessaires à l'assignation. Par ailleurs, il a le taux le plus élevé d'impossibilité de lecture de fichiers PDB.

Les autres formes régulières Les hélices 3_{10} et π sont deux formes largement moins présentes que les hélice α dans les protéines. Le tableau 2.5 récapitule les valeurs classiques de ces deux conformations. Les hélice 3_{10} (cf. figures 2.7a et 2.7b) possèdent un diamètre moins important que les hélices α, et sont par ailleurs souvent incluses dans l'une d'elles. Elles représentent entre 3 et 4 % des résidus des structures protéiques, ce qui en fait la quatrième structure secondaire la plus importante.

Des approches théoriques tendent à montrer que leur taille est limitée à quelques résidus du fait de contraintes énergétiques [157]. Un des facteurs importants dans la création lors

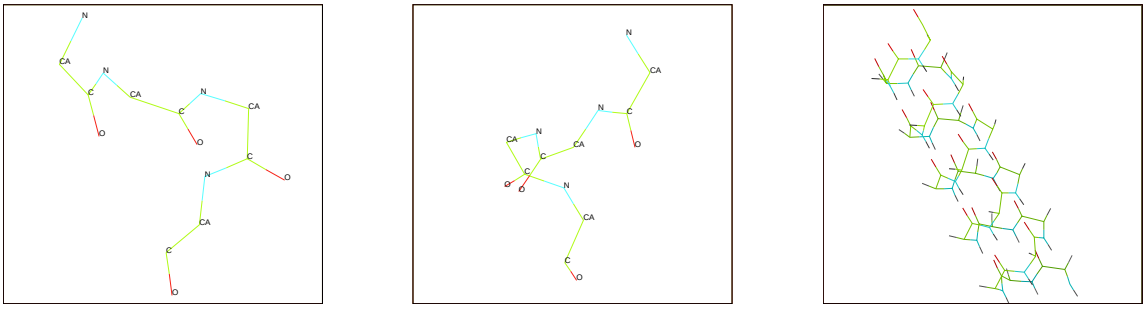

FIG. 2.7 – *Représentations (a) d'une hélice 3_{10} ayant des caractéristiques classiques (AZW chaîne A résidus 51 à 54), (b) une hélice 3_{10} ayant des caractéristiques "parfaites" (protéine 1VPN chaîne C résidus 134 à 137) et (c) une hélice π (modèle théorique construit par le Pr. Etchebest) visualisées avec XmMol [194].*

du repliement d'une hélice 3_{10} est l'absence de liaisons entre les résidus i et $i+4$ [188]. Leur diamètre diffère des hélices α. Leur composition en acides aminés induit des extrémités C- et N-terminales plus étendues [103] que pour les hélices α. Deux définitions existent des hélices 3_{10} (cf. tableau 2.5). La première correspond à la définition classique, moyenne qui est observée dans les protéines. La seconde est dite "parfaite" et représente ce qui serait énergétiquement le plus favorable. Ce second type d'hélice ne correspond qu'à 1 - 2 % des hélices 3_{10} présentes dans la PDB d'après une recherche exhaustive que j'ai menée sur la base de données des structures protéiques (PDB) de juin 2000.

Tout comme les hélices 3_{10}, les hélices π (cf. figures 2.7c) sont fortement impliquées dans les transitions des hélices α vers les boucles. Ces transitions sont de plusieurs types. Les plus importantes sont celles qui partent vers la zone α_L (zone dites des "Glycines" avec un angle dièdre ϕ positif), ensuite ce sont celles qui partent vers la zone α_R (zone classique des hélices α) [150].

Les boucles et les connections entre structures secondaires En dehors des structures répétitives, d'autres structures ont été découvertes, comme les coudes dont le nombre de catégories est d'une huitaine [202]. Les plus classiques sont présentées sur la figure 2.8. Les angles et liaisons les caractérisant sont reportés dans le tableau 2.6. Ils sont retrouvés plus ou moins couramment selon le type de coudes, certains très rares, d'autres spécifiques de fin de protéines [51] ou de conformation spécifique [5]. Ces régions ont toujours toutefois une taille limitée. Le logiciel STRIDE permet de déterminer l'attribution d'un certain type de coudes.

type de structures	Φ	Ψ	sur les résidus	liaison CO-NH
coudes de type I	-60	-30	*i+1*	*i* et *i+3*
	-90	0	*i+2*	
coudes de type II	-60	+120	*i+1*	*i* et *i+3*
	-80	0	*i+2*	
coudes de type III	-60	-30	*i+1*	*i* et *i+3*
	-80	-30	*i+2*	
coudes gamma	+70	-60	*i+1*	*i* et *i+2*

TAB. 2.6 – *valeurs nécessaires à l'assignation des coudes de types I à III et gamma.*

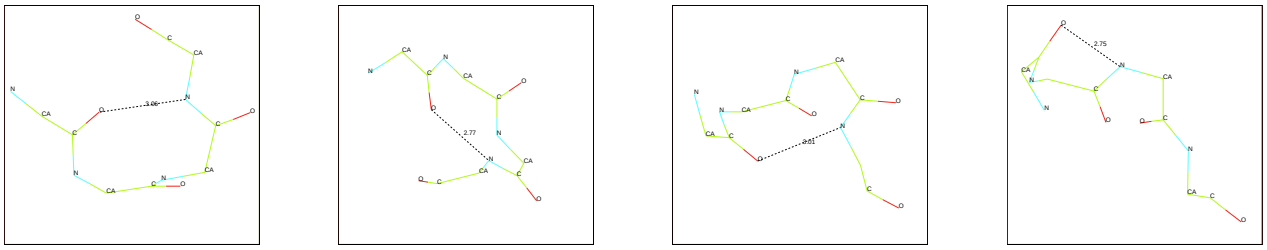

FIG. 2.8 – *Représentations (a) d'un coude de type I (protéine 125L chaîne A résidu 121à 124), (b) d'un coude de type II (1ALL chaîne A résidu 112 à 115), (c) d'un coude de type III (109M chaîne A résidu 13 à 16) et (d) d'un coude de type gamma (1SCY chaîne A résidu 1 à 4).*

Des classifications des boucles donnent lieu à de nombreuses et différentes taxonomies dont des termes identiques recoupent parfois des réalités distinctes [46, 51].

Des classifications des boucles entre structures secondaires α et/ou β ont permis d'observer des spécificités en acides aminés pour certaines boucles [203, 114, 13] et l'établissement de procédures de prédiction [204]. Toutefois, l'ensemble de ces méthodes est basé sur la définition de points de début et de fin des structures secondaires qui sont parfois sujettes à caution et ont des difficultés à considérer l'ensemble des boucles, se limitant à des tailles inférieures à 9 résidus [44, 63].

2.1.2.2 Le nombre de repliements

L'utilisation des structures secondaires permet de catégoriser les protéines en différentes familles, tout-α, tout-β, α/β, $\alpha+\beta$, non classifiable [127]. Elle a permis ainsi de créer de véritables arbres reliant des sous-familles structurales, donnant lieu à des classifications parfois totalement supervisées comme pour SCOP [132] ou automatisées en grande partie comme CATH [136] ou FSSP [86]. L'intérêt de ces approches n'est pas une simple classification mais permet aussi de

trouver de nouveaux liens entre structure, séquence et fonctions [131, 87, 89, 73, 139, 140, 123], et ainsi d'estimer la variété de repliements qui existent dans les protéines [138].

Des études plus théoriques basées sur des méthodes statistiques montrent qu'un certain pourcentage de repliements (70 %) possède une spécificité séquentielle suffisante pour impliquer un repliement local particulier [177]. Le reste est trop peu déterminé et ne pourrait donc pas servir dans le cas d'une prédiction de la structure 3D protéique à partir de la séquence.

Le nombre global de repliements est estimé à un peu plus de 2000 [69]. Toutefois, seuls 900 seraient vraiment différents. Plus récemment, les mêmes auteurs augmentent leurs chiffres à 4000 repliements. 375 de ces repliements représenteraient 70 % des repliements existants dans les protéines, les plus courants, il en faudrait au final 930 pour en représenter 90% [70].

2.2 Prédictions de la structure protéique

2.2.1 La structure secondaire

Deux grandes familles de prédiction de la structure secondaire existent avec les méthodes statistiques, les plus anciennes, et les méthodes "d'apprentissage", les réseaux de neurones. Une nouvelle catégorie est apparue plus récemment, les méthodes consensus qui prennent des résultats de plusieurs algorithmes différents. Plus de milles articles sur la prédiction des structures secondaires ayant été publiés à ce jour, les quelques exemples qui suivent ont donc un caractère arbitraire mais sont, me semble-t-il, les plus représentatifs.

2.2.1.1 Méthodes statistiques

Les méthodes statistiques ont été les premières mises en place. Elles se basent sur une utilisation directe de la composition en acides aminés et se sont complexifiées avec l'utilisation de la théorie de l'information et des alignements de séquences.

- Chou et Fasman [29] : La première méthode de prédiction se base sur la tendance de chaque acide aminé à se trouver associé à une structure hélice α, brin β ou boucle. Cette méthode se base sur le calcul d'un coefficient de corrélation pour chaque type d'acide aminé dans chaque type de structures secondaires. Le taux de prédiction réel de la méthode dépasse 50 %.

- GOR : développé initialement par Garnier et collaborateurs, elle a été améliorée par Gibrat et collaborateurs [62, 65, 66, 61]. GOR I [62] utilise une approche proche de celle de Chou et Fasman, mais met en valeur les incompatibilités dans les distributions. La méthode tient compte des distribution d'angles dièdres [65, 66]. Le taux de prédiction dépasse 64 %. Plus récemment, le taux de prédiction est passé à plus de 67 % en ajoutant une information binaire dépendante juste de l'hydrophobicité du type d'acides aminés [104].

- Alignement de séquences : Du fait de l'augmentation des bases de données aussi bien protéiques que génomiques, il est possible d'utiliser des séquences similaires dans des techniques de prédiction [68, 64, 211]. Le principe est séquentiel et consiste à aligner une séquence à prédire avec d'autres séquences connues, puis à observer dans les zones consensus les compatibilités existantes pour moduler la prédiction finale. Par exemple, une procédure bayésienne qui donnait d'un taux de prédiction moyen de 67 % passe avec cette méthode à plus de 73 % de bonne prédiction [193]. Le programme PREDATOR, lui, passe d'un pourcentage de 68 % à 75 % [60]. Même un nombre limité de séquences dans l'alignement permet des gains significatifs [42].

- Il faut noter que certains algorithmes ne tiennent pas compte de la cohérence séquentielle des résultats. Ainsi une hélice peut être prédite directement aprés un feuillet sans avoir de boucles entre les deux états [210]. DSC [105] en revanche s'en préoccupe. Il utilise un ensemble de paramètres physico-chimiques et en tenant compte des insertions et des délétions présentes dans l'alignement atteint un taux de 70,1 % seul. Pour certaines protéines, il dépasse les méthodes de prédiction par réseau neuronal qui sont décrites dans le paragraphe suivant.

Ainsi, il existe un grand nombre de méthodes fort différentes basées sur des principes statistiques. L'intérêt majeur de ces méthodes, en regard des méthodes d'apprentissages type réseaux neuronaux, est la possibilité de comprendre les facteurs entrant en jeu dans la prédiction, donc les acides aminés prépondérants et leurs dépendances éventuelles.

2.2.1.2 Réseaux neuronaux et alignements de séquence

Avec l'augmentation du nombre de protéines disponibles, l'utilisation de méthodes d'apprentissage plus gourmandes en données comme les réseaux neuronaux était possible. Les premières

études n'ont utilisé comme information que les séquences *brutes* avec un réseau à une couche [148, 85, 173] ou encore avec deux réseaux imbriqués [133]. Pour une explication rapide des réseaux neuronaux, voir l'Annexe 1.

La prise en compte des alignements de séquences a permis une augmentation importante du taux de prédiction :

- PHD [166, 165], développé par Rost et Sander, est l'un des plus connus du fait d'une diffusion importante sur le Web [167]. PHD est un ensemble composé de deux réseaux possédant chacun une couche cachée. Le premier apprend la succession des acides aminés avec en plus des informations arrivant des alignements de séquences et d'échelle d'hydrophobicité. Le second réseau, lui, utilise les mêmes informations avec en plus le résultat de la prédiction et ainsi élimine les incohérences possibles de type "passage direct d'une hélice à un feuillet". Cet ensemble d'information permet de dépasser aisément les 72 % de bonne prédiction.

- NNSSP utilise un réseau avec une seule couche cachée [168]. Il a été conçu pour prendre une séquence seule ou un alignement de séquences. Le pourcentage de bonne prédiction passe alors de 71,0 % à 73,5 % en utilisant des séquences possédant des homologies. Ces résultats sont à mettre en parallèle avec l'ancienne version SSP (purement statistique) dont les taux de prédiction étaient respectivement 65,1 et 68,2 % [182].

- Des travaux montrent l'importance de la taille de la base de données impliquée dans l'apprentissage, et celui du nombre de neurones dans les couches cachées. L'optimisation de tels paramètres permet une augmentation importante du taux de prédiction de 2 à 4 % [27, 28] et, récemment, le taux de 80 % aurait été dépassé [144]. Les valeurs maximales attendues de ce type d'analyse avoisineraient les 85 % dans un futur proche [59].

