Annexe 1 : les cartes de Kohonen

réseaux neuronaux artificiels Ce sont des modèles auto-organisés et connexionnistes. Un réseau d'auto-organisation est un réseau d'éléments de traitements simultanément actifs (nœuds et connexions). Les modèles connexionnistes utilisent une information numérique et sont des systèmes dynamiques qui effectuent des calculs analogues à ceux d'un neurone.

Un modèle connexionniste est caractérisé par trois constituants de base: un réseau, une règle d'activation et une règle d'apprentissage.

Le *réseau* est composé des nœuds (*unités*) connectés par des liens orientés (*connexions*). La règle d'activation d'un modèle connexionniste est une procédure locale que chaque nœud suit en mettant à jour son niveau d'activation en fonction du contexte d'activation des nœuds voisins. Deux aspects sont à voir à ce niveau, tout d'abord le parallélisme massif de l'activation qui implique une diffusion de l'activité et le caractère local de l'information traitée par chaque nœud.

La *règle d'apprentissage* est la propriété du réseau à changer son comportement d'après les résultats de ses activations passées. Localement, le poids de la connexion de chaque nœud est réévaluée en fonction de sa valeur actuelle et des niveaux d'activations des nœuds qu'elles connectent.

Les réseaux linéaires forment une classe particulière de modèles neuronaux (cf fig. 7.2).

On doit distinguer l'apprentissage supervisé où un règle delta qui minimise l'erreur quadratique est appliquée, le couple entrée-sortie est présenté, la réponse est donc pondérée par rapport aux résultats et l'apprentissage non supervisé où le réseau est simplement exposé aux différents exemples sans aucun type de correction.

Explication biologique du modèle de Kohonen Le modèle de Kohonen repose sur l'observation neurophysiologique que les détecteurs de caractéristiques de différentes aires du cortex sensoriel loin d'être indépendants sont en fait regroupées en carte dont la topologie est corres-

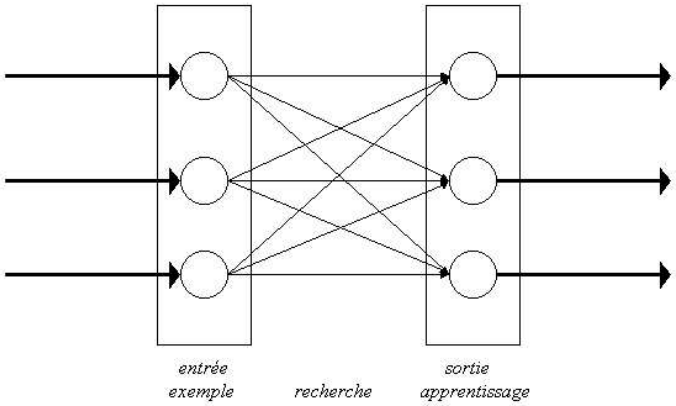

FIG. 7.2 – *Réseau à deux couches et à circulation de l'information dirigée vers l'avant.*

pondance avec la topologie spatiale. Les activités de des détecteurs de caractéristiques sont corrélées suivant la relation de voisinage pour cette topologie corticale. Il se rapproche d'un algorithme de d'apprentissage concurrentiel. Il n'est pas supervisé [108].

Principe d'apprentissage et d'assignation Une couche de neurones d'entrée définit l'espace des entrées possibles. une couche de neurones de sorties sera mise en correspondance avec l'espace des prototypes. Les poids synaptiques reliant les neurones de la couche d'entrée à un neurone de la couche de sortie définissent les coordonnées dans l'espace des entrées du prototype représenté par le neurone de sortie.

L'apprentissage tient compte de la structure topologique de la grille des entrées. L'accroissement de l'algorithme de d'apprentissage concurrentiel est appliqué au plus proche prototype du motif d'entrée mais aussi aux voisins de ce motif.

Dans un premier temps, une phase de *structuration topologique* voit la grille de la couche des sorties se positionne dans l'espace des entrées. Ses points s'ordonnent les uns par rapport aux autres.

Ensuite, une *phase de convergence* où la grille se déforme lentement en conservant sa structuration pour converger vers un échantillonnage régulier de la distribution de probabilités de l'espace des entrées.

Après apprentissage, un motif d'entrée sera représenté par le neurone dont il s'approche le plus. Par conséquent, une fois l'apprentissage achevé, les valeurs des connexions définissent un pavage de l'espace des entrées qui doit échantillonner au mieux la distribution de proba-

bilité des motifs d'entrée. La principale caractéristique est que la carte topologique obtenue n'a aucun rapport avec les dimensions de l'espace des entrées. Elle est stable, robuste et d'une représentation simple.

