Chapitre 5

Dépendance entre les blocs structuraux protéiques

5.1 Objectif

Après avoir développé une méthode d'apprentissage de blocs protéiques (cf. paragraphe 3, *"Apprentissage de la structure locale du squelette protéique"*), nous avons utilisé ces blocs dans une méthode bayésienne de prédiction locale de ces blocs à partir de la séquence (cf. paragraphe 4, *"Prédiction de la structure locale en blocs protéiques"*). Conserver seize PBs au final, permet d'obtenir des BPs purement boucles ainsi que plusieurs types d'extrémités N- et C-terminales des structures périodiques.

Toutefois, cette approche comme beaucoup d'autres approches locales ne ne prend pas implicitement en compte le problème de la continuité implicite de la structure protéique. Ce problème a été déjà soulevé avec la prédiction des structures secondaires [210]. Le réseau neuronal PHD [166] le résout en traitant les résultats de sa prédiction par un second réseau. Actuellement, seules les Chaînes de Markov Cachées [149] utilise directement cette information, comme HMMSTR, pour les structures secondaires, développé par le groupe de Baker [20] et pour les alphabets structuraux par Camproux et collaborateurs [23, 24].

Pour prendre en compte cette continuité, nous avons décidé d'analyser les séries de BPs ayant les occurrences les plus fréquentes, par une approche qui présente certaine similitude avec le travail de Fetrow et collaborateurs [53]. Les séries de blocs ont été établies en tenant en compte des transitions préférentielles. Ces successions de blocs les plus fréquentes se suivent de manière particulièrement régulière. Les chevauchements sont nombreux.

Nous proposons donc de construire un réseau simple formé par la succession de ces mots.

Ce travail est à rapprocher d'autres descriptions des structures par des graphes pour trouver des fragments protéique structuralement proches [170], des groupes de chaînes latérales [101], des domaines protéiques [201], mais surtout des "topologies proches" dans les protéines, en d'autres termes des fragments protéiques de tailles variables ayant des repliements locaux similaires [106, 191], nous regarderons ensuite si la construction d'un tel réseau permet une bonne conservation structurale des fragments protéiques et enfin si la présence de certains acides aminés est dépendante des voies différentes du réseau. Cette approche est intéressante car au lieu de se focaliser sur des zones de longueur constante entre des structures secondaires [203, 204], des zones plus ou moins étendues seront obtenues. Ainsi, les séries de blocs permettent d'observer et d'analyser de manière fine la structure 3D sans avoir les limites liées au choix de longueur fixe.

5.2 Conception du graphe

5.2.1 Construction du réseau

Nous nous sommes intéressés aux mots (i.e. *succession*) de 5 blocs protéiques, ce qui représente 9 C_α soit la taille moyenne d'un feuillet β avec entrée et sortie complète. La base de données est celle utilisée auparavant (342 protéines ayant moins de 25% d'identité de séquence).

L'ensemble des structures protéiques a été recodée en BPs selon la règle du *RMSda* minimal. Le réseau que l'on désire construire est un graphe orienté. Un graphe **G** correspond à un ensemble **T** de nœuds ("vertices") et d'un ensemble **E** de segments ("edges") qui les relient. Le graphe est orienté si chaque segment possède une seule direction, **G**(V,E). Dans notre étude, chaque nœud de l'ensemble **V** est caractérisé par un bloc protéique, et chaque lien orienté une transition entre deux blocs protéiques. L'objectif est de caractériser le graphe orienté des séries de blocs protéiques les plus fréquemment observés dans une base de structures protéiques. Pour réaliser cet objectif, les séries de 5 blocs protéiques les plus fréquents (fréquence supérieure à 150 observations soit 0,18 % de la base étudiée) ont été sélectionnés, puis en se basant sur un principe de séquentialité, le graphe orienté a été construit. Un motif de 5 blocs est représenté par un sous-graphe orienté. Par exemple, le motif *mnopac* est décrit par le sous-graphe m $\rightarrow$ n $\rightarrow$ o $\rightarrow$ p $\rightarrow$ a $\rightarrow$c.

Le principe de séquentialité consiste à trouver dans la liste des motifs ceux dont les 4 derniers blocs d'une série se retrouvent dans les 4 premières d'une autre série, ($a_1 \rightarrow a_2 \rightarrow a_3 \rightarrow a_4 \rightarrow a_5$) et ($a_2 \rightarrow a_3 \rightarrow a_4 \rightarrow a_5 \rightarrow a_6$) devient ($a_1 \rightarrow a_2 \rightarrow a_3 \rightarrow a_4 \rightarrow a_5 \rightarrow a_6$). Par exemple, le motif *mnopa* est suivi par le motif *nopac* par continuité. Par superposition des sous-graphes, le graphe est étendu. Le sous-graphe devient dans ce cas m→ n→ o→ p→ a→c. Le graphe peut présenter des bifurcations, une série étant suivie par plusieurs séries distinctes, ainsi, le motif *mnopa* est suivi par deux motifs *nopac* et *nopaf* dans des proportions respectives de 68 et 32 %. Les structures secondaires répétitives hélices et feuillets sont considérées chacune comme un seul nœud retournant sur lui même ainsi la succession *fklmmmmmmmmmmnop* se traduira par la succession des nœuds *fkl(m*)nop*. La répétitivité des blocs sera notée par une étoile. Cette méthode va nous permettre de bien mettre en relief des transitions à plus longue distance.

5.2.2 Qualité du réseau

Pour s'assurer de la stabilité structurale du réseau, un ensemble de N motifs qui permettent de recouvrir le graphe totalement ont été extraits. Les *RMSds* des fragments extraits de la base de données associés à ces mots ont été calculés pour évaluer leurs sensibilités structurales.