2.2.1.3 Méthode consensus

Elles sont de deux types, la première consiste à utiliser à la fois en un seul algorithme différentes méthodes puis à combiner les informations. En parallèle, des méthodes statistiques et des méthodes de réseau neuronal peuvent être utilisées [187], ou encore faire une procédure plus hiérarchique où certains algorithmes sont pris au départ pour être ensuite affinés par

d'autres résultats de prédiction. Ce dernier type de méthode permet, en utilisant des méthodes statistiques et des réseaux neuronaux avec alignement de séquences, de dépasser des taux de 76 % et surtout d'obtenir des taux excellents sur les feuillets β qui sont normalement les plus difficiles à prédire [142].

Une autre possibilité est largement exploitée : l'utilisation conjointe de logiciels disponibles avec un consensus selon le type de résultats des prédictions. Ainsi SOPMA donne 69,5 % de bonne prédiction, ensuite couplée avec PHD, ce qui permet d'obtenir plus de 82 % de bonnes prédictions pour les résidus ayant été prédits par les deux méthodes dans le même état (soit près des 3/4 des résidus) [64]. La combinaison de 4 des algorithmes les plus connus (DSC [105], PHD [165], NNSSP [168] et PREDATOR [60]) permet ainsi d'atteindre 75,4 % (*http://barton.ebi.ac.uk*) [34, 33]. Actuellement, il est possible de combiner près d'une dizaine de méthodes en parallèle (*http://ibcp.fr*) [40, 72].

Ces dernières méthodes tendent aussi à démontrer que le rôle des interactions à longues distances considérées comme la cause du faible taux initial des prédictions des structures secondaires [82] a souvent été sur-estimé [56].

2.2.1.4 Bases de données

La base de données utilisée est capitale pour la validité de l'expérimentation. Ainsi, une base ne contenant que des protéines avec des taux d'hélices α importants fera que la méthode de prédiction sera particulièrement appropriée pour des protéines ayant de forts taux en hélices α, mais le résultat pour une protéine β sera peu probant. Cuff et collaborateurs récapitulent ainsi les problèmes liés aux premières prédictions. Faits sur des bases de données peu importantes, les taux de prédiction annonçés étaient compris entre 63 et 77 %. Des tests effectués sur de nouvelles protéines montrent en réalité des taux compris entre 50 % et 56 % [33].

Aussi, ne faut-il pas utiliser des bases complètes, mais des bases "nettoyées". Ces bases doivent satisfaire à deux critères : (i) être suffisament peuplées pour représenter la spécificité du plus grand nombre de repliements possibles, sans déséquilibre particulier, et (ii) être dépourvues de biais en terme de séquence, de façon à optimiser la détermination de la relation séquence-structure.

Une approche intéressante a été développée en Finlande [11], mais restée sans suite (sans doute du fait de sa non-disponibilité sur le web). La procédure est de type hiérarchique. Elle

utilise la séquence pour avoir un taux d'identités faible, mais aussi la structure en définissant des taux de structures répétitives par DSSP [100], pour ne pas avoir des protéines trop reliées entre elles. Les deux bases de données "nettoyées" sont PDBselect et HSSP. La première est un ensemble de protéines possédant un taux d'identité de séquences inférieur ou égal à un seuil choisi (*http://swift.embl-heidelberg.de/pdbsel/*) [84, 83]. Aucun lien avec la structure n'est testé. La méthode est basée sur l'exclusion des séquences proches jusqu'à obtention d'un taux maximal. La seconde prend en compte la structure et se base principalement sur des alignements de séquences [174, 43]. Ainsi, des séquences proches sont exclues et donc des repliements proches normalement aussi.

2.2.2 Modèles protéiques

Les structures secondaires donnent une idée de la topologie de la protéine. Il convient néanmoins d'utiliser d'autres techniques pour aboutir à la structure réelle de la protéine. Deux types de techniques existent: celles qui se basent sur l'utilisation de fragments ou de séquences protéiques connus, pour la modélisation par homologie et pour l'enfilage (ou *threading*) et celles qui se basent sur une représentation simplifiée des acides aminés pour la modélisation *ab initio*.

2.2.2.1 Modélisation par homologie

La modélisation par homologie se base sur un concept simple. Des séquences proches ont des repliements proches, comme on le constate pour des protéines ayant des taux de similitude de séquence supérieures à 30 %.

Plusieurs équipes ont montré qu'à partir de structures connues, il est possible de construire pour une séquence, ayant des taux de similitudes de séquences significatifs, de manière supervisée, une nouvelle structure avec de bons résultats [99, 30].

Au vu de la complexité et du nombre croissant de protéines, l'utilisation des procédures automatisées est devenue incontournable [120].

Différents programmes existent maintenant comme COMPOSER [10, 9] ou MODELLER [169], prenant en entrée une séquence et pouvant donner en sortie un certain nombre de modèles. Le principe de ces méthodes se rapproche fortement de la méthode développée par le groupe

de David Eisenberg, qui peut être subdivisée en 4 parties [15, 121] :

1- Classer une base de données en fragments protéiques associés à certaines caractéristiques telles que la structure secondaire, l'exposition au solvant, la position par rapport aux positions N et C-terminales de la protéine,...

2- Un profil est effectué sur des alignements de séquences [71]. Ce profil est alors comparé à la base de données et l'on cherche les compatibilités locales de manière extensive.

3- Les parties associées aux hélices et aux feuillets sont trouvées en premier. La construction des zones variables est plus complexe. Une recherche de compatibilité entre la disposition spatiale et les contraintes des profils est effectuée.

4- Les chaînes latérales sont ajoutées et l'ensemble est minimisé. Des ensembles de rotamères sont ici utilisés. Ils représentent des prototypes moyens des conformations des chaînes latérales. Pour choisir lequel est associé à un résidu, l'interaction avec les voisins et les contraintes géométriques sont déterminées.

2.2.2.2 Technique d'enfilage (ou *threading*)

Les techniques d'enfilage (ou *threading*) sont utilisées principalement dans le cas de très faibles homologies et se basent sur une étude énergétique [54, 95, 17, 126]. Construisant la chaîne polypeptidique résidu par résidu, on cherche à l'allonger au mieux en calculant les incompatibilités existantes entre la séquence utilisée et des bases de données de structures cristallographiques. Diverses fonctions d'énergies sont alors requises. Les résultats étaient initialement fort dépendants de la taille et surtout de la composition en structures secondaires répétitives [96]. Ainsi THREADER 2 obtient des résultats particulièrement intéressants dans la reconnaissance des repliements dans la série CASP (pour *Critical Assessment of Techniques for Protein Structure Prediction*, ensemble de congrès où les méthodes sont testées sur des nouvelles séquences dont la structure réelle n'est révélée qu'à la fin) [97]. Outre une difficulté classique de paramétrisation des champs de force, la difficulté de la méthode vient de l'incertitude des régions d'insertions. Différentes améliorations existent pour moduler ce problème comme les techniques d'enfilage à "champ de force auto-consistant" [55] où le champs de force est ajusté à la séquence examinée et non pas à celle de la séquence cible.

Une autre difficulté est l'obtention de nombreux modèles et donc la recherche du "meilleur" [152]. L'utilisation de séquences homologues aide particulièrement quand il n'y a pas d'insertions dans l'alignement [153], du réseau neuronal PHD en parallèle [7] ou des méthodes statistiques comme les Chaînes de Markov Cachées [8]. Le temps de calcul est souvent loin d'être négligeable [205].

Les méthodes qui semblent les plus prometteuses sont des méthodes hiérarchiques, tel LINUS [183] qui part d'un modèle et le raffine en recherchant avec une minimisation rapide les zones qui sont apparemment associées avec un type de repliements. Il prend en compte différentes méthodes dans un ordre précis [93] ou encore utilise des données phylogénétiques [143].

2.2.2.3 Modélisation *ab initio*

Enfin, une dernière technique est applicable, la modélisation *ab initio*. Elle consiste en une représentation schématique des protéines. Le plus souvent un résidu n'est représenté que par un atome, souvent le C_α, et, avec différents jeux de contraintes le système évolue pour explorer au maximum l'espace conformationnel [41]. Les pseudo-atomes sont déplacés suivant des mouvements tenant compte de processus aléatoires ce qui permet de dépasser certains minima locaux. Il peut y avoir des contraintes énergétiques liées à des contraintes de distances [88], des contraintes liées à la séquence à un niveau local par des méthodes statistiques [179, 180, 178], avec des prédictions de structures secondaires [137] ou encore en recherchant dans des bases de données [38]. Des techniques combinant plusieurs approches permettent d'avoir une idée plus précise du type de topologie de la protéine [171].

Toutefois, les résultats sont fortement liés à la taille de la protéine ainsi qu'au type de protéines [39, 185, 4, 141].

2.2.2.4 Quelques exemples de prédiction jouant un rôle important dans le repliement protéique

La nature des acides aminés fait que certains résidus aiment plus ou moins le solvant, et qu'ils préfèrent interagir avec certains autres acides aminés plutôt que d'autres. Ces propriétés physico-chimiques induisent leur capacité à créer des contacts entre eux, ce qui induit au final la structure 3D de la protéine Ainsi, trois catégories de prédictions importantes dans la modélisation moléculaire sont :

chaînes latérales. Elles se basent sur une recherche locale du minimum énergétique [197, 14] par une utilisation successive des différents rotamères possibles. Les rotamères sont des prototypes moyens de chaînes latérales [196, 45, 195].

accessibilité. Le calcul de l'accessibilité se fait le plus souvent en faisant rouler une bille virtuelle sur la protéine [118, 91, 155]. La prédiction de l'accessibilité permet de rendre compte du rôle prépondérant de l'effet hydrophobe/hydrophyle [129, 12, 25]. Le score maximal de prédiction est de 77,6 % pour un taux relatif de 9 % d'accessibilité [128]. A plus de 90%, c'est l'effet hydrophobe/hydrophyle qui joue le rôle primordial [130]. L'accessibilité est d'un point de vue biologique capital, car l'ensemble des fonctions biochimiques des protéines sont assurées à la surface des protéines. C'est le site de liaison des autres composés protéiques [32] ou nucléiques [92, 134].

contacts. La prédiction des contacts à partir de la séquence primaire est un des défis les plus difficiles actuellement. Les meilleurs algorithmes ne dépassent pas 15 % de bonnes prédictions [128, 48]. Certains algorithmes s'intéressent au point précis de contacts, d'autres à des zones plus étendues [135, 124]. Un des problèmes majeurs est lié à la recherche à "longue distance" (plus de 40 résidus), en dehors des ponts disulfures; les prédictions sont à peine plus fiables qu'une recherche aléatoire [67].

Aussi, d'autres types d'informations plus simples sont exploitées comme prédire le nombre total de contacts dans une protéine [49], ou encore les Cystéines impliquées dans une liaison disulfure. Dans ce dernier cas, les taux de prédiction dépassent les 70 % [50, 57].

2.3 Les alphabets structuraux

2.3.1 But

Les structures secondaires permettent une description structurale précise des deux structures périodiques que sont les hélices α et les feuillets β (cf. paragraphe 2.1.2.1). Toutefois, la classification à l'aide des structures secondaires laisse un peu plus de 50% des structures protéiques non attribuées (les boucles). La description structurale est finalement assez limitée. La construction de bibliothèques de boucles définies par rapport aux structures répétitives apporte

une aide dans l'approximation de la structure et permet d'établir des classifications utilisables dans des méthodes de prédiction [44, 204]. Les boucles de taille supérieure à 8 ne sont cependant jamais prises en compte dans ces approches. La reconstruction à partir de la seule information des structures secondaires ne permet pas une reconstruction du repliement 3D.

Aussi, de nombreuses études ont été menées pour tenter de catégoriser et de classifier des fragments protéiques de petites tailles (inférieures à 20 acides aminés) représentatifs des protéines présentes dans les bases de données de structures cristallographiées, la PDB. Les différentes méthodes présentées ici cherchent à examiner des successions de quelques acides aminés, les structures statistiquement répétées les plus courantes, pour décrire au mieux l'ensemble des repliements des structures protéiques.

Le travail précurseur de Jones et Thirup doit-être noté ici. Ils ont reconstruit une protéine de liaison au rétinol avec des fragments structuraux tirés de 3 protéines ayant peu d'homologie de séquence. Ce travail se basait sur l'étude d'une carte de densité électronique [99]. Depuis, deux approches assez distinctes se trouvent dans la littérature. La première cherche un nombre important de prototypes, une centaine pour représenter au mieux les protéines [198, 146, 200, 175]. La seconde recherche un nombre plus limité de prototypes, normalement inférieur à une vingtaine, afin de permettre leur utilisation dans des méthodes de prédiction plus fiables [162, 163, 53, 19, 23, 24]. Je parlerai de blocs structuraux ou parfois de blocs protéiques dans cette partie, les notions étant équivalentes et correspondant toujours aux petits prototypes, nommés *Buildings blocs (BB)* par Unger et collaborateurs [198], *Short Structural Motifs (SSM)* par Unger et Sussman [200], *Substructure* par Prestrelski et collaborateurs [146], *Local Structural Motifs (LSM)* par Schuchhardt et collaborateurs [175], *Recurrent Local Structural Motifs (RLSM)* par Rooman et collaborateurs [162, 163], *Structural Buildings blocs (SBB)* par Fetrow et collaborateurs [53], *Local Structure (LS)* par Bystroff et Baker [19], et les *Short Structural Building blocs (SSB ou SSBB)* de Camproux et collaborateurs [23, 24].

Dans les pages qui vont suivre, je vais présenter les différentes études qui ont été menées à l'heure actuelle en développant la méthodologie et les résultats propres à chaque équipe. Avec dans un premier temps, les études sur des nombres de blocs importants (plus de 100), ensuite ceux sur les nombres faibles de blocs (entre 4 et 13), puis enfin le travail du groupe de Baker, qui est plus complexe.