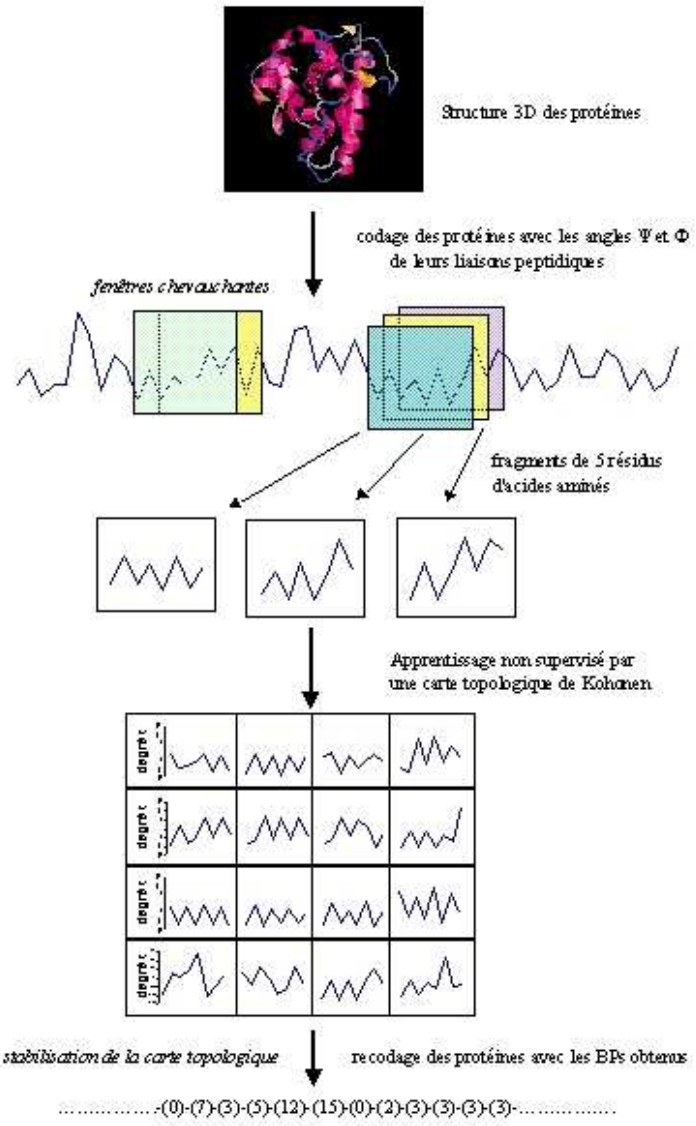

FIG. 7.3 – *Schéma récapitulatif du travail effectué par Schuchhardt et collaborateurs, avec la traduction de la structure tridimensionnelle des protéines en termes d'angles ϕ et ψ, puis leur utilisation dans une carte de Kohonen pour obtenir des blocs structuraux.*

L'intérêt des cartes topologiques est la diffusion latérale qui permet de diminuer les risques de minimum local comme dans une méthode de nuées dynamiques ou de k-means. Une carte topologique est représentée par un réseau de N neurones. Il est initialisé en tirant au hasard des fragments de la base de données pour donner des valeurs initiales aux neurones. Pour expliciter ceci, l'article de Schuchhardt et collabarateurs servira d'exemple (pour la problématique), la

figure 7.3 récapitule les étapes décrites au paragraphe 2.3.2.3.

La figure 7.4 explicite les différentes étapes pour un exemple de 16 blocs :

i)- Tirage au sort d'un fragment (vecteur d'angles dièdres)

ii)- Comparaison de ce fragment avec tous les neurones. Pour connaître le neurone le plus proche du fragment présenté, nous avons cherché le minimum de différence entre le fragment et les neurones de la carte :

$$RMSda(w^{next}, v) = \textbf{minimum}$$

avec w^{next} le vecteur de poids le plus proche de v le fragment présenté

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

avec s et t représentent deux motifs structuraux distincts. Les différences angulaires se font sur 180^o au maximum.

iii)- Le neurone le plus proche (valeur de rmsda la plus faible) est modifié très légèrement pour ressembler au fragment qui lui est présenté. Les poids (vecteur de 8 coordonnées pour chaque neurone) sont pondérés à chaque présentation de fragments :

$$w^k(t+1) = w^k(t) + [v - w^k(t+1)]\nu e^{-\frac{1}{2\rho^2}(r^k - r^{next})}$$

avec ,

$$\nu = \frac{\nu_0}{1 + \frac{t}{\theta}}$$

$$\rho = \frac{\rho_0}{1 + \frac{t}{\theta}}$$

Avec θ le nombre total de motifs à passer, t le nombre de motifs déjà passés, ν le coefficient d'apprentissage pris classiquement entre 0,01 et 0,02 ρ est un coefficient qui définit la distance de propagation (cf. iii bis) autour du neurone le plus proche wnext et r l'amplitude des modifications autorisées.

