Une approche phylo-HMM pour la recherche de séquences

Jean-Baka Domelevo Entfellner, Olivier Gascuel

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.

2 HMM, phylogénies et phylo-HMM

On décrit ici les grands principes de trois modèles probabilistes qui permettent de décrire un alignement multiple. Celui-ci est préalable à la modélisation, le modèle n'étant pas destiné à l'obtenir mais à en représenter les caractéristiques ou à l'exploiter. Ces caractéristiques sont de deux ordres : (1) des caractéristiques longitudinales qui correspondent aux suites d'acides aminés typiquement rencontrés dans la famille (par exemple, un site aliphatique, suivi d'un site hydrophobe, suivi d'une cystéine, suivi d'une zone indéterminée contenant souvent des gaps, etc.), et (2) des caractéristiques verticales ou évolutives, qui décrivent les colonnes de l'alignement et indiquent par exemple que le site est plutôt conservé et contient deux acides aminés correspondant à deux clades distincts. Les HMM sont l'outil standard pour représenter les aspects longitudinaux, tandis les phylogénies modélisent les colonnes de l'alignement et que les phylo-HMM combinent les deux types de description.

2.1 HMM

Un modèle de Markov caché (HMM) est un modèle probabiliste qui décrit en termes d'automate d'états la structure d'une séquence biologique. Il se compose d'états i et de transitions $a_{i \to j}$. À chacun des états correspond une distribution $(P(X|i))$ donnant la probabilité pour l'automate de produire la lettre X lorsqu'il se trouve dans l'état i. Les transitions $a_{i \to j}$ représentent la probabilité de passer de l'état i à l'état j. Les paramètres constitutifs du modèle sont donc de trois sortes : (1) la topologie de l'automate (états et arêtes), (2) les distributions de probabilité dans chacun des états, (3) les probabilités de transition entre états. Les lettres émises par les états seront ici les 20 acides aminés, mais le principe s'étend à d'autres alphabets (ADN, codons, ...).

Une fois un HMM établi, l'utilisation principale consiste à l'utiliser pour lire des séquences et leur donner un score. Ce score est la probabilité de générer la séquence au moyen du modèle considéré. Grâce à des algorithmes efficaces (par exemple l'algorithme "forward-backward", qui relève de la programmation dynamique), ce score s'obtient en temps quadratique.

Construire un HMM, c'est assigner des valeurs à chacun de ses paramètres en se basant sur des statistiques extraites des séquences à modéliser, grâce à des techniques d'apprentissage de type EM [2]. Les HMM profils sont des HMM particuliers, fournissant une architecture canonique. Pour construire un tel modèle, on commence par aligner les séquences d'apprentissage, puis on en déduit une séquence de consensus. C'est la longueur n de cette séquence de consensus qui va déterminer la longueur du HMM profil, composé de n motifs répétés de trois états (match, insertion et délétion). Un exemple d'architecture ainsi contrainte est l'architecture Plan 7 adoptée par les outils HMMER ([4], voir aussi Figure 2). À l'intérieur de chacun des états de match ou d'insertion, on trouve une distribution de probabilités d'émission des 20 acides aminés apprise sur l'alignement. Ainsi, on passe naturellement d'un alignement multiple à un HMM profil, qui implémente de façon directe la vision longitudinale d'un alignement multiple comme succession d'états avec des variations probabilisées.

2.2 Phylogénies

Là où les modèles de Markov fournissent une vision longitudinale des séquences, les modèles phylogénétiques apportent des informations supplémentaires basées sur les liens évolutifs entre séquences. Ces dernières sont reliées par un arbre dont les feuilles constituent les séquences actuelles et les nœuds internes les séquences ancestrales. La longueur des branches correspond au nombre de substitutions par site entre les deux nœuds reliés. Les événements de susbtitution sont modélisés par

un processus markovien en temps continu [3]. Un arbre phylogénétique est défini par plusieurs paramètres : (1) la topologie de l'arbre (généralement binaire), (2) les longueurs de branches, (3) la ou les matrices de substitution qui agissent le long des branches de l'arbre.