5.3 Résultats

5.3.1 Le réseau

Sur l'ensemble des 342 structures protéiques recodées, 72 motifs de taille 5 ont été obtenus. Leurs fréquences varient entre 17,3 % (14729 observations) et 0,18 % (152 observations). Les *RMsd* ont été calculés pour l'ensemble de ces motifs, appellés mots structuraux (les valeurs et la description de ses mots sont disponibles sur le site web de l'Equipe de Bioinformatique Génomique et Moléculaire). Les motifs les plus fréquents sont associés aux formes des structures secondaires (*mmmmm* et *ddddd*). Les 14 premiers ont servi à mettre en place le réseau et 44 des 58 suivants ont servi ŀe compléter. La figure 5.1 montre le réseau obtenu. Ce réseau est ainsi composé de 31 nœuds et utilise 15 des 16 BPs. Les couleurs représentent les occurrences observées, ainsi l'entrée principale en hélice α *fkl* (nœuds *08-09-10*) vers le BP *m* (nœud *01*) a une sortie principale *nop* (nœuds *02-03-04*) et deux sorties plus courtes (et moins observées) *pcc*

(nœuds *20-21-22*) et *c* (*23*). La partie entourant le site du BP *d* (nœud *07*) est plus complexe, l'entrée s'effectue par un seul nœud *c* (nœud *06*), celui-ci étant compris dans une succession *nopac* (nœuds *02-03-04-05-06*) ou *kbc* (nœuds *31-27-06*), sa sortie la plus importante est la série *dehiacd* (nœuds *07-25-16-17-18-19-07*). La seconde sortie du feuillet la plus importante est la succession *dfkl* (nœuds *07-08-09-10*). Ensuite la sortie plus complexe de *df* (nœuds *07-13*) allant vers les sites *b* (nœud *14*) et *k* (nœud *31*).

Des parties moins observées ont était mises comme le dédoublement du BP *e* (nœud *26-15*) en amont du site *h* (nœud *16*), la sortie en *j* (nœud *25*) de ce dernier, la suite alternative de *fkl* (nœuds *08-09-10*) vers *pc* (nœuds *11-12*) ou les deux passages de *dfk* (nœuds *07-13-31*) vers *bc* (nœuds *27-06*) ou *op* (nœuds *03-04*). Le réseau final obtenu est visible dans un plan 2D sans aucune intersection des flèches. Le caractère discontinu du réseau doit être noté, ainsi tous les fragments conservés sont ceux qui peuvent s'inscrire dans le réseau, c'est à dire s'ils ont 5 sites consécutifs. La figure 5.2 donne les fréquences de début et de fin des motifs observés sur le réseau, cad la fréquence relative de chaque nœud pour être le premier ou le dernier résidu d'un fragment protéique compris dans le graphe. Seul le site *c* (nœud *06*) situé juste avant le bloc *d* (feuillet β, nœud *07*) possède des taux largement supérieurs aux autres sites. Ce fait s'explique à la fois par des différences de tailles et de formes de structures associées aux feuillets β et surtout au principe même du graphe où 3 des 4 terminaisons sont des nœuds correspondant au PB *c*: en fin d'hélice α nœuds *22* et *23* des séries *mc* et *mpcc* (en haut à gauche de la figure 5.1) et le nœud *12* de la série *lpc* (en bas du graphe).

5.3.2 Exemple

Pour mieux appréhender cette notion de réseau non continu, la figure 5.3 représente la partie centrale de l'endoribonuclease du virus sarcome aviaire (code PDB: 1vsd) avec sa structure protéique recodée en blocs protéiques et en-dessous les différentes parties du réseau alors utilisées. Trois zones appelées 'fragment' se trouvent incluses dans le réseau. La figure 5.3 montre cet exemple avec en haut la séquence en BPs, et, en-dessous les trois fragments avec les zones du graphe correspondantes. Le fragment **I** est composé de 7 blocs protéiques, c'est un court feuillet *bcddfbf* (partie supérieure droite du réseau). Le fragment suivant **II** se superpose au dernier bloc protéique puis repart vers une forme α entrant par la successions de blocs *fkl* et en res-