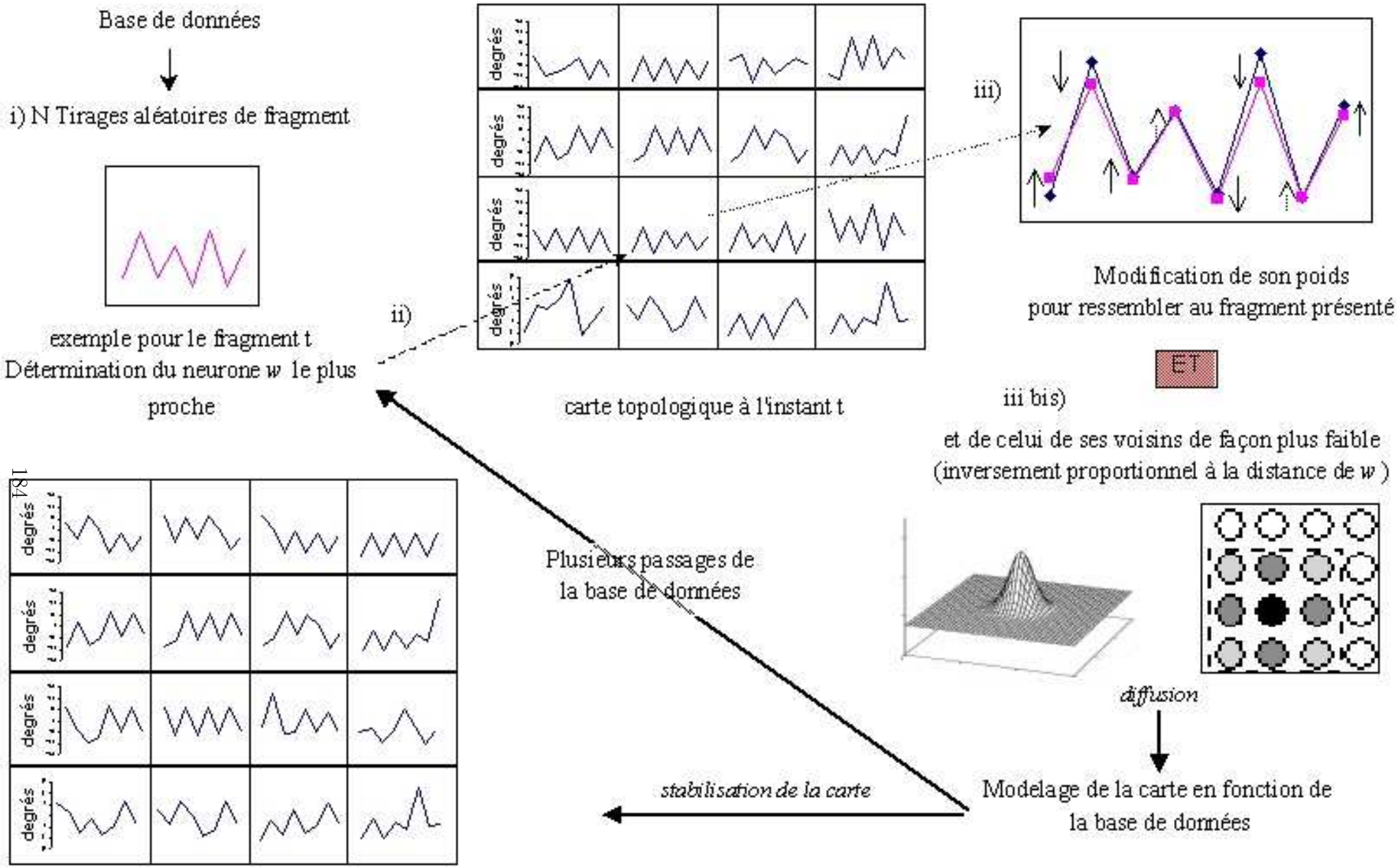

FIG. 7.4 – *Principe de l'apprentissage d'une SOM. (i) Un fragment est tiré aléatoirement dans la base de donnée, (ii) il est comparé à chaque neurone de la carte topologique, (iii) le neurone vainqueur est sélectionné et modifié, (iii bis) les neurones voisins de ce neurone sont modifiés plus faiblement, le processus est répété jusqu'à stabilisation du systéme.*

iii bis)- De même les voisins les plus proches seront très légèrement modifiés, la variation dépend la distance $r^k - r^{next}$ qui a été calculée avec la formule de la distance euclidienne

iv)- Le processus recommence depuis le i)- jusqu'à la stabilisation du système. La valeur des modifications diminuera donc avec le temps. La carte se déformera peu à peu pour atteindre un équilibre.

Les paramètres d'apprentissage sont capitaux dans l'apprentissage ainsi la figure 7.5 montre l'évolution des paramètres . Cette influence est fortement perceptible, des coefficients trop faible tendent à dimininuer les mouvements et provoquent une fixation des neurones trop rapides, d'où une mauvaise classification.

