Classification non supervisée par HMM de sites de fixation de facteurs de transcription chez les bactéries

Sébastien Hergalant, Bertrand Aigle, Bernard Decaris, Jean-François Mari, Pierre Leblond

Classification non supervisée par HMM de sites de fixation de facteurs de transcription chez les bactéries

Sébastien HERGALANT[†‡] Bertrand AIGLE[‡] Bernard DECARIS[‡]
Jean-François MARI[†] Pierre LEBLOND[‡]

[†] LORIA équipe Orpailleur, BP 239, 54506 Vandœuvre-lès-Nancy, France.
[‡] Laboratoire de Génétique et Microbiologie, UMR UHP-INRA 1128, IFR 110, 54506 Vandœuvre-lès-Nancy, France.

Résumé

Nous développons des méthodes de fouille de données basées sur l'utilisation de modèles Markoviens du second ordre adaptés à l'étude des génomes. Ceux-ci réalisent une segmentation pouvant être observée sous la forme d'un signal stochastique traduisant l'organisation et la structure des motifs d'ADN sous-jacents [1]. En d'autres termes, ces modèles de Markov cachés (HMM pour Hidden Markov Model) sont sensibles à la fois aux variations du contexte nucléotidique et aux taux d'observation des mots dans l'ADN. Aucune hypothèse *a priori* n'est effectuée sur le contenu génétique des séquences étudiées. La modélisation du corpus de séquences est réalisée par une étape d'apprentissage automatique qui produit une classification non supervisée, selon des critères probabilistes, des segments nucléotidiques observés sur les différents états des HMM.

Une première étape d'apprentissage sur les séquences chromosomiques complètes des bactéries actinomycètes *Streptomyces coelicolor*, *Streptomyces avermitilis* et *Mycobacterium tuberculosis* permet l'obtention de trois classes de HMM décrivant chacune un génome. Ces modèles sont par la suite utilisés pour segmenter des séquences de 50 à 500 kb provenant des génomes respectifs. Lors de ce processus, certaines chaînes d'états cachés des HMM décrivent des fragments génomiques comme les gènes et les séquences intergéniques alors qu'une autre chaîne se spécialise sur la distribution de motifs d'ADN locaux particuliers (observations *a posteriori*). Ceux-ci correspondent à des mots de 5 à 12 nucléotides présents à des fréquences inhabituelles dans les régions intergéniques. Chez *S. coelicolor*, la classification de 2500 de ces motifs, issus d'une extraction automatique et identifiés dans 1,2 Mb d'ADN génomique, indique que 7% correspondraient à des sites de reconnaissance de facteurs sigma connus (SigR, SigB, WhiG, HrdB), et 5% à des sites de fixation du ribosome ou des terminateurs de transcription potentiels. Concernant le régulon SigR/SigH (réponse au stress oxydatif chez les *Streptomyces*/*M. tuberculosis*) [3, 2], la mise en œuvre de cette approche a permis de détecter tous les promoteurs déterminés biologiquement chez *S. coelicolor* et *M. tuberculosis*. D'autres motifs (88%) correspondraient à des séquences déjà identifiées lors d'études de compilations de mots rares dans les espaces intergéniques et pourraient correspondre à de nouvelles séquences régulatrices (notamment des promoteurs transcriptionnels). Enfin, certains de ces motifs ne peuvent être corrélés à des rôles biologiques connus ou prédits à ce jour. Leur classification pourrait mettre en évidence des groupes à propriétés communes et viserait à définir des motifs promoteurs puis, à terme, des réseaux de gènes co-régulés. Cette perspective originale est particulièrement intéressante compte-tenu du nombre important de facteurs sigma prédits chez les *Streptomyces* (65 chez *S. coelicolor* et 60 chez *S. avermitilis*).

Références

[1] HERGALANT (S.), AIGLE (B.), DECARIS (B.), LEBLOND (P.) et MARI (J.-F.), « Fouille de données à l'aide de HMM : application à la détection de réitérations intragénomiques », dans *JOBIM'02*, 2002, p. 269–73.

[2] MANGANELLI (R.), VOSKUIL (M.), SCHOOLNIK (G.), DUBNAU (E.), GOMEZ (M.) et SMITH (I.), « Role of the extracytoplasmic-function sigma factor sigma(H) in *Mycobacterium tuberculosis* global gene expression », *Mol. Microbiol.*, *45*, 2002, p. 365–74.

[3] PAGET (M. S.), MOLLE (V.), COHEN (G.), AHARONOWITZ (Y.) et BUTTNER (M.), « Defining the disulphide stress response in *Streptomyces coelicolor* A3(2) : identification of the sigmaR regulon », *Mol. Microbiol.*, *42*, n° 4, 2001, p. 1007–20.