SequencesViewer : comment rendre accessible des motifs séquentiels de gènes trop nombreux ?

Arnaud Sallaberry *, Nicolas Pecheur **
Sandra Bringay ***, Mathieu Roche **, Maguelonne Teisseire ****

* LaBRI & INRIA Bordeaux - Sud Ouest & Pikko, arnaud.sallaberry@labri.fr,
**LIRMM - Université Montpellier 2, {pecheur,mathieu.roche}@lirmm.fr
***LIRMM - Université Montpellier 3, bringay@lirmm.fr
**** CEMAGREF - UMR TETIS, maguelonne.teisseire@cemagref.fr

Résumé. Les techniques d'extraction de connaissances appliquées aux gros volumes de données, issus de l'analyse de puces ADN, permettent de découvrir des connaissances jusqu'alors inconnues. Or, ces techniques produisent de très nombreux résultats, difficilement exploitables par les experts. Nous proposons un outil dédié à l'accompagnement de ces experts dans l'appropriation et l'exploitation de ces résultats. Cet outil est basé sur trois techniques de visualisation (nuages, systèmes solaire et treemap) qui permettent aux biologistes d'appréhender de grandes quantités de motifs séquentiels (séquences ordonnées de gènes).

1 Introduction

Ces dernières années, les puces ADN ont été utilisées avec succès pour de nombreuses applications (diagnostic, thérapie...). Elles permettent de comparer l'expression de milliers de gènes dans différents tissus, cellules et conditions physiologiques (Hoerndli et al. (2005)). Exploiter ces données pour obtenir une interprétation biomédicale reste difficile en raison des gros volumes de données. En effet, pour une étude, les biologistes utilisent généralement moins d'une centaine de puces mais chaque puce mesure l'expression de milliers de gènes. Par exemple, les puces Affymetrix U-133 plus 2,0 mesurent 54675 valeurs numériques.
Dans ce contexte, les techniques de fouille de données (Cong et al. (2004); Pensa et al. (2004)) jouent un rôle clé en permettant de découvrir des connaissances jusqu'alors inconnues. Nous utilisons ici l'algorithme DBSAP (Salle et al. (2009)) et obtenons des motifs séquentiels composés de gènes corrélés et ordonnés selon leur niveau d'expression. Selon les paramètres utilisés en entrée de DBSAP, nous obtenons entre 1.000 et 100.000 motifs pour chaque jeu de données. La quantité de résultats obtenus reste alors trop importante pour permettre aux experts de les interpréter facilement.
Outre les problèmes liés aux trop nombreux résultats, les biologistes rencontrent également des difficultés lorsqu'ils interprètent les motifs. Ils accèdent alors à des bases bibliographiques (e.g. PubMed) afin de comparer et évaluer la pertinence des corrélations découvertes. Ces interrogations sont manuelles ce qui rend le processus long et fastidieux. S'il existe désormais