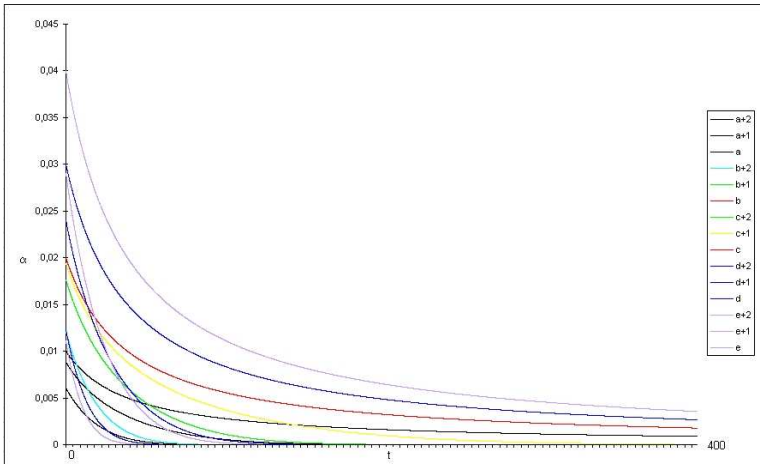

FIG. 7.5 – *Evolutioon des coefficients en fonction des cycles avec diffŕents coeffcients ν et ρ classiques.*

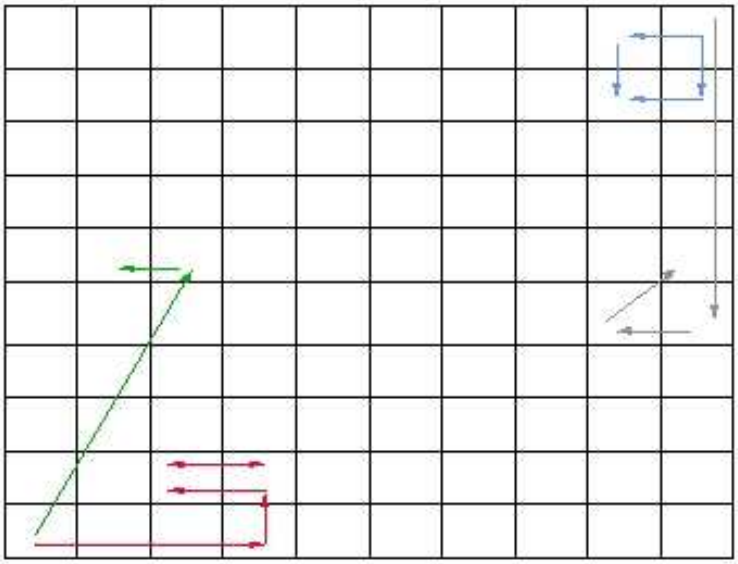

FIG. 7.6 – *Mouvement lors des premiers cycles des neurones les plus proches des structures α et β pour des valeurs d'apprentissage et de diffusion distincte.*

La figure 7.6 montre les différences de mouvement impliqués par des coefficients distincts. Sur l'exemple, le neurone de l'hélice α central a été placée en bas à droite, celui du feuillet en haut à gauche. Selon les coefficients utilisés soit les mouvements sont importants et l'apprentissage semble peu dépendant de l'initialisation, soit ils sont faibles et dépendent fortement de l'intialisation. Ainsi, le neurone β peut ne pas se déplacer et donc se focaliser de maniére "injuste", génant les neurones qui lui sont proches.

k-means La méthode des k-means est proche des cartes auto-organisées de Kohonen [108], mais ne possède aucun processus de diffusion. 5 étapes principales peuvent-être définies :

(1) Il faut définir le nombre b de groupe,

(2) Une observation est associée à chaque groupe de manière aléatoire. et devient le *centre* du groupe.

(3) Le processus dynamique peut alors commencer, il consiste à associer chaque observation de la base de donnée au groupe dont elle est le plus proche, sa distance est minimale avec le centre de ce groupe.

(4) Quand toutes les observations sont assignées à un groupe, Chaque centre est recalculé comme étant la moyenne, le barycentre des observations associées au groupe.

(5) Et le processus recommence depuis l'étape (3) jusqu'à stabilisation du système.

Annexe 2 : Articles

de Brevern AG, Etchebest C, et Hazout, S (2000),"Bayesian probabilistic approach for prediction backbone structures in terms of protein blocks", *Proteins : Structure, Fucntions and Genetics*, 41(3), pp.271-287 [35].

abstract Using an unsupervised cluster analyser, we have identified a local structural alphabet composed of 16 folding patterns of five consecutive C_α ("protein blocks"). The dependence that exists between successive blocks is explicitly taken into account. A Bayesian approach based on the relation protein block-amino acid propensity is used for prediction and leads to a success rate close to 35%. Sharing sequence windows associated with certain blocks into "sequence families" improves the prediction accuracy by 6%. This prediction accuracy exceeds 75% when keeping the first four predicted protein blocks at each site of the protein.

de Brevern AG, et Hazout S (2000),"Hybrid Protein Model (HPM): a method to compact protein 3D-structures information and physicochemical properties", *IEEE - Computer Society : Proceedings of the 7th Symposium on String Processing and Information Retrieval*, pp.49-54 [36].

abstract The transformation of protein 1D-sequence to protein 3D-structure is one of the main difficulties of the structural biology. A structural alphabet had been previously defined from dihedral angles describing the protein backbone as structural information by using an unsupervised classifier. The 16 Protein Blocks (PBs), basis element of the structural alphabet, allows a correct 3D structure approximation [35]. Local prediction had been estimated by a Bayesian approach and shown that sequence information induces strongly the local fold, but stays coarse (prediction rate of 40.7% with one PB, 75.8% with the four most probable PBs).