2.3.2 Un nombre important de blocs protéiques pour une description fine de la structure protéique

2.3.2.1 Par une méthode d'aggrégation en deux étapes (Unger *et al.*, 1989, 1993)

L'objectif de ce travail est d'obtenir un grand nombre de prototypes pour pouvoir ensuite reconstruire l'ensemble des structures protéiques avec ces blocs (1989,[198] et 1993,[200]). Pour leur base de données, Unger et collaborateurs ont conservé sur les 354 chaînes présentes (PDB de janvier 1987), 82 ayant un facteur R dit "correct" et une résolution de moins de 2.8 Å, soit 12 973 résidus.

Leur méthode consiste à calculer l'écart quadratique moyen ou *RMSd* (root mean square deviation) entre deux structures s et t, en ne conservant que les carbones α du squelette dans cette comparaison:

$$RMSd(s,t) = \sqrt{\frac{\sum_{i=1}^{n}(r_i^s - r_i^t)^2}{n-2}}$$

Un exemple intéressant des problèmes liés au calcul du *RMSd* est donné. Le *RMSd* minimal entre deux structures n'est pas obligatoirement une information adéquate. Ainsi la figure 2.3.2.1 montre 3 fragments protéiques. Le *RMSd* entre (a) et (b) est plus grand que celui entre (b) et (c), alors que le type de repliement de (a) est plus proche de celui de (b).

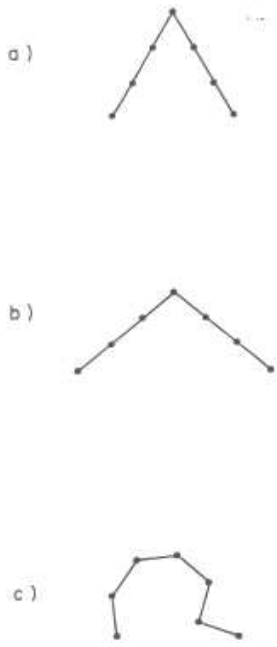

FIG. 2.9 – *3 fragments protéiques d'une longueur de 7 C_α (Figure 1. p.357 [198]).*

Après un calcul préliminaire effectué sur 4 protéines (426 résidus au total), Unger et collaborateurs ont décidé de prendre comme prototypes des hexamères (i.e., des fragments de longueur 6). Ils considérent cette longueur comme la plus petite correcte pour bien différencier les fragments protéiques entre eux. L'ensemble des *RMSd* entre les hexamères des 4 protéines en question (soit 82 215 couples) est calculé et ensuite réuni par une méthode dite "d'annexion".

Cette procédure se passe en trois étapes:

1- Un hexamère est tiré au hasard dans la base de données. Tous les hexamères étant à moins de 1 Å de *RMSd* de cet hexamère lui sont attribués. Ensuite tous les hexamères du groupe sont utilisés pour leur associer les hexamères étant à moins de 1 Å. La procédure s'arrête quand plus aucun hexamère ne peut être attribué au groupe. Un hexamère n'appartenant à aucun groupe est alors tiré au sort et le processus recommence jusqu'à l'attribution de tous les hexamères à un groupe.

2- L'inconvénient de la première étape est de créer des groupes importants dont certains ont plus d'1 Å de RMSd. Aussi, Unger et collaborateurs redivisent les groupes et font plusieurs tentatives pour n'obtenir que des sous-groupes dont tous les hexamères sont à moins de 1 Å de *RMSd* les uns des autres.

3- Le centre de chaque groupe est conservé comme bloc structural type du groupe (bloc le plus proche du barycentre calculé).

La méthode d'annexion est moins conventionnelle que la méthode des nuées dynamiques. La plus grande crainte des auteurs est de n'obtenir que fort peu de groupes dans la première étape. Elle donne 55 groupes, ce nombre étant directement lié au choix de leur valeur limite de 1 Å , qui est assez faible pour des fragments de cette taille. La seconde étape est plus criticable. En effet, rien ne laisse supposer qu'un groupe X doit être subdivisé en sous-groupe X_1 et X_2. Il est logique de penser que des éléments de X après division puissent être alors plus proches d'un autre groupe Y plutôt que X_1 ou X_2. Après la seconde étape, 103 blocs structuraux sont obtenus. Sur la base de données complète :

a- 76% des fragments sont proches d'un des blocs structuraux avec un *RMSd* de moins d'1 Å.

b- 92% avec moins de 1,25 Å.

c- 65% ne sont proches que d'un bloc (à moins d'1 Å).

d- 5% sont proches de plus de 2 blocs (à moins d'1 Å).

e- En ne conservant que des fragments ayant moins de 1 Å de différence avec la réalité, 99% de leur base de données est couverte.

Utilisant 4 autres protéines, la méthode donne à 144 blocs protéiques finaux. En combinant les 8 protéines, elle atteind à 170 blocs structuraux. Les plus fréquents sont retrouvés dans les deux expériences, mais le reste est trés fluctuant. Les structures connues répétitives et leurs liens caractéristiques sont retrouvés.

Une méthode de reconstruction sur des protéines de longueur 60, ou alors sur les 60 premiers acides aminés (extrémité N-terminale), est effectuée en ajoutant à chaque fois "au mieux" un nouveau carbone correspondant au nouveau bloc protéique. La procédure consiste à superposer en une position X les 5 premiers atomes du bloc structural x avec les 5 derniers de la chaîne reconstruite. La reconstruction est correcte dans plus de 64% des cas. Seuls des changements rapides dans le repliement posent un réel problème.

En conclusion, la méthode développée par Unger et collaborateurs est intéressante car elle est basée sur une méthodologie pertinente pour regrouper des fragments protéiques, et surtout, les blocs ont servi lors d'une méthode de reconstruction 3D avec de bons résultats. Toutefois, deux points posent problémes :

(a) Le choix de 4 protéines seulement au départ est trop faible et ne peut rendre compte de la diversité des repliements protéiques, comme le nouvel échantillonnage le montre fort bien. Un choix de 25 protéines soit 1/3 de la base de données aurait été plus judicieux. Pour éviter un temps de calcul élevé, un traitement rapide pour éliminer du regroupement des fragments vraiment trop proches, comme des coeurs d'hélices ou de feuillets, aurait été possible.

(b) La méthode de réallocation est fortement discutable, la taille des groupes devenant rapidement très faible. Par exemple, le seizième groupe le plus important de la base de données représente 1.0%. Sur la base de données de départ de 4 protéines, il ne représente donc que 4 fragments.

L'équipe d'Unger utilisera ces blocs, comme exemple, pour une autre méthode de reconstruction, basée sur les angles dièdres [199]. Ils montrent que la plupart des hexamères trouvés sont créables par une procédure aléatoire dans les zones de Ramachandran classiques.

Enfin en 1993, Unger et Sussman réutilisent les blocs obtenus quatre ans auparavant [200], mais n'en conservent que 81, ceux-ci étant observés plus de 35 fois dans leur base de données, soit une fréquence minimale de 0,3 %. Détail troublant, les chiffres donnés sont les même que ceux obtenus pour les 103 blocs. Les hexamères obtenus se retrouvent dans les zones préférentielles du diagramme de Ramachadran, mais beaucoup sont significatifs car une procédure purement aléatoire crée beaucoup de fragments peu ou non observés dans la base de données. Ainsi, les auteurs espèrent déceler un début d'architecture des protéines avec leurs hexamères. Pour explorer un cas plus localisé, tous les blocs définis en 'brin étendu' ('E' ou feuillet β dans DSSP [100]) sont extraits de la base de données. Ces différents blocs ont une répartition différente en acides aminés. Toutefois, les chiffres donnés sont en pourcentage et on ne peut juger de la valeur statistique des variations observées.

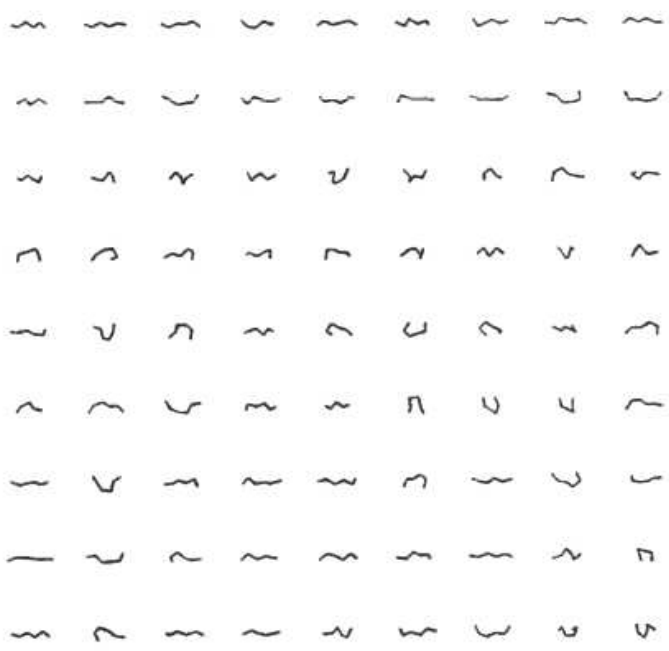

FIG. 2.10 – *Représentation 3D des 82 blocs structuraux (BB) (Figure 1. p.463 [200]).*

2.3.2.2 Pour une librairie de sous-structures (Prestrelski *et al.*, 1992)

Prestrelski et collaborateurs ont créé une librairie de blocs, sans *a priori* sur le type de structures secondaires. Ils désirent les utiliser pour trouver des homologies structurales, d'une

manière similaire à celle de Jones et Thirup [98], mais en concevant une librairie fixe de prototypes et non spécifiquement conçue pour une protéine unique (1992, [146]).

La base de données utilisée est composée de 14 protéines ne possédant pas d'homologie structurale, résolues à moins de 2,5 Å et représentant 2 437 résidus.

La méthode consiste en la génération d'un petit nombres de blocs différents structuralement. Le critère de similarité choisi est moins classique que dans le précédent travail. Il s'agit d'une fonction liant un critère de distance dite distance linéaire et un critère angulaire, l'angle α qui décrit 4 C_α successifs du squelette polypeptidique. La figure 2.11 montre ces deux critères.

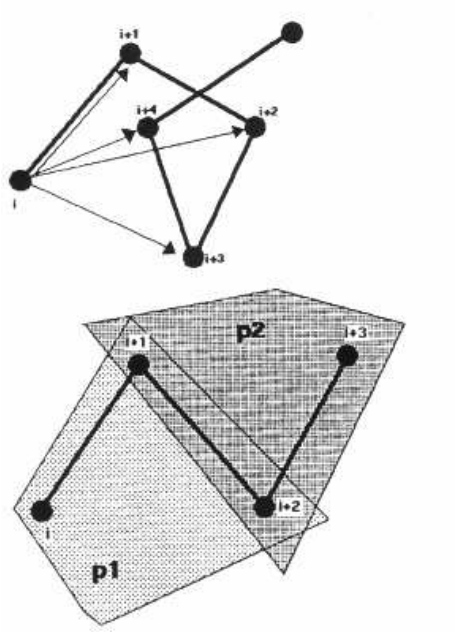

FIG. 2.11 – *Représentation schématique de la distance linéaire et de l'angle α [146].*

La distance linéaire (DL) est la somme des distances C_α-C_α [122]:

$$LD_i = \sum_{j=1}^{4} (\mathbf{d}_{i,i+j})$$

avec $\mathbf{d}_{i,i+j}$ la distance entre le $C\alpha_i$ et le $C\alpha_{i+j}$. Ce critère permet de différencier les structures répétitives. Elle a été utilisée pour observer des insertions-délétions dans des structures cristallographiques. Toutefois, elle ne permet pas de différencier une hélice gauche d'une hélice droite par exemple [122]. Pour contrecarrer ce type de problème, ils ont donc adjoint l'angle α. Pour différencier deux fragments protéiques s et t, une fonction de coût $C_{s,t}$ dépendant de deux paramètres C_1 et C_2 a été définie:

$$C_{s,t} = C_1 \times (|DL_s - DL_t|) + C_2 \times (\tan|\alpha_s - \alpha_t|) + C_2 \times (\tan|\alpha_{s+1} - \alpha_{t+1}|)$$

La valeur maximale de la différence des angles α est limitée à 85^o. Pour travailler sur des fragments de taille supérieure à 5, ils ont utilisé une fenêtre glissante définie par $Max = (\mathbf{L}_f$ - 4 $)\times$ $maxdif$ $\times 0,75$, avec $\mathbf{L}_f$ la taille du fragment considéré, $maxdif$ la valeur maximale de la fonction de coût pour les fragments considérés. La valeur 0,75 sert à comparer des fragments de taille 8 utilisés dans l'étude et ceux de taille 5.

Quand les valeurs des distances linéaires sont proches, l'angle α est très différent. L'utilisation conjugée des deux critères semble donc discriminante.

Avec cette approche, la première étape à réaliser est la calibration des deux coefficients C_1 et C_2. Pour les ajuster, ils ont testé un ensemble de valeurs de C_1 allant entre 0,1 et 1,0 et de C_2 entre 1,0 et 2,2. Pour déterminer la qualité discriminante des coefficients, ils ont testé une dizaine de prototypes d'hélices, de feuillets et de structures périodiques en calculant leurs distances linéaires, leurs angles α et les valeurs de Max associées aux coefficients testés. Ensuite, ils ont choisi comme notion d'équivalence le fait que deux structures ont un $RMSd$ inférieur à 1 Å. Les résultats ont particulièrement pris en compte la distinction entre les hélices α et le reste des exemples. C_1=0,9 et C_2=2,0 ont ainsi été choisis et appliqués comme critère de coût à l'ensemble des fragments issus de la base de données.

