STATISTICAL PERSPECTIVE OF PHYLOGENETIC ANALYSIS

Reet Kamal[1] Manpreet Kaur[2] Neeshu Shrama[3]

[1]*Department of Computer Science Engineering, RIMT-MAEC, Mandi Gobindgarh, India*
[2]*Department of Computer Science Engineering, RIMT-MAEC, Mandi Gobindgarh, India*
[3]*Department of Computer Science Engineering, RIMT-MAEC, Mandi Gobindgarh, India*

Abstract: This research paper outlines various statistical interpretations of the techniques largely employed to generate phylogenetic tree using the Multiple sequence alignment Score. Phylogenetic inferences are of great importance as they represent the evolution of a taxa and its relationship to other taxa's. The problem that this paper addresses is how to mathematically represent the phylogenetic relationship between the taxa given the sequences.

KEYWORDS: STATISTICAL ANALYSIS, PHYLOGENETIC TREE, UPGMA, NEIGHBOR JOINING, MAXIMUM LIKELIHOOD, MAXIMUM PARSIMONY, MULTIPLE SEQUENCE ALIGNMENT.

1. INTRODUCTION

PHYLOGENETIC ANALYSIS OF A FAMILY of related nucleic acid or protein sequences is a determination of how the family might have been derived during evolution. The evolutionary relationships among the sequences are depicted by placing the sequences as outer branches on a tree. The branching relationships on the inner part of the tree then reflect the degree to which different sequences are related. Two sequences that are very much alike will be located as neighboring outside branches and will be joined to a common branch beneath them. The objective of phylogenetic analysis is to discover all of the branching relationships in the tree and the branch lengths [9].

Phylogenetic relationships are usually depicted in the form of binary trees. The structure of the tree illustrates possible ancestor-descendant relationships between unknown variants of a sequence existing in the past, which are ancestral with respect to the contemporary (extant) variants (the external nodes) [8]. Inference about the past branchings in the tree (the internal nodes) can be carried out if a principle or model of evolution of the sequence is assumed.

When performing a phylogenetic analysis, it is important to keep in mind that the genomes of most organisms have a complex origin. Some parts of the genome are passed on by vertical descent through the normal reproductive cycle. Other parts may have arisen by horizontal transfer of genetic material between species through a virus, DNA transformation, symbiosis, or some other horizontal transfer mechanism. Accordingly, when a particular gene is being subjected to phylogenetic analysis, the evolutionary history of that gene may not coincide with the evolutionary history of another [9].

1.1. Definitions

Mathematically speaking, all of the diagrams we shall consider are graphs: they are finite structures built out of vertices (sometimes called nodes) and edges, in which each edge connects two vertices for background. A graph is usually represented by drawing the vertices as dots and the edges as line segments.

We will primarily be concerned with graphs that are trees. Mathematically, a tree is a graph T containing no closed loops; intuitively, if you walk along the edges from vertex to vertex, the only way to return to your starting point is to retrace your steps. If we designate one vertex r as the root of T, then every edge connects a vertex x that is closer to r with a vertex y that is further away. In this case, we say that x is the parent of y, and it is often convenient to regard the edge between them as a directed edge (or arc) pointing from x to y, represented by the symbol $x \rightarrow y$. Every vertex in a tree has a unique parent, except for the root, which has no parent. An immediate consequence is the useful fact that every tree with n edges has n + 1 vertices, and vice versa.

The ancestors of a vertex are its parent, its parent's parent, its parent's parent's parent, and so on. Equivalently, we might say that an edge $x \rightarrow y$ is an ancestor of another edge $a \rightarrow b$ if y is equal to, or an ancestor of a. A lineage of a vertex x is the complete list of vertices that are ancestors of x and are descendants of, or equal to, some other vertex y. If y = root (T), then this list is called the total lineage of x.

A subtree of a tree T is a tree U all of whose vertices and edges are vertices and edges of T as well. This is equivalent to saying that U can be

formed by removing some vertices and edges from T. If in addition T is a rooted tree, then U inherits its "ancestor-of" relation from T as well. A proper subtree of a rooted tree is a subtree that consists of a vertex and all its descendants. A proper subtree is uniquely determined by its root vertex, so there are exactly as many proper subtrees of T as there are vertices. Trees are well suited for modeling phylogenetic relationships between species or taxa, in which each species or taxon has a unique parent.