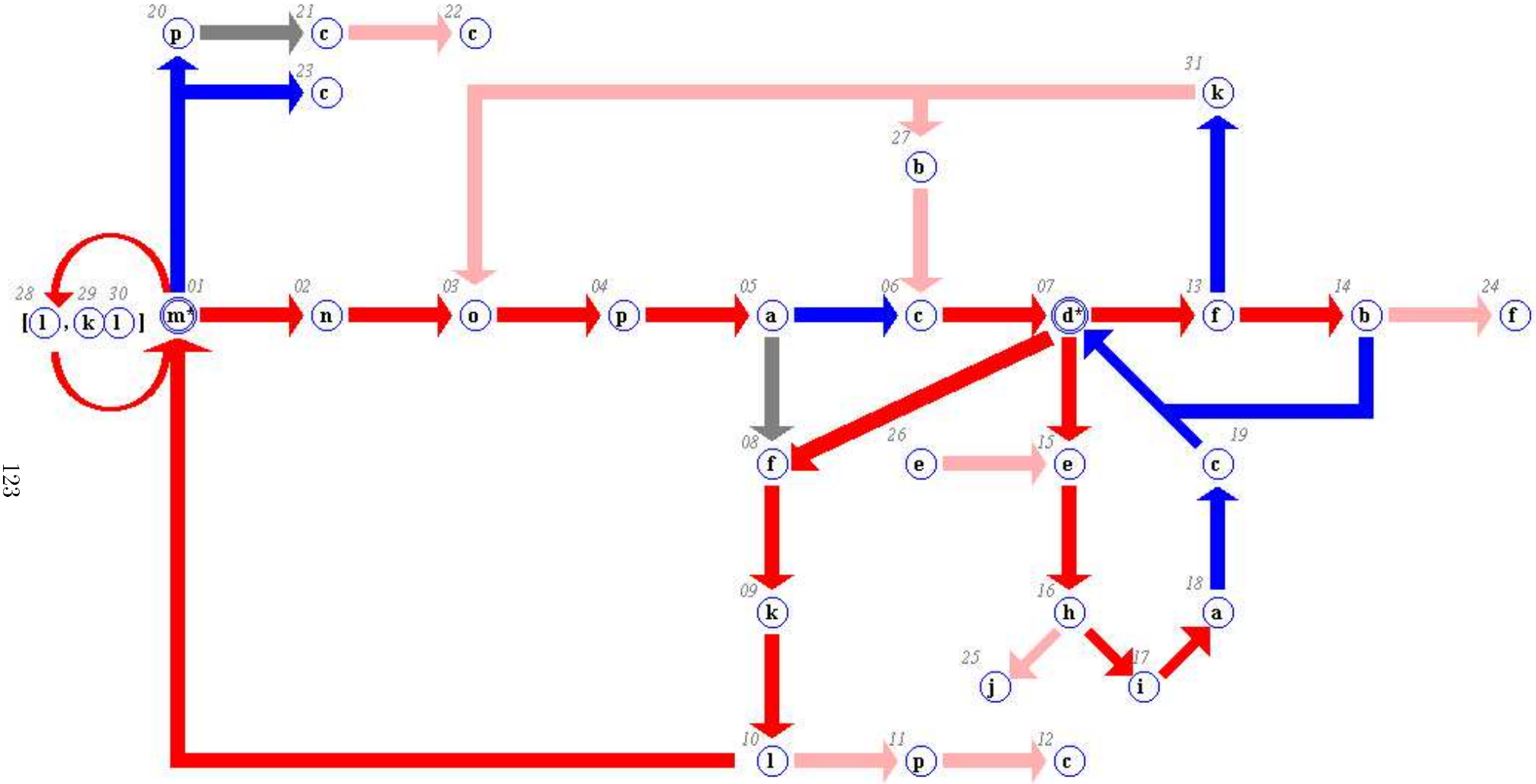

FIG. 5.1 – *Le réseau obtenu contient 31 nœuds, avec au-dessus le numéro du nœud, les couleurs dépendent du nombre d'occurrence. En rouge, des occurrences > 800 observations (10 %), en bleu entre 500 (6 %) et 800 (10 %), en gris entre 500 (6 %) et 300 (3,5 %), enfin en rose en-dessous de 300 (3,5 %). Les sites m* et b* sont des répétitions des blocs protéiques m et b. A gauche, la flèche allant de m vers le site l (nœud 28) et les sites kl (nœuds 29-30) représente des successions mlm et mklm. Le réseau étant discontinu, les sites c dans les séries mpcc, mc, lpc, le site j de la suite hj et le site b de bf sont des terminaisons du graphe (voir figure 5.2).*