Le calcul de la vraisemblance d'une phylogénie s'effectue de façon récursive [3] : les vraisemblances conditionnelles au nœud u sont obtenues à partir des vraisemblances conditionnelles aux nœuds fils v et w. L'information véhiculée par les vecteurs de vraisemblances conditionnelles remonte donc des feuilles (qui sont les séquences observées) à la racine, point où l'on calcule la vraisemblance de l'arbre et donc du modèle phylogénétique. La complexité de l'algorithme est linéaire sur la taille de l'alignement (nombre de séquences $\times$ longueur des séquences).

Les modèles phylogénétiques permettent d'expliquer l'évolution (verticale) des sites, mais ne prennent pas du tout en compte l'aspect longitudinal, puisque les sites sont supposés évoluer de manière indépendante.

2.3 Phylo-HMM

Après les premiers travaux de Thorne, Goldman et Jones en 1996 [11], c'est à Siepel et Haussler [10] que l'on doit la formalisation dans un cadre général d'un nouveau type de modèle de Markov caché incluant des aspects phylogénétiques. Là où chacun des états d'un HMM classique produit un caractère d'une séquence, un phylo-HMM produit une colonne d'un alignement. Avec les phylo-HMM il ne s'agit plus d'affecter un score à une séquence isolée mais à un alignement de séquences. Ainsi, on trouve en chaque état du phylo-HMM non plus une distribution de probabilités sur les vingt acides aminés, mais un modèle phylogénétique sous lequel on peut calculer, grâce à l'algorithme de Felsenstein [3], la probabilité de la colonne de l'alignement en cours d'examen.

Les phylo-HMM présentent un intérêt certain pour l'annotation d'un alignement. On peut par exemple distinguer des états conservés d'états non conservés en ayant deux modèles phylogénétiques distincts, les longueurs de branche de l'état "conservé" étant plus courtes que celles de l'état "non conservé". Ainsi, Siepel et al. [9] ont appliqué un phylo-HMM de ce type (Figure 1) à des alignements de séquences appartenant à quatre groupes d'eukaryotes, établissant une typologie des éléments conservés selon les groupes d'espèces et/ou le degré de conservation. On peut également rendre compte des différentes positions de codon, avec des modèles phylogénétiques (ou états du phylo-HMM) présentant différentes matrices de substitution [7].

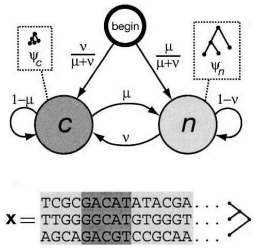

Fig. 1. Un exemple de phylo-HMM à deux états, représentant les positions conservées Ψ_c et non-conservées Ψ_n. La vraisemblance d'un alignement est calculée comme dans les HMM pour ce qui concerne l'aspect séquentiel, tandis que la vraisemblance d'une colonne de l'alignement est obtenue grâce au modèle phylogénétique correspondant à l'état considéré.

3 Une approche phylo-HMM pour la recherche de séquences

3.1 Présentation

Nous expérimentons ici l'idée d'un modèle statistique qui combine la facilité d'usage des HMM et l'intégration de connaissances phylogénétiques. Il s'agit de fournir un score d'alignement d'une séquence *vs* un modèle conçu à partir d'un ensemble de séquences dont on connaît les liens phylogénétiques. Le modèle que nous avons développé permet, à partir de la connaissance de la place de l'espèce cible dans la phylogénie des espèces utilisées pour fabriquer notre modèle, d'aller rechercher des candidats potentiels à l'homologie dans l'espèce cible. L'idée consiste à modifier un HMM construit à partir d'une famille de séquences pour qu'il tienne compte à la fois de la phylogénie des séquences sur lesquels il a été appris et de la position supposée de l'espèce cible dans cette même phylogénie. Pour cela :

1. nous réalisons un alignement multiple des séquences connues de la famille,
2. nous construisons un HMM profil à partir de cet alignement,
3. nous positionnons la séquence cible (inconnue) dans la phylogénie (connue) de la famille,
4. nous calculons grâce à l'algorithme de Felsenstein les probabilités a posteriori de chaque acide aminé en chacun des états match,
5. nous utilisons ces probabilités a posteriori en lieu et place des probabilités d'émission du HMM d'origine, pour rechercher la séquence cible. On a ainsi contextualisé le HMM profil pour l'adapter à la recherche d'homologues chez un taxon dont on connaît la position phylogénétique.

