PHYLOGENIES: AN OVERVIEW

SUSAN P. HOLMES*

Abstract. This is an overview that aims to help statisticians access interesting problems developing in the biological literature on estimating and evaluating phylogenetic trees.

Key words. Bootstrap, cladistics, DNA, molecular evolution, parsimony, phylogeny, systematics, tree.

1. Introduction. Representation of biological families by trees predates Darwin's theory of evolution, although the latter gave such representations a true explanatory justification. For biologists, at each branch of the tree are situated separation events that split orders or families or genera or species. For example, the figure shows a classification by Haeckel, 1870.

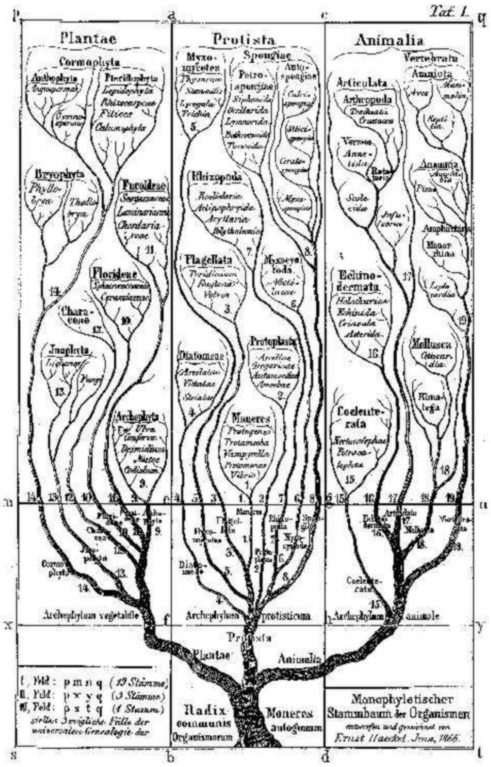

FIG. 1.

*Biometrics Unit, Cornell University, NY14853 Ithaca, sph11@cornell.edu

Neighbors on the tree share the same ancestor. Characters that are derived from this common ancestry are called homologous. Many geneticists doing population studies replace the term homology by identity by descent (IBD). The distinction between homology and similarity is a subtle one. In particular, sisters in the tree defined by a common ancestor are called clades or monophyletic groups, they have more than just *similarities* in common. Finding such groups is one of the goals of phylogenetic studies.

For over 200 years biologists have built trees to classify their species based on **morphological** [1] data. More recently the explosion of genetic data available through molecular biology has made tree-building even more popular. This presentation aims to interest statisticians in bringing their know-how to some of the open issues that currently fill the biological literature. The systematics literature is fraught with a great deal of polemics, much of which are statistical in nature.

Some questions that are raised include:

- Whether parametric methods using models should be preferred to nonparametric methods.
- How the data should be coded, for instance, is one categorical variable preferable to several binary ones?
- Should certain characters suspected of conflicting with the tree structure be down-weighted?
- Which methods should be used to validate a tree that results from the analysis? (This entails recourse to confidence regions and conditional testing).
- Which parameterizations of the problem have the most desirable statistical properties, consistency, identifiability, robustness ?
- How should the information from different genes be combined into an overall species tree?
- How should prior information on the species be incorporated into the analysis? This can be translated into a Bayesian dilemma; how can we code the fact that we know in advance that certain species are very different?

There are many obstacles to reading literature from a new field. Surprisingly, the most difficult hurdles may not be the new words encountered here but the 'faux amis'-the old friends (statistical terms) with new meanings.

I will document these below. Here are a few examples I had difficulties with:

[1] data about presence or absence of wings, sepals, hair, nodules,...

Biological articles		Standard Statistical Terminology
inferring phylogenies	-	estimating phylogenies
biased	-	systematically wrong
consistent	-	robust
consistency	-	existence of an iterative limit
statistical power	-	efficient
repeatability	-	
transition	-	
substitution model	-	transition matrix
independence	-	conditional independence
jackknife	-	cross-validation
statistical method	-	parametric method
likelihood	-	probability
statistician	-	philosopher

Statisticians interested in more details about molecular evolution will find Li (1997) rewarding, it explains clearly many aspects of the problem. There is a collection of chapters on the subject in Hillis, Moritz and Mable (1996) which has the merit and handicap of attempting to be exhaustive. A review written for statisticians 15 years ago can be found in Felsenstein (1983). His programs are publicly available in `phylip` [24]. I will start my review as he did by defining phylogenies and the data used; then our paths separate. Section 3 presents a translation of the problem in statistical terms. Section 4 presents the three main families of tree-building methods: maximum likelihood, distance-based methods and maximum parsimony. Section 5 attempts to outline some of the sources of trouble in the procedures, and why 20 years after this field of research began, four specialized journals and hundreds of books later, no agreement has been reached, either on which method is better or how sure one is of the answer the methods provide.

