Une approche phylo-HMM pour la recherche de séquences

Jean-Baka Domelevo-Entfellner[1,2] et Olivier Gascuel[1]

[1] Méthodes et Algorithmes pour la Bioinformatique
LIRMM, CNRS-UM2, 161 rue Ada, 34392 Montpellier Cedex 5, France
[2] ENS Cachan, Antenne de Bretagne, 35170 Bruz, France (domelevo@lirmm.fr)

Abstract: *We introduce a new type of phylogenetic Hidden Markov Model, combining the strength of usual HMM and the knowledge of the phylogeny of a family of sequences. We use such models to look into the genome of a target species for members of the sequence family. Our results on some 690 protein families show a better sensitivity and a better specificity when compared to standard profile HMM or Blast searchs.*

Keywords: Genomics, sequences, HMM, phylo-HMM, phylogeny, models, homology.

1 Introduction

Les HMM [3] sont des modèles probabilistes qui permettent notamment de décrire une classe de séquences. Dans ce cas, ils sont généralement construits à partir d'alignements multiples, et servent à rechercher si un génome contient des séquences de la classe visée. L'exemple typique d'une telle application est la base Pfam [1], qui décrit à l'aide de HMMs environ 10.000 domaines protéiques dont la structure et/ou la fonction sont bien documentées. L'approche HMM modélise bien les profils biochimiques attachés à chaque site des séquences, et elle permet de distinguer les zones de gaps et les blocs conservés. En revanche, elle perd une part de l'information évolutive contenue dans l'ensemble des séquences appartenant à l'alignement multiple. Un HMM ne décrit pas les proximités évolutives entre séquences et reste le même quel que soit le génome visé, même si dans l'alignement multiple original on a des séquences de génomes très proches du génome cible. En ceci, les HMM diffèrent des approches d'alignement simple de type Blast, qui donnent d'excellents résultats lorsque la requête et la cible sont phylogénétiquement proches l'une de l'autre.

Nous proposons ici une solution pour à la fois bénéficier des caractéristiques globales de la famille des séquences recherchées et bénéficier des proximités évolutives entre la séquence cible et les séquences de référence. Cette approche est basée sur la notion de phylo-HMM [10] qui combine phylogénie et HMM et constitue une modélisation plus complète d'un alignement multiple que les phylogénies ou HMM usuels pris isolément. Notre approche est cependant différente de l'utilisation standard des phylo-HMM [10,9] puisque notre but n'est pas de modéliser un alignement mais de rechercher une séquence particulière au sein d'un organisme dont la position phylogénétique est connue. Dans la suite, nous rappelons rapidement ce que sont les modèles attachés aux HMM, aux phylogénies et aux phylo-HMM. Nous décrivons ensuite notre approche et sa mise en œuvre, et nous étudions ses performances sur un jeu de 690 familles protéiques et la recherche de protéines humaines. Cette étude est un cas d'école, au sens où les séquences cibles sont déjà identifiées. Elle nous permet de comparer une approche HMM classique, une recherche d'homologues de type Blast et notre approche qui vise à combiner le meilleur des HMM et des modèles phylogénétiques. Les résultats montrent que nos phylo-HMM présentent une meilleure spécificité et une plus grande sensitivité que les HMM usuels ou que la recherche d'homologie par Blast.