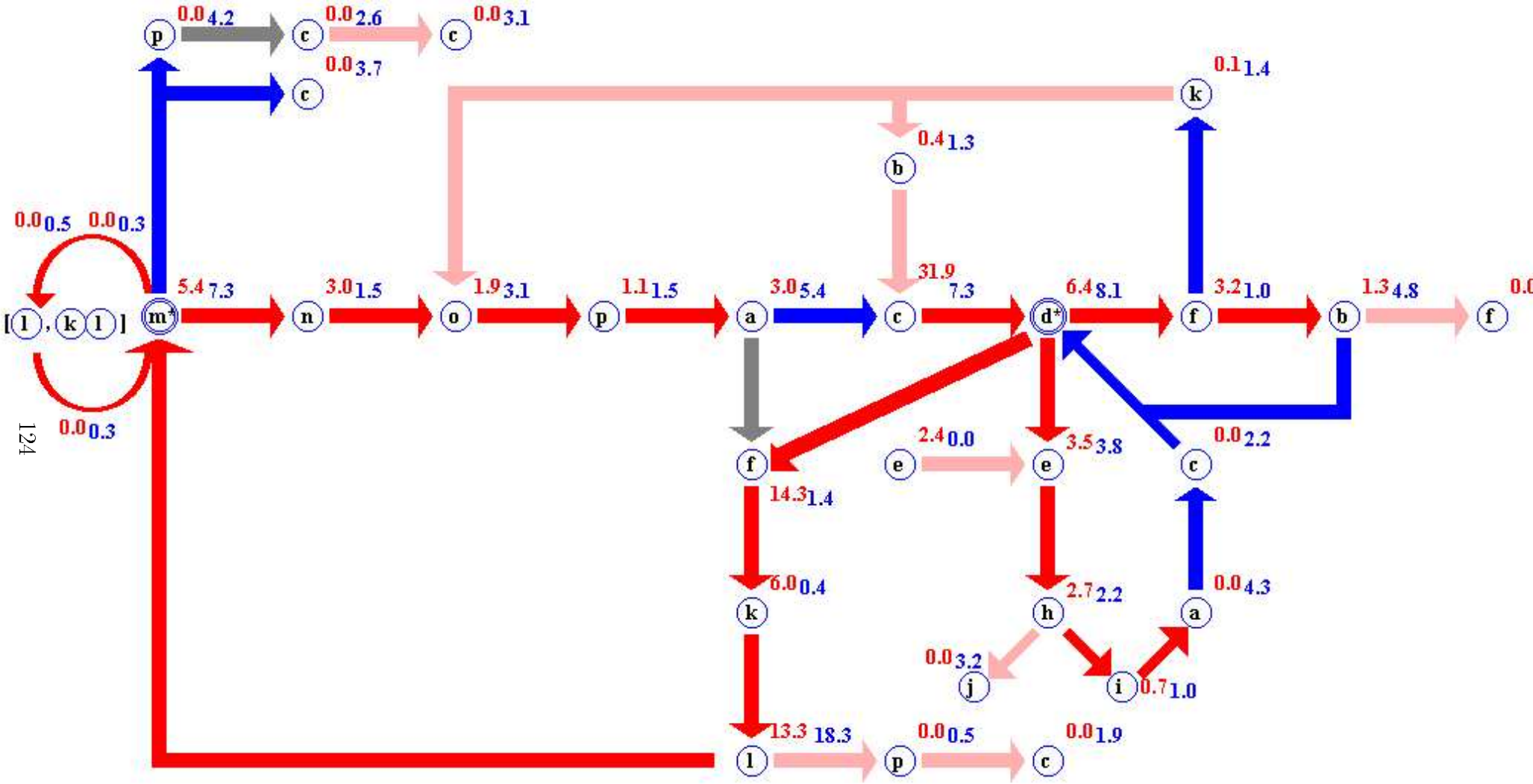

FIG. 5.2 – *Les pourcentages de début et de fin pour les 31 nœuds du réseau. Les fréquences de début sont situés au-dessus des nœud en rouge, et celles de fin juste en-dessous des précédentes en bleues.*

sortant par la série de blocs *nopacde* (partie inférieure gauche du réseau, puis partie centrale). Ensuite une zone composée par une série de blocs *bjghehpap* est non observée parmi les motifs les plus fréquents et donc n'a pas de contre-partie dans le réseau. Enfin le fragment **III** est une longue hélice qui est composé de 19 blocs protéiques *m* consécutifs (partie inférieure gauche du réseau).

5.3.3 Analyse du réseau

Ayant recodé l'ensemble des protéines à l'aide de l'alphabet structural, plusieurs questions se posent quand à la pertinence et à la stabilité structurale du réseau : (1) Quelle est la part des parties non-répétitives dans le réseau? (2) Est-ce que les fragments protéiques contenus au même site du réseau sont bien structuralement identiques?

Pour répondre à la première question, l'ensemble des protéines a été suivi sur le graphe, comme pour l'exemple (cf. figure 5.3) et les taux des résidus appartenant au réseau a été calculé. Avec cette métrique (un minimum de 5 BPs consécutifs compris dans le réseau) et le réseau proposé, plus de 83 % des acides aminés des protéines est inclus dans le graphe, mais le point le plus intéressant n'est pas que les structures répétitives secondaires soient bien retrouvées (96,14 % pour les hélices et 89,11 % pour les feuillets) mais que 70,58 % des boucles soient aussi présentes.

Le premier point ayant été validé, il faut se pencher sur le second car, en effet, l'alphabet structural a été construit pour approximer localement la structure tridimensionnelle sur 5 C_α et le réseau utilise des successions chevauchantes de 5 BPs (soit 9 C_α), il convient donc de vérifier que ces approximations locales n'ont pas généré des fragments protéiques trop distincts.

5.3.4 Extraction des mots et stabilité 3D

17 mots de tailles 4 à 7 BPs (i.e. 8 à 11 C_α) ont été extraits du graphe, ils permettent de le recouvrir intégralement. Le *RMSd* moyen a été calculé sur l'ensemble des fragments protéiques correspondants dans la base de données. Celui des mots structuraux a été effectué préalablement. Le tableau 5.1 récapitule les mots observés et les *RMSds* associés. Dans une seconde étape, nous avons regardé s'il n'était pas possible dans ces mots de discerner des sous-familles structurales, cad des repliements distincts, mais associés aux mêmes successions de

