Projet Long: Cartes topologiques de Kohonen et coudes protéiques

Nicolas Silva

https://github.com/nicolassilva/projetLong_Kohonen-BetaTurn

Table des matières

1 Introduction

La structure tridimensionnelle (3D) d'une protéine a un impact important sur ses propriétés fonctionnelles. Depuis maintenant de nombreuses années, ces propriétés ont été analysées sur la base notamment des structures secondaires de la protéine. Les structures secondaires sont composées de nombreuses parties répétitives, comme les hélices α (Pauling, Corey, & Branson, 1951) (un tiers des résidus) ou encore des feuillets β (Pauling & Corey, 1951) (un cinquième des résidus), le tout lié par des boucles (environ 50% des résidus) (Eisenberg, 2003). De ce point, il est donc possible de passer d'une structure 3D à une structure unidimensionnelle avec chaque résidus associé à une structure secondaire précise (Fourrier, Benros, & De Brevern, 2004). De plus, d'autre conformations locales ont été caractérisées comme les coudes.

Les coudes sont des conformations locales qui comprennent n résidus consécutifs nommés de i à $i+n$. Il y a différents types de coudes en fonction du nombre de résidus impliqués dans le coude dont les coudes-γ ($n = 3$) (Matthews, 1972 ; Milner-White, 1990), les coudes-β ($n = 4$), les coudes-α ($n = 5$) (Nataraj, Srinivasan, Sowdhamini, & Ramakrishnan, 1995 ; Pavone et al., 1996) et les coudes-π ($n = 6$) (Dasgupta & Chakrabarti, 2008 ; Rajashankar & Ramakumar, 1996). Les coudes sont caractérisés par une distance entre les carbones α (Cα) des résidus i et $i+n$ inférieure à 7Å et c'est cette distance qui créer la conformation spéciale du squelette de la protéine en la faisant se retourner sur elle même. Parmi les différents types de coudes, ce sont ceux du type β qui ont été le plus étudiés. Les coudes-β sont définis et classifiés en fonction des valeurs des angles dièdres φ (Phi) et ψ (Psi) des résidus centraux *i+1* et *i+2* (Figure 1).

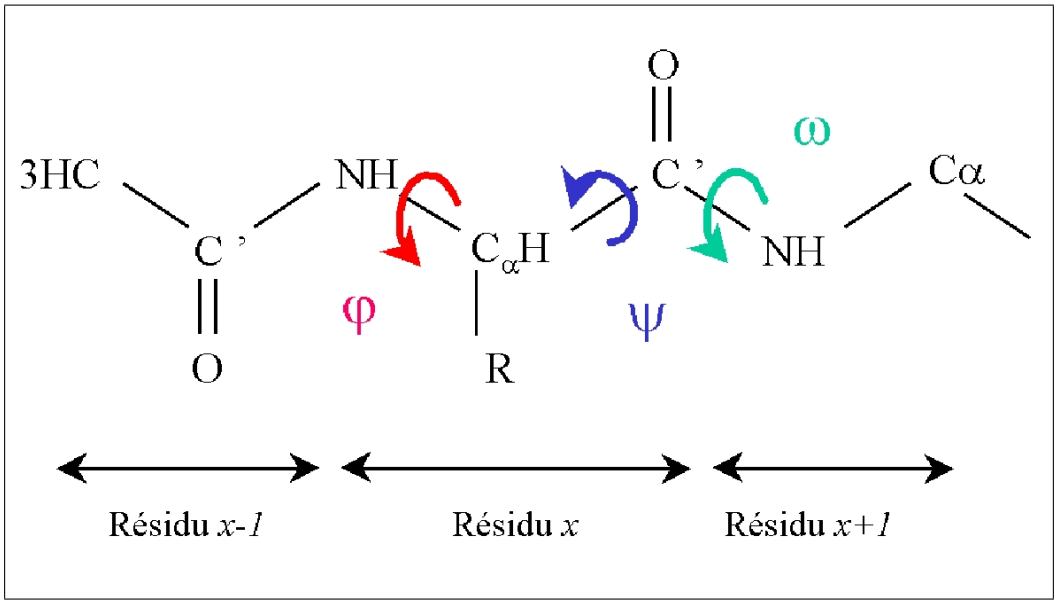


FIGURE 1 – Schéma des angles dièdres φ, ψ et ω (Omega) des résidus centraux *i+1* et *i+2*.

C'est en 1968 que les coudes-β sont pour la première fois étudiés et définis Venkatachalam. Ils sont caractérisés par une liaison hydrogène entre le N-H et C=O des résidus i et *i+3* (Venkatachalam, 1968). Lors de cette étude, il définit différents types de coudes-β, les types I, II et III ainsi que leur équivalents miroirs I', II' et III'. Au cours des années et des recherches, différents types de coudes ont été ajoutés tels les types V et V', le

type VI caractérisés par la présence d'une Proline, le type VII, le type $VIII$ (Wilmot & Thornton, 1990) et le type IV qui correspond aux différents coudes-β qui ne rentrent dans aucuns des autres types (Lewis, Momany, & Scheraga, 1973). Alors que certains sont ajoutés, d'autres sont cependant retirés comme les coudes-β de type III et III' qui sont extrêmement proches des types I et I' ainsi que des hélices 3.10 ; ou encore comme les types V, V' et VII qui sont rares et dont la caractérisation était inexacte. (Richardson, 1981). D'autres types ont quant à eux été sous-divisés, comme le type VI qui a été divisé en deux avec les types VI_a et VI_b, puis en VI_{a1}, VI_{a2} et VI_{a2} (Hutchinson & Thornton, 1994) ; et le type IV sous divisé en cinq parties IV_1, IV_2, IV_3, IV_4 et IV_{misc} (De Brevern, 2016). Le dernier sous-type IV_{misc} comprend tous les coudes qui ne rentrent dans aucunes des autres catégories. Par conséquent, on dénombre aujourd'hui 13 types de coudes-β : les types I, I', II, II', IV_1, IV_2, IV_3, IV_4, IV_{misc}, VI_{a1}, VI_{a2}, VI_b et $VIII$. Pour étudier les coudes, notamment en fonction des valeurs des couples (φ ; ψ) des résidus *i+1* et *1+2*, il est possible de les représenter sur des graphiques dit de Ramachandran (Figure 2) (Ramachandran, 1963 ; Ramakrishnan & Ramachandran, 1965 ; Venkatachalam, 1968).

