Fitting nonreversible substitution processes to multiple alignments

Von Bing Yap [1]

Keywords: DNA base substitution, maximum likelihood, EM algorithm

Consider a DNA base substitution model on a rooted phylogenetic tree with initial distribution x and rate matrix Q. This process is stationary if x is the equilibrium distribution of Q, i.e., $xQ = 0$, and it is reversible if in addition, the detailed balance condition is satisfied:

$$XQ = Q'X,$$

where Q' denotes the transpose of Q and X is the diagonal matrix with x as its diagonal. Thus, this model allows base composition to evolve and is more general than the reversible models used in routine phylogenetic analysis. Given a multiple alignment related by a rooted tree, the EM algorithm introduced by Holmes and Rubin for fitting reversible models [2] turns out to be an efficient tool for fitting the present nonreversible model. This algorithm is applied to two datasets described by Yang [3]: six primate $\psi\eta$ globin pseudogenes, and mitochondrial DNA from nine primates. The estimate of Q from the pseudogenes is similar to previous estimates by Yang and by Arvestad and Bruno [1]. However, on the mtDNA, the new estimate is very different from Yang's. Generally, the new estimates fit the data better in the sense that the predicted base compositions are closer to the observed base compositions.

If it is known that the root lies on a particular branch, then its distance from a reference node can also be estimated jointly with x and Q, by maximum likelihood. On the pseudogene dataset, the most likely root position is at orangutan, followed very closely by spider monkey, the latter being intuitively more sensible. On the mtDNA dataset, the most likely root position is at the node connecting orangutan to the others. Another local maximum is somewhere on the branch conneting the ancestors of gibbon and crab-eating macaque.

References

[1] Arvestad, L. and Bruno, W. J. 1997. Estimation of reversible substitution matrices from multiple pairs of sequences. In: *Journal of Molecular Evolution* 45:696–703.

[2] Holmes, I. and Rubin, G. H. 2002. An expectation maximization algorithm for training hidden substitution models. In: *Journal of Molecular Biology* 317:753–764.

[3] Yang, Z. 1994. Estimating the pattern of nucleotide substitution. In: *Journal of Molecular Evolution* 39:105–111.

[1] Department of Mathematics, University of California. E-mail: `vonbing@math.berkeley.edu`