Fouille de données du génome à l'aide de modèles de Markov cachés

Sébastien Hergalant * **, Bertrand Aigle *
Pierre Leblond*, Jean-François Mari**

*Laboratoire de Génétique et Microbiologie, UMR-UHP-INRA, IFR 110,
54506 Vandœuvre-lès-Nancy, France
{bertrand.aigle,pierre.leblond}@nancy.inra.fr,
**LORIA UMR-CNRS 7503, 54506 Vandœuvre-lès-Nancy, France
{hergalan,jfmari}@loria.fr
http ://www.loria.fr/~jfmari/ACI/

Résumé. Nous décrivons un processus de fouille de données en bioinformatique. Il se traduit par la spécification de modèles de Markov cachés du second-ordre, leur apprentissage et leur utilisation pour permettre une segmentation de grandes séquences d'ADN en différentes classes qui traduisent chacune un état organisationnel et structural des motifs d'ADN locaux sous-jacents. Nous ne supposons aucune connaissance *a priori* sur les séquences que nous étudions. Dans le domaine informatique, ce travail est dédié à la définition d'observations structurées (les k-d-k-mers) permettant la localisation en contexte d'irrégularités, ainsi qu'à la description d'une méthode de classification utilisant plusieurs classifieurs. Dans le domaine biologique, cet article décrit une méthode pour prédire des ensembles de gènes co-régulés, donc susceptibles d'avoir des fonctions liées en réponse à des conditions environnementales spécifiques.

1 Introduction

L'accumulation des séquences issues des projets de séquençage oblige la mise en œuvre de méthodes de fouille de données efficaces pour comprendre les mécanismes impliqués dans l'expression, la transmission et l'évolution des gènes. Nous nous intéressons aux modèles stochastiques et méthodes classificatoires permettant de prédire les séquences promotrices et autres petites séquences régulatrices chez les bactéries. Une manière de cerner notre ignorance vis à vis des motifs et segments d'ADN impliqués dans les mécanismes décrits plus haut est de modéliser l'évolution et la structuration du génome par des processus stochastiques capables d'apprentissage statistique nécessitant un minimum de connaissances *a priori*. Ces modèles stochastiques sont utilisés comme révélateurs d'organisations locales remarquables qu'un expert doit interpréter.

Nous nous intéressons à la localisation de sites de fixation de protéines. Ces sites de fixation – appelés TFBS (*Transcription Factor Binding Sites*) ou encore promoteurs transcriptionnels – sont constitués de trois séquences adjacentes de nucléotides :

$$N_x\text{—}N_y\text{—}N_z \quad \text{avec} \quad \begin{array}{c} N \in \{A, C, G, T\} \\ 3 \leq x, z \leq 9 \\ 0 \leq y \leq 25 \end{array}$$