La méthode utilisée de groupements n'est pas décrite en détail, mais aux vues des résultats donnés qui récapitulent les 30 blocs les plus fréquents, une simple recherche des coûts minimaux aurait pu suffir. Ils obtiennent au final 113 blocs distincts. Pour donner un ordre de grandeur, le deuxième bloc le plus fréquent ne représente que 17 fragments.

En conclusion, la méthodologie utilisée est peu conventionnelle. Les seuls points ambigus concernent la recherche des coefficients C_1 et C_2 et la signification exacte du coût pour les fragments de longueur 8 usités. Si la formule donnée est bien celle utilisée, le seul critère important est $maxdif$ qui représentent en fait $argmax = \{C_{s,t}\}$ pour chacuns des fragments, s et t. Ceci convient parfaitement pour "écarter" les fragments. On ne travaille que sur les différences maximales. Toutefois, il est possible que cette fonction génère un nombre de blocs distincts trop important. Avoir plus de 72% des blocs vus moins de 10 fois paraît un peu

élevé. Unger et collaborateurs ont des chiffres fort différents. Un autre facteur a pu jouer: le choix des protéines de la base de données qui sont sans homologie structurale. Au final, seules les hélices α réapparaissent fortement; les feuillets β sont peu présents. Cela peut être dû à leur caractéristique structurale plus lâche. Par ailleurs, les auteurs désirent faire plutôt une classification par classe de protéines. L'adjonction de 6 nouvelles protéines fait passer leur taux de recouvrement par les 30 prototypes les plus courants de 67% à 77%.

Cette méthode a été mise au point pour aider lors de travaux de spectroscopie à infra-rouge. La librairie générée leur a permis de constater des corrélations impossibles avec une autre méthode et donc de proposer une architecture de sérine-protéase compatible avec les structures connues [147].

2.3.2.3 Par une Carte-auto organisée de Kohonen (Schuchhardt *et al.*, 1996)

L'objectif des travaux de Schuchhardt et collaborateurs est d'obtenir un grand nombre de prototypes pour approximer la structure tridimensionnelle d'une protéine dans un but de classification et d'analyse sans tenter de reconstruire cette structure (1996, [175]).

La base de données est composée de 136 protéines ayant moins de 30% de similitude de séquence [84, 83], ce qui représente un total de 24 239 résidus.

La méthode consiste à caractériser les protéines en utilisant des fragments d'une longueur de 9 résidus décrits par les angles ϕ et ψ. Les protéines sont découpées en fragments de 9 résidus donnant chacun un vecteur d'observation de longueur 16 (le premier ϕ et le dernier ψ n'étant pas pris en compte). L'ensemble comporte donc 23 151 exemples. Schuchhardt et collaborateurs ont utilisé une méthode non supervisée d'apprentissage, les cartes de Kohonen (cf. Annexe 1 pour plus de détail). Cette méthode consiste à créer $N \times M$ neurones (représentés par une matrice), chaque neurone étant une observation moyenne, ici un vecteur d'angles dièdres, à rechercher pour chaque observation de la base de données le neurone le plus proche de cette observation et lui faire apprendre cette observation. Le principe est décrit en détail dans l'annexe I. Son intérêt principal est un aspect visuel particulièrement utile pour l'analyse. La méthode des cartes de Kohonen ne diffère des nuées dynamiques ou des k-moyennes (ou *k-means*) [79] que par l'existence d'un processus de diffusion qui permet de réunir dans un environnement proche des neurones ayant des points communs.

Pour évaluer les différences entre les observations, une mesure de dissemblance, le *RMSda*

ou *root mean square deviation on angular values* est utilisée. Cette distance sur les angles a déjà servi dans une méthode de recherche d'homologie structurale [102]. Il s'agit d'une distance euclidienne entre deux vecteurs d'observations s et t:

$$RMSda(s,t) = \sqrt{\frac{1}{16}\sum_{1}^{8}(\phi_s^i - \phi_t^i)^2 + (\phi_s^{i+1} - \phi_t^{i+1})^2}$$

La figure 2.12 montre la carte de Kohonen obtenue pour une taille de 10×10, soit 100 neurones. Cette carte est "plane", ce qui veut dire qu'elle possède des bords, le neurone (9/9) n'est voisin que des neurones (8/8), (8/9) et (9/8), alors qu'au milieu de la carte, un neurone est entouré de 8 voisins.

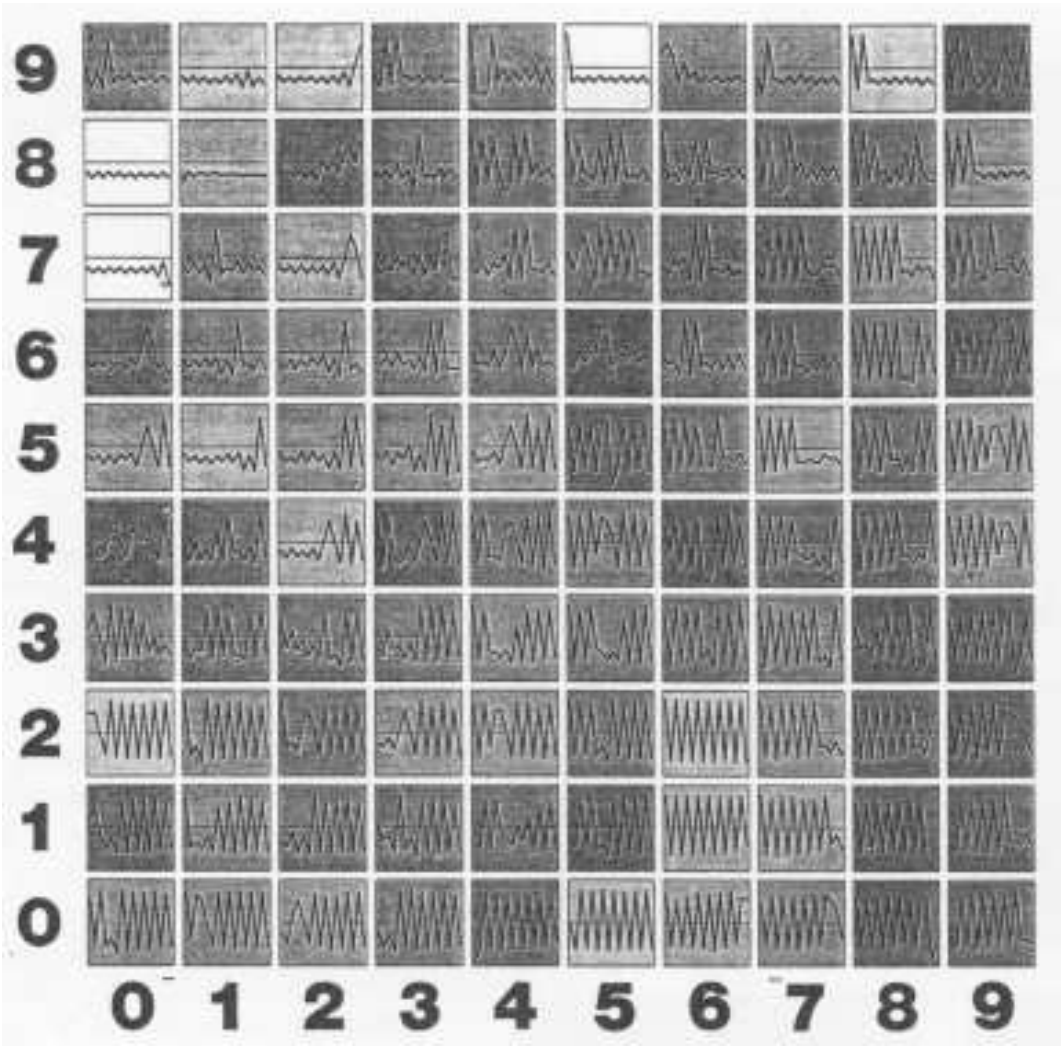

FIG. 2.12 – *Carte auto-organisée de Kohonen obtenue pour 100 neurones (Figure 4a. p.836 [175]). Chaque neurone est décrit par la suite des 16 angles dièdres moyens.*

Les feuillets et les hélices se retrouvent dans des neurones distincts. Les hélices α sont forte-

ments localisées en haut et à gauche de la carte, les feuillets β plus sur la droite. La validation structurale est effectuée exclusivement sur le critère du *RMSda* et les résultats semblent satisfaisants. Des motifs dits de "cassure" classique avec une forte proportion de Glycine sont trouvés. Quelques exemples de suivi de trajectoires de protéines sur la carte sont montrés.

La méthode est particulièrement appropriée à la recherche et à l'analyse de données biologiques, comme le montre le travail de Hanke et Reich sur la reconnaissance de motifs propres à des familles protéiques [77, 78]. L'utilisation des angles dièdres avec le *RMSda* est beaucoup plus souple que le classique *RMSd*. Du fait de sa rapidité, un travail plus important sur les différents paramètres de l'apprentissage est possible. Les résultats sont convaincants. Toutefois, ayant refait leurs expériences, deux points me paraissent importants :

1- L'utilisation d'un réseau plan est peu approprié. Un réseau fermé, où les neurones ont tous 8 voisins (i.e, le neurone (9/9) est voisin des neurones (8/8), (8/9), (9/8) et (0/0), (0/8), (0/9), (9/0) et (8/0)), permet une meilleure diffusion sans effet de bord. Et, en fin d'apprentissage, les hélices et les feuillets sont bien mieux séparés. Les motifs caractéristiques sont aussi plus accentués.

2- Une longueur de neuf résidus est un peu élevée. La recalcul des angles dièdres réels à partir d'une longueur est peu évidente. Cette méthode est donc peu utilisable en particulier pour reconstruire une protéine. De plus, la précision obtenue est parfois trés médiocre pour certains neurones, d'ailleurs aucune notion de variabilité n'est donnée dans le texte. Certains angles ont une variabilité proche de 180^o. Avec une carte plus petite, on conserve un approximation comparable, et une longueur plus faible, elle est plus intéressante, en fait.

2.3.3 Peu de blocs protéiques

2.3.3.1 Utilisation d'un regroupement hiérarchique (Rooman *el al.*, 1990)

Les travaux de Rooman et collaborateurs ont consisté en l'obtention d'un petit nombre de blocs structuraux, représentant pertinement la base de données. Ces blocs sont de taille fixe, mais plusieurs longueurs ont été étudiées (1990, [162]). Ils ont été ensuite utilisés dans une méthode de prédiction de la structure protéique à partir de la séquence; les auteurs ont surtout

	N.obs.	freq. rel.	*RMSd* moyen (Å)
η_4	4858	38,1	0,33
ϵ_4	1795	14,1	0,24
ζ_4	3151	24,7	0,39
λ_4	2949	23,1	0,60
η_5	3917	31,0	0,48
ϵ_5	2639	20,9	0,44
ζ_5	3377	26,7	0,82
λ_5	2700	21,4	1,07
η_6	3645	29,1	0,76
ϵ_6	4845	38,7	1,00
ζ_6	1246	10,0	1,08
λ_6	2779	22,2	1,32
η_7	3997	32,2	1,14
ϵ_7	3469	27,9	1,06
ζ_7	2652	21,5	1,67
λ_7	2284	18,4	1,56

TAB. 2.7 – *Nombre d'observations (N.obs.) pour chaque bloc obtenu, avec sa fréquence relative (freq. rel.) et le* RMSd *moyen (Figure3. p.332-333 [162])*

mis en parallèle ces blocs et les structures secondaires (1990, [163]).

La base de données utilisée se compose de 75 protéines ayant une bonne résolution et possédant peu de similitude de séquence. Cela représente à 12 978 résidus.

La méthode d'apprentissage se base sur une classification hiérarchique qui utilise comme critère de distance entre fragments protéiques le *RMSd* (cf section 2.3.2.1). Dans un premier temps, les fragments de taille L sont tous comparés deux à deux grâce au *RMSd* sur les C_α du squelette polypeptidique. Ensuite, une classification hiérarchique est effectuée.

Des longueurs L allant de 4 à 7 C_α ont été testées, et, pour chaque longueur, 4 groupes distincts ont été déterminés. La table 2.7 récapitule le nombre d'observation et le *RMSd* moyen de chaque groupe. J'ai ajouté la fréquence relative pour avoir une idée de la contribution de chaque bloc.

Sur les 4 classes observées pour chaque longueur, la famille λ représente toujours un ensemble exclusivement boucles, la famille η étant composée d'hélices α, la famille ζ de boucles et de feuillets β et la famille ϵ de feuillets β. Cette classification hiérarchique est utilisée ensuite pour discriminer différentes familles protéiques. La figure 2.13 donne les valeurs des angles ϕ et ψ obtenues. Je les ai caculées à partir de la figure 3 p.333 [162] pour pouvoir comparer mon alphabet avec le leur.

