Développement de méthodes bioinformatiques pour l'extraction de motifs structurés au sein des boucles protéiques

Regad L., Guyon F., Bahrini H., Hazout S., et Camproux A.C.

Equipe de Bioinformatique Génomique et Moléculaire (EBGM), INSERM U726,

Université Denis Diderot - Paris 7, case 7113,

2 place Jussieu, 75251 Paris - FRANCE (camproux@ebgm.jussieu.fr)

BUT

La caractérisation de la fonction d'une protéine à partir de sa séquence primaire est un des plus importants défis actuels de la biologie. Cette tâche est souvent facilitée par l'obtention de la structure tri-dimensionnelle (3D) de la protéine d'intérêt. Des régions particulièrement importantes sont exposées à la surface des protéines et présentent une grande fléxibilité et diversité structurale. Ces régions, dites en boucles, sont non répétitives et connectent les structures secondaires entre elles, menant à des repliements topologiques spécifiques. De plus, elles interviennent dans la fonction et la spécificité des protéines (Heuser, 2004, Fernandez, 2004, Fetrow, 1995).

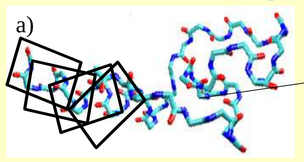
Bien que les boucles contiennent des motifs structuraux locaux notamment des motifs courts et répétés comme les coudes et coudes multiples ou plus longs (Hutchinson, 1994), la large variabilité structurelle de ces régions fait de la caractérisation et de la prédiction des conformations en boucles un des problèmes de modélisation moléculaire les plus difficiles.

S'appuyant sur cette observation, et prenant en compte que l'on ne dispose pas d'assez d'information pour arriver à une prédiction satisfaisante des boucles, l'idée est de proposer dans ce travail, est de mettre au point une prédiction partielle et précise des boucles. L'approche consiste non pas à proposer une prédiction complète des boucles en classes floues mais à identifier au sein de ces régions des motifs structurés répétés prédictibles.Ces motifs permettent de décomposer et de simplifier la modélisation et la prédiction des boucles qui pourra être étendues ensuite ou complétée par une prédiction plus globale.

Notre objectif a été d'extraire des motifs répétés de 5 à 8 résidus dans les régions en boucles puis de caractériser leurs propriétés de structure et de séquence. L'originalité de notre approche repose sur une description simplifiée des conformations 3D, en se basant sur le concept d'alphabet structural (AS).

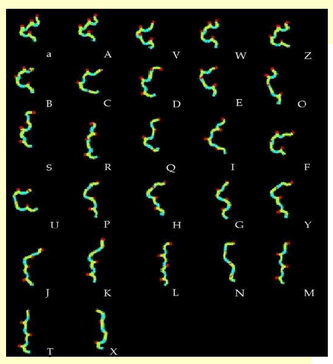
MATERIEL

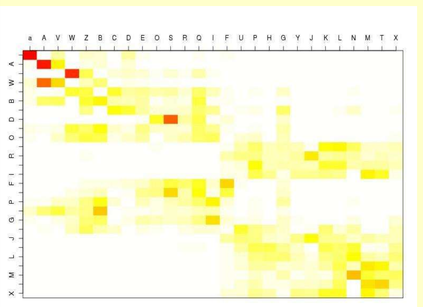
Définition de l'Alphabet Structural (AS)



c) Géométrie des 27 lettres structurales

d) Matrice de transition





L'AS a été obtenu en décomposant la chaîne polypeptidique en fragments chevauchant de 4 carbones α consécutifs (cf a) (Camproux, 2004).

Chaque fragments est décrit par un vecteur de 4 descripteurs (b) : les 3 distances entre 2 carbones α non consécutifs et la projection orientée du 4ème carbone dans le plan formé par les 3 premiers carbones

A partir de chaînes de Markov cachées (HMM), il a été obtenu un AS optimal de 27 lettres structurales (c). L'utilisation de HMM permet de prendre en compte la dépendance entre les lettres structurales (d).