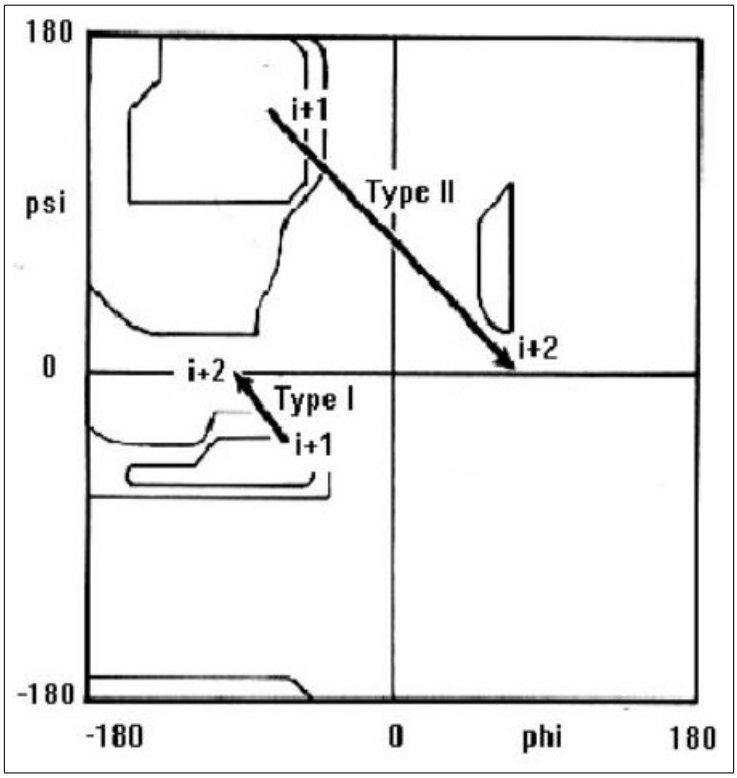


FIGURE 2 – Graphique de Ramachandran des angles φ, ψ pour les coudes-β de type I et II.

Un bon moyen d'étudier la classification des coudes-β est les cartes de Kohonen (SOM ou Self-Organize Maps). Les SOM permettent de classifier les données sous formes de cartes qui sont dîtes auto-organisées. Elles sont parfois également appelées cartes topologiques car elles présentent la particularité de conserver la topologie des données. Il s'agit de réseaux de neurones avec chacun des neurones reliés les uns aux autres. Le processus se base sur un apprentissage compétitif et une diffusion de l'information aux autres neurones de la carte. Le but final étant d'obtenir des neurones spécialisés dans la classification de types de coudes-β.

L'objectif principal de ce projet est donc d'étudier la classification des coudes protéiques

de type β et d'amener un apprentissage différent à l'aide de cartes topologiques de Kohonen. Dans un premier, il s'agira d'étudier un jeu de données et de déterminer si les fréquences d'acides aminés, de structures et de types de coudes-β sont correctes et en accord avec les fréquences attendues dans la littérature. Dans un second temps, il s'agira d'effectuer un apprentissage à l'aide de SOM afin de classer les différents types de coudes-β. La totalité du projet (code source, résultats, ...) peut être retrouvé sur le dépôt git : https://github.com/nicolassilva/projetLong_Kohonen-BetaTurn.

2 Matériels et Méthodes

La totalité de la programmation a été effectuée sous environnement Conda récupérable sur le git. Le langage de programmation utilisé est Python3 et les graphiques des fréquences ont été réalisés avec Rstudio (version 1.2.5001)

2.1 Le jeu de données

Le jeu de données est constitué de 4 984 chaînes protéiques. Il a été généré basé sur du Culled-PDB (Culled-Protein Data Base) et serveur PISCES (Wang & Dunbrack Jr, 2003). Seulement un jeu de protéines est utilisé ici. Ce jeu contient pas plus de 20% d'identité de séquence par paire. Les chaînes sélectionnées ont une résolution de cristallographie par rayons X inférieure à 2,5Å et un facteur-R inférieure à 1. Le fichier de données est composé de la manière suivante : une première ligne composé du nom de la chaîne protéique précédée de '>' suivit des acides aminés qui composent la chaîne. Plusieurs informations sont disponibles concernant chaque acide aminé. On a accès à son nom avec une lettre code (Tableau 1), une assignation ProteinBlock (De Brevern, Etchebest, & Hazout, 2000 ; Joseph et al., 2010), une assignation de structure secondaire, les valeurs des angles dièdres φ, ψ et ω, et les positions dans l'espace (X,Y,Z) en Ångström du carbone α. L'assignation des structures secondaires a été effectuée avec DSSP (Kabsch & Sander, 1983) et prévoit 9 états de structures secondaires : (0) hélice α, (1) hélice 3.10, (2) hélice π, (3) coude hydrogène, (4) coude non-hydrogène, (5) boucle, (6) pont β, (7) feuillet β et (8) structure non-assignée. Cependant, dans un objectif de simplification, les catégories ont été ici fusionnées en "hélices" ou 'H' (catégories 0, 1 et 2), en 'boucles ou 'C' (catégories 3, 4, 5 et 6), en 'feuillets' ou 'E' (catégories 7) et en 'gaps' ou '-' (catégorie 8).

Tableau 1 – Code de lecture des acides aminés.