Dans un second temps [163], le lien existant entre ces structures et les séquences a été recherché. Pour cela, un travail important de formalisme de la significativité d'une séquence a

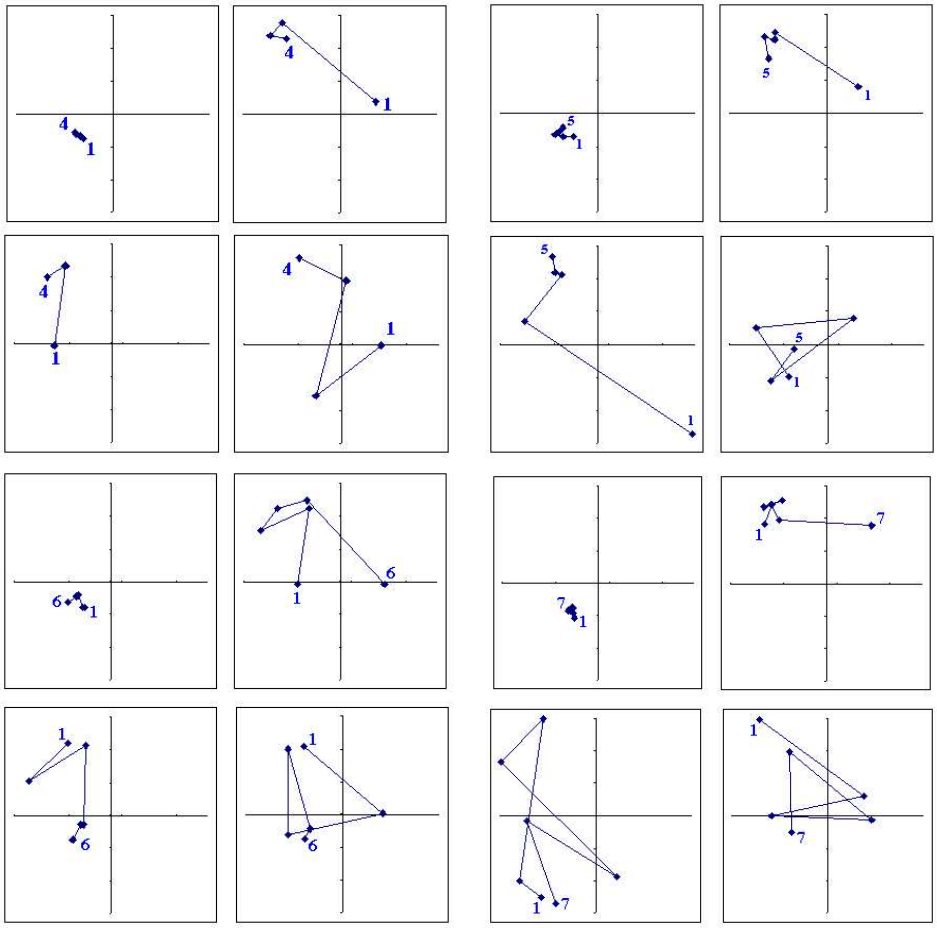

FIG. 2.13 – *Représentation des angles des 4 blocs de (a) longueur 4, (b) longueur 5, (c) longueur 6 et (d) longueur 7 (recalculés à partir de la figure3. p.331-332 [162]).*

été effectué. Chaque succession de 7 résidus a été traitée sous la forme $x-my-nz$, avec x, y et z des acides aminés déterminés compris dans une séquence de longueur 7, avec $n+m$=4 (7-3), n et m variant donc, entre 1 et 3, et représentant n'importe quel acide aminé. Les occurrences conservées sont représentatives à la fois de la séquence (un nombre d'occurrence supérieur à trois) et la structure (associée majoritairement à un type de structure donnée). Ce type de représentation avec des acides aminés non déterminés vient de la taille limitée de la base de données [164] et d'un précédent travail sur certaines formes de coudes [158].

Avec cette approche, les auteurs trouvent à partir de la séquence, un taux de prédiction compris entre 41% et 47%. Ce taux est à mettre en comparaison avec les taux des structures secondaires de l'époque qui avoisinent 60% pour 3 états [62, 65], alors que les blocs en proposent 4. En parallèle de cette observation, ils trouvent un plus grand déterminisme dans les séquences observées avec leurs 4 blocs pour une longueur L=6, mais en contre-partie un bruit de fond, lui aussi accentué. Aucune amélioration de la prédiction n'est donc observée.

Ainsi, les blocs obtenus sont fortement liés aux structures secondaires répétitives. Le choix du nombre final de blocs peut porter à critiques. Il ne permet pas, comme pour les structures secondaires, de reconstruire directement une structure protéique à partir de cette information. La taille de la base de données est pour beaucoup dans le taux final de prédiction.

Par la suite, seule la méthode de prédiction, mais appliquée aux structures secondaires, sera réutilisée [159] et mise en oeuvre dans une recherche basée plus spécifiquement pour une recherche à partir de coordonnées extraites de 7 régions dans la carte de Ramachandran [159, 160] et avec utilisation de protéines homologues [161].

2.3.3.2 Utilisation d'un réseau de neurones (Fetrow *et al.*, 1993, 1997)

L'objectif de ce travail est de reclassifier les structures protéiques en "super" structures secondaires avec des blocs obtenus à l'aide d'une méthode de compression de l'information, puis de regroupements des données ainsi traitées en quelques prototypes (1993, [208] et 1997, [53]).

Plusieurs bases de données ont été utilisées. La plus ancienne [208, 209] est composée de 74 chaînes ayant un taux d'identité de séquence inférieur à 50%, une résolution de moins de 2,5 Å, et, un facteur R inférieur à 0,3. La base est composée de 13 114 acides aminés. La plus importante, et plus récente [53], possède moins de 25 % d'identité de séquence pour 116 chaînes

protéiques représentant 23 355 résidus.

La méthode repose sur un réseau de neurones dit réseau auto-associatif (autoANN). Le principe, exposé dans la figure 2.14, est de prendre un vecteur d'information en couche d'entrée, de le comprimer dans la couche cachée comme pour tout réseau neuronal artificiel classique, mais au lieu d'avoir en couche de sortie une information nouvelle tirée des observations mises en couche d'entrée, on recherche la même information que les observations mises en entrée. En résumé, le réseau doit restituer ce qu'il voit.

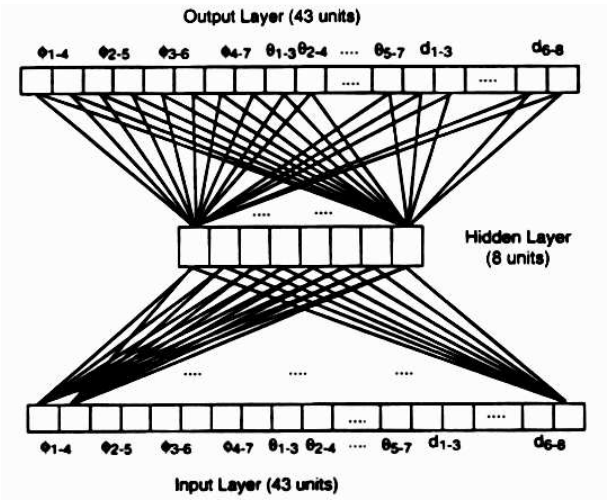

FIG. 2.14 – *Schéma représentant le réseau auto-associatif avec sa couche d'entrée à 43 neurones, la couche cachée de 8 neurones et la couches de sorties avec 43 neurones.*

Le groupe de Fetrow a utilisé 3 types d'informations structurales mises dans un réseau autoANN et ils se sont servis de la couche cachée comme nouveau descripteur des informations. Ils ont appliqué sur les données "compressées" par la couche cachée un algorithme classique de classification, les k-moyennes (ou *k-means*) qui est décrit dans l'Annexe 1.

Pour décrire les protéines, ils ont travaillé sur des fragments de 7 résidus consécutifs, qu'ils ont décrits par :

(-) 15 distances liant les C_α (sauf pour les C_α contigus). Ayant observé que la distribution des distances est bi-modale, chaque distance a été codée sur deux valeurs (bits). La valeur entre les deux pics de la courbe bi-modale a été prise comme seuil (τ). Si la distance d_i est inférieure à τ, le second bit est mis à zéro, et, dans le premier le rapport d_i/τ est introduit. Si la distance d_i est supérieure à τ, le premier bit est mis à un, et, dans le second le rapport $d_i - \tau/v_m - \tau$ est introduit. v_m est la valeur maximale des distances

observées. Ce système donne pour 15 distances 30 valeurs, toutes normalisées entre 0 et 1.

(-) les 4 angles dièdres, qu'ils ont codé avec leurs sinus et cosinus, soit 8 valeurs au final, ce qui élimine les problèmes de rotation.

(-) 5 angles de valences ($C\alpha_i$,$C\alpha_{i+1}$,$C\alpha_{i+2}$,$C\alpha_{i+3}$).

Ainsi 7 résidus sont traduits en un vecteur de 43 composantes, toutes comprises entre zéro et un. Pour le premier autoANN [208, 209], la moitié de leur base de données (6 422 fragments) a été utilisée pour l'apprentissage, le reste (6 242) servant pour sa validation. Ensuite, les auteurs utilisent les vecteurs recodés par la couche cachée (8 valeurs au lieu de 43) et les séparent en 6 groupes grâce à la méthode des k-moyennes ou *k-means* (cf. Annexe 1).

Leur première étude [208] (il n'y a aucune évolution notable dans leur seconde publication [209]), montre surtout une recherche de la corrélation entre leurs 6 blocs obtenus et les structures secondaires. Ainsi, plusieurs initialisations ont été testées et le maximum de ressemblance entre les résultats dans les différentes expériences a été recherché, les blocs obtenus devant être les plus proches possibles.

Différentes remarques sont à faire sur cette étude. Aucune notion de ressemblance 3D n'est mentionnée. Les auteurs décident de conserver 6 groupes soit 6 blocs. Ces blocs ne sont décrits que comme une entrée en hélice, une partie centrale, une extrémité C-terminale. Il en va de même pour les feuillets. La répartition précise dans les différents blocs des boucles (près de 50% de la base de données) n'est pas explicitée. Avec une absence totale de vérification de la fiabilité structurale et de la variabilité des blocs, des questions restent en suspens. Dans la même perspective, il aurait été intéressant de savoir ce que donne directement une méthode de classification sur les observations recodées en 43 composantes.

Fetrow et collaborateurs ont repris ce travail en 1997 avec une nouvelle base et aboutissent à des résultats similaires. Un travail intéressant est fait concernant les successions de doublets de blocs (*i,i+1*) et des triplets (*i,i+1,i+2*).

En conclusion, aucune précision n'est donnée quant à l'approximation structurale qu'engendrent les blocs. Cette information n'est retrouvée indirectement que dans la base de données mise sur le web *ftp://ftp.cs.albany.edu/pub/compbio/db/2.2A/cat/*, et, l'absence d'explication quant aux chiffres donnés la rend difficilement utilisable. Les doublets et triplets de blocs ne

sont donnés qu'en fonction de leurs fréquences attendues et observées. Une étude statistique plus élaborée aurait permis de voir les plus significatifs (qui ne sont pas obligatoirement les plus fréquents). Les quelques exemples montrant l'intérêt de la méthode sont peu nombreux et reposent principalement sur des successions très rares (entre 1 et 4 observations dans la base de données).

Plus récemment, Fetrow et Berg [52] ont tenté de mettre en évidence une relation existant entre leurs blocs et les chaînes latérales. Ils ne présentent que quelques graphiques sans véritable conclusion.

2.3.3.3 Utilisation des Chaînes de Markov Cachées (Camproux *et al.*, 1999)

Le but de cette étude est de définir un ensemble de prototypes pour approximer la structure tridimensionnelle des protéines en tenant compte principalement des transitions entre ces prototypes. Ils sont appelés blocs structuraux pour la reconstruction (Structural Building Blocs ou SBBs) [23]. La base de données est composée de 100 protéines ayant moins de 25 % de similitude de séquences [84, 83], soit 19 317 résidus.

La méthode se base sur l'utilisation des Chaînes de Markov Cachées (CMC, ou Hidden Markov Model, HMM) [149, 22] et utilise comme descripteur de la structure des protéines des distances entre C_α et une projection dans un plan. Les fragments de la base de données ont été découpés en fragments chevauchants d'une longueur de 4 résidus. La figure 2.15 montre les 3 distances utilisées: distance d_1 entre le $C\alpha_i$ et le $C\alpha_{i+2}$, distance d_2 entre le $C\alpha_i$ et le $C\alpha_{i+3}$ et distance d_3 entre le $C\alpha_{i+1}$ et le $C\alpha_{i+3}$. Une quatrième distance est calculée. Il s'agit de la distance d_4 qui est une projection du $C\alpha_{i+3}$ dans le plan défini par les $C\alpha_i$, $C\alpha_{i+1}$ et $C\alpha_{i+2}$. Cette dernière distance est normalisée par le produit des deux premières distances. Elle permet de donner la direction du fragment. Si d_4 est proche de 0, le fragment est allongé, sans volume, sinon il décrit une certaine torsion.

Le modèle est complexe. Il tend à maximiser le passage d'un état i à un nombre d'états limités. Il demande l'établissement au préalable d'un type de loi. Ici, la dispersion des observations associées à chaque état (SBB) a été considérée comme gaussienne.

Après avoir testé plusieurs possibilités, Camproux et collaborateurs [23] ont décidé de conserver un jeu de 12 blocs protéiques.