Statisticians do have tools for comparing methods and Section 6 reviews some of the qualities of the various methods as measured with these statistical yardsticks. Section 7 presents methods for evaluating a tree, once it has been estimated. The question answered by the methods of this section are similar to those answered by the computation of a confidence region, unfortunately in a space with neither a natural distance nor a natural probability measure. The bootstrap is the most popular method among biologists for evaluating a tree, and we will try to underline some of its features and drawbacks in this context. Finally we will propose some more exploratory indices for evaluating how tree-like the data are to provide a scale of plausible error for the trees.

Finally the more practically minded may find the appendix a good starting point; it contains some exemplary runs of some freely available programs that can be easily down-loaded from the internet.

2. What is a phylogeny?. ¿From a mathematical point of view a phylogeny is a rooted binary tree with labeled leaves.

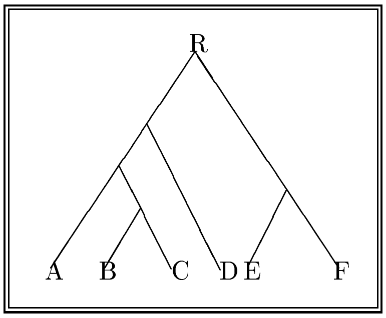

FIG. 2. *A rooted binary semi-labeled tree.*

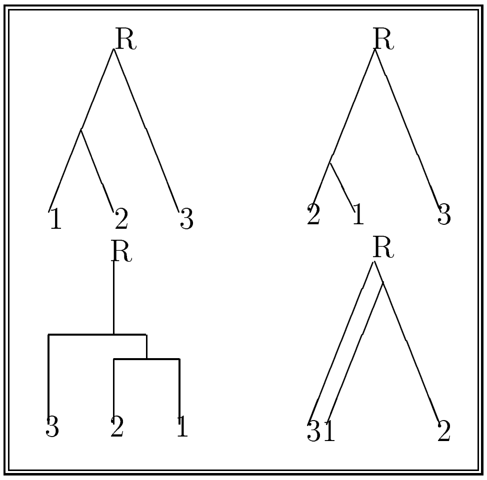

FIG. 3 *Four representations of a same tree topology.*

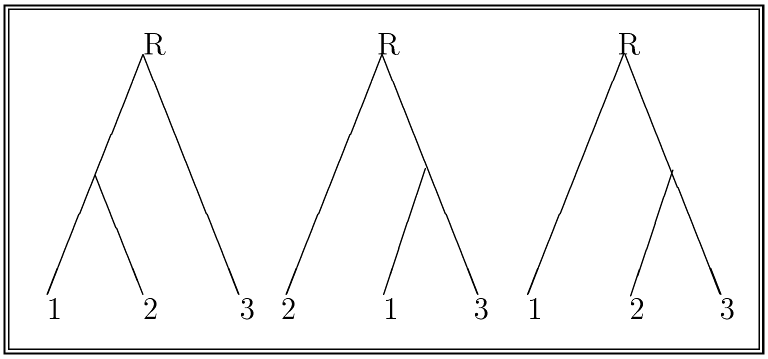

FIG. 4. *All possible distinct rooted binary topologies with 3 leaves.*

Unrooted trees are graphs in which all $N-2$ inner vertices (nodes) are of degree 3, and the N outer vertices (leaves) of degree 1, are labeled.