Nom	**Code**	**Nom**	**Code**	**Nom**	**Code**	**Nom**	**Code**
Sérine	S	*Leucine*	L	*Acide glutamique*	E	*Tyrosine*	Y
Arginine	R	*Glycine*	G	*Acide aspartique*	D	*Cystéine*	C
Alanine	A	*Valine*	V	*Phénylalanine*	F	*Histidine*	H
Thréonine	T	*Proline*	P	*Asparagine*	N	*Méthionine*	M
Isoleucine	I	*Lysine*	K	*Glutamine*	Q	*Tryptophane*	W

2.2 Assignation des types de coudes-β

L'assignation des types de coudes a été effectuée en suivant les règles classiques (Bornot & de Brevern, 2006 ; Fuchs & Alix, 2005 ; Venkatachalam, 1968). La distance entre les

résidus *i* et *i+3* du coude doit être inférieure à 7Å ; les deux résidus centraux ne peuvent être des hélices et les quatre résidus ne peuvent être tous des feuillets. La règle des deux résidus centraux qui ne doivent pas être des hélices vient du fait que de part la conformation des hélices, une suite d'hélices peut parfois être confondue avec des coudes. L'assignation des types (I, I', II, II', IV_1, IV_2, IV_3, IV_4, IV_{misc}, VI_{a1}, VI_{a2}, VI_b et $VIII$) a ensuite été effectuée en fonction des valeurs des angles dièdres φ et ψ des résidus centraux *i+1* et *i+2* (Tableau 2), où 3 des angles peuvent varier au maximum de ±30° de la valeur autorisée et un des angles a le droit de varier au maximum de ±45°.

Tableau 2 – Valeurs des angles dièdres (en degré) des résidus *i+1* et *i+2* utilisées pour assigner les types de coude-β.

Type	**Résidu *i+1***		**Résidu *i+2***		**Type**	**Résidu *i+1***		**Résidu *i+2***	
	φ	ψ	φ	ψ		φ	ψ	φ	ψ
Type I	-60	-30	-90	0	*Type I'*	60	30	90	0
Type II	-60	120	80	0	*Type II'*	60	-120	-80	0
Type IV_1	-120	130	55	41	*Type* IV_2	-85	-15	-125	55
Type IV_3	-71	-30	-72	-47	*Type* IV_4	-97	-2	-117	-11
Type VIII	-60	-30	-120	120	*Type* VI_{a1}	-60	120	90	0
Type VI_{a2}	-120	-120	-60	0	*Type* VI_b	-135	135	-75	160
Type IV_{misc}	*Aucunes des autres catégories*								

2.3 Analyses des fréquences

Afin de vérifier si le jeu de données de base ne comprend pas d'erreurs, il est important de calculer les fréquences d'acides aminés, de structures secondaires et de types de coudes présents dans les chaînes protéiques. Pour les fréquences d'acides aminés, on utilise des fréquences calculées à partir de 5 492 acides aminés dans 53 vertébrés polypeptides (Tableau 3) (King & Jukes, 1969). On effectue la même chose pour les fréquences de structures secondaires attendues (Tableau 4) et les fréquences de types de coudes-β attendues (Tableau 5) tirés de l'étude de De Brevern (2016).

Tableau 3 – Fréquences (%) attendues d'acides aminés (King & Jukes, 1969).

Nom	%	**Nom**	%	**Nom**	%	**Nom**	%
Sérine	8,10	*Leucine*	7,70	*Acide glutamique*	5,80	*Tyrosine*	3,30
Arginine	4,20	*Glycine*	7,40	*Acide aspartique*	5,90	*Cystéine*	3,30
Alanine	7,40	*Valine*	6,80	*Phénylalanine*	4,00	*Histidine*	2,90
Thréonine	6,20	*Proline*	5,00	*Asparagine*	4,40	*Méthionine*	1,80
Isoleucine	3,80	*Lysine*	7,20	*Glutamine*	3,70	*Tryptophane*	1,30

Tableau 4 – Fréquences approximatives (%) attendues de structures secondaires (De Brevern, 2016).

Structure	*Hélices*	*Feuillets*	*Boucles*
Fréquence	33	20	50

Tableau 5 – Fréquences (%) attendues de type de coudes-β (De Brevern, 2016).

Type	**Fréquence**	**Type**	**Fréquence**	**Type**	**Fréquence**
Type I	38,21	*Type* IV_1	5,10	*Type* IV_{misc}	16,44
Type I'	4,10	*Type* IV_2	3,95	*Type* VI_{a1}	0,73
Type II	11,81	*Type* IV_3	3,53	*Type* VI_{a2}	0,20
Type II'	2,51	*Type* IV_4	2,70	*Type* VI_b	0.88
Type VIII	9,84				

2.4 Self-Organize Map

L'objectif principal d'une SOM est de créer un réseau composé de neurones dits spécialisés, c'est-à-dire pour l'étude menée ici que ces neurones seront spécialisés dans le classement d'un type de coude-β. L'apprentissage via des SOM se fait en plusieurs étapes. Tout d'abord il y a l'initialisation du réseau de manière aléatoire. On sélectionne ensuite une valeur du jeu de donnée que l'on va comparer au réseau. Le neurone le plus proche de cette valeur verra sa valeur à lui modifiée par la valeur aléatoire tirée et il en va de même pour les neurones voisins. Les étapes de sélection aléatoire d'une valeur à comparer au réseau et de modification de la valeurs des neurones sont réitérées n fois.

2.4.1 Initialisation

L'initialisation du réseau commence tout d'abord par la détermination de sa taille et de sa forme. Il existe plusieurs type de réseau de Kohonen avec des tailles qui peuvent varier. Ici, la taille du réseau est fixée à 16 neurones (4x4). La forme des neurones peut également être variable avec des formes hexagonales ou encore octogonales. Ici les neurones auront la forme d'un quadrilatère. Chacun des neurones aura donc 8 voisins (les neurones haut, bas, droite, gauches et les diagonaux). Il s'agira d'un système fermé afin de supprimer les effets de bord et d'avoir le même nombre de voisins pour chaque neurones. Une fois la taille et la forme du réseau décidée, l'initialisation s'effectue en attribuant à chacun des neurones une valeur du jeu de donnée de manière aléatoire. L'initialisation du réseau ne se fait qu'une seule fois au début et à ce stade le réseau est prêt pour l'apprentissage.