L'Analyse en Composantes Principales (ACP) effectuée sur la base de données codée suivant

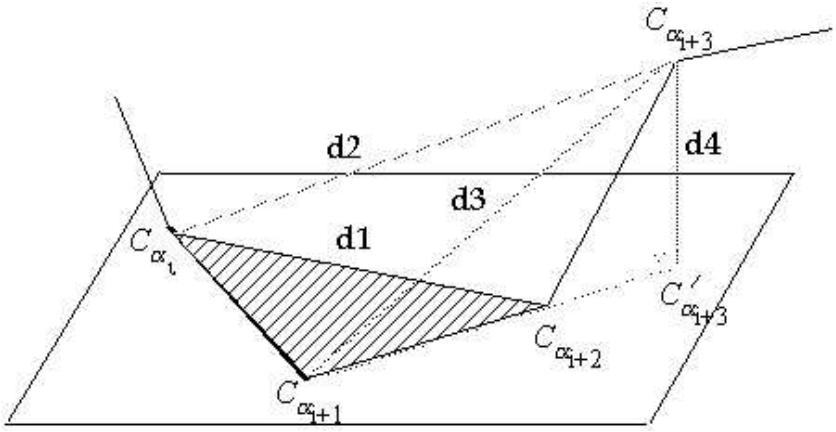

FIG. 2.15 – *Les 3 distances inter-C_α et la projection décrivant les fragments de longueurs 4, avec d_1 distance entre $C\alpha_i$ et $C\alpha_{i+2}$, d_2,$C\alpha_i$ et $C\alpha_{i+3}$ et d_3,$C\alpha_{i+1}$ et $C\alpha_{i+3}$, la distance d_4 est une projection du $C\alpha_{i+3}$ dans le plan $C\alpha_i$, $C\alpha_{i+1}$, $C\alpha_{i+2}$.*

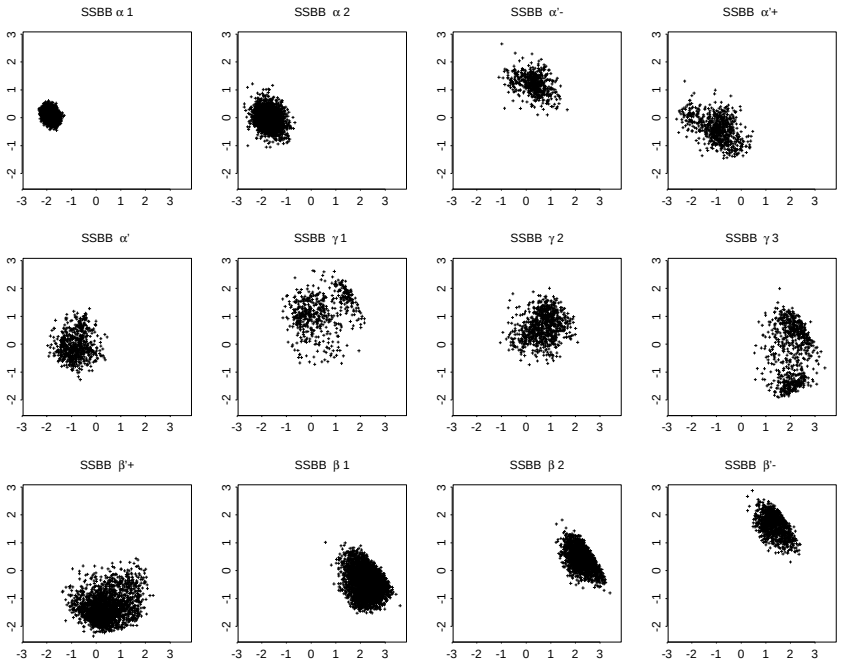

FIG. 2.16 – *Représentation des 12 SBBs suivant les deux premières composantes principales de l'ACP effectuée sur les 4 distances décrites (cf. figure 2.15).*

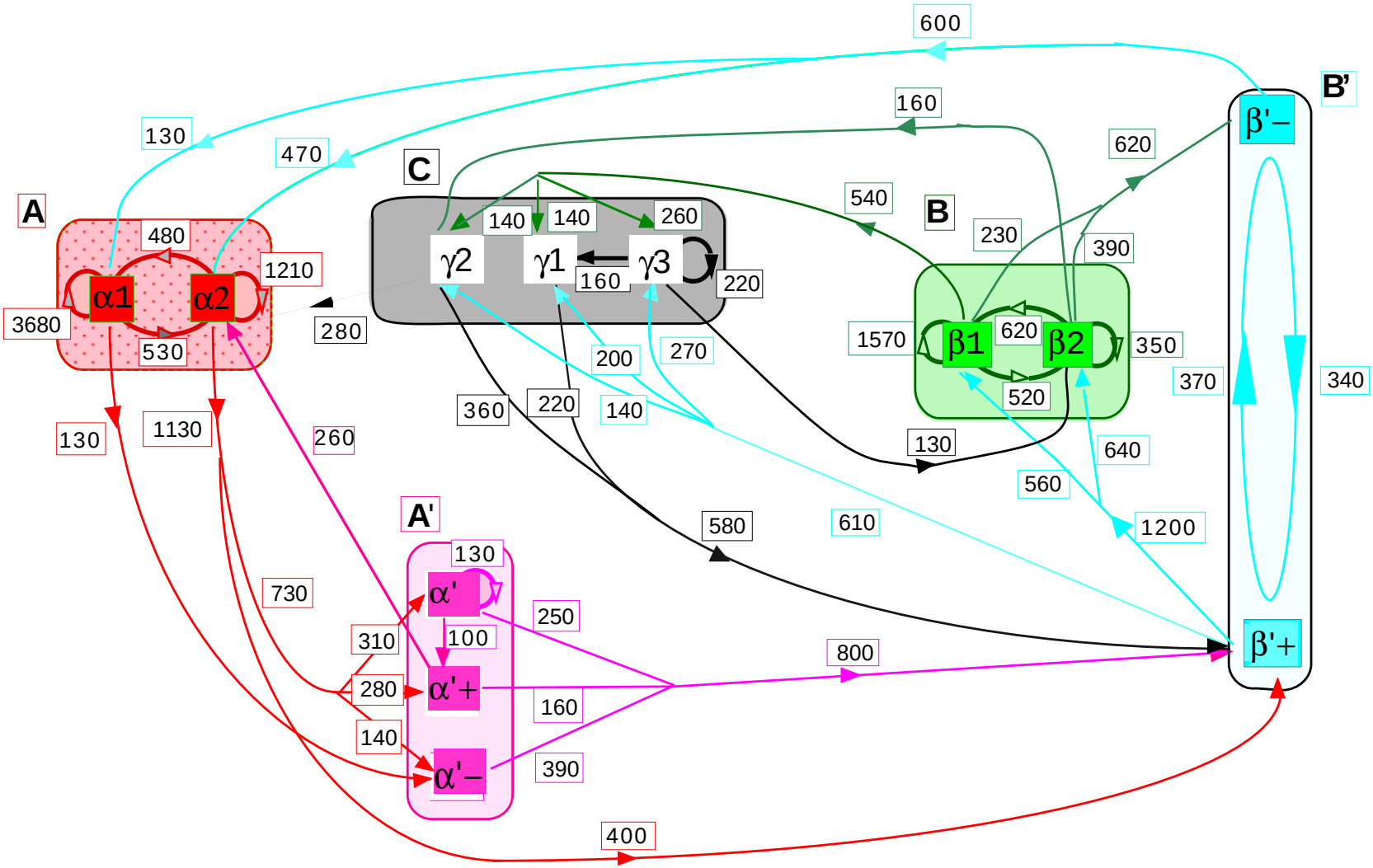

FIG. 2.17 – *Graphe de transition entre les 12 SBBs. Seules les occurrences supérieures à 100 sont représentées, avec en rouge, les SBBs des hélices α, en rosé, les SBBs Nterminaux des hélices α, en gris les SBBs impliqués dans les boucles, en vert, les SBBs des feuillets β et en bleu les SBBs aux extrêmités des feuillets β.*

les 4 distances montre certaines tendances distinctes suivant les types de structures secondaires. La figure 2.16 montre le nuage de point associé à chaque bloc obtenu. On voit clairement l'efficacité de la méthode pour bien discriminer les différents groupes qui sont tous bien compacts.

Deux SBBs sont caractéristiques des hélices α, SBB α_1 et SBB α_2, le premier représente 63,3 % des hélices α de la base de données, le second 28,3 %. Ils ne restent donc que 8,4 % des hélices α assignées à un autre bloc. De même, deux SBBs sont caractéristiques des feuillets β, SBB β_1 et SBB β_2. Le premier représente 57,6 % des feuillets β et le second 21,1 %. Il reste donc 21,6 % des feuillets associés à un autre bloc. Deux SBBs sont particulièrement liés aux boucles, les SBBs γ_1 et γ_2. Les autres sont liés à des entrées et/ou des sorties de structures répétitives. Les SBBs obtenus sont particulièrement stables d'un point de vue structural avec des *RMSd* compris entre 0,15 Å et 1,4 Å.

Un des intérêts des CMC est l'apprentissage basé sur les transitions. Ainsi la figure 2.17 montre le graphe des transitions existantes entre les 12 SBBs, pour des occurrences supérieures à 100.

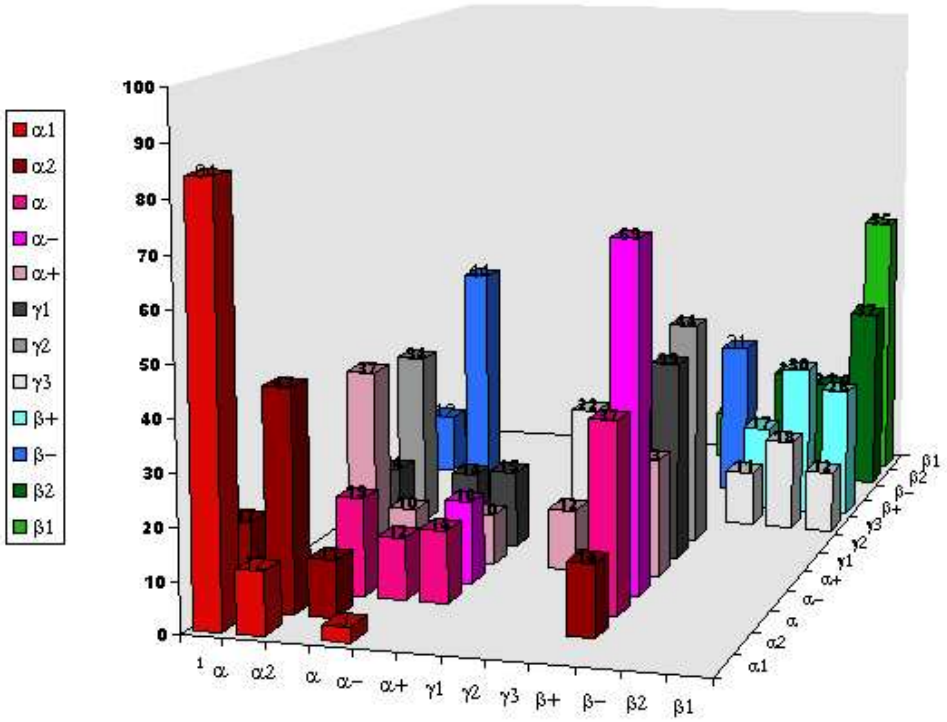

FIG. 2.18 – *Matrice de transition entre les 12 SBBs.*

La polarité existante entre les SBBs est extrêmement marquée. La matrice de transition représentée sur la figure 2.18 montre clairement le regroupement qui existe entre chaque groupe de SBBs qui suit bien les structures secondaires mais avec une certaine flexibilité. Ainsi le bloc β_+ (encore noté γ_β) possède 5 transitions possibles vers un autre type de SBB. Ce type d'apprentissage permet d'apprendre en suivant les transitions, mais sans focalisation excessive quand cela n'est plus possible.

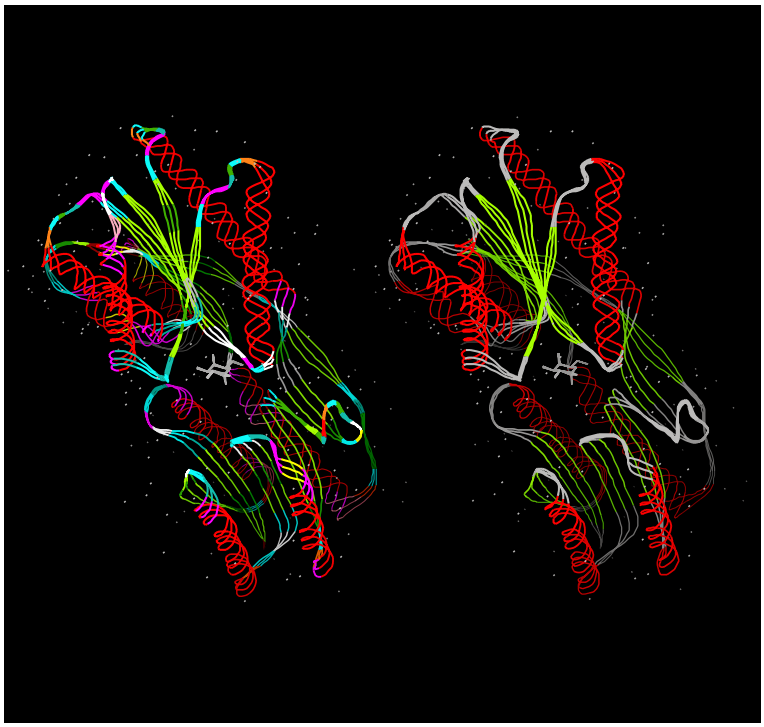

FIG. 2.19 – *Comparaison entre l'attribution par les 12 SBBs à gauche et les structures secondaires à droite avec l'aide du logiciel XmMol [194].*

La figure 2.19 montre la protéine de liaison au L-arabinose (code PDB: 8abp) colorisée à gauche suivant les SBBs, à droite suivant la définition classique des structures secondaires. On observe ainsi nettement la bonne correspondance entre les deux types d'information avec des changements ponctuels. Cette figure montre l'intérêt de ce type d'approche qui permet de bien distinguer, par exemple, le début d'une hélice assigné à un bloc α'_+ colorié en jaune.