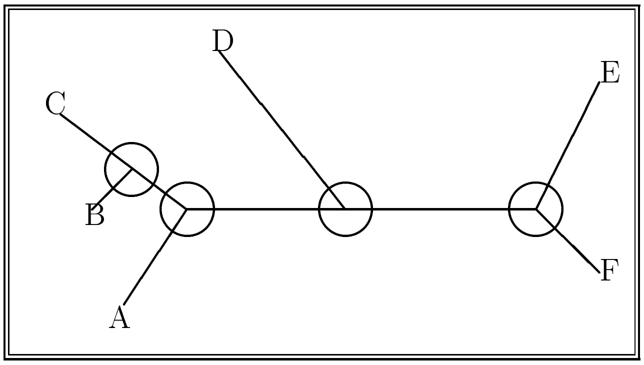

FIG. 5. *An unrooted tree.*

The number of unrooted semi-labeled trees with N leaves is known since Schröder (1870) to be:

$$(2N-5)!! = (2N-5) \times (2N-7) \times (2N-9) \ldots \times 3 = \frac{(2(N-1))!}{2^{N-1}(N-1)!}$$

where $n!!$ is the double factorial where the difference between the successive factors is 2 instead of 1 in the classical factorial $n!$.

As there are $2N-3$ possible branches on which to place the root the number of rooted semi-labeled trees is: $(2N-3)!!$. For $N=10$ there are $2,027,025$ unrooted trees and $34,459,425$ rooted ones. These numbers grow rapidly. Using Stirling's formula, we have an asymptotic approximation

$$(2N-3)!! \sim (\frac{2}{e})^{(N-1)}(N-1)^{(N-1)}\sqrt{2}$$

This tells us that for $N=20$, there are around (2.10^{20}) unrooted and $.8 \times 10^{22}$ rooted topologies from which to choose. Even if there is a lot of data, we can see that the choice is going to need some more outside information. We will develop this problem in section 6.

Many useful facts about such trees can be gleaned from books on graph theory and combinatorics: Stanley (1996) is particularly useful. It contains an elegant proof of Schröder's formula.

The leaves of these phylogenetic trees (called trees from here on) are called **Operational Taxonomic Units** or OTU's by the biologists and called simply units below. They can be:

- genes[2], for instance hemoglobins were some of the first to be sequenced and used for phylogenetic purposes.

[2]segment of DNA that codes for a polypeptide chain or specifies a functional RNA molecule, see Li (1997,page 9)

- individuals, represented by part of their genome. They are usually from within a population and can actually be connected by classical family relations.
- populations from within the same species but from different areas.
- species of which there is usually one representing sequence from a particular individual.
- families or larger classes of species.

The data I have seen up to now usually has the following features:

- The leaves are all contemporaries.[3] This is why the trees are represented with the leaves all falling level 'on the ground level' rather than a more mathematical representation which would inspire the right hand part of Figure 6.

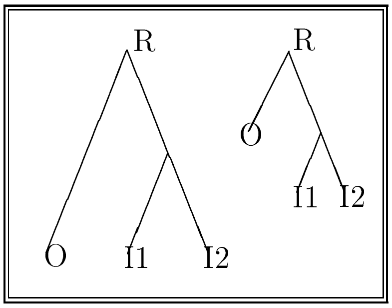

FIG. 6. *The left tree shows contemporary leaves.*

- Up to now sequencing has been slow and phylogenetic studies concentrated on relating species using one representative unit. With the rapid growth of Polymerase Chain Reaction (PCR) usage, data will be so abundant that it will be possible to study conjointly whole samples of sequences from different individuals of a same species, thus introducing interesting statistical information on variability.
- Usually it is the topology of the tree that is essential. The units (OTU's) are the only parts that are observed, the internal nodes have to be guessed at. (They are the ancestors, and they are the part of the tree that inspired the term *inferred.*) However, as we shall see later, the estimated tree is often augmented by branch-lengths. These present additional subtleties in the evaluation of the quality of the estimated tree.

3. What are the data from which the trees are built?. We will suppose that we want to study N units, this number of species or genes studied at the same time is usually between 10 and 50.

In all that follows the examples are of molecular data, either amino

[3]Excepting for paleontological data.

acids or nucleotides obtained sometimes from fragments, restriction sites, or whole DNA/RNA sequences[4].

The first part of the analysis consists of alignment of the sequences. There is a lack of coherence between the methods chosen to align the sequences and methods that are used after to build the tree. A method that would improve the coherence of the methodology would allow simultaneous alignment and tree-fitting. For recent work, see Schwikowski and Vingron (1996), and Wheeler (1994).

There are three main schools of tree-building methods:

- Maximum likelihood methods.
- Distance methods.
- Maximum parsimony methods.

