Résumé

La variabilité des vitesses d'évolution entre sites est un phénomène très répandu au sein des séquences génétiques. Celui-ci est généralement modélisé par une loi gamma dont le paramètre de forme doit être estimé. Nous proposons ici une méthode visant à déterminer la valeur efficace de ce paramètre, c'est-à-dire la valeur la plus adaptée à l'estimation de topologies d'arbres. Nous montrons que (1) les valeurs efficaces conduisent généralement à sous-estimer la variabilité des vitesses et (2), celles-ci offrent une amélioration significative en termes de précision topologique comparé aux cas ou l'inférence d'arbre est réalisée à partir des vraies valeurs (inconnues) du paramètre.

Outre la paramétrisation adéquate des modèles de substitution, l'exploration de l'espace des topologies d'arbres est un point sensible pour bon nombre de méthodes d'inférences de phylogénies. Nous proposons ici une nouvelle méthode pour l'estimation d'arbres de vraisemblances maximales, basée sur la modification simultanée de la topologie de l'arbre et de ses longueurs de branches. Nous montrons que cette approche autorise la construction de topologies particulièrement fiables à partir de l'analyse de plusieurs centaines de séquences.

Mots-clefs : phylogénie, séquences génétiques, loi gamma, maximum de vraisemblance.

Methods and algorithms for a statistical approach in phylogenetics

Variation of substitution rates across sites is widespread among biological sequences. A gamma distribution, defined by a shape parameter, is generally used to model this phenomenon. We propose here a new method to estimate an efficient value of the gamma shape parameter, i.e., the value that is best suited for tree topology estimation. We show that (1) efficient values lead to underestimate the rate variability and (2) the tree topologies that are obtained are more accurate than those deduced from the true (unknown) values of the parameter.

Exploring the tree space is another important issue in phylogenetic inference. We propose a new approach for building maximum likelihood phylogenies. The core of this method is a simple hill climbing algorithm that adjusts tree topology and branch lengths simultaneously. We show that this approach reconstructs very accurate tree topologies from data sets containing hundreds of taxa. The speed of this method also greatly facilitates bootstrap analysis.

Keywords : phylogeny, genetic sequences, gamma distribution, maximum likelihood.

Discipline : biologie

Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier
161, rue Ada - 34392 Montpellier-Cedex 5 France