1.2. Statistical Representation

The construction of optimal evolutionary trees is a very challenging problem, since most versions of the problem are NP complete.

DEFINATION 1. A phylogenetic tree T = (V, E) is a binary connected acyclic graph, where V are the vertices (nodes) and E denotes the edges of the graph. A leaf in T has degree 1 and L is used to denote the subset of V which contain the leaves of T. we use T(S) to denote a tree with leafset S.

DEFINATION 2. A Tree scoring function is a function F: T→R

DEFINATION 3. Let T be the set of all possible trees that can be generated for a given set of sequences S= {s1,s2,…,sn}. The optimal tree $T' \in T$ is a tree such that F(T′) = min F(T).

By the term phylogenetic tree, we mean a tree that models (hypothesized) phylogenetic relationship among taxa by depicting taxa by edges, and speciation events by vertices. For instance, in the phylogenetic tree in Fig. 1a, the terminal edges, labeled A, B, and C, represent named taxa; that is, large groups of individual organisms represented by sampled specimens. The internal edges, labeled y and z, represent ancestral lineages needed to account for the terminal taxa under the paradigm of descent with modification [1]. The vertices represent speciation events, in which the edge below the vertex is the common ancestor and the edges above it are descendants. Mathematically, the edge y is the youngest common ancestor of the edges B and C. Biologically, moving up the tree represents moving forward in time, so the edge y represents a lineage of common ancestors of the sampled taxa B and C, occurring before the speciation event that distinguishes B and C and after any previous speciation events. Thus the total lineage of a species (or, more properly, a hypothesis of its lineage) is represented by a chain of edges starting with the species itself and moving down the tree towards the root vertex, which necessarily has only one edge emanating from it—representing the common ancestor of all sampled taxa.

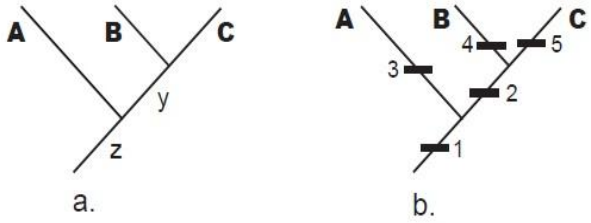

Figure 1. (a) An example of a phylogenetic tree, indicating the evolutionary relationship among the sampled taxa A, B, C and their unsampled ancestral species y and z. (b) The same tree with character data shown. The names of the internal edges have been omitted for clarity. In each case, taxon names are displaced from the leaf position to emphasize that the edge is the taxon.

In Fig. 1b, we have added more information to the phylogenetic tree. Each numbered black rectangle represents an evolutionary character hypothesized to be fixed somewhere in the lineage represented by the edge to which the rectangle is attached.

1.3. Applications of phylogenetic methods

1. **Detection of orthology and paralogy**

Phylogenetics is commonly used to sort out the history of gene duplications for gene families. This application is now included in even preliminary examinations of sequence data; for example, the initial analysis of the mouse genome included neighbour-joining trees to identify duplications in cytochrome P450 and other gene families[6].

2. **Estimating divergence times**

Bayesian implementations of new models allowed to estimate when animal phyla diverged without assuming a molecular clock.

3. **Finding the residues that are important to natural selection**

Amino-acid sites on the surface of influenza that are targeted by the immune system can be detected by an excess of non-synonymous substitutions. This information might assist vaccine preparation.

4. **Determining the identity of new pathogens**

Phylogenetic analysis is now routinely performed after polymerase chain reaction (PCR) amplification of genomic fragments of previously unknown pathogens. Such analyses made possible the rapid identification of both Hantavirus and West Nile virus [6].

2. METHODS USED

Traditionally, phylogenies have been constructed from morphological data, but following the growth

of genetic information it has become common practice to construct phylogenies based on molecular data, known as molecular phylogeny. The data is most commonly represented in the form of DNA or protein sequences, but can also be in the form of e.g. restriction fragment length polymorphism (RFLP) [7].
The methods to construct phylogeny are distance based methods and character based methods.

