FaBR-CL : méthode de classification croisée de protéines

Walid Erray*, Faouzi Mhamdi**

*Laboratoire ERIC, Université Lumière Lyon 2
69500 Bron France
walid.erray@eric.univ-lyon2.fr,
http://eric.univ-lyon2.fr
**URPAH, faculté des Sciences de Tunis,Université d'El Manar
1060 Tunis Tunisie
faouzi.mhamdi@ensi.rnu.tn,
http://www.mes.tn/fst/index.html

Résumé. Dans cet article, nous proposons une méthode de classification croisée permettant de classer des protéines, d'une part, et de classer des descripteurs (3-grammes) selon leurs pertinences par rapport aux groupes de protéines obtenus, d'autres part.

1 Classification croisée de données biologiques

Afin d'étudier les séquences d'acides aminés représentant les protéines, nous avons utilisé des techniques de text mining afin d'extraire des descripteurs. Ces descripteurs nous permettrons de construire un tableau de données Protéines $\times$ Descripteurs. L'une des techniques les plus utilisées est l'extraction des x-grammes (Miller et al. (1999), Mhamdi et al. (2004)), x étant la taille d'un descripteur.

Plusieurs méthodes de classification croisée ont été proposées (Govaert (1977), Ritschard et Nicoloyannis (2000)). Récemment, des méthodes de classification croisée ont été appliquées aux données biologiques (Cheng et Church (2000)). Cependant, plusieurs de ces méthodes restent très coûteuses en temps de calcul.

2 FaBR-CL : méthode de classification croisée

Afin d'effectuer une classification croisée, nous nous sommes basé sur une méthode de classification peu coûteuse en temps de calcul (Erray (2005)). La méthode proposée, FaBR-CL, utilise FaUR dans une approche "Combinaison itérative de regroupement des lignes et des colonnes" afin d'obtenir un regroupement complet des protéines et des 3-grammes. Ainsi, nous effectuons le regroupement des protéines, dans un premier temps, et le regroupement des 3-grammes dans un deuxième temps. La complexité de cette méthodes est en O(l log l + p log p), l étant le nombre de protéines et p le nombre de descripteurs.

3 Classification croisée de protéines

Nous avons travaillé sur les trois familles de protéines PAD, TLR et AD afin de valider notre approche. Les trois études portant à chaque fois sur des protéines appartenant à deux familles, montrent que la méthode FaBR-CL donne un classement très proche de la réalité (PAD, TLR et AD). Aussi, nous obtenons des groupes de 3-grammes fortement pertinents par rapport à chaque classe de protéines. L'étude de toutes les protéines des trois familles, confirme ces résultats.

4 Conclusion

La méthode proposée permet, avec un coût calculatoire très faible, de classer les protéines. Aussi, cette méthode permet de mettre en évidence des groupes de 3-grammes, de faibles effectifs, et qui permettent d'identifier une classe de protéines par leurs présences ou par leurs absences.

Références

Cheng, Y. et G. Church (2000). Biclustering of expression data. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pp. 93–103.

Erray, W. (2005). Faur : Méthode de réduction unidimensionnelle d'un tableau de contingence. In *SFC05 : 12ème rencontres de la Société Francophone de Classification*, Montreal, Canada.

Govaert, G. (1977). Algorithme de classification d'un tableau de contingence. In *First international symposium on Data Analysis and Informatics*, INRIA, Versailles, pp. 487–500.

Mhamdi, F., M. Elloumi, et R. Rakotomalala (2004). Textmining, features selection and datamining for proteins classification. In *In Proceedings of IEEE/ICTTA'04*, Damascus, Syria.

Miller, D., D. Shen, et C. N. J. Liu (1999). Performance and scalability of a large-scale n-gram based information retrivial system. *Journal of digital information 1*(5).

Ritschard, G. et N. Nicoloyannis (2000). Aggregation and association in cross tables. In D. Zighed, H. Komorowski, et J. Zytkow (Eds.), *Principles of Data Mining and Knowledge Discovery, 4th European Conference, PKDD 2000, Lyon, France, September 13-16, 2000, Proceedings*, pp. 593–598. Springer.

Summary

In this article we propose a bi-clustering method able to classify proteins in one hand, and to classify descriptors (3-grams) according to their pertinence to the obtained groups of proteins in another hand.