Recherche de motifs structuraux répétés

Application of relational motifs to multiple structural alignment

Mathilde Carpentier[1], Nadia Pisanti[2] Joël Pothier[3], Henry Soldano[4]

(1) Laboratoire de modélisation en biologie intégrative,IJM, CNRS, Universités Paris 6 et 7, France.

(2) Dipartimento di Informatica, Universita di Pisa, Italy

(3) Atelier de BioInformatique, Université Paris 6, France

(4) Laboratoire d'Informatique de l'Université Paris-Nord,UMR-CNRS 7030, France.

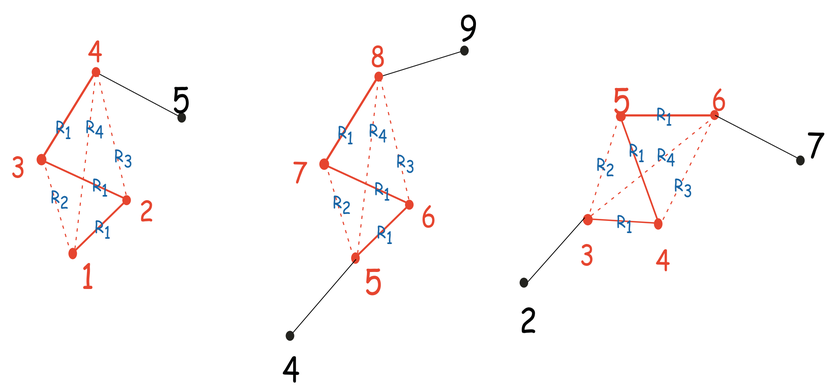
Courriel : mathilde@abi.snv.jussieu.fr

Beaucoup d'algorithmes ont été développés pour aligner deux structures protéiques, beaucoup moins l'ont été pour aligner plusieurs structures. Or il est intéressant de trouver les similarités entre plusieurs structures, par exemple pour définir des régions structuralement conservées ("coeurs") pour les méthodes de reconnaissance de repliements ("threading"), ou pour la classification automatique des structures (il existe plusieurs classifications des structures protéiques: SCOP [1], CATH [2], CE database [3], FSSP [4], VAST [5]).

Certains programmes d'alignement structuraux multiples utilisent au départ un alignement des séquences[6], [7] ; d'autres initient l'alignement multiple par des alignements par paires des structures [8], [9], [6] ; d'autres encore utilisent un pivot [10]. Enfin, l'alignement peut être multiple dès le départ [11,12,13]. Dans notre méthode, la description de la structure tridimensionnelle des protéines est réalisée soit par les angles (f,y), soit par les angles (a, t) (cette description est linéaire). Certaines des meilleures méthodes de comparaison de deux structures, utilisent une description en distances internes des structures. Or il n'existe aucune méthode de comparaison multiple de structures travaillant avec les distances internes. Nous avons donc développé un algorithme, inspiré de l'algorithme de KMRC, permettant de comparer de manière multiple les structures selon leurs distances internes. L'implémentation de cette méthode est nommée Triades.et son algorithme a été publié. [14,15]

Présentation générale

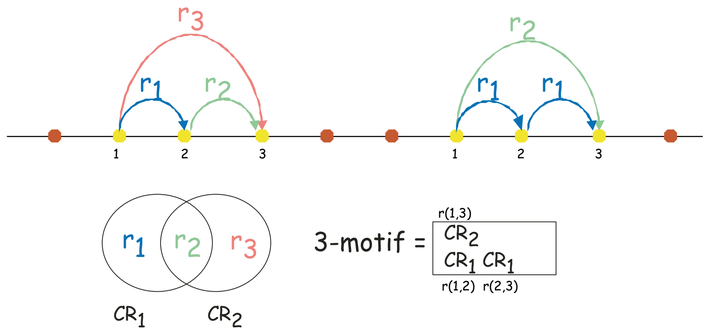
Les distances internes, comme les angles, peuvent être échantillonnées suivant un pas ou **maille** (en Angström) et donc être représentées par des symboles associés à des classes ("classes d'intervalles"). Comme la distance est calculée entre deux résidus. il s'agit bien d'une **relation.** Chercher des fragments structuraux similaires représentés par leurs distances internes revient à rechercher les résidus contigus entre lesquels les distances internes - les relations - sont identiques ou similaires (voir figure ci-contre). Ces fragments sont appelés **motifs relationnels**..Notre algorithme permet de trouver des motifs composés à la fois de symboles similaires et de relations similaires. Dans l'application au cas des structures protéiques, seules les relations sont prises en compte, le même symbole est donc affecté à tous les résidus. Cependant, d'autres applications de cet algorithme utilisent les



Exemple de fragments structuraux dont les distances internes sont proches : ils ont donc des relations identiques (relations R1 à R4). Les points représentent< les Ca. Les fragments structuraux proches sont en rouge, les lignes pointillées représentent les distances internes autres que celles entre Ca contigus La distance entre deuxCa successifs est quasiment fixe, la relation est toujours R1. Par contre, les autres distances sont variables et les relations sont R2, R3, R4 (en bleu).