2.1 Distance Based Methods

Two common algorithms, both based on pairwise distances, are the UPGMA and the Neighbor Joining algorithms. Thus, the first step in these analyses is to compute a matrix of pairwise distances between OTUs from their sequence differences.

2.1.1 UPGMA

UPGMA stands for Unweighted Pair Group Method using Arithmetic average. Given a distance matrix, it starts with grouping two taxa with the smallest distance between them according to the distant matrix. A new node is added in the midpoint of the two, and the two original taxa are put on the tree. The distance from the new node to other nodes will be the arithmetic average. We then obtain a reduced distance matrix by replacing two taxa with one new node. Repeat this process until all taxa are placed on the tree. The last taxon added will be the root of the tree. More precisely, for any two clusters C_i and C_j we define the distance between the clusters as

$$d_{ij} = \frac{1}{|C_i||C_j|} \sum_{p \in C_i, q \in C_j} d_{pq},$$

where |Ci| and |Cj | denote the number of sequences in cluster i and j, respectively [5].

Algorithm: UPGMA

Initialization:
    assign each sequence i to its own cluster C[i];
    define 1 leaf for each sequence, place @ height 0
Iteration:
    determine the two clusters i, j with d[ij] min;
    define a new cluster k by C[k]=C[i][C[j];
    define node k with daughter nodes i and k;
    place the new node at height d[ij]/2;
    add k to current clusters and remove i and j;
Termination:
    when only two cluster i, j remain;
    place root at d[ij]/2.

UPGMA is actually a very simple algorithm. The search for the smallest distance takes complexity n log n, with totals to n2 log n complexity. There have been some variation of UPGMA, taking minimum or maximum distance of constituent sequences, instead of taking average, but none of those actually improves the performance [4].

2.1.2 NEIGHBOR JOINING

The neighbor joining algorithm, on the other hand, builds a tree where the evolutionary rates are free to differ in different lineages, i.e., the tree does not have a particular root. The method works very much like UPGMA. The main difference is that instead of using pairwise distance, this method subtracts the distance to all other nodes from the pairwise distance. This is done to take care of situations where the two closest nodes are not neighbors in the "real" tree. The neighbor join algorithm is generally considered to be fairly good and is widely used.

Algorithm: Neighbor Joining

Initialization:
    define T the set of leaf nodes,
    one for each sequence
    L=T
Iteration:
    pick pair i, j in L for which D[ij] is minimal; D[ij] is defined in (4.1);
    define new node k;
    set d[km]=(d[im]+d[jm]-d[ij])/2 for all m in L;
    add k to T;
    set d[ik]=(d[ij]+r[i]-r[j])/2;
    set d[jk]=d[ij]-d[jk]-d[ij];
    join k to i, join k to j;
    remove i and j from L and add k;
Termination:
    when L consists of only 2 leaves i and j;
    add the remaining edge between i and j;
    set the length d[ij]

2.2 Character Based Methods

Whereas the distance based methods compress all sequence information into a single number, the character based methods attempt to infer the phylogeny based on all the individual characters (nucleotides or amino acids) [5].

2.2.1 PARSIMONY

In parsimony based methods a number of sites are defined which are informative about the topology of the tree. Based on these, the best topology is found by minimizing the number of substitutions needed

to explain the informative sites. Parsimony methods are not based on explicit evolutionary models.

2.2.2 Maximum Likelihood

Maximum likelihood and Bayesian method are probabilistic methods of inference. Both have the pleasing properties of using explicit models of molecular evolution and allowing for rigorous statistical inference [3]. However, both approaches are very computer intensive. A stochastic model of molecular evolution is used to assign a probability (likelihood) to each phylogeny, given the sequence data of the OTUs. Maximum likelihood inference then consists of finding the tree which assign the highest probability to the data.

2.2.3 Bayesian inference

The objective of Bayesian phylogenetic inference is not to infer a single "correct" phylogeny, but rather to obtain the full posterior probability distribution of all possible phylogenies. This is obtained by combining the likelihood and the prior probability distribution of evolutionary parameters[2]. The vast number of possible trees means that bayesian phylogenetics must be performed by approximative Monte Carlo based methods.