Une comparaison entre les structures secondaires et les lettres structurales a été faite en terme de Zscores:

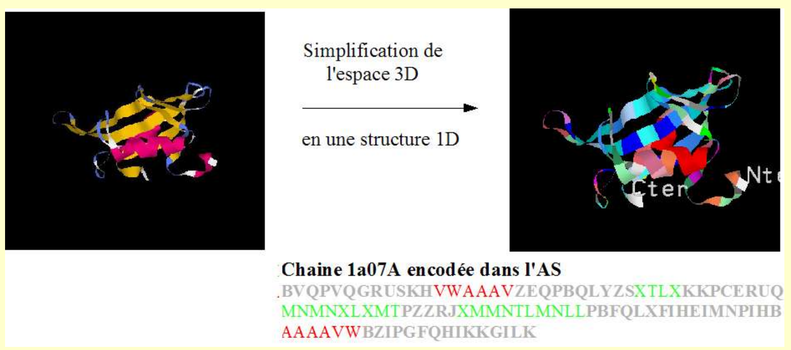
A,a,V, W : groupe de lettres hélicales

L, M, N, T, X : groupe de lettres en feuillets

Z, B, C, D, E, O, S, R, Q, I, F, U, P, H, G, Y, J, K : groupe de lettres en boucles

l'AS encode les chaînes protéiques, ce qui simplie leur conformation 3D en une série de lettres structurales un espace 1D.

Banques protéiques



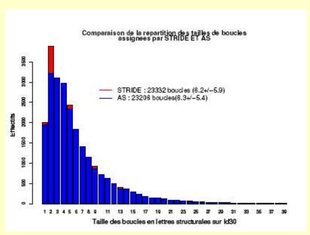
2 banques de protéines non redondante avec moins de 50% d'identité de séquence

Banque d'apprentissage : 3552 chaînes protéiques (763810 lettres structurales après encodage dans l'espace de l'AS)

Banque de validation : 256 chaînes protéiques (61871 lettres structurales après encodage dans l'espae de l'AS)

METHODES

Définition des boucles protéiques en terme d'AS



Hélice : succession d'au moins 3 lettres structurales hélicoïdales séparées au maximum par 2 lettres non hélicoïdales. Les lettres [Z, B, H, C] sont acceptées en entrée ou en sortie d'hélice

Feuillet : succession d'au moins 2 lettres structurales en feuillet séparées au maximum par 2 lettres non hélicoïdales. Les lettres [I, K] sont acceptées en entrée ou en sortie d'un feuillet

Boucle : fragment qui relient les hélices et/ou feuillet

Définition des structures secondaires en terme d'AS cohérente avec STRIDE

Définition permettant de détailler les boucles (18 lettres structurales)

Définition des mots en boucles (de taille k)

k>2 : succession de k lettres comportant au minimum 2 lettres n'appartenant pas au groupe boucle et e excluant les lettres de structures secondaires pures [A,a,L,M,N,T,X]

k=2 : succession de 2 lettres structurales appartenant au groupe boucle

Dépendance à la séquence

Zscores

$$Z_{a,l,w} = \frac{N_{a,w,l} - \frac{N_{a,l}.N_w}{N_k}}{\sqrt{\frac{N_{a,l}.N_w}{N_k}}}$$

Mesure asymétrique de Kullback-Leibler

$$Kld_{lw} = \sum_{a=1}^{n_{aa}} p_{a,l,w}.ln\left(\frac{p_{a,l,w}}{p_{a,l}}\right)$$

Dépendance à la séquence :

Zscore : position par position en acides aminés en acides aminés

Kld : position par position

EXTRACTION DES MOTS REPETES AU SEIN DES BOUCLES

Cohérence géométrique et dépendance à la séquence

Taille des mots (Effectifs)	2 (190)	3 (745)	4 (200)	5 (8)
$\overline{Kld_{l,w}}$ (sd)	1.01 (0.48)	2.10 (0.72)	3.37 (0.89)	5.82 (0.74)
% de mots avec un $Kld_{l,w}$ significatif	97%	88%	96%	100 %
Effectifs après sélection des mots[a] : sous-banque $S(\%)$	158 (83)	567 (83)	183 (91)	8 (100)
$\overline{RMSd_w}$ des mots sélectionnés (sd) $[\AA]$	0.53 (0.14)	0.57 (0.13)	0.6 (0.12)	0.57 (0.14)
Taux de recouvrement[b]	83%	71%	21%	1.3%
% de mots avec une dépendance aux flancs [c]	94%	88%	95 %	100%