The Hybrid Protein Model presented in this study learns both sequence and structure of the proteins. The analysis made along the hybrid protein has permitted to appreciate more precisely the spatial location of some types of amino acid residues in the secondary structures and their flanking regions. This study leads to a fuzzy model of dependence between sequence and structure.

de Brevern AG, et Hazout S (2001), "Compacting local protein folds with a Hybrid Protein", *Theoretical Chemistry Accounts*, *sous presse* [37].

abstract The "Hybrid Protein Model" (HPM) is a fuzzy model for compacting local protein structures. It learns a non-redundant database encoded in a previously defined structural alphabet composed of 16 protein blocks (PBs) [35]. The hybrid protein is composed of a series of distributions of the probability of observing the PBs. The training is an iterative unsupervised process that for every fold to be learnt consists of looking for the most similar pattern present in the hybrid protein and modifying it slightly. Finally each position of the hybrid protein corresponds to a set of similar local structures. Superimposing those local structures yields an average root mean square of 3.14 Å. The significant amino acid characteristics related to the local structures are determined. The use of this model is illustrated by finding the most similar folds between two cytochromes P450.

Camproux AC, de Brevern AG, Tufféry P, et Hazout S, "Exploring the use of a structural alphabet for a structural prediction of protein loops", *Theoretical Chemistry Accounts*, *sous presse* [21].

abstract The prediction of loop conformations is one of the challenging problems of homology modeling, due to the large sequence variability associated with these parts of protein structures. In the present study, we introduce a search procedure that evolves in a structural alphabet space deduced from a hidden Markov model to simplify the structural information. It uses a Bayesian criterion to predict, from the amino acid sequence of a loop region, its corresponding word in the structural alphabet space.

Results show, that our approach ranks 30% of the target words with the best score, 50% within the 5 best scores. Interestingly, our approach is also suited to accept or not the prediction performed. This allows to rank 57% of the target words with the best score, 67% within the 5 best scores, accepting 16% of learned words and rejecting 93% of unknown words.

