Identification de complexes protéine-protéine par combinaison de classifieurs. Application à *Escherichia Coli*

Thomas Bourquard[*,**], Damien M. de Vienne[***,****], Jérôme Azé[♯]

[*] BIOS group, INRA, UMR 85, Unité Physiologie de la Reproduction et des Comportements, Nouzilly, France
[**]CNRS, UMR 6175, Nouzilly, France ; Univ. François Rabelais, Tours, France
Thomas.Bourquard@tours.inra.fr
[***]Centre for Genomic Regulation, C/Dr. Aiguader 88, 08003 Barcelona, Spain
[****]Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain
Damien.de-Vienne@crg.es
[♯]LRI, CNRS UMR 8623, AMIB INRIA Team, Univ. Paris-Sud, Orsay, France
Jerome.Aze@lri.fr, http://www.lri.fr/∼aze

Résumé. Nous proposons une approche permettant de prédire des complexes impliquant trois protéines (appelés trimères) à partir de combinaison de classifieurs appris sur des complexes n'impliquant que deux protéines (dimères). La prédiction de ces trimères repose sur deux hypothèses biologiques : (i) deux protéines orthologues présentent des caractéristiques fonctionnelles similaires; (ii) deux protéines interagissant sous la forme d'un complexe sous-tendent une fonction biologique essentielle à l'espèce concernée. Ces deux hypothèses sont exploitées pour décrire chaque paire de protéines par l'ensemble des espèces pour lesquelles elles possèdent un orthologue. Un ensemble de mesures de qualité classiquement utilisées pour évaluer l'intérêt des règles d'association est utilisé pour évaluer la force du lien entre les deux protéines. L'organisme modèle *Escherichia Coli* a été utilisé pour évaluer notre approche.

1 Introduction

L'étude des génomes nous apprend beaucoup sur le vivant. Chaque organisme possède un génome dont la composition est le résultat de l'histoire évolutive propre à cet organisme. Cette information biologique codée par l'ADN est divisée en unités discrètes, **les gènes**. Ces gènes codent pour **les protéines**, véritables "rouages" des réseaux biologiques au niveau cellulaire. Le projet de séquençage du génome humain ($3,4$ Gb, 1 Gb représentant un milliard de paires de bases ou acides nucléiques A, C, G et T), initié en 1990, a duré 13 ans pour un coût total de $2,7$ milliards de dollars. Les techniques récentes (2nd Generation Sequencing) ou futures (3rd Generations sequencing) permettent des vitesses de séquençage bien plus élevées, de l'ordre de 50 Gb/ jour, soit un génome comme celui de l'Homme entièrement séquencé en moins de 3 heures (HiSeq Systems, Illumina Inc.).

On comprend aisément le déluge de données brutes qu'il faudra savoir stocker et analyser. Nous nous intéressons à l'exploitation de ces données pour en extraire des connaissances,