DEMON : DEcouverte de MOtifs séquentiels pour les puces adN

Paola Salle* Sandra Bringay *,** Maguelonne Teisseire *

* LIRMM, Univ. Montpellier 2, CNRS, 161 rue Ada, 34392 Montpellier, France
prenom.nom@lirmm.fr,
** Dpt MIAP, Université de Montpellier 3, Route de Mende, 34199 Montpellier Cedex 5

Résumé. Prometteuses en terme de prévention, de dépistage, de diagnostic et d'actions thérapeutiques, les puces à ADN mesurent l'intensité des expressions de plusieurs milliers de gènes. Dans cet article, nous proposons une nouvelle approche appelée DEMON, pour extraire des motifs séquentiels à partir de données issues des puces ADN et qui utilise des connaissances du domaine.

Présentation

Dans le cadre du projet "Gene Mining" mené en collaboration avec le laboratoire "Mécanismes moléculaires dans les démences neurodégénératives" (MMDN) de l'Université Montpellier 2[1], nous nous intéressons à un type de données particulières issues de l'analyse de puces ADN. Il s'agit de biotechnologies récentes qui reposent sur le principe suivant : dans un tissu cellulaire, dans des conditions différentes, le niveau d'expression des gènes est différent. Les méthodes de fouille de données sont alors pertinentes pour les biologistes qui sont à la recherche de relations entre ces expressions. Hélàs, les méthodes classiques, basées sur une énumération de colonnes, ne peuvent être utilisées sur ce type de bases qui sont composées de milliers de colonnes. Ainsi proposer des méthodes de fouille, capables de traiter ces données pour une interprétation efficace par les experts, est donc un véritable challenge.

Dans la littérature, différentes techniques permettent aux biologistes d'exploiter les données issues de l'analyse de puces ADN. Par exemple, Eisen et al. (1998) proposent une méthode faisant référence. Ils appliquent un clustering hiérarchique sur les données préalablement discrétisées selon la distinction "sur" et "sous" exprimé. Les biologistes identifient des groupes de gènes dont l'intensité varie de manière similaire selon les conditions biologiques. Pan et al. (2003), Rioult et al. (2003) proposent d'extraire des motifs fermés en réalisant une énumération sur les lignes. Pensa et al. (2004) réalisent une extraction de règles d'associations sous contraintes en utilisant des propriétés sur les gènes issues de Gene Ontology. Nous proposons d'extraire des motifs séquentiels à partir des données issues des puces ADN. L'originalité de notre approche est que ces motifs permettent d'identifier des séquences de gènes qui tendent à s'exprimer de manière similaire selon les conditions biologiques (malades, jeunes, etc.). Nous introduisons une source de connaissances du domaine pour réduire l'espace de recherche en ciblant la recherche aux séquences fréquentes dans lesquelles on retrouve des gènes appartenant à la liste de gènes cibles.

[1] www.mmdn.univ-montp2.fr

L'extraction de motifs séquentiels en utilisant des classes permet aux biologistes d'identifier des motifs discriminants. En effet, il est plus pertinent pour l'expert de savoir qu'une séquence s = <(G1)(G2)(G3)> est fréquente chez tous les jeunes mais non fréquente chez tous les âgés plutôt que de savoir qu'il existe s' = <(G2)(G1)(G3)> fréquente chez tous les âgés. Par exemple, chez tous les jeunes, s=< (—) (FZR1) (NIF3L1BP1 A2M) >[100%] et chez tous les âgés s'=< (—) (FZR1) (A2M) (NIF3L1BP1) (WWOX) (TXN) >[100%]. Le gène NIF3L1BP1 a une expression toujours similaire à A2M dans une classe mais toujours supérieure à A2M dans l'autre classe.

Nous proposons ainsi une nouvelle approche pour extraire des motifs séquentiels à partir de données issues des Puces ADN. Nous avons intégré une source de connaissances extérieures (liste de gènes cibles) lors de la génération des connaissances pour réduire l'espace de recherche.

Références

Eisen, M., P. Spellman, P. Brown, et D. Botstein (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Science 85*(25), 14863–14868.

Pan, F., G. Cong, A. K. H. Tung, J. Yang, et M. J. Zaki (2003). Carpenter : finding closed patterns in long biological datasets. In L. Getoor, T. E. Senator, P. Domingos, et C. Faloutsos (Eds.), *KDD*, pp. 637–642. ACM.

Pensa, R. G., J. Besson, et J.-F. Boulicaut (2004). A methodology for biologically relevant pattern discovery from gene expression data. In E. Suzuki et S. Arikawa (Eds.), *Discovery Science*, Volume 3245 of *Lecture Notes in Computer Science*, pp. 230–241. Springer.

Rioult, F., J.-F. Boulicaut, B. Crémilleux, et J. Besson (2003). Using transposition for pattern discovery from microarray data. In M. J. Zaki et C. C. Aggarwal (Eds.), *Proc. 8th ACM SIGMOD Workshop on research issues in Data Mining and Knowledge Discovery DMKD'03,*, pp. 73–79. ACM.

Summary

DNA microarrays are used to measure the expression levels of thousands of genes. Therefore, it is a promising technology to improve prevention, detection, diagnosis and therapeutic actions. In this paper, we propose a new algorithm called DEMON to extract sequential patterns from microarray datasets using knowledge domain such as Metabolic Pathway.