Annexe 3 : Les I-sites

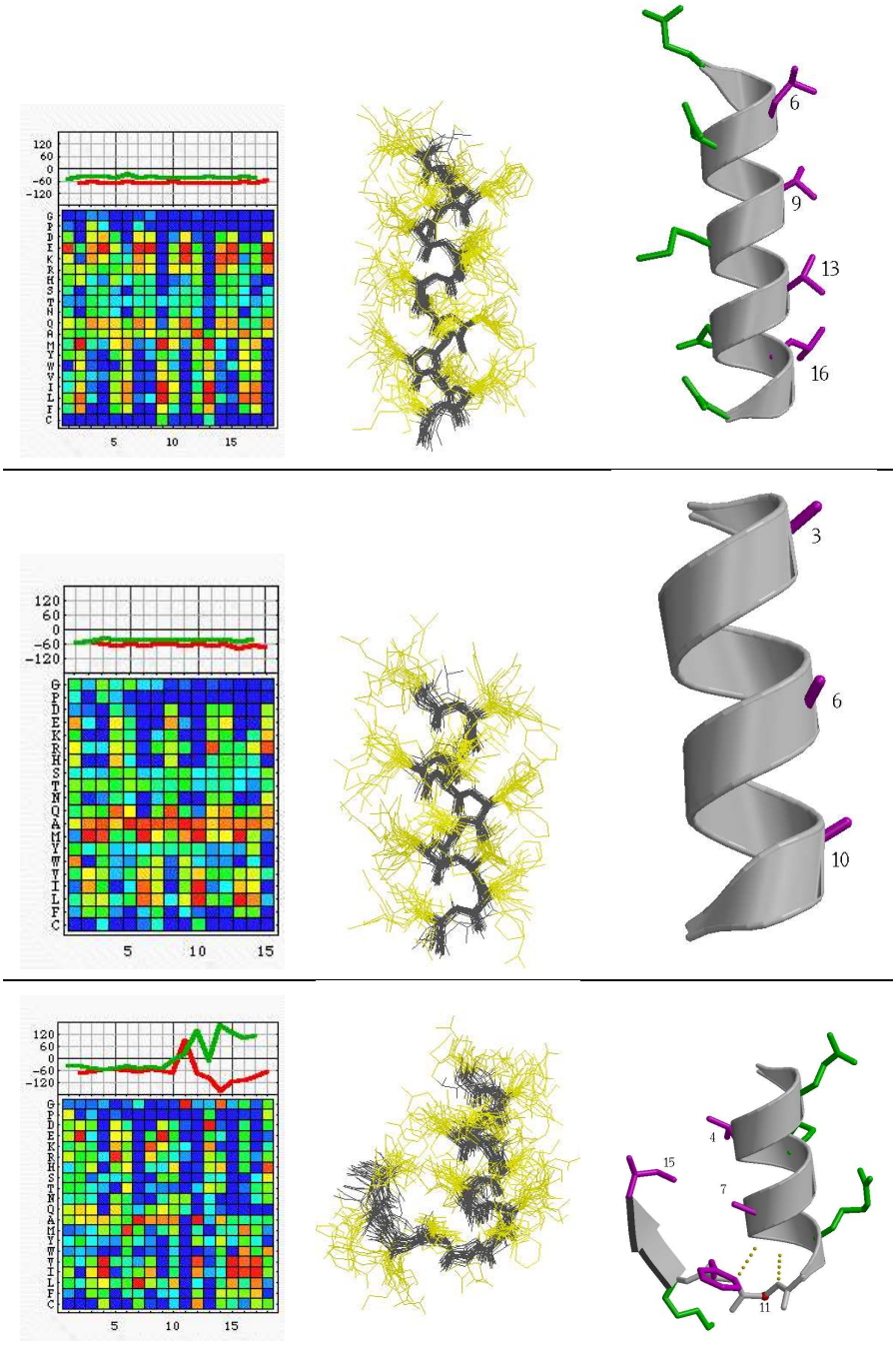

FIG. 7.7 – *I-sites 1 à 3. 1- hélice α amphipatique. 2- hélice α non polaire. 3- extrémité C-terminale d'hélice α Glycine Type 1. (la légende détaillée est sous la figure 7.11).*

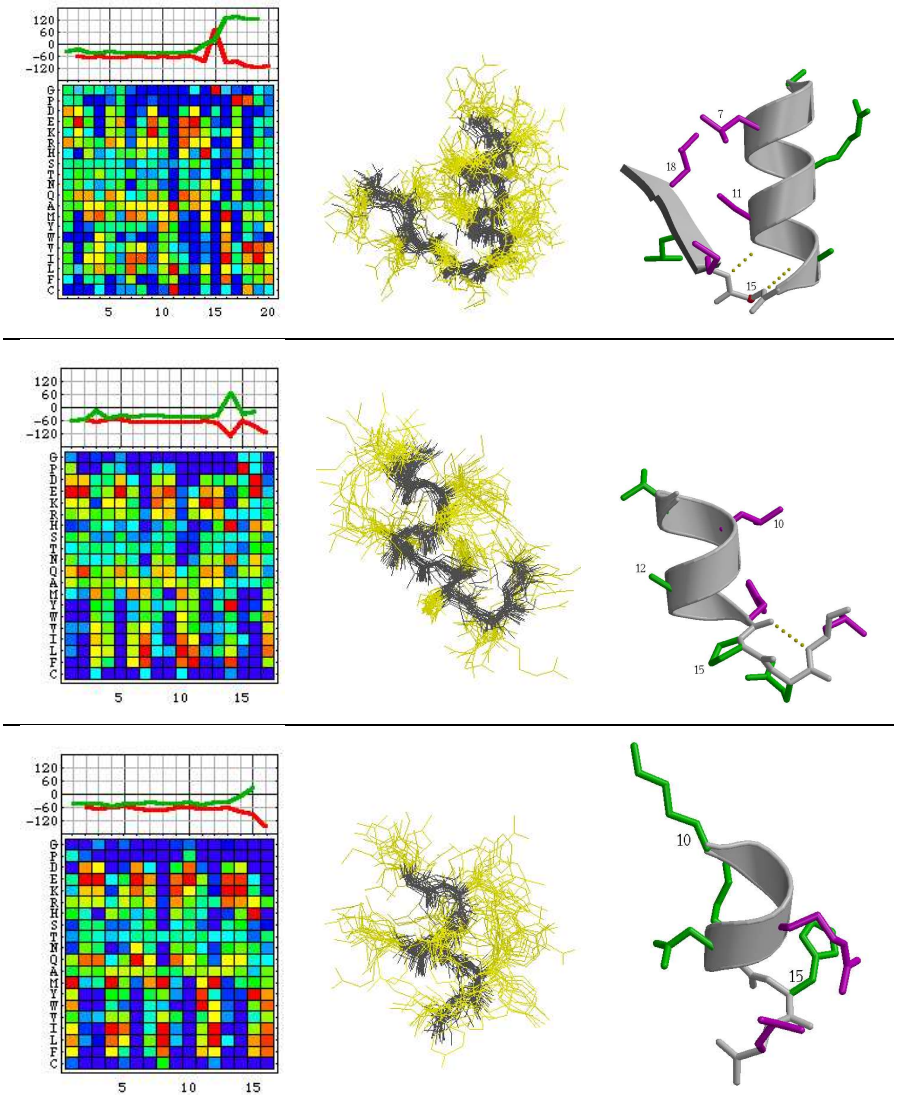

FIG. 7.8 – *I-sites 4 à 6. 4- extrémité C-terminale d'hélice α Glycine Type 2. 5- extrémité C-terminale d'hélice α Proline. 6- hélice α mélée. (la légende détaillée est sous la figure 7.11).*