Définition des motifs

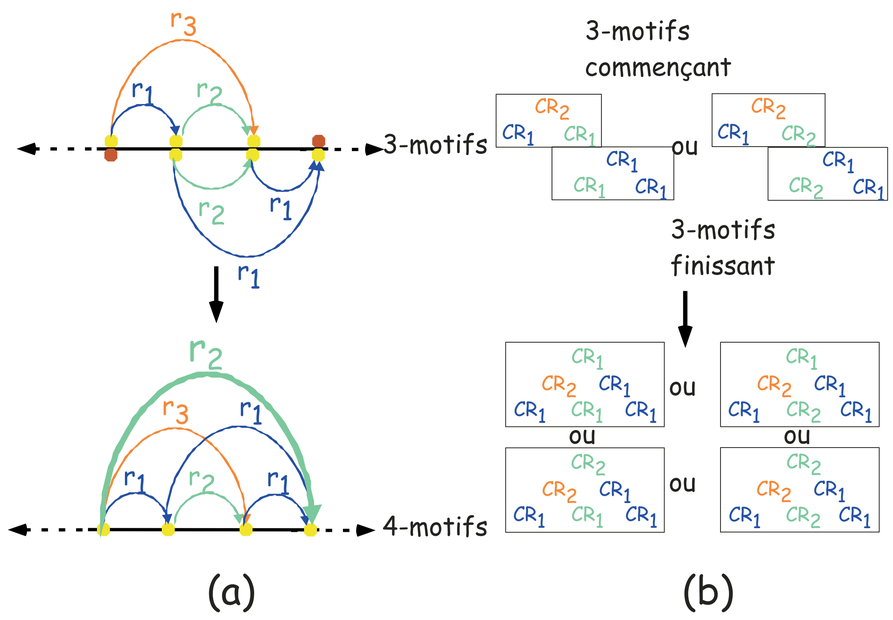
Pour permettre la recherche de motifs non pas exactement identiques mais approchés, les relations "proches" sont considérées comme faisant partie du même "pavé relationnel". Deux occurences d'un motif dont les relations appartiennent aux mêmes pavés relationnels sont considérées comme similaires.



Exemple de deux motifs relationnels de taille 3.
Trois relations sont définies : r1, r2, et r3,regroupées en deux pavés relationnels CR1 = {r1, r2} et CR2 = {r2, r3}. Les symboles (non relationnels), représentés par des points jaunes ou marrons, sont omis car ils sont tous identiques dans ces exemples. Les deux occurences de motifs relationnels différent uniquement au niveau des relations. Comme la relation r2 appartient à deux pavés relationnels, un motif de taille 3 (3-motif) est répété.

Construction des motifs

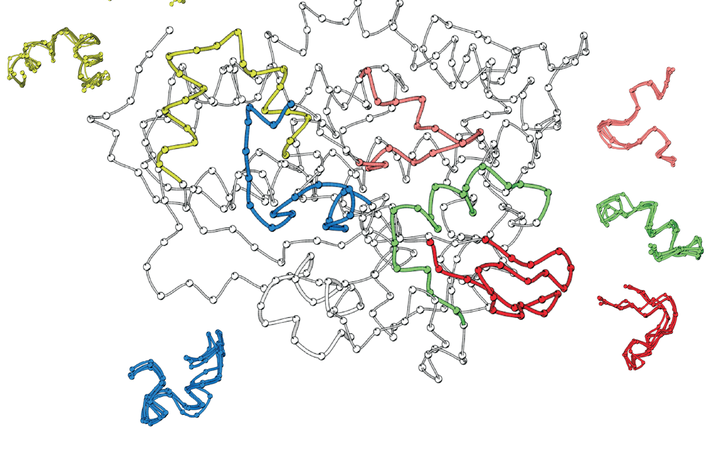
Un motif de taille k est construit à partir de deux motifs chevauchants de taille k-1. Il faut ensuite vérifier que les relations entre le 1er et le dernier symbole sont identiques ou proches. A chaque étape, tous les motifs de même taille sont construits et seront utilisés pour construire les motifs de taille +1 à l'étape suivante.



Construction de 4-motifs relationnels à partir de 3-motifs relationnels chevauchants.
Seule une instance de chaque motif est représentée mais on suppose que tous les motifs ont assez d'instances pour être répétés.
(a) Toutes les relations sont connues sauf la relation entre la première et la dernière position (en gras) qu'il faut vérifier.
(b) Comme la relation r2 appartient à deux pavés relationnels, chaque instance de longueur 3 correspond à deux 3-motifs. Les 3-motifs sont combinés lorsque les pavés de relations se chevauchant sont les mêmes. Les premiers 3-motifs sont nommés « motifs commençant» (k—moti fC) et les seconds « motifs finissant» (k—moti f F). Deux 4-motifs sont donc construits. Cependant, lors de la vérification de la relation entre la première et la dernière position, deux pavés relationnels sont trouvés, et ce

Résultats pour les structures protéqiues

Cinq motifs structuraux trouvés dans quatre cytochromes P450 : deux de 22 résidus, deux de 19 et un de 16. Seule une structure est affichée (3CPP). Détail des cinq motifs structuraux trouvés dans les 4 cytochromes P450 : le second motif -en jaune - est montré deux fois (7.16(b) et 7.16(c)) car il possède deux occurrences qui se chevauchent dans la protéine 1 (le premier débute en position 248 et le second au résidu 258; ces occurrences apparaissent en couleur plus foncée.



