Apprentissage de signatures de facteurs de transcription à partir de données d'expression

Mohamed Elati*, Céline Rouveirol*[1], François Radvanyi**

* LRI, UMR CNRS 8623 ,
Université Paris Sud, bât 490
91405 ORSAY cedex
elati, celine@lri.fr
** Institut Curie, UMR CNRS 144,
26 rue d'Ulm
75248 Paris cedex 05
francois.radvanyi@curie.fr

Résumé. L'inférence de signatures de facteurs de transcription à partir des données puces à ADN a déjà été étudié dans la communauté bioinformatique. La principale difficulté à résoudre est de trouver un ensemble d'heuristiques pertinentes, afin de contrôler la complexité de résolution de ce problème NP-difficile. Nous proposons dans cet article une solution heuristique alternative à celles utilisées dans les approches bayésiennes, fondée sur la recherche de motifs fréquents maximaux dans une matrice discrétisée issue des données numériques de puces ADN. Notre méthode est appliquée sur des données de cancer de vessie de l'Institut Curie et de l'Hôpital Henri Mondor de Créteil.

1 Introduction

Un des principaux objectifs de la biologie moléculaire consiste à comprendre la régulation des gènes d'un organisme vivant dans des contextes biologiques spécifiques. Les facteurs de transcription (notés Tfs dans la suite) sont les régulateurs de la transcription qui vont réagir avec les promoteurs de la transcription des gènes cibles. Ils ont deux modes d'action : ils peuvent *activer* ou *inhiber* l'expression d'un gène. Les mécanismes d'interaction facteurs de transcription/gènes cibles sont complexes. Plusieurs facteurs de transcription peuvent être nécessaires pour l'induction (resp. la répression) d'un gène cible et, dualement, un facteur de transcription peut induire ou réprimer plusieurs gènes. Les techniques récentes d'analyse du transcriptome, telles que les puces à ADN permettent de mesurer simultanément les niveaux d'expression de plusieurs milliers de gènes. Un ensemble de puces permet donc de connaître l'expression de ces milliers de gènes dans plusieurs conditions expérimentales d'intérêt. En général, ces mesures (appelés *données d'expression* dans la suite) sont représentées dans une matrice dont les lignes représentent les gènes et les colonnes représentent les différentes puces disponibles. Certains travaux d'analyse de puces font l'hypothèse que l'observation de corrélations dans les données d'expression va permettre d'inférer des relations

[1]Ce travail a été effectué pendant la délégation CNRS de Céline Rouveirol à l'Institut Curie.