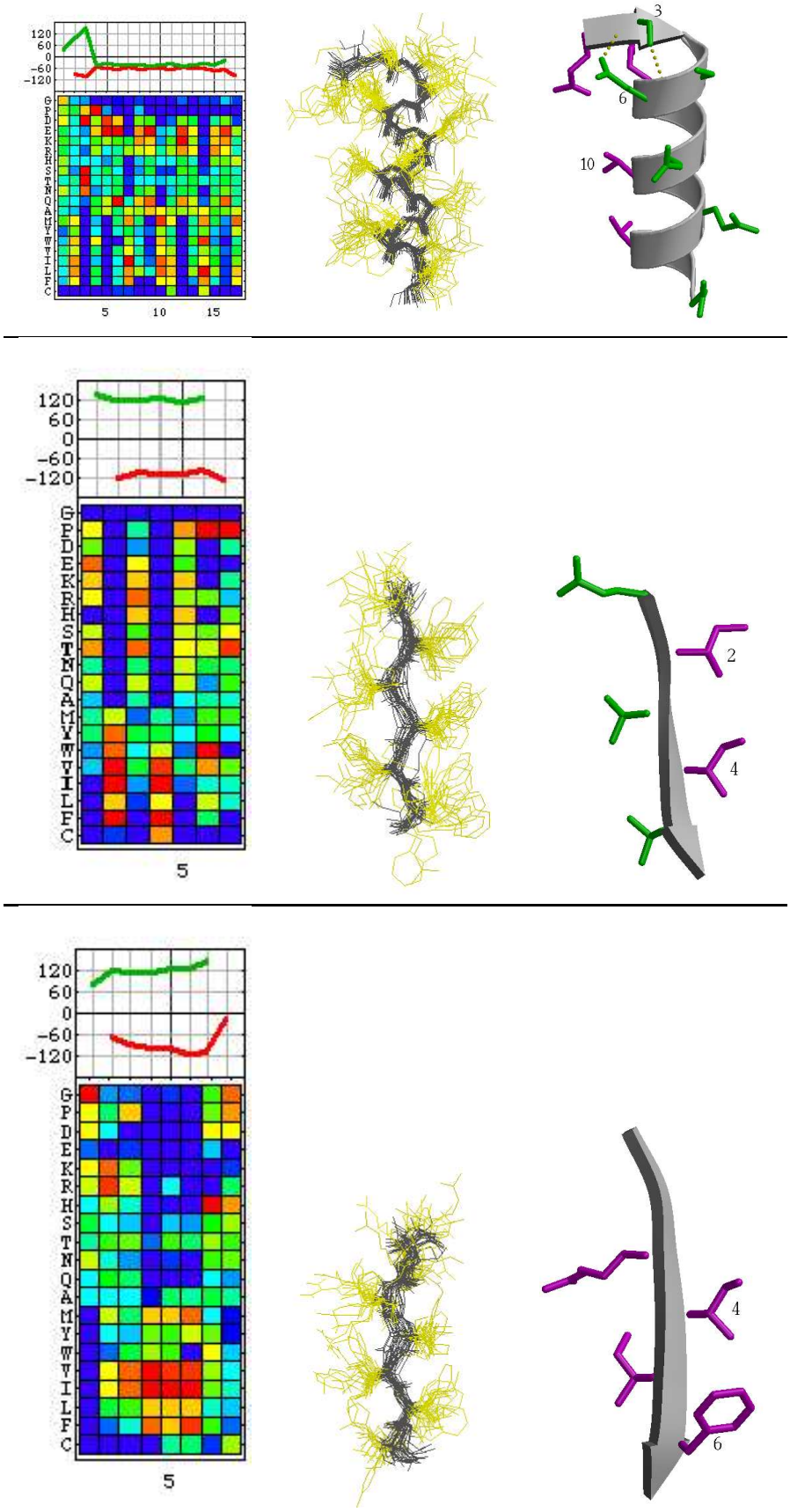

FIG. 7.9 – *I-sites 7 à 9. 7- extrémité N-terminale d'hélice α Sérine. 8- feuillet β amphipatique. 9- feuillet β hydrophobe. (la légende détaillée est sous la figure 7.11).*