For the time being I will consider that we are given sequences in which gaps have been inserted to enhance their **alignment** so that a typical sequence looks something like:

```
      21   383
 VVi     M-SGTAGQVICCKAAVAWEAGKPVIEEVEVAPPQAMEVRLKILYTSLCH
 Zma1    M--ATAGKVIKCKAAVAWEAGKPSIEEVEVAPPQAMEVRVKILFTSLCH
 Zma2    M--ATAGKVIKCRAAVTWEAGKPSIEEVEVAPPQAMEVRIKILYTALCH
 Hvu1    M--ATAGKVIKCKAAVAWEAGKPTMEEVEVAPPQAMEVRVKILFTSLCH
 Hvu2    M--ATAGKVIKCKAAVAWEAGKPSMEEVEDAPPQAMEVRDKILYTALCH
 Hvu3    M--ATAGKVIKCKAAVAWEAGKPSIEEVEVAPPQAMEVRVKILYTALCH
 Tae     M--ATAGKVIECKAAVAWEAGKPSIEEVEVAPPHAMEVRVKILYTALCH
 Osa1    M--ATAGKVIKCKAAVAWEAGKPSIEEVEVA--KEMEVRVKILFTSLCH
 Osa2    M--AT-GKVIKCKAAVAWEAGEASIEEVEVAPPQRMEVRVKILYTALCH
 Ath     M-S-TTGQIIRCKAAVAWEAGKPVIEEVEVAPPQKHEVRIKILFTSLCH
 Psa     M-SNTVGQIIKCRAAVAWEAGKPVIEEVEVAPPQAGEVRLKILFTSLCH
 Fan     M-SSTEGKVICCRAAVAWEAGKPVIEEVEVAPPHPNVVRVKILYTSLCH
 Tre     M-SNTAGQVIKCRAAVAWEAGKPVIEEVEVAPPQAGEVRLKILFTSLCH
 Stu     M-STTVGQVIRCKAAVAWEAGKPVMEEVDVAPPQKMEVRLKILYTSLCH
 Pgl     M-A-TAGKVIKCKAAVAWEAGKPSIEEVEVAPPQAMEVRVKILYTSLCH
 Phy     MSSNTAGQVIRCKAAVAWEAGKPVIEEVEVAPPQKMEVRLKILFTSLCH
 Pde     M-SSTVGKVIRCKAAVAWEAAKPSIEEVEVAPPQANEVRLRILFTSLCH
 Pta     MASSTAGQVIKCKAAVAWAAGEPKIEEVEVAPPQAMEVRVKIHYTALCH
 Fra     M-SSTEGKVICCRAAVAWEAGKPVIEEVEVAPPQANVVRVKILYTSLCH
 Mal     M-SNTAGQVIRCRAAVAWEAGKPVIEEVEVAPPQANEVRIKILFTSLCH
 Lyc     M-STTVGQVIRCKAAVAWEAGKPVMEEVDVAPPQKMEVRLKILYTSLCH
```

This table is a subset of a larger data matrix that was downloaded from GENBANK. It was originally submitted by Yokoyama (1995) for all the sequences except `Vitis Vinifera` which comes from Sarni-Manchado,

[4] For detailed technical explanations see Li ,(1997) or Hillis et al, (1996)

Verriès and Tesnière (1997) and is from an *adh* gene.

The first two numbers indicate the dimensions of the data matrix X. The first number here is 21 because there are $N = 21$ species being studied, the second integer indicates that there are $k = 383$ characters. (they represent either one of the 20 amino-acids or an insertion '-').

The actual tree-building analysis will be run on different subsets of the data depending on which of the methods is used to build the tree.

1. For maximum likelihood, the complete matrix of sequences are used. Even columns with no difference at all between the units contain information on the relative frequency of various characters.
2. On the other hand, parsimony methods only use 'informative' columns, (for a complete definition see Li (1997), page 113). Informative sites are those that enable differentiation between possible trees, in particular either monotypical sites, or sites that have all the same value except for one unit are **not** informative and so are left out of the data set.
3. Distance methods have an in-between strategy. In a first step all the data are processed to estimate the relevant parameters for the distance formula, then the distances are computed between units. Only the distance matrix is used after that.

Other types of data can be used for tree building instead of DNA sequences. These can be either presence/absence of characters coded in binary and morphological characters coded as categorical data. Gene frequency data were used in the past, but recent molecular studies at a more precise scale seem to have replaced them.