2.4.2 Neurone gagnant

L'objectif de cette étape est de tirer une valeur aléatoire dans le jeu de donnée et de regarder quel neurone est le plus semblable à cette valeur. Cette ressemblance a été calculée ici à l'aide de distance, le but étant de minimiser la distance entre la valeur tirée et un potentiel neurone. Ici l'apprentissage basé sur un critère de distance euclidienne n'est pas adapté à des données angulaires. En effet, une rotation de 180° et de $-180°$ équivaut au même. Ainsi la différence entre un angle de 170° et de $-170°$ est seulement de 20° et non pas de 340°. La difficulté est donc d'arriver à générer des différences angulaires en prenant en compte ce phénomène. On calcule donc la différence angulaire (D_a) de la manière suivante (1) avec V_i la valeur aléatoire tirée et w_i la valeur du neurone :

$$D_a = |V_i - W_i| \begin{cases} Si \;\; D_a < 180, \;\; alors \;\; D_a = D_a \\ Si \;\; D_a > 180, \;\; alors \;\; D_a = 360 - D_a \end{cases} \quad (1)$$

2.4.3 Diffusion

Une fois que le neurone le plus proches de la valeur aléatoire tirée est identifié, il est possible de déterminer les 8 neurones voisins et leur coordonnées dans le réseau ([numéro-de-ligne ; numéro-de-colonne]). Ensuite, la valeur du neurone vainqueur est modifiée afin que sa valeur soit plus similaire à la valeur tirée. La modification est ensuite étendue aux neurones voisins mais de manière plus modérée. Pour ce faire on utilise la fonction suivante (2) avec W_{new} la nouvelle valeur du neurone ; V_i la valeur aléatoire tirée ; W_i la valeur initiale du neurone à modifier et γ le coefficient d'apprentissage :

$$W_{new} = W_i + (V_i - W_i) \times \gamma(t) \quad avec \;\; \gamma(t) = \alpha(t) \times \beta(t) \tag{2}$$

Les coefficients $\alpha(t)$ et $\beta(t)$ correspondent respectivement au coefficient d'amplitude et au coefficient de diffusion. Le coefficient de diffusion va permettre de prendre en compte la distance de voisinage entre les neurones et modifier leur valeurs en conséquences. Les neurones les plus proches du neurone gagnant (haut, bas, droite et gauche) verront leur valeur plus modifiée que les neurones moins proches (les neurones diagonaux). Le coefficient est défini par l'équation suivante (3) avec r_{winner} les coordonnées du neurone vainqueur dans la carte de Kohonen, r les coordonnées du neurone à modifier dans la carte de Kohonen, η le coefficient de voisinage, η_0 le coefficient initial (fixé à 4), n le nombre total d'observations et t le nombre d'observation déjà vues :

$$\beta(t) = \exp -\frac{(r - r_{winner})^2}{2\eta^2} \quad avec \;\; \eta = \frac{\eta_0}{1 + \frac{t}{n}} \tag{3}$$

Le coefficient d'amplitude $\alpha(t)$ va permettre au début de l'apprentissage une modification importante des neurones et une modification de plus en plus faible à la fin (Figure 7). Cela permet une stabilisation du réseau à la fin de l'apprentissage. Ce coefficient est défini par l'équation suivante (4) avec α_0 le coefficient initial (fixé à 0,8), n le nombre total d'observation et t le nombre d'observations déjà vues :

$$\alpha(t) = \frac{\alpha_0}{1 + \frac{t}{n}} \tag{4}$$

Les résultats de l'apprentissage seront présentés sous forme de graphiques de Ramachandran. Chaque neurone de la carte sera associé à un graphique composé de 2 points, un pour le résidus $i+1$ et un pour le résidu $i+2$. Les graphiques seront présentés avant/après apprentissage avec les types de coudes-β associés pour chaque neurones.

3 Résultats

3.1 Fréquences

Etudier les fréquences d'acides aminés et de structures secondaires dans des séquences permet notamment de vérifier que le jeu de données semble correcte et que la lecture de celui-ci s'effectue bien. Les proportions d'acides aminés présents dans les chaînes protéiques correspondent environ aux proportions attendues (Figure 3). Sur les 20 types d'acides aminés seuls les Sérines, les Cystéines et les Leucines semblent être légèrement différents (respectivement 5,89%, 1,27% et 9,24% pour des valeurs attendues de 8,10%, 7,70% et 3,30%). Les proportions de structures secondaires présentes dans les séquences sont elles

aussi semblables aux fréquences attendues dans des séquences protéiques (Figure 4). En effet, les feuillets présentent une différence inférieure à 1% (20% attendus pour 20.78% observés). Pour ce qui concerne les hélices, les proportions observées sont légèrement supérieures aux fréquences attendues (38.88% au lieu de 33%). Les boucles elles sont environ 10% moins présentes que dans les résultats attendues. Cependant cette différence reste faible.

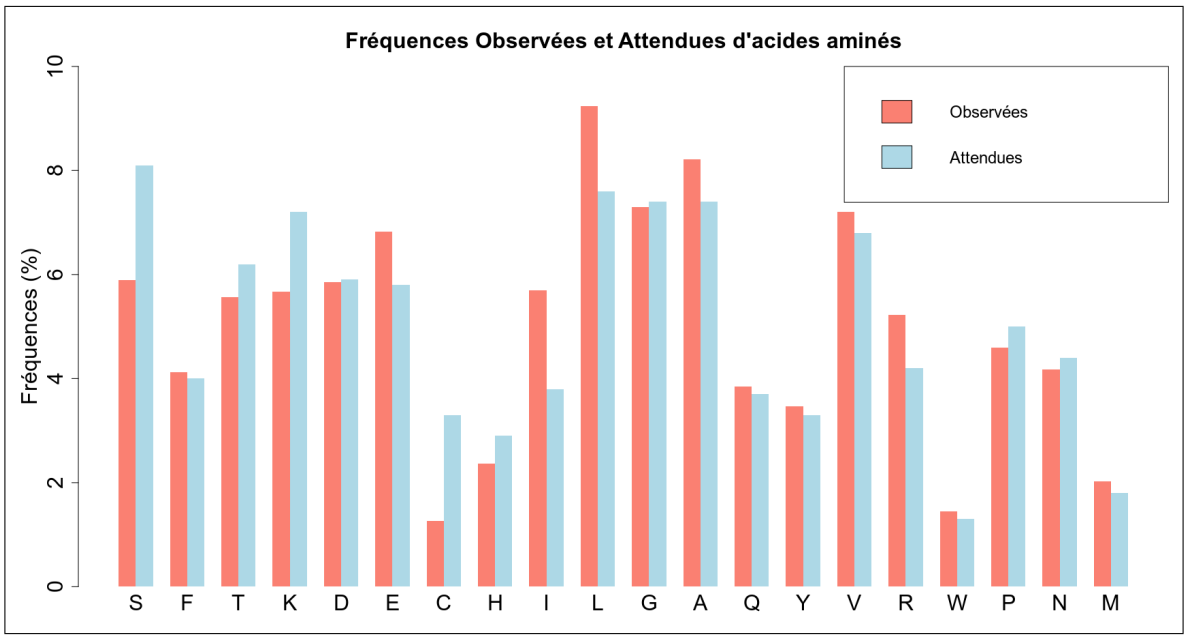


FIGURE 3 – Fréquences d'acides aminés observées dans les séquences comparées aux fréquences attendues.