Conclusion

Les avantages de cette méthode est qu'elle est réellement multiple, exhaustive (tous les motifs maximaux sont construits), et générique. Dans le cas des structures, il faut cependant accepter que certaines distances ne soient pas vérifiées, car la dégénérescence est trop grande. Néanmoins les motifs structuraux trouvés sont tout à fait similaires. En effet, les contraintes géométriques de l'enchaînement des résidus rend ces vérifications en fait inutiles. La principale limite de la méthode réside dans le mode de construction "en largeur" des motifs. Lorsque le nombre total de résidus n dans toutes les structures augmente, le nombre de fragments structuraux candidats augmente ce qui peut poser des problèmes de mémoire (beaucoup de motifs non maximaux sont générés lors des étapes intermédiaires) De plus, quelques traitements finaux seront nécessaires pour que les motifs structuraux trouvés forment un vrai alignement structural multiple. Pour l'instant, tous les motifs maximaux d'une taille donnée et de taille inférieure sont fournis par la méthode. Pour avoir un alignement multiple de structures, il faudra trouver la meilleure combinaison de motifs de tailles variables (opération pour l'instant possible uniquement sur les motifs de même taille). Il sera aussi possible et intéressant d'obtenir des motifs de taille variable selon la protéine. En effet, lors de de la construction des motifs de taille k à partir de motifs de taille k—1, il est possible qu'une protéine possédant l'extension du k—1-motif ne possède plus l'extension du k-motif. Connaissant les motifs de taille k et ceux de taille k—1 ayant permis de les construire, il est possible de construire un bloc structural avec les motifs de taille k pour certaines protéines et ces motifs de taille inférieure.

Bibliographie

[1] Murzin A. G., Brenner S. E., Hubbard T., Chothia C.. Scop : a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995 ;247 :536-40.
[2] Orengo C. A., Michie A. D., Jones S., Jones D. T., Swindells M. B., Thornton J. M.. Cath: a hierarchic classification of protein domain structures. Structure 1997 ;5 :1093-108.
[3] Shindyalov I. N., Bourne P. E.. An alternative view of protein fold space. Proteins 2000 ;38 :247-60.
[4] Holm L., Sander C.. Mapping the protein universe. Science 1996 ;273 :595-603.
[5] Gibrat J. F., Madej T., Bryant S. H.. Surprising similarities in structure comparison. Curr Opin Struct Biol 1996 ;6 :377-85.
[6] Gerstein M, Altman RB. Using a measure of structural variation to de□ne a core for the globins. Comput Appl Biosci 1995 ;11 :633-644.
[7] Gelfand I., Kister A., Kulikowski C., Stoyanov O.. Geometric invariant core for the v(l) and v(h) domains of immunoglobulin molecules. Protein Eng 1998 ;11 :1015-25.
[8] Orengo C. A., Taylor W. R.. Ssap : sequential structure alignment program for protein structure comparison. Methods Enzymol 1996 ;266 :617-35.
[9] Guda C., Scheeff E. D., Bourne P. E., Shindyalov I. N.. A new algorithm for the alignment of multiple protein structures using monte carlo optimization. In Pacific Symposium on Biocomputing. Hawaii. 2001.
[10] Leibowitz N., Fligelman Z. Y., Nussinov R., Wolfson H. J.. Multiple structural alignment and core detection by geometric hashing. Proc Int Conf Intell Syst Mol Biol 1999 ;169-77.
[11] Jean P., Pothier J., Dansette P. M., Mansuy D., Viari A.. Automated multiple analysis of protein structures : application to homology modeling of cytochromes p450. Proteins 1997 ;28 :388-404.
[12] M. F. Sagot, A. Viari, J. Pothier, and H. Soldano. Finding flexible patterns in a text: an application to threedimensional molecular matching. Comput Appl Biosci, 11(1):59–70., 1995.
[13] H. Soldano, A. Viari, and M. Champesme. Searching for flexible repeated patterns using a non-transitive similarity relation. Pattern Recognition Letters, 16:243–46, 1995.
[14] N. Pisanti, H. Soldano, M. Carpentier, and J. Pothier. Implicit and explicit representation of approximated motifs. In C. Iliopoulos, K. Park, and Steinhofel K., editors, Algorithms for Bioinformatics, C, volume in press. King's College London Press, London, 2005.
[15] Nadia Pisanti, Henry Soldano, and Mathilde Carpentier. Incremental inference of relational motifs with a degenerate alphabet. In Kunsoo Park Alberto Apostolico, Maxime Crochemore, editor, Combinatorial Pattern Matching: 16th Annual Symposium, CPM 2005, Jeju Island, Korea,, 2005. Proceedings, volume 3537 of Lecture Notes in Computer Science, pages 229–240, Jeju Island, Korea, 2005. Springer-Verlag GmbH.