Une information pertinente à prendre en compte est alors la distribution en acides aminés. Les distributions classiques associées aux structures répétitives sont retrouvées, comme

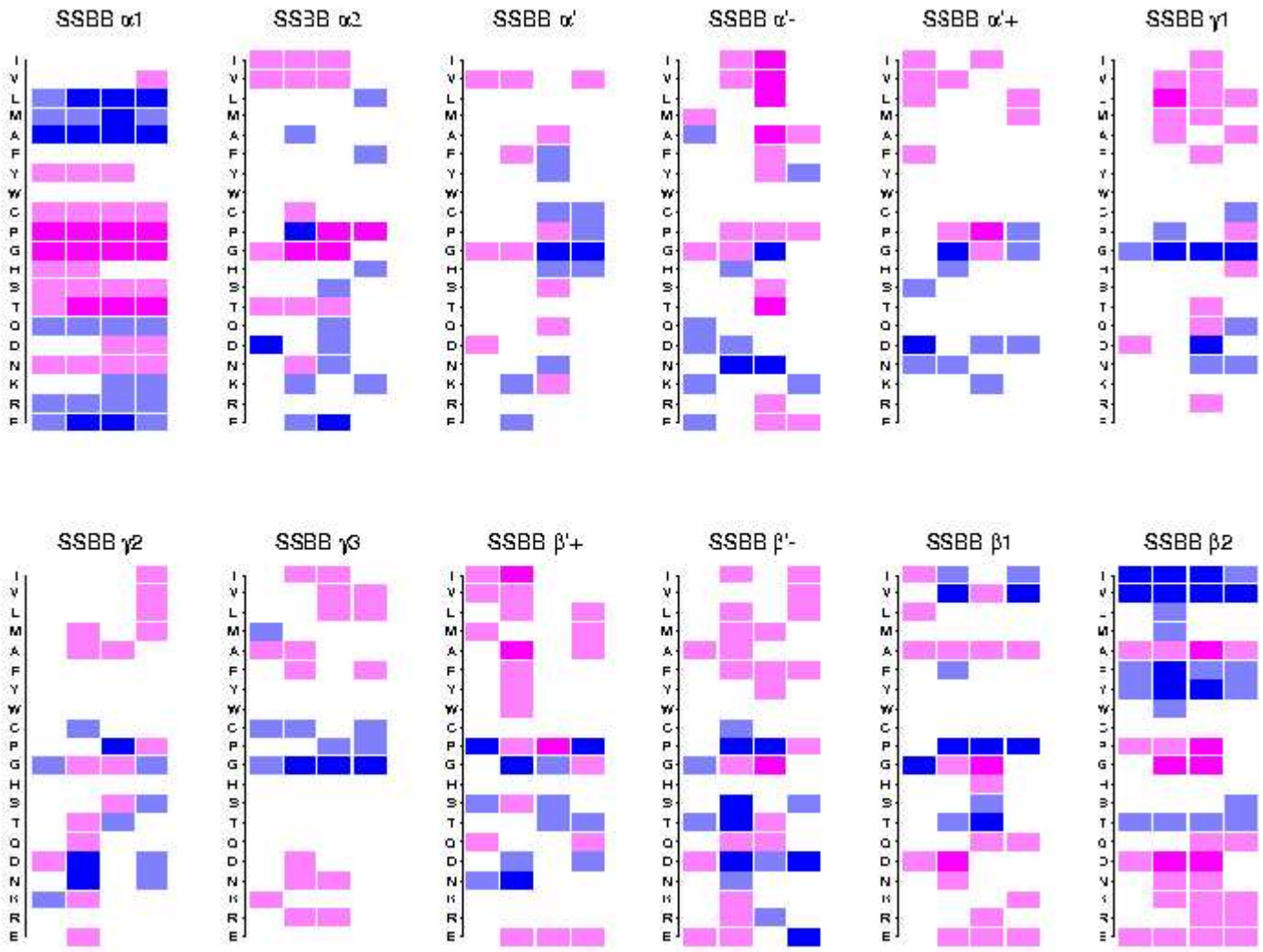

FIG. 2.20 – *Fréquences des acides aminés en chaque SBB pour les 4 positions. Les fréquences sont normalisées en Z-scores (bleu sur-représentation, en rosé sous-représentation). Les couleurs pâles représentent des Z-scores dont la valeur absolue est comprise entre 2 et 4, les autres sont supérieures à 4. Pour une définition du Z-score cf. paragraphe 3.3.5.2*

des acides aminés hydrophobes associés au bloc α_1. Le SBB α_2 du fait de sa composition en Acide Aspartique, Proline et Acide Glutamique semble être plutôt un coude α. Le seul point véritablement critique est la longueur des blocs. 4 C_α ne permet pas l'établissement de liaisons hydrogénes.

Deux travaux distincts ont suivi l'élaboration de cet alphabet structural, tout deux portant sur une analyse des parties boucles entre deux structures répétitives α et/ou β [24, 21].

2.3.4 Une méthode d'apprentissage sur la séquence et la structure (Bystroff et Baker, 1998)

Le travail de Bystroff et Baker concerne l'obtention d'un certain nombre de blocs structuraux protéiques (I-sites) types pour faire de la prédiction. Ils cherchent à optimiser la relation séquence-structure (1998, [19]). Les 471 protéines sont issues de la base de données HSSP [172], possèdent moins de 25% d'identités de séquences, et sont groupées en familles d'homologues.

Ce travail est à l'heure actuelle le plus abouti et celui qui a généré le plus grand nombre d'applications. Les I-sites débutent en réalité trois ans avant leur réalisation par un travail de Han et Baker. Ils mettent au point une méthode de regroupement de séquences ayant de fortes similitudes de séquences par une méthode proche des k-means [74]. Dans un second temps, ils mettent en relation le fait que les groupes qu'ils ont obtenus peuvent correspondre à un ou plusieurs types de repliements protéiques [75]. Ils se limitent à une simple description de répartitions en structures secondaires. Ils définissent alors directement à partir de leur groupes certaines structures plus ou moins précises [76].

La figure 2.21 récapitule l'ensemble du processus de création des I-sites. Les 471 protéines issues de HSSP sont alignées en une centaine de groupes par un alignement multiple. Certaines protéines sont exclues si leur résolution est trop mauvaise, s'il y a plusieurs ponts disulfures ou si la protéine est membranaire. Les longues boucles semblent avoir été exclues aussi.

Ensuite un regroupement de fragments est effectué à l'aide de la méthode des k-moyens (ou *k-means*, cf. Annexe 1) avec l'utilisation de profils sur les acides aminés. K groupes dont les longueurs sont variables sont obtenus.

Ces K groupes sont ensuite regroupés suivant leur similitude de repliement. Cette similitude

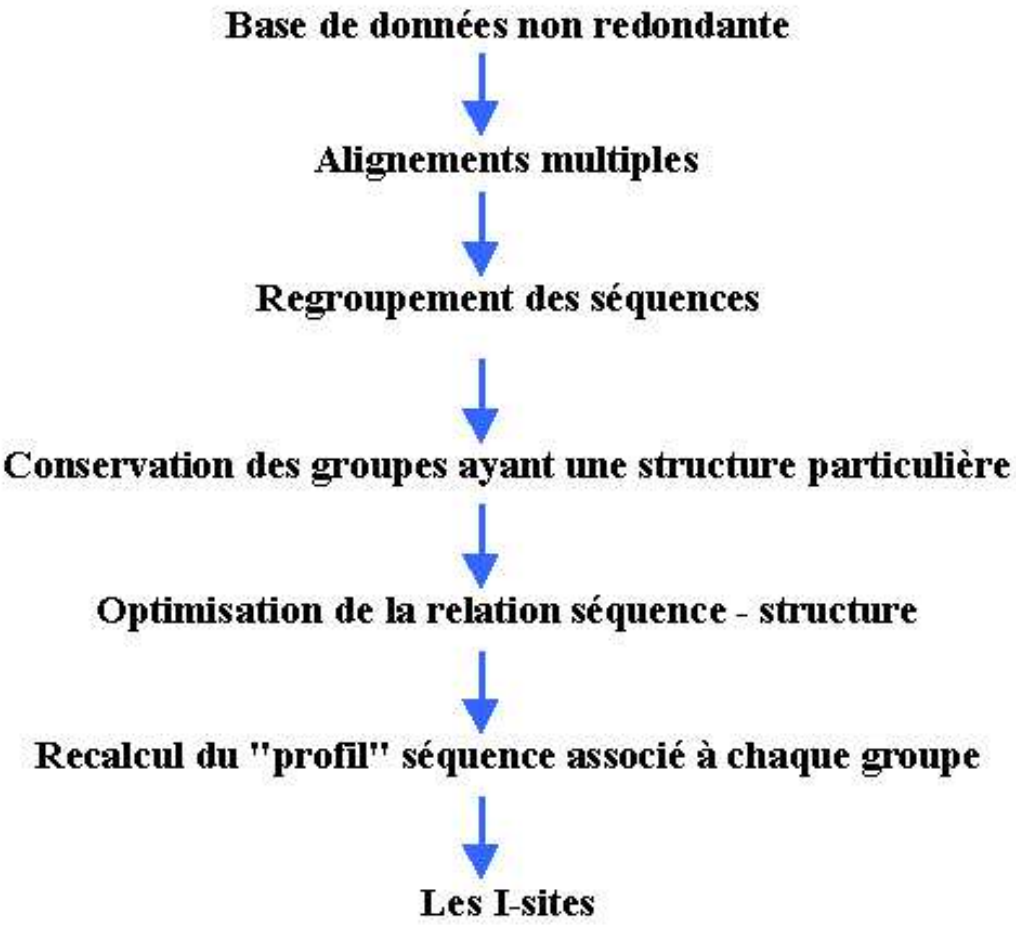

FIG. 2.21 – *Principe de la création des I-sites.*

est mesurée par deux critéres :

- dme :la "distance matrix error" qui revient au classique *RMSd* sur les C_α.
- dma :la déviation maximale angulaire : l'angle dièdre (ϕ ou ψ) ayant l'écart le plus important entre les deux structures :

$$dma(L) = max_{i=1,L-1}(\delta\phi_{i+1}, \delta\psi_i)$$

Pour créer un groupe, on sélectionne les "meilleures" séquences de chacun des K groupes. Ces séquences sont définies comme les plus proches du consensus (*centre*) de chaque groupe. Deux fragments protéiques sont considérés structurellement similaires si leur *dme* est inférieure à 1.4 Å et leur *dma* est inférieure à 120°. Cette méthode permet de regrouper leurs K groupes-séquences en K' groupes-structures. Une nouvelle optimisation est alors effectuée: elle consiste à ne conserver que les fragments les plus proches sur le plan de la séquence et à recalculer le profil séquence moyen. Les 400 séquences les plus proches servent de nouveau centre et le processus recommence. Il suffit de 3 à 5 cycles d'apprentissage pour stabiliser les groupes. Une autre étape d'optimisation suit, mais elle est peu décrite, et sert de dernier filtrage entre la séquence et la structure.

Les figures 7.7 à 7.11 montrent les $K' = 13$ groupes finaux obtenus à partir des $K = 82$ groupes obtenus sur les séquences. Les figures présentant les I-sites sont mises en Annexe 3. Les tailles varient au départ entre 3 et 17 résidus. Après le second regroupement, les longueurs moyennes observées ne sont pas données.

Les groupes ou I-sites sont fortement liés aux structures secondaires répétitives avec pour les hélices α 7 sites : le I-site 1, une hélice α amphipatique (cf. figure 7.7a), le I-site 2, hélice α non polaire (cf. figure 7.7b), le I-site 3, extrémité C-terminale d'hélice α Glycine dite de type 1 (cf. figure 7.7c), le I-site 4, extrémité C-terminale d'hélice α Glycine dite de type 2 (cf. figure 7.8a), le I-site 5, extrémité C-terminale d'hélice α Proline (cf. figure 7.8b), le I-site 6, hélice α mélée (cf. figure 7.8c), le I-site 7, extrémité N-terminale d'hélice α Sérine (cf. figure 7.9a), pour les feuillets β 2 sites : le I-site 8, feuillet β amphipatique (cf. figure 7.9b) le I-site 9, feuillet β hydrophobe (cf. figure 7.9c), pour les boucles 4 sites : le I-site 10, coude β Aspartate (cf. figure

7.10a), le I-site 11, épingle β Sérine (cf. figure 7.10b), le I-site 12, épingle type I étendu (cf. figure 7.10c), le I-site 13, coude type II divergeant (cf. figure 7.11).

Plusieurs points donnent matière à discussion sur cette première partie. Tout d'abord, l'importance des I-sites en relation avec les hélices α avec 3 I-sites pour la partie centrale, 1 pour les N-terminales et 3 pour les extrêmités C-terminales avec les deux types d'acides aminés classiques Proline et Glycine. Seuls deux I-sites sont clairement attachés aux feuillets β. Les 4 I-sites associés aux boucles représentent des changements brusques et courts. Ce dernier fait est peut-être dû à l'élimination des boucles les plus longues de l'apprentissage.

Le second point est l'absence de données concernant la structure des I-sites. Leur variabilité n'est pas clairement exprimée et ne concerne que le résultat de la prédiction.

La méthode se base sur les profils moyens obtenus précédement. Les scores obtenus sont pondérés en fonction de ceux obtenus dans la base d'apprentissage. Il convient de s'arréter particulièrement sur la méthode de calcul du taux de prédiction. Elle consiste à calculer le I-site le plus probable pour un fragment de 8 résidus consécutifs. Ensuite il convient de comparer les angles dièdres du fragment avec ceux du I-site et, si la *dma* est inférieure à 120^o, les 8 sites sont considérés comme correctement prédits :

$1 \leftarrow$ si $dma < 120^o$

$0 \leftarrow$ si $dma > 120^o$

Une utilisation conjointe de PHD [165] avec les I-sites est possible, principalement pour trouver des séquences homologues. L'évaluation par PHD n'est pas une simple prédiction par type de structures secondaires, mais une redéfinition des angles des I-sites selon la concordance entre la prédiction de PHD et des I-sites.