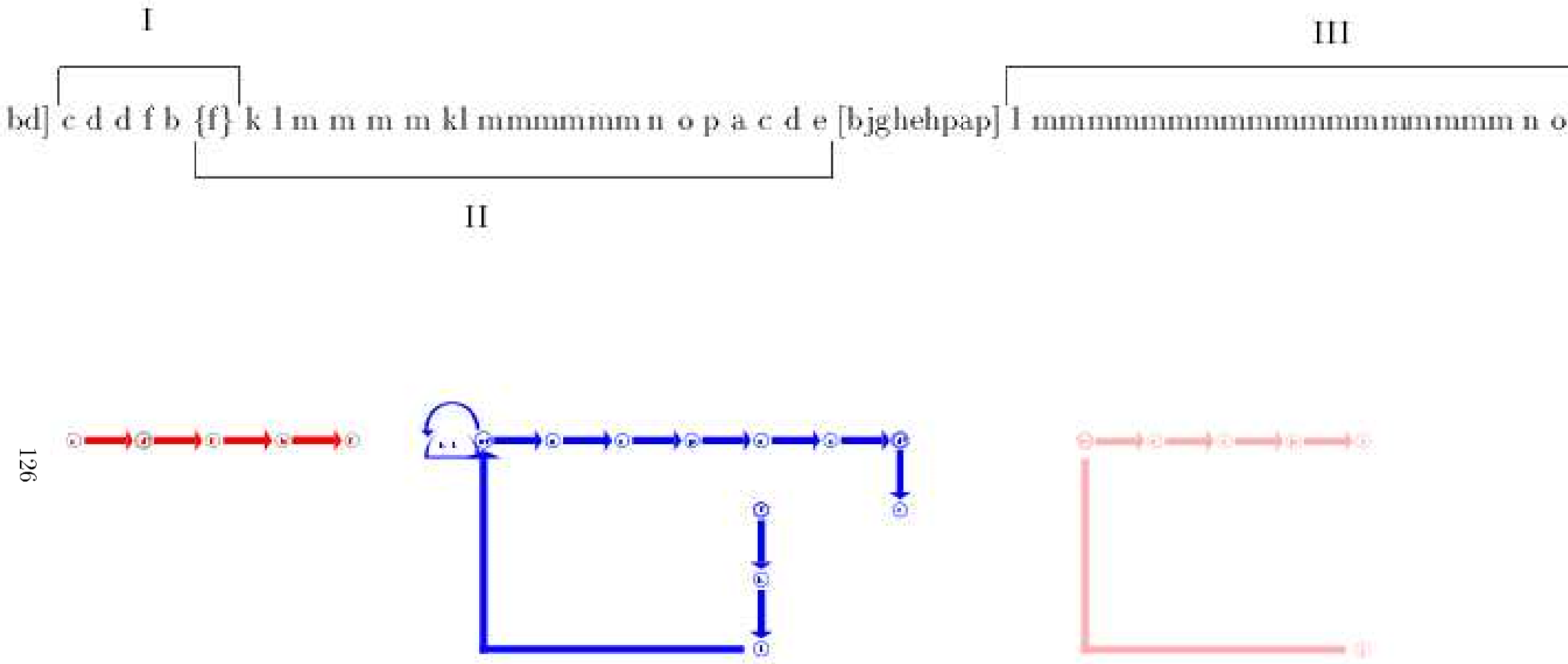

FIG. 5.3 – *Exemple de recodage d'une séquence protéique (virus sarcome aviaire, code PDB: 1vsd) par le réseau avec en haut la structure protéique recodée en blocs et en dessous les différentes parties du réseau alors impliquées pour les trois fragments* ***I****,* ***II*** *et* ***III****.*

num	motif	BPs	C_α	*RMSd*m *Å*	nœuds
1	mnopacd	7	11	1,94	01 - 02 - 03 - 04 - 05 - 06 - 07
2	mnopa	5	9	1,76	01 - 02 - 03 - 04 - 05
3	ehiac	5	9	2,68	15 - 16 - 17 - 18 - 19
4	dehiacd	7	11	3,07	07 - 15 - 16 - 17 - 18 - 19 - 07
5	dfbf	4	8	1,74	07 - 13 - 14 - 24
6	dfkopa	6	10	1,37	07 - 13 - 31 - 03 - 04 - 05
7	fklm	4	8	1,69	08 - 09 - 10 - 01
8	afkl	4	8	1,81	05 - 08 - 09 - 10
9	dfbd	4	8	2,02	07 - 13 - 14 - 07
10	mklmm	5	9	1,49	01 - 29 - 30 - 01 - 01
11	mlmm	4	8	1,42	01 - 28 - 01 - 01
12	mpcc	4	8	1,74	01 - 20 - 21 - 22
13	fklpc	5	9	1,93	08 - 09 - 10 - 11 -12
14	eehiac	6	10	1,34	26 - 15 - 16 - 17 - 18 - 19
15	dehj	4	8	1,98	07 - 15 - 16 - 25
16	dfkbc	5	9	1,94	07 - 13 - 31 - 27 - 06
17	dfklm	5	9	2,09	07 - 08 - 09 - 10 - 01

TAB. 5.1 – *Pour chaque motif est donné la succession des blocs protéiques compris dans le motif testé, le nombre de Blocs Protéiques les composant et le* RMSd *moyen (en Å) associé.*

BPs.

5.3.4.1 Test de l'homogénéite structurale au sens du *RMSd*

La superposition des C_α des fragments montre une constance remarquable avec un *RMSd* moyen pour la plupart des "mots" aux alentours de 2 Å. La figure 5.4 illustre le cas de la succession de blocs *mnopacd* qui possède un *RMSd* de 1,94 Å pour une longueur de 11 C_α (partie gauche et centrale du réseau, nœuds *01-02-03-04-05-06-07*). Ce ne sont pas obligatoirement des motifs liés à des hélices α qui sont le plus conservés, ainsi *dfkopa* (nœuds *07-13-31-03-04-05*) et *dfbf* (nœuds *07-13-14-24*) ont des *RMSd* de, 1,37 et 1,74 Å, respectivement.

Afin d'évaluer l'homogénéité du groupe de structures associé à un mot, les *RMSd*s entre les couples de fragments ayant été calculés, une matrice de distance a été constituée et a servi lors d'un regroupement par une méthode hiérarchique (paquage *hclust* du logiciel *S-Plus*). Dans chaque arbre obtenu par cette méthode 2 et/ou 3 groupes ont été formés, pour voir, s'ils n'existaient pas dans certaines séries des sous-familles structurales discernables. Ces groupes de fragments ont été donc de nouveaux superposés. Ils s'avérent qu'ils ne sont pas mieux