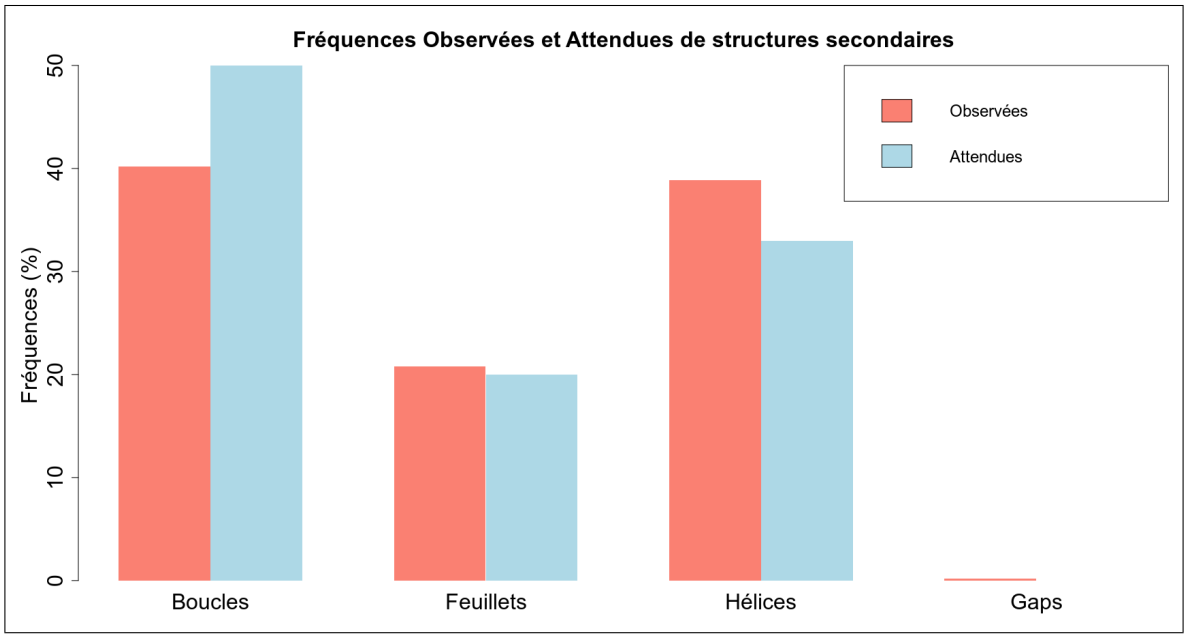


FIGURE 4 – Fréquences de structures secondaires observées dans les séquences comparées aux fréquences attendues.

Si les fréquences observées d'acides aminés et de structures secondaires dans les chaînes protéiques sont similaires aux fréquences attendues, ce n'est pas le cas pour les proportions

de types de coudes-β (Figure 5). En effet, les données de la littérature prévoient une majorité de type I. Dans l'ensemble des chaînes protéiques analysées, très peu de type I sont retrouvés (38,21% attendus contre 0,08% observés). A l'inverse, il est attendu peu de type IV_3 (3,53%) alors qu'il s'agit ici du type le plus présent dans les données (27,78%). Au final seuls la moitiés des types (II, II', IV_1, VI_{a1}, VI_{a2}, VI_b et $VIII$) semblent être présents dans les proportions attendues.