Ainsi (Figure 2), l'approche revient à chercher au sein du génome cible les séquences qui maximisent le score d'un phylo-HMM dont sont connues la structure, la phylogénie, et toutes les séquences à l'exception de la cible. On décrit donc bien les aspects longitudinaux tout en préservant l'information phylogénétique qui rend plus probables les résidus des voisins de la cible que ceux correspondant à des espèces éloignées. L'utilisation d'une loi gamma discrète des vitesses d'évolution des sites [12] permet de rendre compte du fait que certains sites sont très conservés, tandis que d'autres le sont beaucoup moins et "étalent" la distribution attendue des acides aminés.

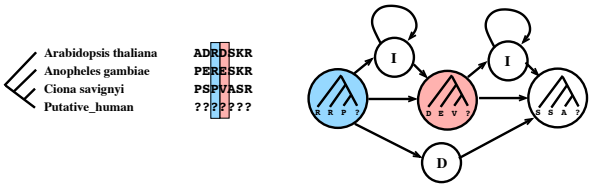

Fig. 2. Principe de nos phylo-HMM : représentation partielle de trois nœuds successifs.

3.2 Mise en œuvre et implémentation

Pour tester notre modèle, nous avons utilisé des famille protéiques alignées, issues du projet TreeFam [8]. Ces familles contiennent des gènes appartenant à diverses espèces animales ainsi qu'à deux levures (*S. cerevisiae* et *S. pombe*) et deux plantes (*A. thaliana* et *O. sativa*). Les familles dites "A" ont été revues et annotées manuellement. Ce sont elles que nous avons utilisées ici.

Notre travail pour chacune de ces quelque 860 familles a consisté à :

1. sélectionner les 690 familles ne contenant qu'un seul gène humain (afin de simplifier cette phase de test),
2. élaguer l'arbre correspondant pour supprimer tous les gènes appartenant aux espèces de mammifères, d'oiseaux, de batraciens ou de poissons. Après cet élagage, l'espèce la plus proche de l'homme est dans certaines familles *Ciona savignyi*, une ascidie. Dans les autres c'est un insecte, un ver ou une espèce plus distante encore. Au cours de cet élagage, nous préparons l'introduction d'un taxon *Putative_human* que l'on enracine sur sa position originelle, avec les mêmes longueurs de branche que dans la phylogénie initiale. Nous tirons donc parti du fait que la phylogénie est connue pour les espèces en question.
3. réaligner les séquences restantes grâce à `muscle` [5],
4. construire un HMM profil à partir de cet alignement avec `hmmbuild` [4],
5. modifier les états *match* de ce HMM grâce à notre approche phylogénétique, basée sur les probabilités a posteriori des acides aminés, calculées par l'algorithme de Felsenstein implémenté dans PhyML [6]. Nous appelons "phylo-HMM" le HMM obtenu à l'issue de cette étape, bien qu'il s'agisse d'un HMM profil contextualisé grâce à notre approche ; ceci afin de le distinguer du HMM standard construit en 4.
6. tenter de retrouver dans le protéome humain (Ensembl v.48), à l'aide du HMM profil construit à l'étape 4 et du phylo-HMM, le gène initialement supprimé. Nous utilisons pour cela `hmmsearch` de la suite HMMER,
7. faire une recherche d'homologie classique en tentant d'aligner sur le protéome humain la séquence restante phylogénétiquement la plus proche de l'espèce humaine. Cette recherche d'homologie a été effectuée à l'aide de Blast.

3.3 Résultats expérimentaux sur 690 familles Treefam et le génome humain

Afin de mesurer l'efficacité de notre approche, nous avons comparé le nombre de prédictions de protéines obtenues par les trois méthodes (HMM, Blast, Phylo-HMM), en considérant comme prédictions correctes les protéines listées dans Ensembl v.48 (décembre 2007) correspondant aux transcrits du gène humain originellement présent dans la famille de gènes considérée. Lorsqu'on en attend k, on décompte les prédictions correctes dans les k premiers hits renvoyés par les trois méthodes. Nous obtenons une augmentation du pourcentage de réussite de 86,0% à 90,5% lorsqu'on passe du HMM profil standard au phylo-HMM contextualisé, avec 1280 protéines prédites par le premier contre 1348 pour le second, sur un total de 1489 prédictions attendues. Le nombre de prédictions réalisées par Blast est sensiblement équivalent à celui donné par le HMM profil standard (1293).