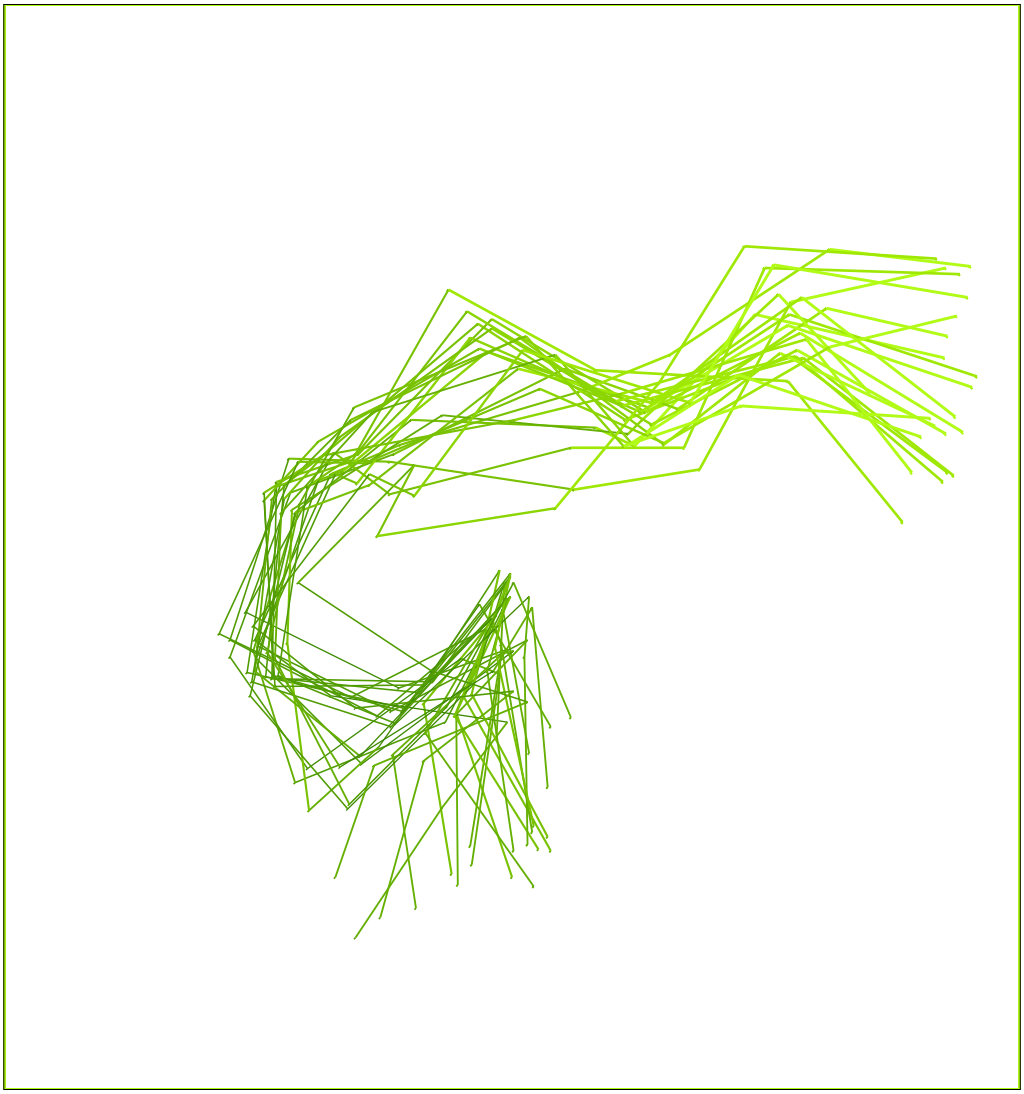

FIG. 5.4 – *Superposition de fragments représentant la succession de blocs mnopacd.*

définis que les motifs globaux et qui ne donnent pas des *RMSd* moyens meilleurs. Seul le motif "long" *dehiacd* (nœuds 07-15-16-17-18-19-07) possède une flexibilité qui semble importante; ses extrémités N- et C- terminales parraissent plus mobiles (distance de 13 Å+/-5 Å), toutefois le repliement interne étant caractéristique, aucun découpage utilisant le *RMSd* n'est possible.

Ainsi, les motifs pris dans le réseau ont une stabilité structurale correcte sans être aussi bonne que dans le cas des BPs seuls (0,46 Å à 1,04 Å pour une longueur de 5 C_α). Le repliement observé apparaît globalement semblable pour des fragments associés à un motif donné.

5.3.4.2 Test de l'homogénéite structurale au sens du RMSda

Les blocs n'ont pas été définis sur la base du *RMSd* mais du *RMSda* (*root mean square on angular values*) qui par définition est beaucoup plus sensible. Les *RMSda* moyens ont été calculés pour les même couples de fragments que pour le *RMSd*. Leurs valeurs moyennes étaient toutes proches de 30^o (avec une distribution gaussienne classique). Il ont donc une approximation exactement similaires à ceux des BPs seuls, ce qui est intéressant car la même précision est conservée pour un nombre d'angles passe de 8 à 14 et jusqu'à 20 angles, la spécificité semble donc assez grande.

Toutefois, il est vrai que le *RMSda* est une mesure plus difficile à apprécier. Une simple superposition des angles permet un aspect plus "visuel", la recherche d'angles précis ayant une variablité plus grande est arbitraire, mais pour la majorité des mots un seul type de signal est visible. Les séries *dehiacd* (nœuds 07 -15-16-17-18-19-07) et *mlmm* (nœuds 01-28-01-01) sont les seules séries ayant un angle qui montre une plus grande variabilité. Cependant en utilisant cet unique critère pour rechercher de possibles sous-familles, les résultats ne sont pas améliorés. En conclusion, sur un plan structural, le réseau permet d'approximer la structure 3D locale de manière assez correcte et ceci même pour des longueurs assez importantes.

5.3.5 Relation avec les acides aminés

D'un point de vue composition en acides aminés, la distribution des Z-scores en certains sites est ainsi beaucoup plus prononcée que pour les BPs seuls. Il a fallu tenir compte du fait qu'un bloc protéique est présent en plusieurs sites et donc que certains sites nettement moins observés que le BP seul ne sont pas statistiquement représentatifs. Certains BPs sont peu influencés d'un

point de vue compositionnel par leur environnement local, d'autres le sont fortement. La figure 5.5 montre un exemple les différentes répartitions en acides aminés pour les nœuds du réseau correspondant aux blocs protéiques *f*, *k*, *l*. Ainsi les nœuds *b*, *14* de la suite *df<u>b</u>f* et *27* de *fk<u>b</u>c*, ont peu de différence entre eux, le premier représentant 4/5 des sites *b* du graphe. L'absence de Proline dans le second nœud, alors qu'elle est fortement sur-representée pour le BP *b*, est compensée par une présence de Glycine, absente elle du BP *b*. Le cas des nœuds *e* est assez similaire pour le nœud *15* de la suite *d<u>e</u>h* (89 %) et *15* de *<u>e</u>eh* (11 %) avec peu de différence avec le BP *e* original, sauf pour la Proline et la Glycine totalement absente du nœud *15* de la série *eeh*. Les nœuds *c 06* de la série *a<u>c</u>d* (68 % des *c* du graphe) et *21* de *mp<u>c</u>* (13 %) ont peu de différence sauf pour des hydrophobes moins présents dans le dernier cas. Les sites *k 09* de la série *f<u>k</u>l* (75 %) et *29* de *m<u>k</u>l* (18 %) ont peu de différence avec le BP *k* : une absence de Thréonine dans le premier et une sur-représentation de Sérine dans le second. Le principal nœud *l 10* de la série *k<u>l</u>m* (85 %) n'apporte que peu de spécificité avec juste une légère augmentation du nombre d'Alanine et une sous-représentation accrue de Glycine par rapport au BP *l*.