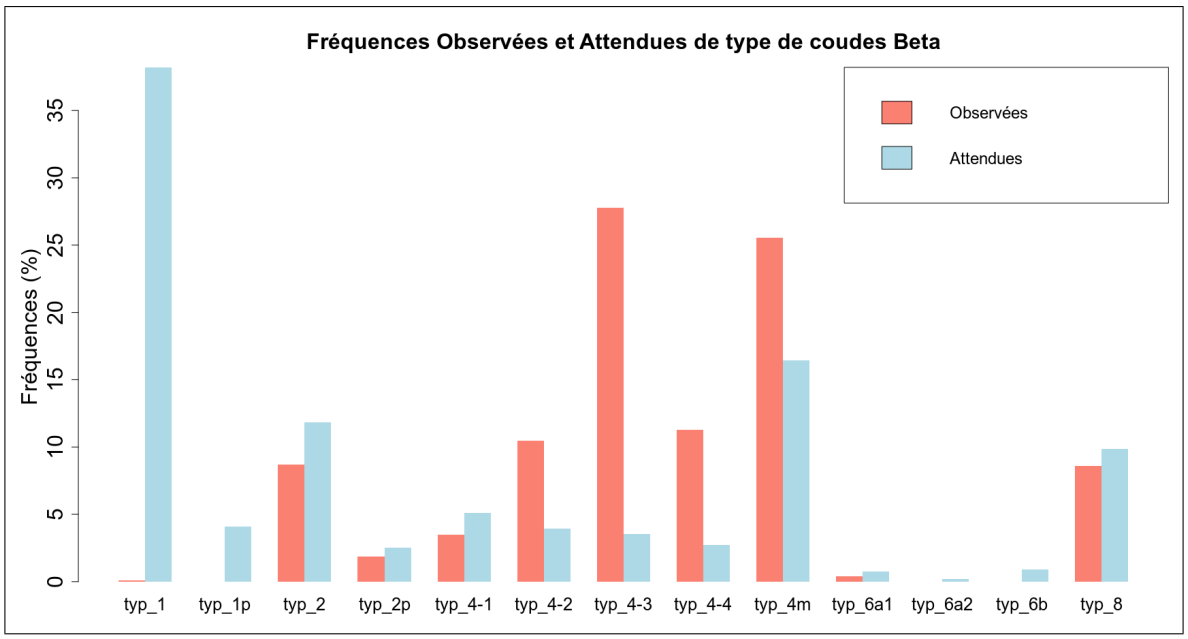
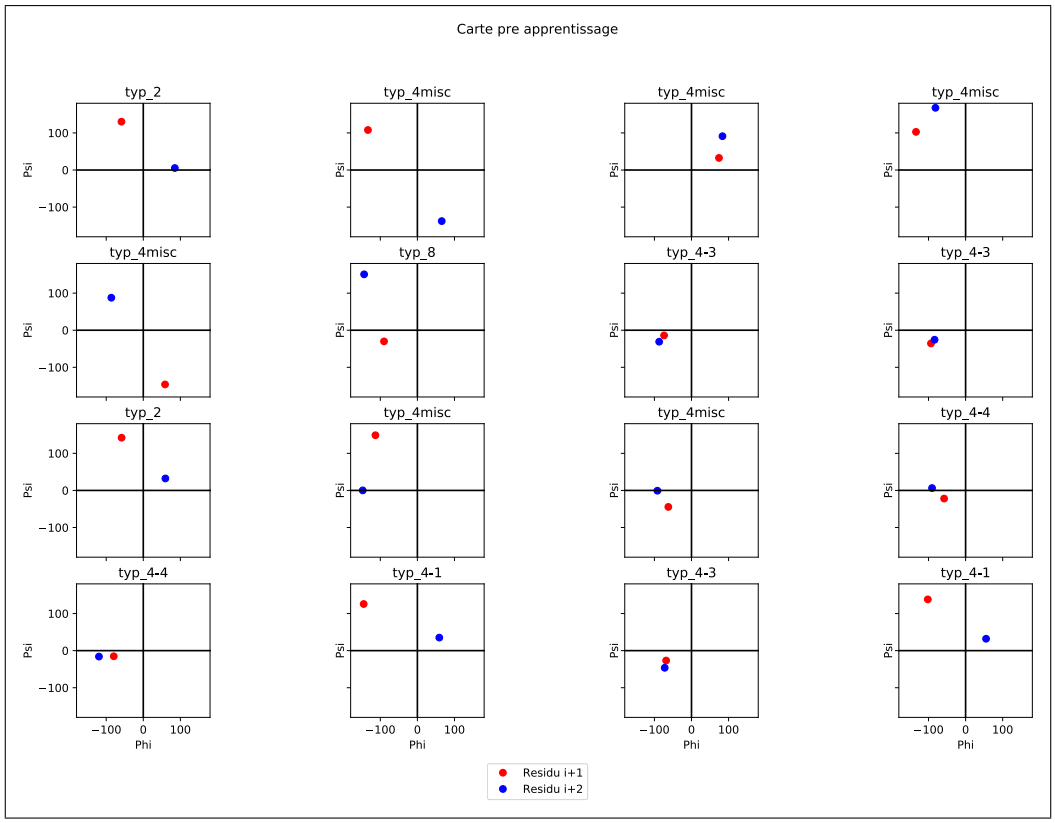


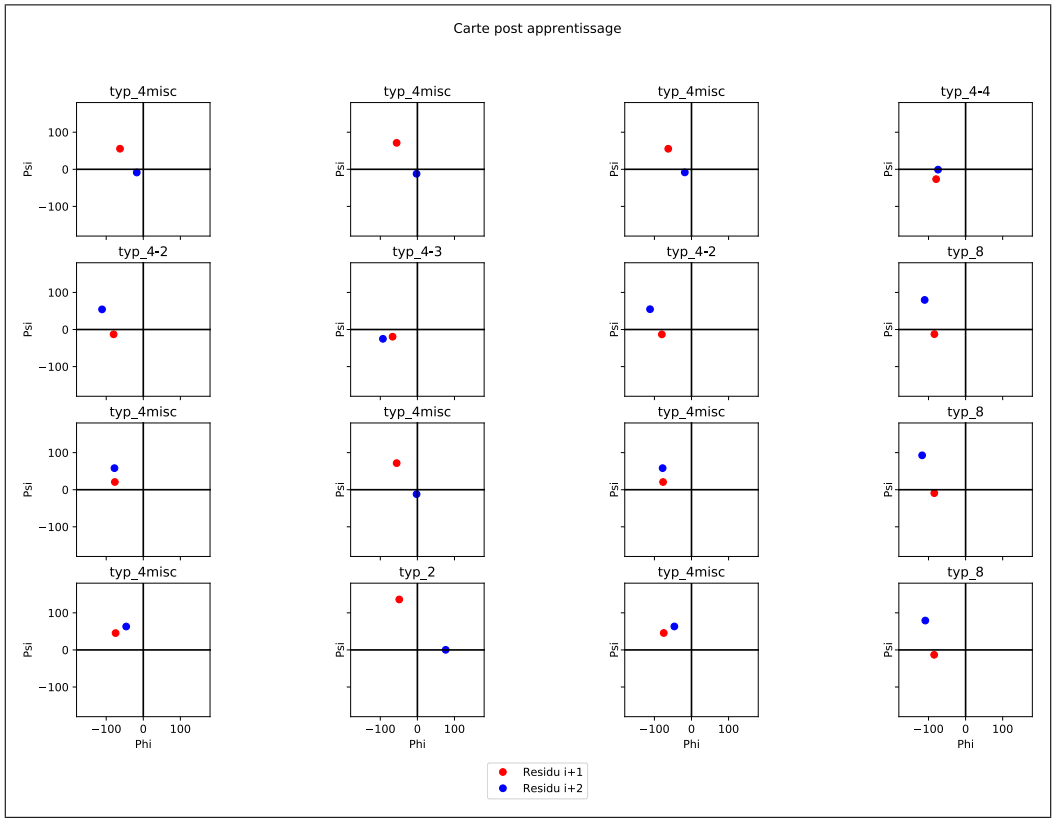
FIGURE 5 – Fréquences de types de coudes-β observées dans les séquences comparées aux fréquences attendues.

