

# Branch-and-Bound Reconstruction of Ancestral Sequences

Nir Friedman<sup>1</sup>, Itsik Pe'er<sup>2</sup>, Tal Pupko<sup>3</sup>

**Keywords:** ancestral sequence, branch and bound, phylogeny, sequence analysis

## 1 Introduction

The problem of ancestral sequence reconstruction is the statistical inference of sequences that correspond to internal nodes in a phylogenetic tree [1]. *Joint reconstruction* is the task of seeking the most likely set of ancestral states corresponding to all the ancestral taxa, while *marginal reconstruction* aims at inferring the sequence in a specific internal node. In simple probabilistic models of evolution, both tasks can be performed efficiently using dynamic programming [3, 1].

The situation is more complicated in more detailed models of evolution, such as models with *among-site-rate-variation* (ASRV). In these models, one assumes that the rate of evolution can vary among different sites. This is modeled by introducing a latent quantity that models the rate at each site. Maximum likelihood (ML) models incorporating ASRV are statistically superior to those assuming among site rate homogeneity [2]. For example, it was shown that strong support for rodent nonmonophyly results from systematic error associated with the oversimplified assumption of homogeneity [4].

Currently, no efficient algorithm exists for joint ancestral reconstruction in ASRV models. In particular, dynamic programming approaches fail in these models. In this work we devise a branch-and-bound algorithm for joint ancestral reconstruction under ASRV and show that it can find the most likely reconstruction for large phylogenies.

## 2 The Algorithm

As usual, we assume independence of the stochastic process among sites and, hence, restrict the following description to a single site. The input to our problem consists of *phylogeny* (a bifurcating tree annotated with branch lengths), a prior distribution over possible rates, and observations of characters at the leaves (which correspond to the observed letters at this site in current-day taxa). Our aim is to find a joint assignment of characters to internal nodes, whose likelihood is maximal given the observations.

We start by considering dynamic programming solutions to these problems. Such solutions are based on a “divide and conquer” property of standard phylogenetic trees: once we assign a character to an internal node, we break the problem into two independent sub-problems. When we introduce rate variation, this “divide and conquer” property fails—in order to separate the tree into two parts, we need to assign a value to an internal node *and* also fix the rate. Indeed, a dynamic programming for computing the likelihood of observation in ASRV models, uses exactly these joint assignments (to an internal node and to the rate) to recursively decompose the likelihood computation. However, if we want to perform joint reconstruction we cannot use this decomposition. The joint reconstruction requires finding the assignment to the internal nodes that will be most likely for all the rates. This reconstruction can differ from the maximal reconstruction given any particular rate.

Our approach is to *search* the space of potential reconstructions. Given a putative reconstruction or *partial* reconstruction (that assigns values only to some of the internal nodes), we can compute its likelihood given the observation (using the dynamic programming procedure discussed above). Thus, we can

---

<sup>1</sup>School of Computer Science & Engineering, Hebrew University, Jerusalem 91904, Israel. E-mail: nir@cs.huji.ac.il

<sup>2</sup>School of Computer Science, Tel-Aviv University, Ramat-Aviv 69978, Israel. E-mail: izik@post.tau.ac.il

<sup>3</sup>The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo, Japan. E-mail: tal@ism.ac.jp

define a *search space* that consists of partial reconstructions  $\sigma$ . We can move from one reconstruction to another by assigning values to an additional internal nodes. Our aim is to systematically traverse this space and find the full reconstruction with maximum likelihood.

Of course, since there are exponential number of reconstructions, we cannot hope to traverse all of the space. Instead, we use *branch and bound* search. The key idea of such a procedure is to prune regions of the search space by computing bounds on the quality of solutions within the region. In our case, given a partial reconstruction  $\sigma$ , we compute a bound  $B(\sigma) \geq \max_{\sigma' \in \mathcal{C}(\sigma)} P(\sigma' | o)$ , where  $\mathcal{C}(\sigma)$  is the set of all completions of  $\sigma$ . We use the bound as follows: if we already found a reconstruction  $\sigma^*$  whose likelihood is higher than  $B(\sigma)$ , then we do not need consider any completion of  $\sigma$  (since they are provably worse than the best reconstruction). The details of the procedure involve two key components: (a) methods for computing bounds, and (b) strategy for determining the order in which to traverse the space of reconstructions that are still “alive” given the current bounds.

In this work, we examine two types upper bounds. The first is based on the observation that the probability of a partial reconstruction is the sum of the probabilities of the complete reconstructions that are consistent with it. More precisely,

$$\max_{\sigma' \in \mathcal{C}(\sigma)} P(\sigma' | o) \leq \sum_{\sigma' \in \mathcal{C}(\sigma)} P(\sigma' | o) = P(\sigma | o)$$

The second bound, is based on the following simple inequality:

$$\max_{\sigma' \in \mathcal{C}(\sigma)} P(\sigma' | o) = \max_{\sigma' \in \mathcal{C}(\sigma)} \sum_r P(\sigma' | r, o) P(r) \leq \sum_r \max_{\sigma' \in \mathcal{C}(\sigma)} P(\sigma' | r, o) P(r)$$

Observe, that  $\max_{\sigma'} P(\sigma | r, o)$  is the maximum likelihood of an ancestral reconstruction with a constant rate of evolution,  $r$  which can be computed efficiently [3].

The second issue is the strategy for expanding the search. We use marginal reconstruction to find internal nodes for which the marginal distribution choose a definite value. We then assign values to these nodes first. This strategy focuses the search in promising directions, and helps prune out larger regions in subsequent moves.

### 3 Results

Using this algorithm we reconstructed the ancestral amino-acid sequences of the cytochrome *b* gene from 34 taxa. Andrews [5] reconstructed the ancestral amino-acid sequences of Cytochrome *b* in order to detect mutations leading to function changes in the lineage leading to Simian primates under the assumption of rate homogeneity. However, the rate variation in this gene is substantial. Indeed, using our branch and bound method, we show that the ASRV reconstruction is different in 59 positions. These differences reaffirm the need for ASRV reconstruction methods.

### References

- [1] Yang, Z., Kumar, S., and Nei, M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141:1641–1650.
- [2] Yang, Z. 1996. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods *Journal of Molecular Evolution* 39:306–314.
- [3] Pupko, T., Pe'er, I. Shamir, R. and Graur, D. 2000. A fast algorithm for joint reconstruction of ancestral amino-acid sequences. *Molecular Biology and Evolution* 17:890–896.
- [4] Sullivan, J., and Swofford, D. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. *Journal of Mammalian Evolution* 4:77–86.
- [5] Andrews, T. D., Jermiin, L. S., and Eastel, S. 1998. Accelerated evolution of cytochrome *b* in Simian primates: Adaptive evolution in concert with other mitochondrial proteins. *Journal of Molecular Evolution* 47:249–257.