At least one of the taxonomic units has a special function. For a statistician it would be seen as a simple outlier: the biologists voluntarily include what they call an **outgroup** to locate the root of the tree. The root is situated by creating an unrooted tree and the edge that joins the outgroup to the other species will be the support for the root. This is a clever use of prior information that simplifies the problem considerably, (by a factor of $(2N - 3)$). What is less obvious to the outsider is why, once the root's position is decided upon, the biologists keep the outgroup in the data set - it seems to distort the image of the closer group (called the **ingroup**), in fact outgroups also provide information on the root's characters, and so on the ancestral states of the character. This seems to be a security check, if in fact the outgroups become misplaced or lost in the tree, then there are signs of trouble. Many methods have trouble as soon as 2 very different outgroups are present (this is named the **long branch attraction problem**), just as in regression two opposite outliers can completely redefine the regression line.

In fact molecular data from one particular gene will only provide information about a certain 'gene tree' and not necessarily about the more general unit (such as the whole species). Combining information from all these different gene trees remains an interesting statistical open problem that could be addressed with conjoint methods such as those developed by Lavit, Escoufier, Sabatier and Traissac (1994) or by meta-analysis type methods. Some work has been started on the subject by Doyle (1992) and more recently Page and Charleston(1997).

3.1. What are molecular-based phylogenies used for?. Lists of possible answers as extracted from Hillis et al. (1996), include gene evolution, population subdivision, analysis of mating systems and heterozygosity, paternity testing, as well as studies of individual relatedness, geographic variation, hybridization, species boundaries. Details of these can be found in the useful textbooks: Li (1997), Hillis et al. (1996).Comparative methods such as those explained in Harvey and Pagel (1993) also use phylogenies along with other, possibly quantitative, information.

Many modern genetic studies aimed at mapping diseases use the notion of **identity by descent**. This is the same as the concept of homology in the case of a study restricted to a small population of individuals for which gene trees are constructed. Thus information about relationships between far cousins can enhance understanding of homologies.

4. Statistical translation of the problem. All tree-building methods are based on the assumption that an evolutionary tree is a relevant representation of the data, an assumption that we will need to make more precise as we advance.

The first goal is **estimation**, producing a tree $\hat{\mathcal{T}}$ on the basis of a data matrix $X_{N\times k}$ that estimates an unknown true tree $\mathcal{T}$. This is strangely called an **inference** problem by biologists whereas the statisticians would call it an estimation problem.

The second goal is to provide a **confidence statement** to associate to the estimator $\hat{\mathcal{T}}$. Currently this is done most often by bootstrapping-type methods that we summarize in section 6. Although definitely related to branch lengths, this aspect of the tree receives less attention, since most goals of phylogeny seem to be more qualitative than quantitative.

The schools of tree-building methods: maximum likelihood methods, distance methods, and maximum parsimony methods can be compared using the statistical paradigm in a way which clarifies their similarities and differences. From a statistical viewpoint these methods can be understood as being ordered by the **dimension** of the underlying parameter space:

- Maximum likelihood uses a parametric model containing from 1 to 12 parameters for the substitution rates and usually $(N-2)$ parameters for the branching times. (See section 4 for a detailed description of the method.)
- Distance based methods use the same parametric model for the substitutions and deduce from these rates 'evolutionary' distances between units. The distance matrix is then analyzed by hierarchical clustering type methods such as neighbor-joining (single linkage clustering) or unweighted pair-group with arithmetic mean (aver-

age clustering). Distance-based methods can be seen as intermediary containing both parametric and nonparametric components.

- As we will see shortly, basic maximum parsimony methods are actually based on building a binary Steiner tree with regards to Hamming distance. They are nonparametric methods where the main assumptions are :
  - The existence of a true evolutionary tree.
  - The independence of characters (columns of the X matrix).
  - Comparable substitution rates across characters.

Connections between the methods follows from recent work of Tuffley and Steel (1997) who show that when the number of parameters in the model is increased to incorporate different mutation rates along sites and different rates along branches the maximum likelihood method becomes equivalent to the maximum parsimony method.

Note that as the number of parameters becomes larger than the number of estimates that the data can usefully provide, the method passes the limit of a parametric model and becomes nonparametric (or infinite dimensional).