3.2 Classification via les SOM

Lorsque les cartes avant et après apprentissage sont comparées, il y a le même nombre de type de coude différents avant et après (Figure 6a et Figure 6b). En effet, la carte de Kohonen est initialisé avec 6 types de coude (II, IV_1, IV_3, IV_4, IV_{misc} et $VIII$) et on retrouve également des neurones spécilisé dans la classification de 6 types différents. Cependant, il ne s'agit pas des même types. En effet, le type IV_1 a disparu et le type IV_2 qui n'était pas présent à l'initialisation est désormais présent à la fin de l'apprentissage. Cependant ce changement n'est pas la seule différence. Il est également possible de voir des différence quand à quel neurones est spécialisé dans la classification des types. Le neurone $[1;1]$ initialisé avec un type II se trouve spécialisé dans les type IV_{misc} à la fin de l'apprentissage. Au final seulement 4 neurones présente une initialisation et un état post apprentissage avec le même type (les neurones $[1;2]$, $[1;3]$, $[3;2]$ et $[3;3]$ n'ont pas changé de type de spécialisation pour les types IV_{misc} après l'apprentissage).



(a) Valeurs des neurones avant apprentissage.



(b) Valeurs des neurones après apprentissage.

FIGURE 6 – Type de coude pour chacun des neurones avant et après la phase d'apprentissage de la carte de Kohonen. Les cartes se lisent de gauche à droite et de haut en bas. En abscisse les valeurs des angles φ et en ordonnée les valeurs des angles ψ. L'apprentissage a été réalisé sur 200 itérations.

4 Discussion

L'étude des fréquences d'acides aminés et de structures secondaires présents dans les chaînes protéiques du jeu de données utilisé permet de montrer que les données d'acides aminés et les données structurales et surtout la lecture des données sont correctes. Les différences entre les fréquences observées et les fréquences attendues de types de coudes-β peuvent elles être significatives de différentes choses. En effet, ces différences pourraient potentiellement venir tout simplement d'une mauvaise assignation de types de coudes. Si les conditions d'assignations sont erronées, alors les fréquences associées aux types seront également fausses. Cependant, une erreur d'implémentation n'est peut être pas le seul problème potentiel. Après vérification, les assignations des coudes suivent les bonnes conditions et également les bonnes valeurs d'angles φ/ψ. Les différences de fréquences proviennent peut être alors d'erreurs de valeurs d'angles φ/ψ dans le fichier de données. Enfin la dernière hypothèse est que les chaînes protéiques étudiées présentent un biais comparé aux fréquences moyennes et ne comportent pas les fréquences attendues. Cependant une différence entre les fréquences observées et les fréquences attendues de type de coudes ne sont pas un problème pour permettre l'étude des coudes-β à l'aide de cartes de Kohonen.

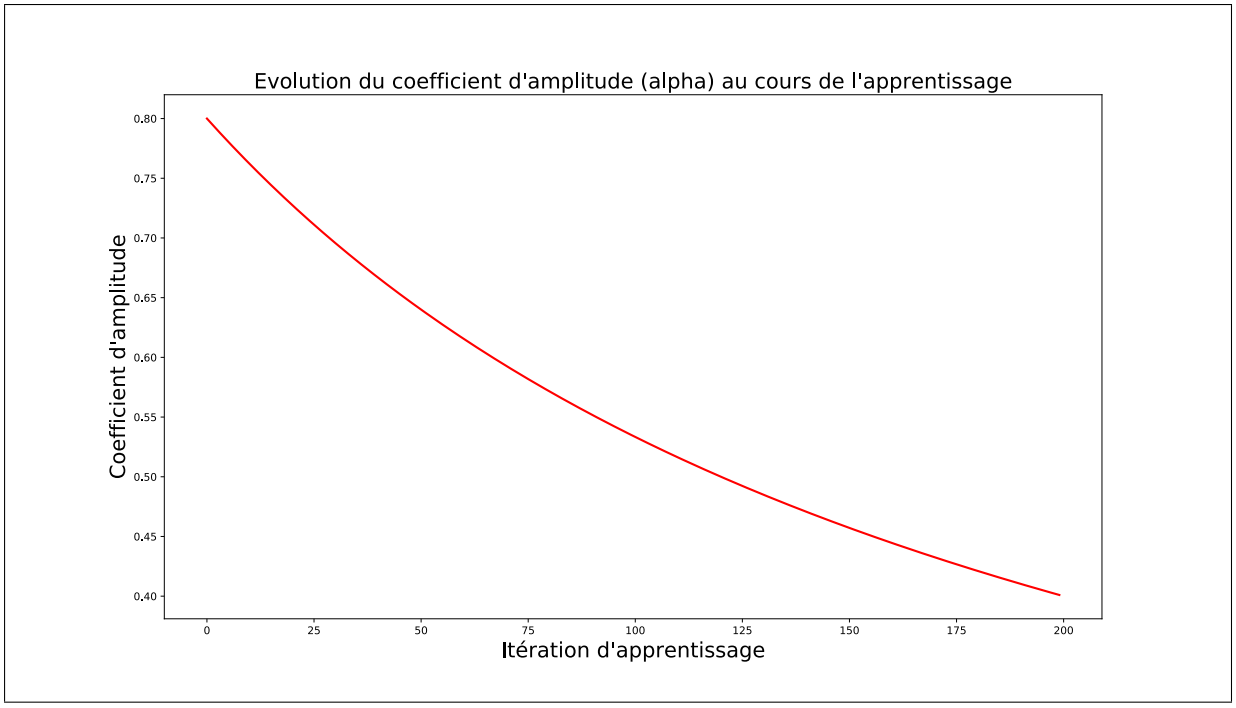


FIGURE 7 – Evolution du coefficient d'amplitude α durant l'apprentissage.

Durant l'apprentissage, le coefficient d'amplitude α(t) diminue (Figure 7). Cette diminution permet une stabilisation de la carte de Kohonen. Ici, seul quelques types de coudes ont été placés dans la carte à l'initialisation avec une majorité de types IV_{misc} et IV_3. Cette majorité s'explique tout simplement par le fait qu'il y a une majorité de ces coudes présents dans le fichier de données. L'évolution de la carte de Kohonen avant et après apprentissage permet de statuer d'un apprentissage au cours des itérations. En effet, même si les types de coudes-β présents dans la carte avant et après la phase d'apprentissage sont quasiment les mêmes, ils ne sont pas situés sur les mêmes neurones. Ce phénomène montre que malgré l'initialisation d'un neurones pour un type de coude-β, l'apprentissage

avec d'autres types de coudes a permis la spécialisation des neurones. Après la phase d'apprentissage, on retrouve encore une fois une majorité de coudes de types IV_{misc}, IV_4 et IV_3, ce qui s'explique encore une fois par une majorité de ces coudes dans le jeu de données.