Dans un second temps, nous avons comparé la spécificité de ces différentes approches, en calculant pour chaque famille la différence de log-vraisemblance entre le moins bon des vrais positifs et le meilleur des faux positifs. Lorsque la méthode intervertit vrais et faux positifs, la différence est négative ; un écart positif important montre à l'inverse que la méthode discrimine bien le vrai du faux. Nous avons obtenu de bons résultats (Figure 3), montrant une meilleure discrimination entre vrais et faux positifs ainsi qu'un plus faible taux global de mélange dans le cas de nos phylo-HMM. En effet, la distribution des écarts issus de la méthode "phylo-HMM" est décalée vers la droite par rapport à la méthode utilisant des HMM profils standard, signe d'un plus grand pouvoir séparateur. De plus, ces bons résultats sont robustes à l'éloignement phylogénétique de l'espèce cible par rapport aux espèces présentes dans l'alignement, comme on le voit dans les graphiques présentés en figure 3. Bien que les résultats donnés par les HMM profils et l'approche Blast ne soient pas directement comparables entre

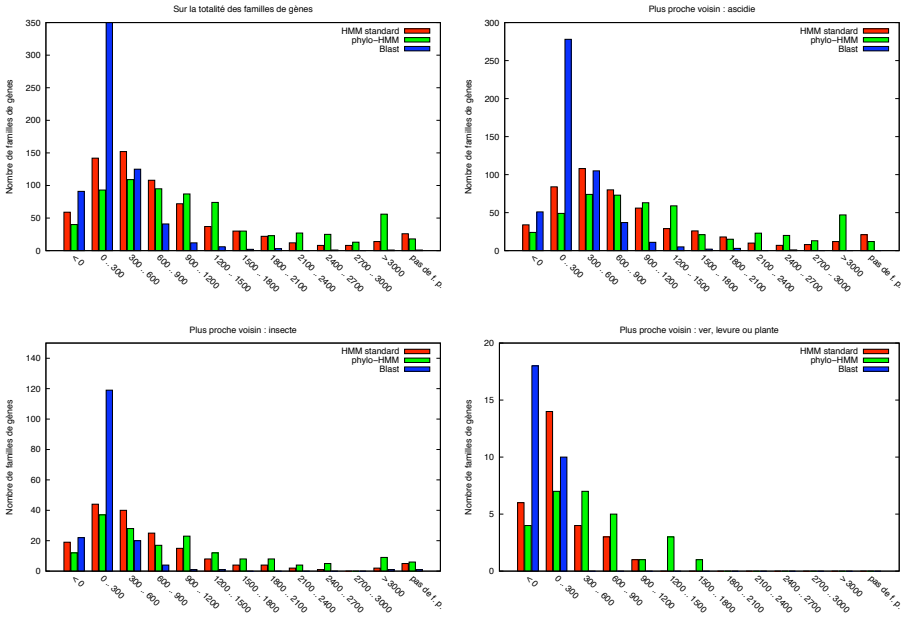

Fig. 3. Spécificité : distribution des gaps de score entre vrais et faux positifs, lors de la recherche de hits dans le protéome humain à partir de 690 familles Treefam. En haut à gauche, toutes les familles sont comptabilisées. Les trois autres histogrammes présentent différentes situations, avec un éloignement phylogénétique croissant.

eux (la formule de calcul du score n'étant pas la même), les résultats montrés en figure 3 illustrent bien le comportement attendu de Blast. Lorsque la séquence cible est suffisamment proche de la séquence de référence (ascidie ou insecte, ce qui correspond à environ 50% d'identité), Blast fonctionne bien avec des performances comparables aux HMM profils. Des résultats non montrés ici indiquent que Blast et phylo-HMM sont quasi-identiques (et meilleurs que les HMM profils) avec des séquences plus proches encore ($>$ 70% d'identité). Mais lorsque l'écart augmente (ver, levure ou plante, soit $\sim$ 25-45% d'identité), les performances de Blast se dégradent avec un grand nombre de confusions entre vrais et faux positifs, et des résultats bien en deçà des HMM et phylo-HMM.

