Maximum Likelihood Reconstruction of Ancestral Amino-Acid Sequences

Tal Pupko[1] **Itsik Pe'er**[2]
tal@kimura.tau.ac.il izik@math.tau.ac.il

[1] Department of Zoology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel
[2] Department of Computer Science, School of Mathematical Sciences, Tel Aviv University, Ramat Aviv 69978, Israel

Keywords: phylogeny, maximum likelihood, ancestral reconstruction, Chebyshev polynomials

1 Introduction

Maximum-likelihood methods are used extensively in phylogenetic studies [3]. In particular, amino-acid sequences of ancestral species have been inferred using these methods [7]. Such *ancestral reconstruction* tasks aim at identifying either the most likely sequence in a specific ancestor species (*marginal* reconstruction), or the most likely set of ancestral states corresponding to all the ancestral taxa in a given phylogeny (*joint* reconstruction [6]). Joint reconstruction is motivated by studies of phenomena involving several independent lineages, like [8], and is implemented in [6]. However, existing algorithms for this task are exhaustive, and take exponential time. Furthermore, these algorithms assume a naive model of evolution, i.e., a constant substitution rate, whereas [5] shows that models incorporating rate variation among sites are statistically superior.

In this work we: (a) Devise a dynamic programming algorithm for joint reconstruction. The complexity of this algorithm is linear in the number of sequences, but assumes no rate variation among sites.* (b) Present a greedy heuristic for joint reconstruction assuming rate variation among sites. (c) Introduce a speed-up for calculating the replacement probabilities between any two states.

2 Methods

2.1 Dynamic program for constant substitution rate

We use the term $P_{i\to j}(t)$ for the replacement probability of amino-acid i by amino-acid j along a branch of length t. We perform the following procedure for each position, independently. For each tree node x, let $t(x)$ be the length of the branch connecting it to its father, $f(x)$. For each state a, consider the joint reconstruction of the subtree of x, assuming that $f(x)$ has been assigned the state a. Let $C(x,a)$ be the likelihood of this reconstruction. $C(x,a)$ is computed for all x,a by traversing the tree from the leaves to the root, according to the following recursion equation:

$$C(x,a) = \max_{\text{state } b}\{P_{a\to b}(t(x)) \times \prod_{y \text{ is a son of } x} C(y,b)\}$$

Initialization of this dynamic program is simple. If one saves the values attaining the maximum while computing $C(x,a)$, then, by traversing the tree back to the leaves, one can readily reconstruct the state of each node.

*This part of the research will appear in [4].

2.2 Ancestral reconstruction assuming the rate is Γ-distributed among sites

The *ancestral vector* is the vector V of all character assignments at the internal nodes of a tree (in a specific position). By numerically integrating over the range of possible rates, weighing them according to the Gamma distribution, one can evaluate the likelihood of a given V. We search for the most likely V using the greedy hill-climbing heuristic: we iteratively perturb an entry in V, and accept the new entry if the resulting V is more likely. Independent restarts are employed to avoid local maxima.

2.3 Numerical approximation for $P_{i \to j}(t)$

The matrix $\mathcal{P}(t) = [P_{i \to j}(t)]_{\text{states } i,j}$ is theoretically computed according to the formula $\mathcal{P}(t) = \exp(\mathcal{Q}t)$, for some fixed matrix $\mathcal{Q}$ [1]. This implies computing a linear combination of 20 exponents in the eigenvalues of $\mathcal{Q}$, for evaluating $P_{i \to j}(t)$. This computation is the running time bottleneck in CPU-intensive applications for phylogenetic inference, like [6], or the algorithm in section 2.2. We accelerate the computation of $P_{i \to j}(t)$ by numerical approximation using the Chebyshev polynomial series [2]. Each of the 20×20 functions $P_{i \to j}(t)$ of the single variable t, is approximated by a Chebyshev polynomial of degree d, for the range of reasonable t values. All the $400(d + 1)$ polynomial coefficients are precomputed, allowing rapid evaluation of $P_{i \to j}(t)$, which empirically attains a 62-fold speedup. In theory, resulting $P_{i \to j}(t)$ values are only approximated, but in practice the likelihood values computed are essentially identical to those estimated by exact computation.

3 Results - Application for Demonstrating Positive Selection

Reconstruction results when the rate parameter is Γ-distributed among sites is shown to be statistically superior to those obtained under the assumption of rate homogeneity. Using this method, we reevaluated putative parallel and convergent amino-acid replacements in the evolutionary history of 43 lysozyme sequences. Fifteen homoplasic events were inferred, consistent with the hypothesis of positive selection in four lineages leading to foregut fermenters [8].

References

[1] Adachi, J. and Hasegawa, M., Model of amino acid substitution in proteins encoded by mitochondrial DNA, *J. Mol. Evol.*, 42:459–468, 1996.

[2] Burden, R.L. and Faires, J.D., *Numerical analysis*, PWS-KENT, 1989.

[3] Felsenstein, J., Phylogenies from molecular sequences: inference and reliability, *Annu. Rev. Genet.*, 22:521–565, 1988.

[4] Pupko, T., Pe'er, I., Shamir, R. and Graur, D., A fast algorithm for joint maximum likelihood reconstruction of ancestral amino acid sequences, *Mol. Biol. Evol.*, 2000, in press.

[5] Yang, Z., Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods, *J. Mol. Evol.*, 39:306–314, 1996.

[6] Yang, Z., *PAML: A Program Package for Phylogenetic Analysis by Maximum Likelihood*, University College London, 1999. Version 2.0.

[7] Yang, Z., Kumar, S. and Nei, M., A new method of inference of ancestral nucleotide and amino acid sequences, *Genetics*, 141:1641–1650, 1995.

[8] Zhang, J. and Kumar, S., Detection of convergent and parallel evolution at the amino acid sequence level, *Mol. Biol. Evol.*, 14:527–536, 1997.