5. The tree-building methods. Here I will give a brief introduction to the three main families of tree-building techniques. Details may be found in Li (1997) for instance. Distance-based methods and maximum likelihood use a special model for describing the process by which changes between sequences occur. This is the substitution model that I will describe first.

5.1. The substitution model. To be more precise I will only show the case where the data are DNA nucleotides: **purines ('A', 'G')** and **pyrimidines ('T', 'C')**. There are many types of substitution models, the simplest model is called the Jukes-Cantor model and supposes that any change of the nucleotides occurs at the same rate, whether from one type to another, (**transversion**), for instance from purines to pyramidines within each type, (**transition**), for instance from purines to purines. The rate matrix Q is of the form:

$$Q = \begin{array}{c|cccc} & A & T & C & G \\ \hline A & -3\alpha & \alpha & \alpha & \alpha \\ T & \alpha & -3\alpha & \alpha & \alpha \\ C & \alpha & \alpha & -3\alpha & \alpha \\ G & \alpha & \alpha & \alpha & -3\alpha \end{array}$$

The 12 parameter model is of the form

$$Q = \begin{array}{c|cccc} & A & T & C & G \\ \hline A & - & \alpha_{1,2} & \alpha_{1,3} & \alpha_{1,4} \\ T & \alpha_{2,1} & - & \alpha_{2,3} & \alpha_{2,4} \\ C & \alpha_{3,1} & \alpha_{3,2} & - & \alpha_{3,4} \\ G & \alpha_{4,1} & \alpha_{4,2} & \alpha_{4,3} & - \end{array}$$

The substitution matrix gives the probability of the change of a nucleotide during a time t as:

$$P(t) = e^{Qt}$$

In the case of the amino acids we would have bigger matrices (20×20 instead of 4×4), but most of the other computations carry through.

5.2. Distance based methods. These methods are variants of cluster analysis, probably more familiar to statisticians. The aim is to reconstruct the distances as computed between the two sequences of the two species x and y by distances along the edges of the tree forming a path between x and y.

First a distance matrix is constructed between the N units in some way. These distances d_{xy} are supposed to estimate the unknown 'true evolutionary' distances between x and y as they would be measured along the unknown true tree $\mathcal{T}$.

For the Jukes-Cantor model which assumes equal rates of substitution between all base pairs provides the estimate of distances between sequences x and y as:

$$d_{xy} = -\frac{3}{4}\log(1 - \frac{4}{3}(1 - (\frac{\#AA}{k} + \frac{\#CC}{k} + \frac{\#GG}{k} + \frac{\#TT}{k})))$$

where k denotes the number of characters (columns) in the data matrix, and $\#AA$ denotes the number of times there is an A in x matched with an A in y.

Once the distances are decided upon, the parametric model is left behind and a clustering technique such as hierarchical clustering with average groups is used to find the tree from the distances.

It seems that this method has declined somewhat in popularity over recent years among biologists. It was the method that made the trees the easiest to compute, but improved facilities have made maximum likelihood and maximum parsimony more tractable. Those who don't believe in the parametric substitution models don't use it because of the assumptions underlying the distance computations and those who don't trust heuristic tree-building algorithms don't use it because of the tree-building phase. Historically it was the first method available on the computer, and people still use it for reasons of computational ease.

Remarks:

If we knew the true evolutionary distances between species, we could build an additive tree that reproduced the distances along the tree in a unique way. The existence of an additive tree reproducing the distances faithfully is not always ensured, a sufficient condition for this to be possible is called the **four point condition**:

$$d_{AB} + d_{CD} \leq max(d_{AC} + d_{BD}, d_{AD} + d_{BC}), \text{for all quadruples } (A, B, C, D)$$

This means that one of the two sums is minimum and the other two are equal. Notice that this is not the same as the ultrametric property which says that for any three points: A, B, C:

$$d_{AC} \leq max(d_{AB}, d_{BC})$$

If the distances obey the ultrametric property the distances can be fit to a binary tree with leaves equally distant from the root. Unfortunately distances computed from real data never obey this property. We will give details later in section 5, but additivity is destroyed by:

- Homoplasy (reversal, parallelism and convergence) which is caused by superimposed changes.
- An uneven distribution of change rates.
- Measurement error.
- **Paralogous** sequences[5].

