Incremental Update of Phylogenetic Trees Using Hierarchical Modeling

Bing Li, James Cole, and Eric Torng

One of the challenges of applying molecular evolution methods to very large sets of sequence data is that of constructing and *updating* the corresponding phylogenetic tree. Construction of a phylogenetic tree typically requires two steps. First, a multiple sequence alignment of all the sequences must be computed. Second, given this multiple sequence alignment, the phylogenetic tree is constructed. Unfortunately, all known optimal algorithms for performing either of these two operations require excessive amounts of time for typical data sets encountered in practice. This makes recreating such trees and alignments de novo each time new sequences are discovered impractical. Thus, adding new sequences requires an incremental approach building on the pre-existing trees and alignments.

In this abstract, we describe our ongoing investigation of a novel methodology for incrementally updating phylogenetic trees. Our methodology attempts to perform the alignment of a new sequence and its insertion into the existing phylogenetic tree simultaneously. Our approach can best be explained by first describing two alternative approaches for sequence alignment that it generalizes.

One heuristic for adding a new sequence to an existing alignment of a set of sequences is to pairwise align the new sequence with a "best match" sequence in the existing set. This approach is attractive because an optimal pairwise alignment can be computed relatively quickly. However, this approach ignores all the information provided by the remaining sequences and the existing phylogenetic tree. In addition, identifying the "best match" sequence quickly is nontrivial.

An alternative alignment approach is to develop a *single* model of all the sequences in the database, possibly taking into account such additional information as secondary and tertiary structures. This model is then used to align the new sequence. Examples of candidate models include Hidden Markov Models (HMM), profiles, and stochastic context-free grammars (SCFG). While some studies indicate these models can be used to construct good alignments for certain sets of data, this single model approach faces the drawback that a single model is unlikely to accurately represent all members of a diverse set of sequences.

We propose a hierarchical modeling approach as a means for overcoming these problems. In hierarchical modeling, rather than constructing a *single* model of the entire set of sequences, we construct a *hierarchy* of models where each model represents some cluster of sequences. See figure 1 for a graphical depiction. We then align a new sequence with a "best match" model, and we use the "best match" model to identify a candidate location for insertion of the new sequence into the phylogenetic tree. Thus we achieve simultaneous alignment and insertion.

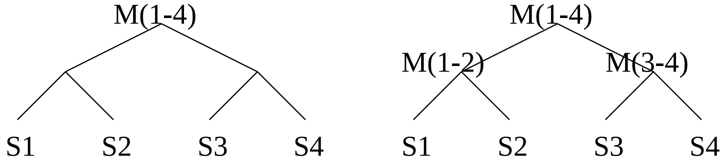

Figure 1: On the left, only a single model is used to represent all 4 sequences. On the right, three models are used to represent clusters of sequences as defined by the phylogenetic tree.

We are currently investigating the feasibility of a hierarchical modeling approach using profiles as the underlying modeling mechanism. Our testbed database, The Ribosomal Database Project at MSU, contains a prokaryotic phylogenetic tree with almost 7,000 leaf nodes. Future work will include experimental evaluation of hierarchical modeling using HMMs and SCFGs as the underlying modeling mechanism.