3. COMPARISON OF METHODS

The most common traditional approaches to reconstructing phylogenies are the neighbor-joining (NJ) algorithm and tree searches that use an optimality criterion such as PARSIMONY or maximum likelihood (ML). TABLE 1 shows a summary of the advantages and disadvantages of these methods, as well as a list of the software packages that implement them.

Table 1. Comparison of Methods

Methods	**Advantages**	**Disadvantages**	**Software's**
Neighbor Joining	Fast	Information is lost in compressing sequences into distances; reliable estimates of pairwise distances can be hard to obtain for divergent sequences	PAUP MEGA PHYLIP
Parsimony	Fast enough for analysis of hundred's of sequences; robust if branches are short	Can perform poorly if there is substantial variation in branch lengths	PAUP NONA MEGA PHYLIP
Maximum Likelihood	The likelihood fully captures what the data tells us about the phylogeny under a given model	Can be prohibitively slow	PAUP PAML PHYLIP
Bayesian	Has a strong connection to the maximum likelihood method; might be a faster way to assess support for trees than maximum likelihood	The prior distribution for parameters must be specified; it can be difficult to determine whether the Markov Chain Monte Carlo(MCMC) approximation has run for long enough	MrBayes BAMBE

4. RESULT AND DISCUSSION

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 [8]. Each of these methods uses a different type of analysis as described below. These methods may find that more than one tree meets 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 [10].

1. The sequences chosen can be either DNA or protein sequence: Different programs and program options are used for each type. RNA sequences are analyzed by covariation methods and by analyzing changes in secondary structure. The selected sequences should align with each other along their entire lengths, or else each should have a common set of patterns or domains that provides a strong indication of evolutionary relatedness.

2. The alignment of the sequence pairs should not have a large number of gaps that are obviously necessary to align identical or related characters. A phylogenetic analysis should only be performed on parts of sequences that can be reasonably aligned. In general, phylogenetic methods analyze conserved regions that are represented in all 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 the multiple sequence alignment 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. Corrections may be applied for such variation, thereby increasing the observed amount of change to a more reasonable value. These corrections assume a uniform rate of change at all sequence positions over time. Gaps in the multiple sequence alignment are usually not scored because there is no suitable model for the evolutionary mechanisms that produce them.