Some distances are obtained directly by hybridization techniques. We will not include these here. We concentrate on distances that are computed from substitution models such as Jukes and Cantor's one-parameter model, Kimura's two-parameter model, or even the complex 12-parameter model for the substitution matrices. These models provide estimates of differences between sequences computed from the frequencies of various changes in the sequences.

5.3. Parsimony method. The foundations of this method have been long discussed, as always with heuristic nonparametric procedures. A detailed account can be found in Farris (1983), his justification for parsimony is that this method "minimizes requirements of ad hoc hypotheses of homoplasy[6]". This is easier to understand for a statistician if the analogy is made between homoplasies and residuals, these are the part of the data that the tree does not explain, minimizing homoplasies is an approach akin to minimizing residuals in regression for instance.

Roughly this method can be seen as based on the assumption that "evolution is parsimonious" which means that there should be no more evolutionary steps than necessary. Thus the best trees are the ones that minimize the number of changes between ancestors and descendants. We will see that under the assumption of independence of each of the characters, this has a clear combinatorial translation.

5.3.1. The parsimony tree as a combinatorial problem. For the time being, we will only consider the construction of unrooted parsimony trees. As we saw in the section on data, the rooting of the tree is done before the construction of the unrooted tree.

[5]Consequences of lineages being created separately after a gene duplication.

[6]These are the transformations caused by reversal, parallelism and convergence that will be explained in section 5.

Recall that the Hamming distance between two units is the number of changes needed to bring one to the other. This assumes that all changes in a categorical character are counted as one step.

$$d_H(AACTGGG, AACTGGC) = d_H(AACTGGG, AACTGGA) = 1$$

Here, given N points in a metric space, the Steiner problem is that of finding the shortest tree connecting the N points where one is allowed to add extra vertices. Thus, with 4 points arranged at the vertices of a unit square, one would add a fifth point in the center to form the Steiner tree.

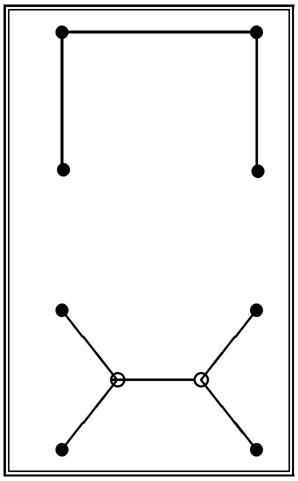

FIG. 7. *The minimum spanning tree and the Steiner tree of the 4 vertices of a rectangle.*

Although statisticians are not familiar with minimal Steiner trees, they may have encountered minimal spanning trees as used by Friedman and Rafsky (1985). The relation between the two is well explained in Gardner's wonderful chapter on Steiner trees (Chapter 22, Gardner (1997)). He explains how minimal spanning trees are good "starting points" since in the plane for instance they can only be 13% longer than Steiner trees.

As a combinatorial problem, the maximum parsimony tree is the problem of finding the Steiner points or Steiner tree for Hamming distance between the units, under the constraint that the tree be binary. The problem of finding a minimal Steiner tree is that of finding the Steiner points (representing ancestors) that minimize the complete length of the tree. Steiner points are points that are added to a graph so that its minimal spanning tree becomes shorter. The minimal Steiner tree problem is NP-hard, meaning that no algorithm is known that will compute an optimal tree in polynomial time in the number of species N.

Much work has been done to implement good heuristic algorithms for finding approximately optimum trees. Swofford's `PAUP`, Felsenstein's

`Phylip`, and Goloboff's `NONA` all contain clever use of branch and bound techniques and branch swapping to find acceptable answers. No explicit analyses of the complexity of the algorithms involved have been published, but recent empirical tests show enormous progress in terms of CPU time, even for large (N=500) problems (Goloboff, personal communication).

Theoretical computer scientists on the other hand have produced many papers on methods to solve the problem, with detailed complexity analysis, but no code is available as yet. See in particular recent work such as Erdös, Steel, Szèkely, Warnow(1996,1997) and Rice, Steel, Warnow and Yooseph (1997).

5.3.2. Parsimony as a statistical procedure. Felsenstein (1983) lists parsimony in a section entitled a section on parsimony as "non-statistical approaches". Farris says (1983) says the "statistical approach to phylogenetic inference was wrong from the start, for it rests on the idea that to study phylogeny at all one must first know *in great detail* how evolution has proceeded". Both these authors identify statistics with parametric modeling. This is unfortunate as it has led many clever cladists to stop reading the statistical literature, thus depriving them of many useful tools. Parsimony methods are well within the boundaries of non-parametric statistical procedures that have been developed over the last twenty years. Methods are no longer considered statistical only if they are justified by an underlying stochastic model.