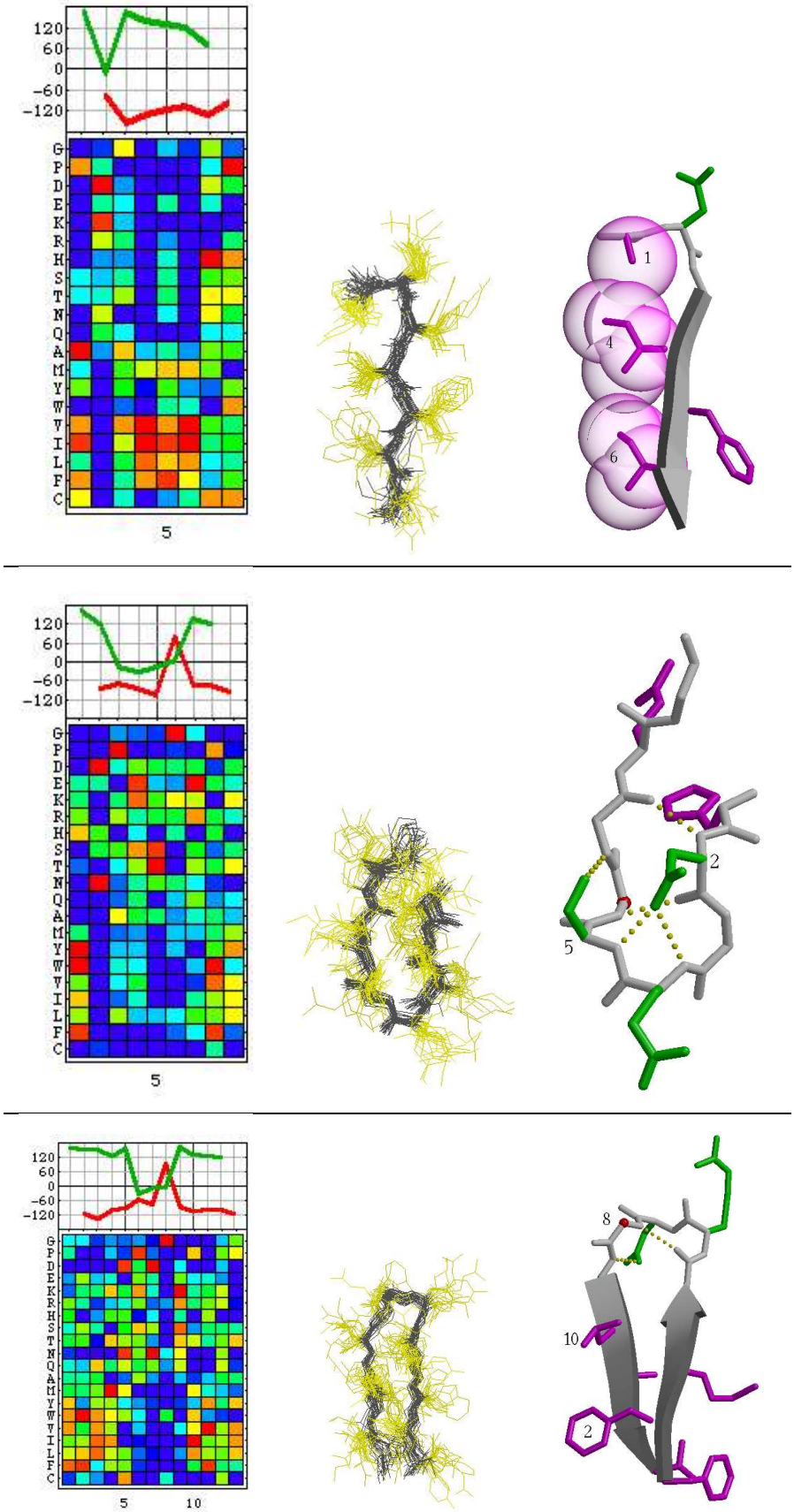

FIG. 7.10 – *I-sites 10 à 12. 10- coude β Aspartate. 11- épingle β Sérine. 12- épingle type I étendu. (la légende détaillée est sous la figure 7.11).*

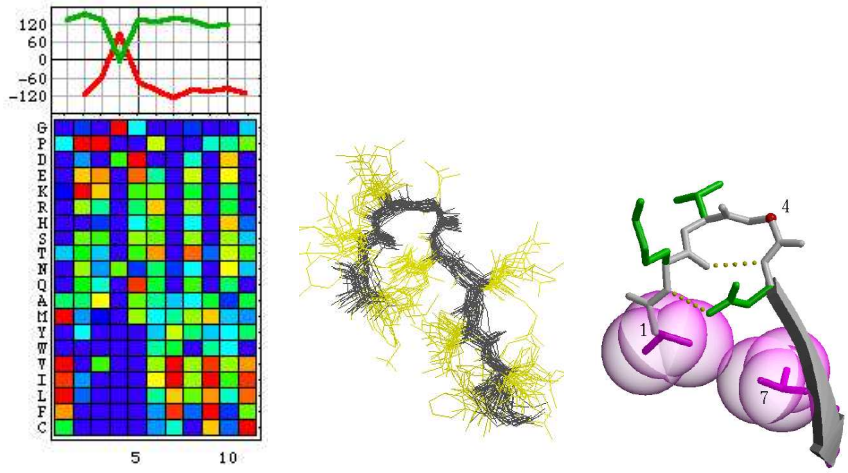

FIG. 7.11 – *I-site 13 coude type II divergeant. underlinelégende: à gauche les angles du squelette polypeptidique (en rouge l'angle ϕ, en bleu l'angle ψ) et en dessous la matrice de contingence du I-site coloré selon l'occurrence de la fréquence de l'acide aminé (en rouge, 3 fois plus que la moyenne de l'acide aminé dans la base de donnée, orange, entre 2 et 3 fois plus, jaune, entre 1 et 2 fois plus, vert, équivalent à la base de donnée, cyan, entre 1 et 2 fois moins, bleu entre 2 et 3 fois moins, bleu marine, plus de 3 fois moins que dans la base de donnée), l'ordre des acides aminés est le suivant en partant du haut: G, P, D, E, K, R, H, S, T, N, Q, A, M, Y, W, V, I, L, R, C; au centre, la superposition des 30 meilleurs représentant du I-site et à droite, une représentation "cartoon" des chaînes latérales (en vert, les polaires, en pourpre, les non-polaires, les glycines sont en rouge).Ces images viennent du site des I-sites http://isites.bio.rpi.edu/Isites/index.html*