Theoretical aspects of bioinformatics application in phylogenetics and genomic analysis

Marcelo R.S. Briones, Gerdine Sanson, Adriana Brunstein, Rogério F Mauad and Paulo B. Paiva
DMIP and DIS - Escola Paulista de Medicina
UNIFESP - Howard Hughes Medical Institute
Rua Botucatu 862, 3° andar CEP 04023-062, São Paulo, S.P., Brazil.
marcelo@ecb.epm.br

Abstract

Here we briefly present an overview of our research in two topics: 1. Experimental and theoretical aspects of phylogeny inference and 2. Development of software for molecular phylogenetic inference and phylogenetic genome annotation.

1. *Experimental and theoretical aspects of phylogeny inference*. To evaluate the effects of non-reversibility on compositional base changes and the distribution of branch lengths along a phylogeny, we extended, by means of computer simulations, our previous sequential PCR *in vitro* evolution experiment [1]. In that study a 18S rRNA gene evolved neutrally for 280 generations and a homogeneous non-stationary model of base substitution based on a non-reversible dynamics was built from the *in vitro* evolution data to describe the observed pattern of nucleotide substitutions. Although models based on non-reversible probability matrices were previously proposed, the long-term effects of non-reversibility in neutral evolution remains to be tested. Here, the process was extended to 840 generations without selection, using the model parameters calculated from the *in vitro* evolution experiment, by means of computer simulations. A computer program was developed in C language, for this purpose, and is available at http://compbio.epm.br/ievol. We observed that under a non-reversible model the G+C content of the sequences significantly increases when compared to simulations with a reversible model. The values of mean and variance of the branch lengths also reduce under a non-reversible dynamics although they follow a Poisson distribution. We conclude that the major implication of non-reversibility is the overall decrease of branch lengths, although no transition from a stochastic to an ordered process is observed. According to our model the result of this neutral process will be the increase in the G+C content of the descendant sequences with an overall decrease in the frequency of substitutions.

2. *Development of software for molecular phylogenetic analysis and phylogenetic genome annotation*. We developed two computer programs: iDate and Phylocomp, which will be available soon in web versions and open source code from installation in UNIX computers, respectively. IDate is a Web Tool to Estimate Divergence Dates with Likelihood Ratio Test for Molecular Clock [2]. This program is a Perl script used to estimate the rate of evolution and/or divergence dates on a phylogenetic tree that supports the molecular clock hypothesis. This tool can also deal with partitioned alignment files. IDate take as input two tree files, one with molecular clock enforcement and the other without it (both of them must be supplied by the user and must be in NEXUS/PAUP format). A likelihood ratio test is performed in order to test the molecular clock hypothesis. Assuming that the molecular clock hypothesis is accepted, IDate calculates the patristic distances (i.e., the sum of branch lengths on a path connecting a pair of taxa) for all taxa in the clock tree. After patristic distances calculation, IDate estimates the divergence dates (given a global rate of evolution for the clock tree) *or* the rate of evolution, (given a fossil register *and* its localization in the tree). Results can be visualized on an ordinary Internet browser and downloaded to the user's own computer as a text file.

Phylocomp is a viewer that enables direct visualization of the Gene Phylogeny versus the Species tree for a given query sequence or PFAM family. It presents phylogenetic trees inferred from gene products and the different species that originated them. Sequence data is user entered, by accession or complete record and extracted from the seed sequences for a family of protein domain fragments (PFAM: http://pfam.wustl.edu). The species data is built from a list of species present in the sequence data using the Ribosomal RNA for the listed species retrieved from the RDB (Ribosomal Database). The application was developed in C++, uses the Qt multiplatform C++ GUI application framework and MySQL database. Communication with NCBI for fetching sequence information uses XML. The HMM processing based on HMMer (http://hmmer.wustl.edu/) and the phylogenetic inference based on Tree Puzzle maximum likelihood [3].

Application of molecular phylogenetics methods here described and implemented have enabled our group to contribute to ongoing research on molecular epidemiology of two important human pathogens *Trypanosoma cruzi* [4] and *Candida spp.* [5], by providing a quantitative means to associate the evolutionary history of these organisms and their hosts. This might in the long run contribute to strategies of vaccination and chemotherapy to diseases caused by these microorganisms.

References

[1] Sanson et al., 2002, Mol. Biol. Evol., 19:170-178.
[2] Felsenstein, 1988, Ann. Rev. Genet., 22:521-565
[3] Strimmer and von Haeseler, 1996, Mol. Biol. Evol., 13:964-969.
[4] Kawashita et al., 2001, Mol. Biol. Evol.,18:2250-2259.
[5] Sanson and Briones, 2000, J. Clin. Microbiol., 38: 227-235.