Résultats

Existence dans les boucles des motifs:

très répétés (Occurrences > 100)

très structurés (RMSd < 0.6 Å)

avec une forte dépendance à la séquence

avec une dépendance aux flancs

recouvre 89% des régions en boucles

PREDICTION DES MOTIFS

Définition des taux de prédiction

La prédiction a été effectuée à l'aide d'un score R en utilisant une approche bayésienne

$$R_j = \frac{\prod_{i=1}^{n_p} p(a_i/w_j) \times p(w_j)}{\prod_{i=1}^{n_p} p(a_i/S)}$$

P_{rg1} : Taux de bien prédit au rang 1

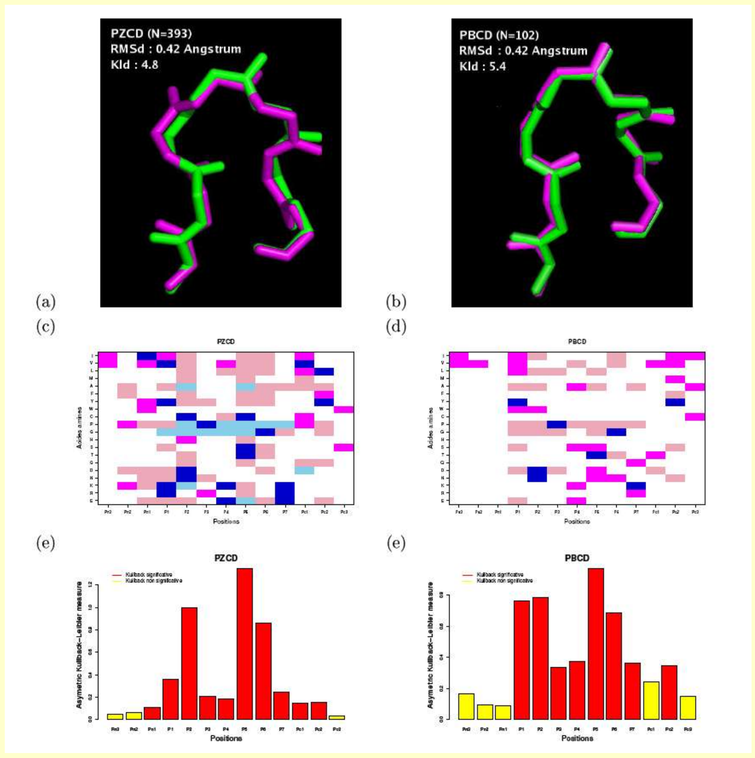
P_{rg4} : Taux de bien prédit au rang 4

P_{rmsd1} : Taux de bien prédit au rang 1 en terme de RMSd acceptable (1Å)