55 nouvelles protéines sont testées et un taux de prédiction global de 48% avec les I-sites est trouvé, et un taux de 54% en combinant leur méthode avec la prédiction des structures secondaires par PHD. Le tableau 2.8 montrent les différents taux de prédiction selon le type de protéines avec les I-sites seuls, avec la méthode PHD et les deux combinées.

Ainsi, Bystroff et Baker proposent un certain nombre de blocs structuraux (13) et font la prédiction des I-sites les plus probables à partir de la séquence [19]. Un point fort positif est l'existence d'un indice de confiance à 5 niveaux vis-à-vis de la prédiction. Toutefois, leur méthode de calcul de la prédiction est particulière et peut induire en erreur. Leur mesure fait

groupe protéique	nombre de protéines	I-sites	PHD	Combiné
tout α	8	40	55	55
tout β	6	51	32	51
$\alpha + \beta$	41	50	41	54
Total	55	48	43	54

TAB. 2.8 – *Prédiction sur la série de 55 protéines non utilisées durant l'apprentissage.*

qu'un fragment de 8 résidus est intégralement considéré comme bon, si chaque angle du I-site correspondant est à moins de 120^o de la réalité. En prenant le cas extrême d'un peptide de longueur 16, si le premier et le dernier I-site sont bons, le taux de prédiction est de 100%. Ceci même quand les positions suivantes sont intégralement incompatibles.

J'ai recalculé un pourcentage de prédiction modifié *P'* pour voir la sensibilité de leur métrique :

$$P' = \frac{(N_p - (N_z \times 7))}{N_t}$$

avec N_p le nombre de sites correctement prédits, N_z, le nombre de zones continues de bonne prédiction et N_t le nombre total de résidus dans la protéine. Le principe est fort simple, si tous les sites correctement prédits sont continus, une seule zone existe donc. Alors, N_z=1, et seuls 7 sites sur les bords doivent être soustraits. La figure 2.22 montre le résulat pour une protéine de 100 résidus de longueur. La descente est moins accentuée avec une protéine de taille supérieure, mais l'exemple me paraît frappant. Sur la figure, le dernier point hyp(1) représente l'hypothèse selon laquelle tous les I-sites correctement prédits sont indépendants et aucun chevauchement n'existe.

Le calibrage de ce type de fonction est complexe, car aucune idée du nombre de zones de I-sites correctement déterminées n'est accessible, en se réferant aux structures secondaires, il serait possible d'évaluer un nombre approximatif. Un calcul rapide montre un nombre de zones au minimum supérieur à 7.

Les résultats récapitulés dans le tableau 2.8 montrent que les hélices α malgré leur importance dans les I-sites sont moyennement prédites, mais que les feuillets β ont un taux supérieur à celui généralement observé.

Avant même la publication de l'article des I-sites [19], Baker et Bystroff les ont utilisés pour

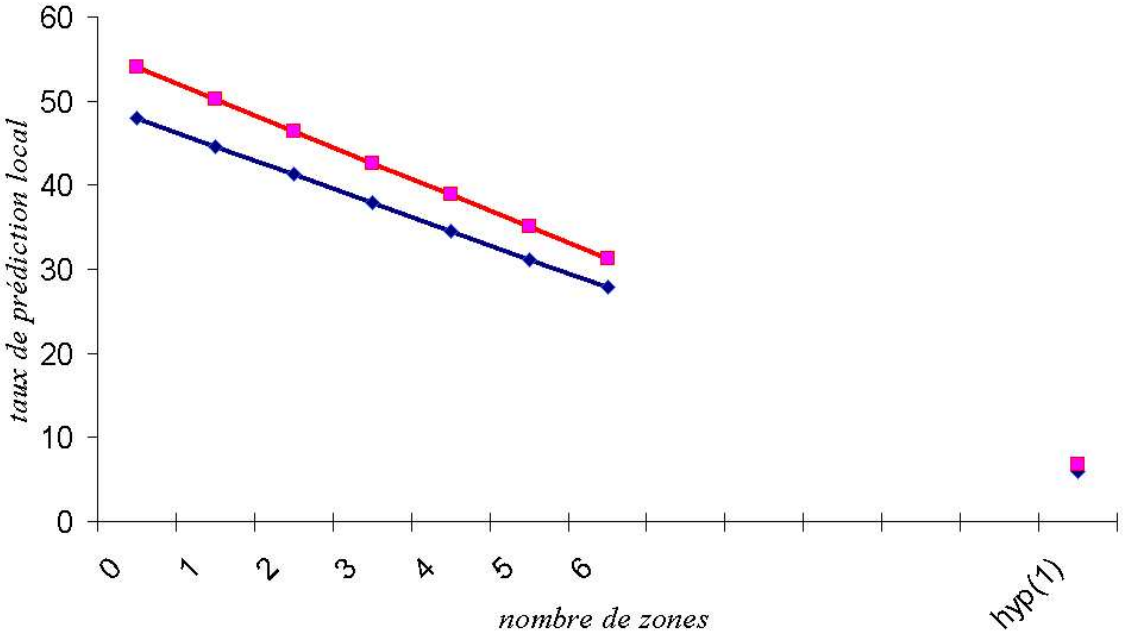

FIG. 2.22 – *Exemple du calcul du taux de prédiction pour une protéine ayant $N_t = 100$ résidus, avec en rouge les taux associés à la prédiction des I-sites combinés avec PHD et en bleu les I-sites seuls. hyp(1) représente le taux associé à l'hypothèse selon laquelle aucun site correctement prédit ne serait chevauchant avec un autre.*

prédire des structures protéiques lors de CASP 2 [18]. Sur les 8 séquences, seules 2 ont été traitées par les I-sites combinés avec PHD. Les autres n'ont utilisé que I-sites. Cela ne change pas les taux qui sont de 39% pour les protéines avec les I-sites et 41% avec la méthode combinée.

Ils appliquent, ensuite, leur méthode à la modélisation d'une zone dans un domaine SH_3 [206]. Les scores de confiance sont particulièrement élevés dans cette zone. De plus, des données RMN sur la structure corroborrent leur application. Cet exemple pose néanmoins un autre problème: l'influence de l'alignement. Leurs I-sites fort bien prédits sont les parties les plus conservées, d'un point de vue séquence.

Dans le même laboratoire, Simons et collaborateurs ont mis au point une méthode de modélisation *ab initio* sur un principe de statistique bayésienne avec recuit simulé [179]. Dans leur modèle bayésien le principe de relation séquence-structure vu par Han et Baker [75] est utilisé. Comme dans d'autres méthodes *ab initio*, seules les protéines de petites tailles ont un repliement acceptable. Ils essayent par la suite une série d'améliorations dont l'utilisation des I-sites, comme filtre au même titre que les structures secondaires, ce qui n'augmente pas l'efficacité de la méthode [180]. Pour CASP 3, l'utilisation de d'une nouvelle approche utilisant les I-sites et la méthode *ab initio*, travaillée cette fois-ci indépendament, a donné quelques très bons résultats

[178].

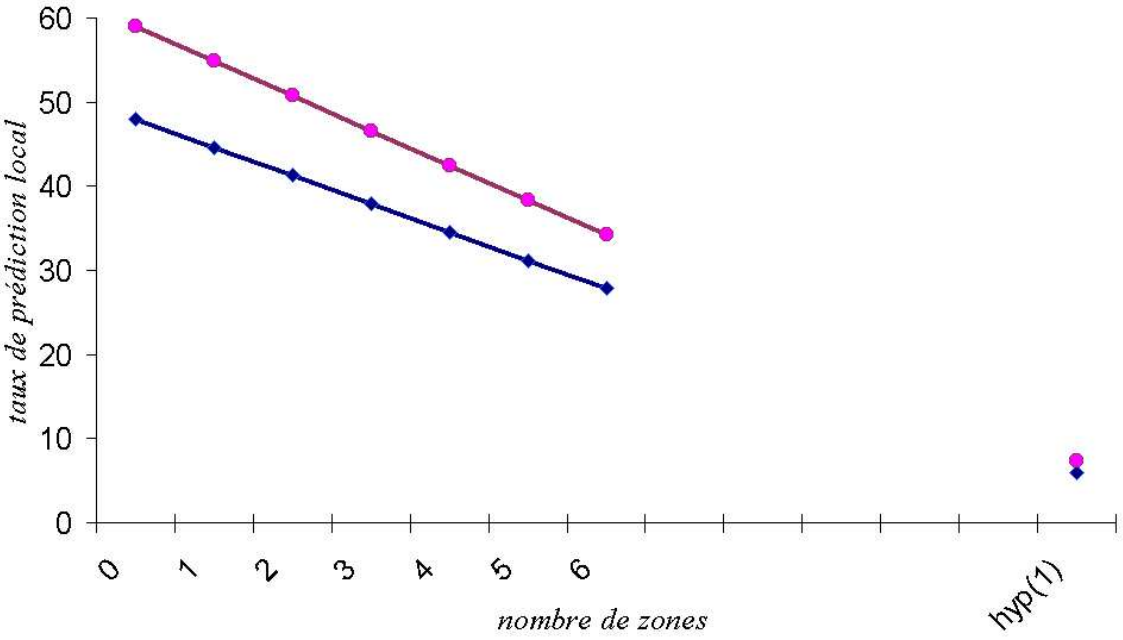

FIG. 2.23 – *Exemple du calcul du taux de prédiction pour une protéine ayant N_t = 100 résidus, avec en rouge les taux associés à la prédiction des I-sites utilisés dans HMMSTR [20] et en bleu les I-sites du papier original [19]. hyp(1) représente le taux associé à l'hypothèse selon laquelle aucun site correctement prédit ne serait chevauchant avec un autre.*

Enfin, Bystroff et collaborateurs ont utilisé les I-sites pour prédire par une méthode de Chaînes de Markov Cachées les structures secondaires (nommé HMMSTR) avec un taux final de 74,3% [20]. Pour ceci, ils ont utilisé les 13 I-sites [19], plus trois nouveaux et ensuite ont effectué leur recherche en utilisant les angles dièdres pour catégoriser le type de structure prédite. Un détail intéressant est le passage du taux de prédiction des I-sites de 48 à 59%, alors que le nombre de sites augmente, ce qui est surprenant. La figure 2.23 montre le calcul du taux de prédiction modifié P' dans ce nouveau cas. En réalité, avec quelques zones distinctes, le taux de prédiction est partiquement identique, ce qui peut expliquer ce taux "élevé" de prédiction.

2.3.4.1 Récapitulatif des alphabets structuraux et conclusion

Un petit résumé des différentes méthodes et techniques utilisées montrent bien l'avancée réalisée en une dizaine d'années. Elle est présentée dans le tableau 2.9. Commençant avec un nombre très limité de protéines, les approches sont devenues plus complexes et plus sûres, utilisant un nombre conséquent de résidus. Toutefois, elles montrent aussi l'hétérogénéité des techniques et par la suite nous verrons les difficultés que cela engendre quand on désire comparer

Equipe	année	Nombre de protéines	Nombre d'acides aminés	Méthodes d'apprentissage Méthodes d'apprentissage	distance sur	Nombre de Blocs	longueur des blocs
Unger *et al.*	1989	4 / 82	426 / 12 973	k-means	*RMSd* C_α	103	6
Rooman *et al.*	1990	75	12 978	regroupement hiérarchique	*RMSd*C_α	4	4,5, 6 et 7
Prestrelski *et al.*	1992	14	2 437	fonction	distance linéaire et angle α	113	8
Zhang *et al.*	1993	74	13114	autoANN	distance C_α	6	7
Schuchhardt *et al.*	1996	136	24239	carte de Kohonen	angles dièdres et de valence angles dièdres ϕ et ψ	100	9
Fetrow *et al.*	1997	116	23 355	autoANN	distance C_α	6	7
Bystroff et Baker	1998	471	*(NP)*	k-means	angles dièdres et de valence profil de séquence et *RMSd / dma*	13	structure : 3-15, séquence :
Camproux *et al.*	1999	100	19317	CMC	distance C_α	12	4

TAB. 2.9 – *Récapitulatif des différents travaux, avec le nom de léquipe, l'année de publication de l'article principal, le nombre de protéines utilisées et le nombre de résidus correspondants, la méthode d'apprentissage et le type de distance utilisée, puis le nombre de blocs protéiques et leurs longueurs. (NP : non précisé).*

entre eux ces différents alphabets.

2.4 Conclusion

Les différentes méthodes d'analyse de la structure protéique montrent l'intérêt des structures secondaires, alphabet structural simple. Les progrès dans leur prédiction sont plus que prometteurs. Toutefois, même en supposant qu'ils puissent être parfaits, ils ne sont qu'un premier pas dans l'établissement d'une structure protéique 3D. En plus de l'analyse classique des boucles, les alphabets structuraux sont un outil précieux d'analyse pour comprendre l'architecture protéique et la répartition des acides aminés.

Les alphabets structuraux récents montrent de constants progrès et leur plus grande finesse, liée à la fois à un contrôle du nombre de blocs structuraux en jeu et à des méthodes d'obtention plus complexe. Toutefois des améliorations restent à faire, pour aboutir à un juste compromis entre une méthode d'obtention de prototypes structuraux simples utilisables dans une méthode de prédiction efficace qui permettent, en outre, une compréhension des différents acides aminés impliqués.