Sur les 13 types de coudes-β présents dans le jeu de données, moins de la moité sont dans les neurones à la fin de l'apprentissage (6 types sur 13). Une solution qui pourrait être évidente pour augmenter la diversité de types serait d'augmenter le nombre de neurones, alors la probabilité d'obtenir des coudes différents augmenterait. Cependant, ici la diversité resterait identique car les neurones même initialisés avec une plus grande diversité de coude subiraient tout de même un apprentissage avec les mêmes types de coudes majoritaires dans le jeu de données. La solution serait donc de créer un sous jeu de données contenant le même nombre de coudes pour tout les types afin que la probabilité de tirer un type au hasard pour l'apprentissage soit la même pour tout les types.

L'apprentissage de la classification des coudes-β grâce aux cartes de Kohonen implémenté ici reste simple. En effet, plusieurs adaptations peuvent être effectuées. Une des variantes d'intérêt est d'implémenter la partie de la diffusion avec la règle dîte de Fisherman (Lee & Verleysen, 2002). Dans un apprentissage de Kohonen classique, le vainqueur est déterminé et celui-ci ainsi que les neurones voisins voient leur valeur modifiée par la valeur tirée au hasard. Dans l'adaptation avec la règle de Fisherman, les neurones voisins sont modifiés par la valeur du neurone vainqueur et non par la valeur tirée au hasard. Il serait donc intéressant de comparer ces deux manières d'apprentissage.

Références

Bornot, A., & de Brevern, A. G. (2006). Protein beta-turn assignments. *Bioinformation*, *1*(5), 153.

Dasgupta, B., & Chakrabarti, P. (2008). pi-turns : types, systematics and the context of their occurrence in protein structures. *BMC structural biology*, *8*(1), 39.

De Brevern, A. G. (2016). Extension of the classical classification of β-turns. *Scientific reports*, *6*, 33191.

De Brevern, A. G., Etchebest, C., & Hazout, S. (2000). Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins : Structure, Function, and Bioinformatics*, *41*(3), 271–287.

Eisenberg, D. (2003). The discovery of the α-helix and β-sheet, the principal structural features of proteins. *Proceedings of the National Academy of Sciences*, *100*(20), 11207–11210.

Fourrier, L., Benros, C., & De Brevern, A. G. (2004). Use of a structural alphabet for analysis of short loops connecting repetitive structures. *BMC bioinformatics*, *5*(1), 58.

Fuchs, P. F., & Alix, A. J. (2005). High accuracy prediction of β-turns and their types using propensities and multiple alignments. *Proteins : Structure, Function, and Bioinformatics*, *59*(4), 828–839.

Hutchinson, E. G., & Thornton, J. M. (1994). A revised set of potentials for β-turn formation in proteins. *Protein Science*, *3*(12), 2207–2216.

Joseph, A. P., Agarwal, G., Mahajan, S., Gelly, J.-C., Swapna, L. S., Offmann, B., ... others (2010). A short survey on protein blocks. *Biophysical Reviews*, *2*(3), 137–145.

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure : pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers : Original Research on Biomolecules*, *22*(12), 2577–2637.

King, J. L., & Jukes, T. H. (1969). Non-darwinian evolution. *Science*, *164*(3881), 788–798.

Lee, J. A., & Verleysen, M. (2002). Self-organizing maps with recursive neighborhood adaptation. *Neural Networks*, *15*(8-9), 993–1003.

Lewis, P. N., Momany, F. A., & Scheraga, H. A. (1973). Chain reversals in proteins. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, *303*(2), 211–229.

Matthews, B. (1972). The γ turn. evidence for a new folded conformation in proteins. *Macromolecules*, *5*(6), 818–819.

Milner-White, E. J. (1990). Situations of gamma-turns in proteins : Their relation to alpha-helices, beta-sheets and ligand binding sites. *Journal of molecular biology*, *216*(2), 385–397.

Nataraj, D., Srinivasan, N., Sowdhamini, R., & Ramakrishnan, C. (1995). a-turns in protein structures. *Curr. Sci*, *69*, 434–437.

Pauling, L., & Corey, R. B. (1951). The pleated sheet, a new layer configuration of polypeptide chains. *Proceedings of the National Academy of Sciences of the United States of America*, *37*(5), 251.

Pauling, L., Corey, R. B., & Branson, H. R. (1951). The structure of proteins : two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences*, *37*(4), 205–211.

Pavone, V., Gaeta, G., Lombardi, A., Nastri, F., Maglio, O., Isernia, C., & Saviano, M. (1996). Discovering protein secondary structures : Classification and description of isolated α-turns. *Biopolymers*, *38*(6), 705–721.

Rajashankar, K., & Ramakumar, S. (1996). π-turns in proteins and peptides : Classification, conformation, occurrence, hydration and sequence. *Protein science*, *5*(5), 932–946.

Ramachandran, G. N. (1963). Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, *7*, 95–99.

Ramakrishnan, C., & Ramachandran, G. (1965). Stereochemical criteria for polypeptide and protein chain conformations : Ii. allowed conformations for a pair of peptide units. *Biophysical journal*, *5*(6), 909–933.

Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. In *Advances in protein chemistry* (Vol. 34, pp. 167–339). Elsevier.

Venkatachalam, C. (1968). Stereochemical criteria for polypeptides and proteins. v. conformation of a system of three linked peptide units. *Biopolymers : Original Research on Biomolecules*, *6*(10), 1425–1436.

Wang, G., & Dunbrack Jr, R. L. (2003). Pisces : a protein sequence culling server. *Bioinformatics*, *19*(12), 1589–1591.

Wilmot, C., & Thornton, J. (1990). β-turns and their distortions : a proposed new nomenclature. *Protein Engineering, Design and Selection*, *3*(6), 479–493.