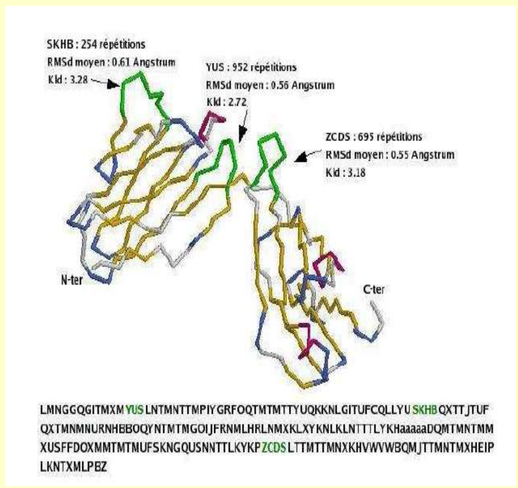
Résultats : taux de prédiction

Taille des mots (Effectifs) (Id-50-Val)	2 (158)	3 (567)	4 (183)	5 (8)
Nombre de mots prédits	21215	14244	2552	8
P_{rg1} (aléatoire)	14 (0.63)	9 (0.18)	19 (0.55)	80 (30)
$P1, rmsd1$ (aléatoire)	32 (8.9)	26.7 (3.4)	41 (4)	80 (6)
$P1, rmsd1.4$ (aléatoire)	44 (22)	36(8.8)	49 (7)	100 (6)
P_{rg5} (aléatoire)	35 (3.2)	24 (0.9)	47 (2.7)	82 (62)
$P5, rmsd1$ (aléatoire)	65(39)	52(14)	64 (12)	96(23)
$P5, rmsd1.4$ (aléatoire)	79 (71)	63(34)	72 (20)	100 (35)

Les taux de prédiction obtenus pour les tailles de 2 à 5 lettres (fragments de 5 à 8 résidus) sont proche de 35% au rang 5 et supérieur à 50% au rang 5 pour un RMSd <1 A avec un fort gain par rapport à l'aléatoire.



Exemple e 2 boucles protéiques PZCD et PBCD en terme de RMSd , de Z-score et de mesure du kullback-Leibler



CONCLUSION AND PERSPECTIVES

Dans la première partie de ce travail, nous avons extrait des motifs répétés au sein des boucles protéiques encodées dans l'espace de l'AS. L'analyse des propriétés de stuctures et de séquences a permis de conclure que ces motifs étaient très structurés et fortement dépendant à la séquence. Ces résultats rejoignent ceux obtenus par Pachenko et Madej (2004) et permettent de conclurent que les boucles ne sont pas des régions coplètement aléatoires, mais qu'elles contiennent des régions structurées. Les résultats de la prédiction donnent de bon scores de prédiction très supérieurs à l'aléatoire.

Perspectives :

Pour améliorer la qualité de la prédiction, nous envisageons d'ajoutons à notre approche de prédiction la dépendance aux flancs et d'intégrer une pondération des positions en fonction de la dépendance à la séquence. A plus long terme, des clases de mots proches en terme de structure et de séquence seront tester pour proposer des mots plus longs. Ces classes nous permettront d'étendre la recherche de mots à la recherche de mots flous, afin d'explorer plus de régions dans les boucles.

De plus, une prochaine étape sera l'analyse de la dépendance entre fonction et motifs.

REFERENCES

[2] Fernandez-Fuentes N., Hermoso A., Espadaler J., Querol E., Aviles F.X., Oliva B. (2004) Classification of common functional loops of kinase super-families. *Proteins*. 15;56(3):539-55.

[3] Fetrow J.S. (1995) Omega loops: nonregular secondary structures significant in protein function and stability. *FASEB J.* 9:708-17.

Heuser P., Wohlfahrt G.& Schomburg D. (2004) Efficient methods for filtering and ranking fragments for the prediction of structurally variable regions in proteins. *Proteins*. 15;54(3):583-95.

Fernandez-Fuentes N., Hermoso A., Espadaler J., Querol E., Aviles F.X. & Oliva B. (2004)Classification of common functional loops of kinase super-families. *Proteins*. 15;56(3):539-55.

Camproux, A. C., Gautier.R. & Tuffery, P. (2004) A Hidden Markov Model derived structural alphabet for proteins. *J Mol Biol*, 339, 591-605.

Hutchinson, E., & Thornton, J. (1994) A revised set of potentials for beta-turn formation in proteins. *Protein Sci*, 3, 2207-2216.Camproux, A. C., Gautier.R. & Tuffery, P. (2004) A Hidden Markov Modle derived structural alphabet for proteins. *J Mol Biol*, 339, 591-605.

Panchenko, A. R., & Madej, T. (2005) Structural similarity of loops in protein families: toward the understanding of protein evolution. *BMC Evol Biol*, 3,5(1):10

Fetrow, J.S. (1995) Omega loops: non regular secondary structures significant in protein function and stability. FASEB J, 9(9):708-717.