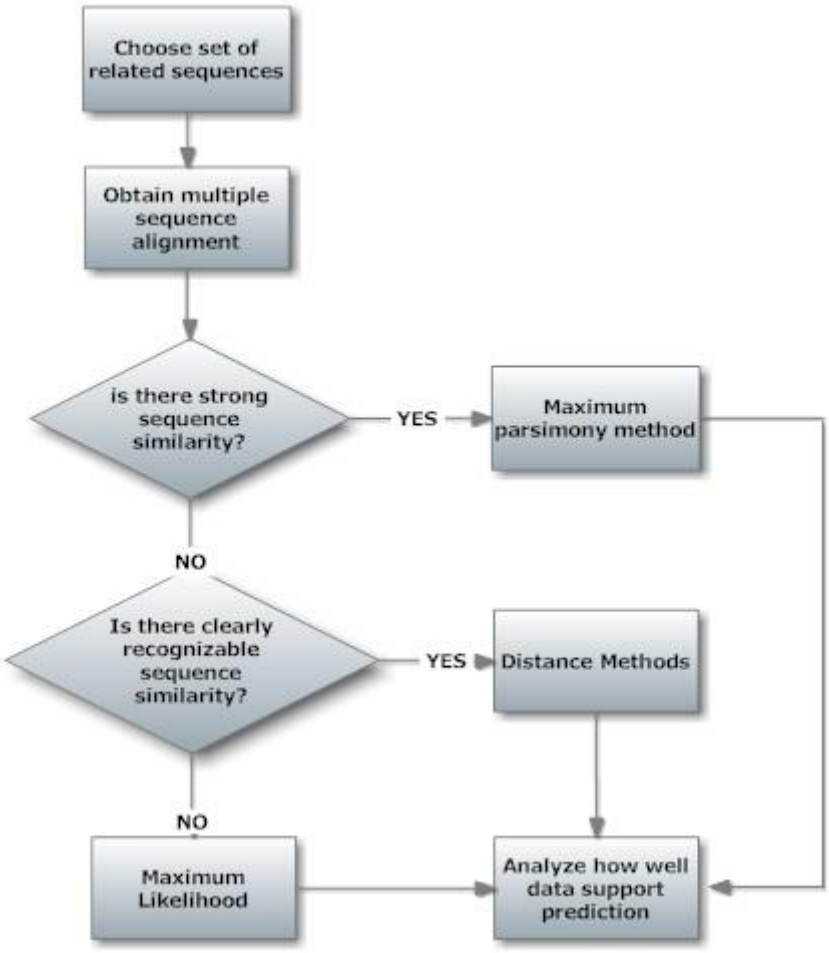

Figure 3. Flowchart for phylogenetic method selection

3. This question is designed to select sequences suitable for maximum parsimony analysis. Other methods may also be used with these same sequences. For parsimony analysis, 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 the amount of variation is small. Some columns in the multiple sequence alignment will have the same residue in all sequences; other columns will include both conserved and non conserved residues. There should be a clear-cut majority of certain residues in some columns of the alignment but also some variation. These more common residues 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. Because the maximum parsimony method has to attempt to fit all possible trees to the data, the method is not suitable for more than 11 or 12 sequences because there are too many trees to test. 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.

4. The purpose of this question is to select sequences for phylogenetic analysis by distance methods. Distance methods are able to predict an evolutionary tree when variation among the sequences is present (some sequences are more alike than others) and when the amount of variation is intermediate. The number of changed positions in an alignment between two sequences divided by the total number of matched positions is the distance between the sequences. As distances increase, corrections are necessary for deviations from single-step changes between sequences . Of course, as distances increase, the uncertainty of alignments also increases , and a reassessment of the suitability of the multiple sequence alignment 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.

5. 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. An advantage of these methods is that they provide evolutionary models to account for the variation in the sequences.

6. The data in the multiple sequence alignment columns is resampled to test how well the branches on the evolutionary tree are supported (boot-strapping).

REFERENCES

[1] C. Ané, O. Eulenstein, R. Piaggio-Talice, M.J. Sanderson (2009), "Groves of phylogenetic trees", Annals of Combinatorics, Vol. 13, No. 2, pp 139-167.

[2] Chantal Korostensky, Gaston H. Gonnet (2000) "Using Travelling salesman problem algorithms for evolutionary tree construction", BIOINFORMATICS, Vol. 16, No. 7, pp 619-627.

[3] Goldman, N., Anderson, J. P. & Rodrigo, A. G.(2000) " Likelihood based tests of topologies in phylogenetics." Syst. Biol., Vol. 4, No. 9, pp 652–670.

[4] J.J. Wiens,(2006) "Missing data and the design of phylogenetic analyses", Journal of Biomedical Informatics ,Vol. 3,No. 9,pp 34-42.

[5] Lemmon, A. R. & Milinkovitch, M. C. (2002) "The meta population genetic algorithm: an efficient solution for the problem of large phylogeny estimation." Proc. Natl Acad. Sci. USA, Vol. 9, No. 9, pp 10516–10521.

[6] Mark Holder and Paul O. Lewis, (2003) "Phylogeny Estimation: Traditional and Bayesian approaches", Nature reviews: Genetics, Vol. 3, pp 275-284.

[7] Siepel, A. and Haussler, D. (2004). "Combining phylogenetic and hidden Markov models in biosequence analysis." J Comput Biol, Vol. 11, No.3, pp 413--428.

[8] Swofford, D. L.(2001) " Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods", Syst. Biol., Vol. 5, No.10, pp 525–539.

[9] Jin Xiong. Essential Bioinformatics, Cambridge University Press, (2006).

[10] How to make a phylogenetic tree: http://hiv-web.lanl.gov

[11] BAMBE: http://www.mathcs.duq.edu/larget/bambe.html

[12] BLAST: http://www.ncbi.nlm.nih.gov/BLAST

[13] MEGA: http://www.megasoftware.net

[14] MrBayes: http://morphbank.ebc.uu.se/mrbayes

[15] PAUP: http://paup.csit.fsu.edu/index.html

[16] PHYLIP: http://evolution.genetics.washington.edu/phylip.html

[17] Phylogeny programs: http://evolution.genetics.washington.edu/phylip/software.html