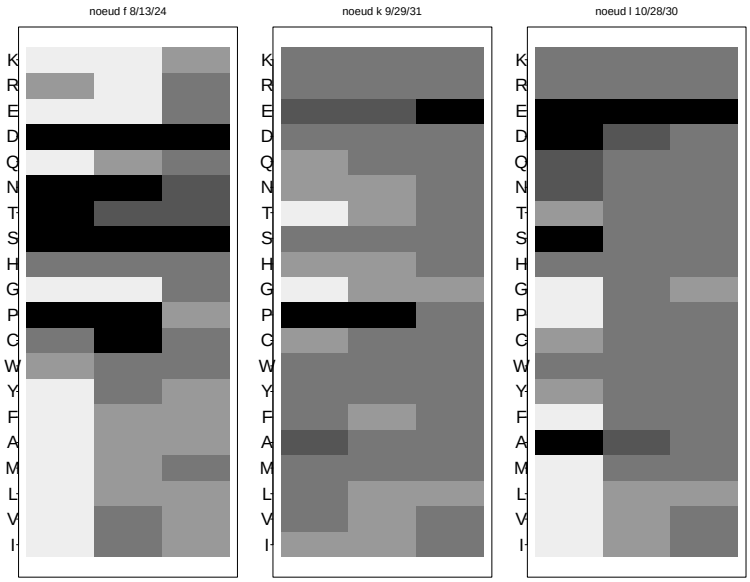

FIG. 5.5 – *Comparaison entre les distributions en acides aminés normalisés en Z-scores des nœuds correspondants aux blocs protéiques f (8, 13 et 24), k (9, 22 et 31) et l (10, 28 et 30).*

Les sites *p* des nœuds *04* de la série *o<u>p</u>a* (65 %) et *20* de *m<u>p</u>c* (28 %) montrent des caractéristiques assez éloignées. Le premier est très proche de la composition du BP *p*, avec juste une accentuation de sous-représentations pour la Méthionine et le Tryptophane. Le second se différencie fortement par une présence d'acides aminés Isoleucine, Leucine, Alanine, Trypto-

phane et Cystéine inexistants dans le BPp, ainsi qu'une sur-représentation accrue de Lysine. Les nœuds *f 08* et *13* montrent dans les suites *fklm* (56 %) et *cdf* (42 %) une forte dissemblance en acides aminés. Le premier est assez proche du BP *f* seul, le second nettement moins avec une fréquence plus élevée en Isoleucine, Valine, Tyrosine, Tryptophane et Glutamine. Enfin, les nœuds *a 05* et *18* dans les suites *opa* (65 %) et *hia* (35 %) sont totalement antagonistes, le premier assez proche du BP *a* montre une surreprésentation en Isoleucine, Valine et Phénylalanine plus importante que le BP *a*, le second lui ne possède presque plus aucun acide aminé Isoleucine, Valine, Leucine, Méthionine, Tyrosine, Thréonine, Sérine, Glutamate, Lysine et Arginine et montre une forte propension en Glycine et Aspartate.

Longueur des protéines n'influence en rien le taux de résidus compris dans le graphe par protéines qui est assez constant. La longueur moyenne des fragments inclus dans le graphe n'en dépend d'ailleurs pas.

5.3.6 Les boucles

Quelques faits classiques, caractérisant les coudes, peuvent être notés. 50 % des coudes de type I de la PDB sont représentés par un bloc *k* et toujours à proximité de blocs de type *m*, ce qui se retrouve dans la partie inférieure gauche du graphe par les successions *fklm* (nœuds *08-09-10-01*) ou *mklm* (nœuds *01-29-30-01*). De même 80 % des coudes de type II se termine par *ia*, et plus de 50 % sont présents dans le réseau dans la succession *ehia* (nœuds *15-16-17-18*). Le coude de type III est comme le coude de type I toujours à proximité du bloc *m*, ainsi 80 % de ces structures sont caractérisées par la succession *lm* (nœuds *10-01/29-01*) et plus de 60 % par *klm* (nœuds *09-10-01/29-30-01*). Le graphe permet donc de retouver ces coudes classiques, mais sans se limiter à une définition trop stricte.

5.3.7 En dehors

Une grande partie de la base de données est comprise dans le réseau : que reste-t-il en dehors ? En analysant les doublets non compris dans le réseau, deux principaux en ressortent: les paires *bd* et *fb*. Les BPs situés aux extrémités N- et C-terminales complétant ces paires sont fortement variées. Le triplet *fbd* est compris dans le réseau" (partie droite du graphe avec les nœuds *13-14-07*), mais se retrouve à 413 reprises hors du réseau. Cette non-*prise en compte* dans le

réseau est dûe à la recherche effectuée : une taille minimale de 5 blocs consécutifs sur le réseau.

Il faut noter l'absence totale du bloc g qui est caractéristique de boucles de grandes tailles, plus variables. Il faudra donc noter que l'établissement d'un réseau prenant en compte un maximum de chemins a depuis été finalisé avec des chemins de fréquence plus faible, ainsi que l'utilisation de ces motifs répétés dans la prédiction de la structure à partir de la séquence.

