Information Panel

Choosing a Method for Phylogenetic Prediction

David W. Mount

Adapted from “Phylogenetic Prediction,” Chapter 7, in *Bioinformatics: Sequence and Genome Analysis*, 2nd edition, by David W. Mount. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, USA, 2004.

INTRODUCTION

Three methods—maximum parsimony, distance, and maximum likelihood—are generally used to find the evolutionary tree or trees that best account for the observed variation in a group of sequences. Each of these methods uses a different type of analysis. Programs based on distance methods are commonly used in the molecular biology laboratory because they are straightforward and can be used with a large number of sequences. Maximum likelihood methods are more challenging and require a greater understanding of the evolutionary models on which they are based. Because they involve so many computational steps and because the number of steps increases dramatically with the number of sequences, maximum likelihood programs are limited to a smaller number of sequences. They can be implemented on a supercomputer in order to analyze a greater number of sequences. This article presents an overview for the researcher who has a set of related sequences and wants to analyze them to predict the best trees that depict the phylogenetic relationships among the sequences.

RELATED INFORMATION

Maximum parsimony, distance methods, and the maximum likelihood approach are explained in more detail in the following *CSH Protocols* articles: **Maximum Parsimony Method for Phylogenetic Prediction**, **Distance Methods for Phylogenetic Prediction**, and **The Maximum Likelihood Approach for Phylogenetic Prediction** (all this issue).

A SCHEMA FOR CHOOSING A PHYLOGENETIC PREDICTION METHOD

The flowchart below (Fig. 1) describes the types of considerations that need to be made in choosing a phylogenetic prediction method, but is not intended as a strict guide. It can be useful to try at least two of these methods, which can add confidence to the resulting analysis if the same results are obtained. These methods may find that more than one tree meet the criterion chosen for being the most likely tree. The branching patterns in these trees may be compared to find which branches are shared and therefore are more strongly supported. Phylogenetic analysis using parsimony (PAUP) provides methods for finding consensus trees, and such trees are also calculated by the CONSENSE program in the phylogenetic inference package (PHYLIP). Trees are stored as a tree file that shows the relationships in nested-parentheses notation. Sometimes, branch lengths are also included next to the names; e.g., A:0.05. From this information, a tree-drawing program may be used to produce a tree representation of the data.

1. The sequences chosen for phylogenetic analysis can be either DNA or protein sequences. Different programs and program options are used for each type. The sequences should align with each other along their entire lengths, or else each should have a common set of patterns or domains

FIGURE 1. A flowchart for assessing the best method(s) to use when making phylogenetic predictions. Numbers correspond to the following notes.

that provides a strong indication of evolutionary relatedness. RNA sequences are analyzed by covariation methods and changes in secondary structure.

2. A phylogenetic analysis should be performed when the sequences produce a multiple sequence alignment (msa) in which sequence similarity is apparent by the presence of conserved positions in the columns of the alignment. Some variation in these columns is necessary to produce the phylogenetic analysis, but too much makes the msa itself uncertain and the resulting phylogenetic analysis more difficult. The alignments should not include a large number of gaps that are obviously necessary to align identical or related characters. Some aligned regions may be better conserved than others and the analysis can then be restricted to these conserved regions. In general, phylogenetic methods analyze conserved regions that are represented in all of the sequences. The more similar the sequences are to each other, the better. The simplest evolutionary models assume that the variation in each column of an msa represents single-step changes and that no reversals (A→T→A) have occurred. As the observed variation increases, more multiple-step changes (A→T→G) and reversions are likely to be present. Most phylogenetic analysis programs can correct for such variation according to predicted patterns of mutational change over time. These corrections usually assume a uniform rate of change at all sequence positions over time. Gaps in the msa are usually not scored because there is no suitable model for the evolutionary mechanisms that produce them.
3. This question is designed to select sequences in which there is a clear-cut majority of certain residues in some columns of the msa but also some variation. Some columns in the msa will have the same residue in all sequences; other columns will include variation. The more common residues in the variable columns are taken to represent an earlier group of sequences from which others were derived. If there is too much variation, there will be too many possible ancestral relationships. If the amount of variation is small but definitely present, these sequences are then suitable for maximum parsimony analysis. For parsimony analysis, the trees that best fit the observed variations in the columns of the alignment are found. The best results are obtained when the amount of variation among all pairs of sequences is similar (no very different sequences are present) and when that amount of variation is small. Because the maximum parsimony method has to attempt to fit all possible trees to the data, the method is not suitable for more than 12 sequences because there are too many trees to test. During a maximum parsimony analysis, more than one tree may be found to be equally parsimonious. A consensus tree representing the conserved features of the different trees may then be produced. A maximum likelihood analysis that also produces a tree which best predicts the sequence variation in each alignment column may also be used.

4. The purpose of this question is to select sequences for phylogenetic analysis by distance methods. These methods do not depend on the presence of limited variation in each column of the msa like the maximum parsimony method. In distance methods, the amount of variation between each pair of sequences in the alignment is measured as the fraction of aligned characters that change (the genetic distance). As a result, the method is not as sensitive as maximum parsimony to variation in the aligned columns. Distance methods predict an evolutionary tree based on the degree of difference among pairs of sequences in the msa and can be used when the amount of variation is sufficient to distinguish the sequence pairs on this basis. As distances increase, corrections are necessary for deviations from single-step changes between sequences (see note 2 above). In addition, as distances increase, the uncertainty of alignments also increases, and a reassessment of the suitability of the msa method may be necessary. Sequences with this type of variation may also be suitable for phylogenetic analysis by maximum likelihood methods. Distance methods may be used with a large number of sequences and usually are not significantly affected by variations in rates of mutation over evolutionary time.
5. If there is considerable variation among the sequences in the msa, then the options given in this box should be considered. The msa itself may be improved by using programs that have been shown to produce alignments of more variable sequences, or the more similar regions may be extracted and used for the phylogenetic analysis. Maximum likelihood methods may be used for any set of related sequences, but they are particularly useful when the sequences are more variable. These methods are computationally intense, and computational complexity increases with the number of sequences, since the probability of every possible tree must be calculated (see **The Maximum Likelihood Approach for Phylogenetic Prediction**, this issue). An advantage of maximum likelihood methods is that they include evolutionary models to account for the variation in the sequences.
6. This box addresses how well the sequence variation that is present in the msa supports the tree or trees predicted by the phylogenetic analysis. If the computed tree is based on variation in only a few columns of the alignment or between certain pairs of sequences, it is not representative of the variation in all of the sequences. To address this possibility, the columns in the msa are resampled randomly to produce many new alignments, and a new phylogenetic analysis is then performed on these resampled alignments—a procedure known as bootstrapping. The frequency with which a particular branch in the original tree appears in these new alignments is then given. The more often the branch appears, the better the data in the original alignment support that particular branch in the predicted tree.

Cold Spring Harbor Protocols

Choosing a Method for Phylogenetic Prediction

David W. Mount

Cold Spring Harb Protoc; doi: 10.1101/pdb.ip49

Email Alerting Service

Receive free email alerts when new articles cite this article - [click here](#).

Subject Categories

Browse articles on similar topics from *Cold Spring Harbor Protocols*.

[Alignment of Sequences](#) (46 articles)
[Bioinformatics/Genomics, general](#) (192 articles)
[Computational Biology](#) (100 articles)
[Multiple Sequence Alignment](#) (23 articles)
[Phylogenetic Prediction](#) (19 articles)

To subscribe to *Cold Spring Harbor Protocols* go to:
<http://cshprotocols.cshlp.org/subscriptions>