MODEL_PHYLO

(Projet de Recherche 2004)

ACI IMPBIO

Modélisation stochastique en phylogénie moléculaire

http://kimura.univ-montp2.fr/ModelPhylo

CONTEXTE

La phylogénie moléculaire est la discipline visant à reconstruire l'histoire des gènes et des espèces par l'analyse comparative de séquences d'ADN ou de protéines. Elle joue un rôle central pour l'étude de la biodiversité et l'analyse comparative des gènes et des génomes. L'approche statistique en phylogénie moléculaire, impliquant la modélisation de l'évolution des séquences par des processus stochastiques, a connu un succès croissant au cours des 10-20 dernières années, et est maintenant largement reconnue comme la plus appropriée [1,2]. Le présent projet vise au développement de méthodes statistiques d'analyse phylogénétique utilisant des modèles complexes, et réalistes, de l'évolution des séquences.

OBJECTIFS

Le développement et l'implémentation de modèles Markoviens réalistes pour représenter l'évolution des molécules répond à deux objectifs principaux:

1. **Reconstruire les phylogénies de manière plus fiable**.

L'usage de modèles se rapprochant des processus évolutifs réels permet de diminuer le biais des estimateurs d'arbres phylogénétiques. C'est notamment une nécessité pour les phylogénies très anciennes, impliquant des niveaux de divergence élevés entre séquences [3,4] .

2. **Appréhender la fonction via l'évolution**.

Modéliser explicitement les phénomènes évolutifs liés à la structure et le fonction des molécules, tels que le niveau de contrainte fonctionnelle ou les interactions entre sites, permet d'éclairer le rôle actuel des molécules dans les différentes espèces grâce à la reconstruction des processus ayant gouverné leurs diversification.

Les applications biologiques envisagées sont nombreuses, tirant partie des bases de données de séquences maintenues au Pôle Bioinformatique Lyonnais. Ce projet vise en particulier à résoudre la question de l'origine et de l'évolution de la thermophilie, c'est-à-dire la capacité à vivre à très haute température, depuis l'origine des procaryotes, bactéries et archae. Celà implique notamment de prendre en compte les variations de compositions en nucléotides (pour les ADN/ARN) ou en acides-aminés (pour les protéines), qui sont des marqueurs de l'écologie thermophile vs. mésophile.

RESULTATS

• **Modèles bayésiens hétérogènes**: Un modèle non-homogène de l'évolution des séquences protéiques a été développé dans le formalisme bayesien [5,6]. Ce modèle prend en compte les différences de mode évolutif, représenté par la composition en amino-acides stationnaire, de différents sites de la molécule analysée. Une extension actuellement en cours de développement autorise ede plus la composition en amino-acides d'une protéine a changer au cours du temps et entre lignées. Des évènements ponctuels de changements de mode évolutif se produisent, rendant le processus non-stationnaire.

• **Maximum de vraisemblance et non-stationnarité**: Un algorithme rapide pour l'estimation de phylogénies au maximum de vraisemblance sous un modèle non-stationnaire de l'évolution des séquences a été implémenté. Il fait la synthèse des méthodes PHYML[7] et NHML précédemment développées par notre groupe. Il permettra d'appréhender l'évolution de la thermophilie au travers des variations des compositions en bases ancestrales de l'ARN ribosomique des procaryotes.

• **Cartographie des substitutions et coévolution:** Une méthode statistique pour la cartographie sur un arbre des évènements de substitution a été développée. La comparaison des cartographies de deux sites permet ensuite de détecter d'éventuels processus de coévolution, l'idée étant que deux sites qui coévoluent fourniront des cartes superposables. Une validation de la méthode sur des données d'ARN ribosomiques bactériens a été réalisée [8], et une extension de la méthode au cas des protéines est en cours.

• **Les covarions en évolution moléculaire**: Les variations de vitesse d'évolution site-spécifiques nous renseignent sur les variations de contraintes fonctionnelles qui s'y appliquent, et donc indirectement sur l'adaptation des molécules. Un modèle précédemment proposé fait actuellement l'objet de dévelopements algorithmiques [9], et une extension des méthodes de cartographie est envisagée pour les événements de changements de vitesse.

LABORATOIRES IMPLIQUES

GPIA / UMR 5171 / Montpellier
LIRMM / UMR 5506 / Montpellier
BBE / UMR 5558 / Lyon
University College London, Royaume Unis

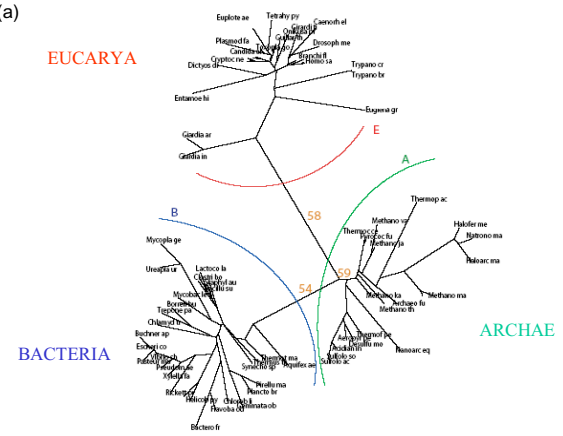
PUBLICATIONS PRINCIPALES

•[1] N. Galtier, O. Gascuel, A. Jean-Marie. An introduction to Markov models in molecular evolution, In R. Nielsen, editor, *Statistical Methods in Molecular Evolution,* Springer, *2005*
•[2] D. Bryant, N. Galtier, M.A. Poursat. Likelihood calculation in molecular phylogeny. In O. Gascuel, editor, Mathematics of Phylogeny and Evolution, Oxford University Press, *2005.*
•[3] F. Thomarat, C.P. Vivarès, M. Gouy. Phylogenetic analysis of the complete genome sequence of Encephalitozoon cuniculi supports the fungal origin of Microsporidia and reveals a high frequency of fast-evolving genes. *Journal of Molecular Evolution 59:780-791, 2004*
•[4] H. Philippe, N. Lartillot, H. Brinkmann. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa and Protostomia. *Molecular Biology and Evolution 22:1246-1253, 2005*
•[5] N. Lartillot, H. Philippe. STAR : A bayesian mixture model for across-sites heterogeneities in the aminbo-acid replacement process, *Molecular Biology and Evolution, 21:1095-1109, 2004*
•[6] N. Lartillot, H. Philippe. Computing Bayes factor using thermodynamic integration. *Systematic Biology, sous presse, 2005*
•[7] S. Guindon, F. Le Thiec, P. Duroux, O. Gascuel. PHYML online: a web server for fast maximum likelihood phylogenetic inference. *Nucleic Acid Research, sous presse, 2005*
•[8] J. Dutheil, T. Pupko, A. Jean-Marie,N. Galtier. A model-based approach for detecting coevolving positions in a molecule, *Molecular Biology and Evolution, sous presse, 2005*
•[9] N. Galtier, A. Jean-Marie. Markov-modulated Markov chains and the covarion process. *Journal of Computational Biology 11:272-733, 2004*

ARNr ET THERMOPHILIE

Une phylogénie universelle reconstruite sur la base de l'ARN ribosomique sous un modèle non-stationnaire de l'évolution a permis d'estimer le taux de GC ancestral de cette molécule sous trois hypothèses concernant la position de la racine (a). Comparés à la distribution actuelle des taux de GC et des températures optimales de croissance, ces estimateurs plaident pour une origine non-thermophile du vivant (b), et de multiples apparitions de la vie à très haute température chez les procaryotes.

(a)



(b)