5.4 Conclusion

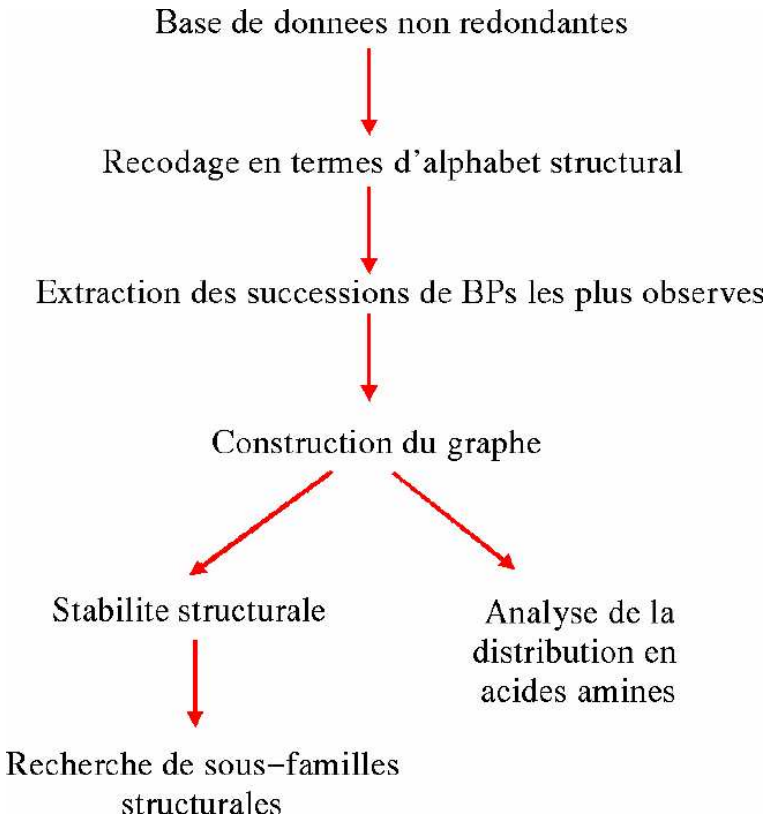

FIG. 5.6 – *Schéma récapitulatif de la validation du réseau discontinu.*

La prédiction de la structure tridimensionnelle d'une protéine à partir de sa séquence primaire est un défi constant de la biologie moderne. Pour apprécier l'hétérogénéité des structures protéiques, nous avons défini un ensemble de 16 prototypes (ou blocs protéiques) qui décrivent "au mieux" la structure locale des protéines [35]. En analysant une base de données de structures protéiques recodées à l'aide de ces blocs, certaines successions de blocs, i.e. des mots structuraux, apparaissent particulièrement fréquentes et fortement dépendantes. Elles permettent la construction d'un graphe orienté simple reliant de manière logique l'ensemble des motifs de 5 blocs les plus fréquents. Le réseau obtenu est composé de 31 nœuds distincts, correspondant

chacun à un des 16 blocs protéiques. Il contient du fait de sa conception, plus de 83% des résidus des protéines et point intéressant plus de 70% des boucles, régions considérées "variables". Pour évaluer la pertinence du réseau à reproduire des structures 3D homogènes, 17 motifs décrivant la totalité du réseau en ont été extraits. Ils ne correspondent qu'à un type unique de repliement. L'écart quadratique moyen observé *RMSd* est assez bon, il varie entre 1.37 à 3.07 Å, pour des tailles allant de 8 à 11 C_α. D'un point de vue angulaire, ces motifs sont aussi stables que les BPs seuls et ceci pour des tailles supérieures. L'utilisation de l'alphabet permet de se débarasser en fait des problémes liés à l'attribution des structures secondaires [31], principalement les feuillets β.

50 % de "structures secondaires répétitives se retrouvent en deux blocs. Le réseau pemet la construction des voies "logiques" qui entrent et sortent des BP m et du BP d qui vont suivre aussi longtemps que possible ces chemins "préférentiels". En fait, ce réseau ne montre pas que les boucles sont mieux déterminés que ce que l'ont pensent, mais qu'en fait les catégorisation actuelles basées sur l'existence d'une classification en 3-états (hélices α, feuillets β et boucles ou plutôt non-α/non-β) entraînent une classification biaisée. Des approches classiques se basent sur des structures de taille constante comprises entre deux structures répétitives, alors que la définition des bornes est loin d'être optimale, et, donc en conséquence des parties importantes de ces boucles semblables (sur le début ou sur la fin principalement) sont déclarées dans des groupes de longueur différentes. Ce réseau montre, en outre, des chemins préferentiels ,mais pas toujours continu. Les taux de départ/fin des fragments dans la réseau montrent clairement que les fragments compris dans le réseau commencent un peu partout et finissent un peu partout (cf. figure 5.2), conservant toutefois un fort déterminisme structural (cf. tableau 5.1). Les taux importants liés au nœud *c 05* présent dans la succession *mnopacd* (nœuds 01-02-03-05-06-07) sont dûs à la présence des autres blocs *c* (nœuds *12*, *22* et *23*) qui servent de terminateurs obligatoires. Cette notion de discontinuité doit être signalée pour ne pas prendre ce réseau comme un graphe continu.

Ainsi, on peut voir que l'alphabet même imparfait permet de retrouver des zones structuralement très stables et connnectées entre elles, dans les protéines. De plus, elles montrent localement des différences en répartition d'acides aminés ce qui est important pour la prédiction. L'analyse des fréquences des acides aminés associées à chaque partie du réseau permet de voir des différences séquentielles importantes. Ces résultats devraient être mis en relation avec la

méthode de prédiction bayésienne. Cette approche permettra une amélioration des recherches *ab initio* de la structure 3D à partir de la séquence. D'ailleurs, nous avons depuis développé une nouvelle méthode de prédiction basée sur l'utilisation de ces voies. Cette nouvelle stratégie nommée stratégie d'*épinglage* permet de prédire localement un mot structural et ensuite de l'étendre. Le taux de prédiction global passe de 40,7 à plus de 45 %. De plus, une amélioration de cette technique nommée *multi-épinglage* permet de générer plusieurs longs fragments en chaque position de la séquence. L'ensemble de ces deniers résultats permet de valider les concepts que nous avons développé.