Les résultats de cette étude indiquent donc que les phylo-HMM satisfont bien les attentes en étant aussi performants que Blast lorsque l'éloignement phylogénétique est faible, et sensiblement plus performants que les HMM profils dans tous les cas de figure.

4 Discussion

Notre approche phylo-HMM est une synthèse de deux modèles bien établis : les modèles de Markov cachés et les arbres phylogénétiques. Elle se distingue des phylo-HMM usuels en ce que l'objectif n'est pas de décrire un alignement à l'aide d'un phylo-HMM de petite taille, mais de chercher une séquence au sein d'un génome de position phylogénétique connue, à l'aide d'un phylo-HMM de grande taille représentant à la fois la classe de séquences et les relations avec la cible. Un des intérêts de notre approche est qu'elle produit un HMM classique, pouvant ensuite être traité par les outils

classiques (par exemple HMMER) comme que tout autre HMM obtenu directement à partir d'un alignement.

Les résultats que nous exposons ici ne sont que des résultats de type "banc d'essai" que nous étendons actuellement à la recherche de séquences dans le patrimoine d'une espèce cible mal connue (levure, Plasmodium, ...) et relativement distante des espèces ayant servi à construire l'alignement à l'origine du HMM. Les premiers résultats obtenus ici (en particulier la robustesse à l'éloignement phylogénétique, cf. Fig. 3) laissent augurer de bonnes performances de la méthode sur ce type de cas réel difficile.

Remerciements

Nous remercions Samuel Blanquart, Laurent Bréhélin, Jean-François Dufayard, Stéphane Guindon, Heng Li et Nicolas Terrapon pour leur aide et les nombreuses discussions que nous avons eues ensemble. Ce travail est soutenu par le projet ANR PlasmoExplore (MD&CA - 06).

Références

[1] A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, D. J. Studholme, C. Yeats, and S. R. Eddy. The pfam protein families database. *Nucleic Acids Res*, 32 Database issue, January 2004.

[2] A. Dempster, N. Laird, and D. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1) :1–38, 1977.

[3] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis*. Cambridge University Press, 2002.

[4] S. Eddy, M. W. U. S. Louis, S. of Medicine, D. of Genetics, and N. H. G. R. I. (US. *HMMER Profile Hidden Markov Models for Biological Sequence Analysis*. Washington University School of Medicine, 1992.

[5] R. C. Edgar. MUSCLE : multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.*, 32(5) :1792–1797, 2004.

[6] S. Guindon and O. Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, 52(5) :696–704, October 2003.

[7] S. Mccauley and J. Hein. Using hidden markov models and observed evolution to annotate viral genomes. *Bioinformatics*, 22(11) :1308–1316, 2006.

[8] J. Ruan, H. Li, Z. Chen, A. Coghlan, L. J. M. Coin, Y. Guo, J.-K. Heriche, Y. Hu, K. Kristiansen, R. Li, T. Liu, A. Moses, J. Qin, S. Vang, A. J. Vilella, A. Ureta-Vidal, L. Bolund, J. Wang, and R. Durbin. TreeFam : 2008 Update. *Nucl. Acids Res.*, 36(suppl_1) :D735–740, 2008.

[9] A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, 15(8) :1034–1050, 2005.

[10] A. Siepel and D. Haussler. Combining phylogenetic and hidden markov models in biosequence analysis. In *RECOMB '03 : Proceedings of the seventh annual international conference on Research in computational molecular biology*, pages 277–286, New York, NY, USA, 2003. ACM.

[11] J. Thorne, N. Goldman, and D. Jones. Combining protein evolution and secondary structure. *Mol Biol Evol*, 13(5) :666–673, 1996.

[12] Z. Yang. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol Biol Evol*, 10(6) :1396–1401, 1993.