Many data-analytic procedures such as correspondence analysis, projection pursuit, neural nets, classification and regression trees (CART) and minimal spanning trees have proved that complex situations can be satisfactorily understood by heuristic procedures before any theoretical framework supposing a probabilistic model justifies their properties (Diaconis and Efron (1984)).

On the other hand it can be an interesting challenge for theoretical statisticians to do for parsimony what Rubin and Anderson (1956) did for factor analysis, that is find a model for which the heuristic method was providing the correct estimate, as also was the case for partial likelihood. However I doubt from a practical point of view that this would be of any interest to those who use parsimony as their standard tree-building technique.

5.4. Maximum likelihood trees. For a statistician this is the easiest of the methods to understand. A parametric model (θ, T) is postulated, θ is a η-dimensional vector that we explain below and $\mathcal{T}$ is the tree's topology. Under this model the likelihood for each possible tree $\mathcal{T}$ is separately computed for each character, the independence of characters then allows the total likelihood of the tree for all data to be computed by taking the product.

The first part of the vector of parameters θ comes from the substitution model as explained in section 4.1. The number of other parameters that have to be specified depends on the complexity of the model. If a molecular clock [7] is postulated, speciation times $\{t_1, t_2, ...t_{N-2}\}$ (splitting events) are the other parameters. Otherwise both the branch lengths $\{v_1, v_2, ...v_{N-2}\}$ and the different rates along those branches have to be parametrized.

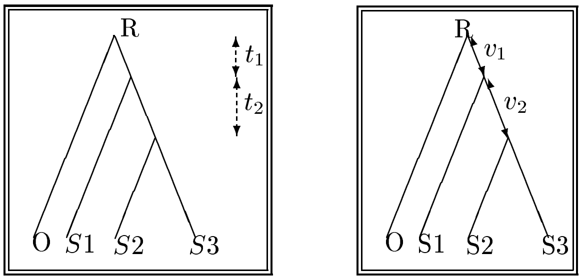

FIG. 8. *Two parametrizations of the tree.*

The substitution parameters are estimated from the data. A complete model including distributions of separation events is postulated and the likelihood can be computed for each possible tree by computing the likelihood of the tree given each site $X_{.j}$:

$$f(X_{.j}|\theta_1, \theta_2, \ldots, \theta_\eta, \mathcal{T}).$$

This actually requires computing the likelihood of all the subtrees, so the method is recursive.

$$\mathcal{L}(\theta_1, \theta_2, \ldots, \theta_\eta | X_{.1}, X_{.2}, \ldots, X_{.k}, \mathcal{T}) = \prod_{j=1}^{k} f(X_{.j}|\theta, \mathcal{T})$$

As the assumptions are essential, I present them here:

1. Each site in the sequence evolves independently.
2. Different lineages evolve independently.
3. Each site undergoes substitution at an expected rate which is chosen from a series of rates with a given distribution.

Fancier versions of the procedure enable different sites to have different evolution rates.

Many biologists won't use maximum likelihood because of the computational expense, each tree's likelihood computation is NP hard. This is a surprising exception to the usual rule that parametric methods are advantageous by their lesser computational needs. Others don't use the MLE because there seems to be little evidence that the assumptions are actually realistic in real biological applications.

6. Where does the trouble come from?. Here are a few details about the hurdles the tree-making algorithms have to deal with.

[7]branch lengths in evolutionary change depend linearly on time

6.1. Homoplasy. A character change may become invisible through time, because there has been a **reversal** or **back-substitution** for instance:

$$A \longrightarrow G \longrightarrow A.$$

There are also changes of exactly the same type that appear in different parts (clades) of the tree, giving a false impression of similarity. This is called **parallelism**.

Another variant is substitutions that occur in different clades but have the same results:

$$\left.\begin{array}{r} A \longrightarrow G \longrightarrow A \\ A \longrightarrow C \longrightarrow T \longrightarrow A \end{array}\right\} \text{ these are called } \textbf{convergent} \text{ substitutions.}$$

The effect on the resulting measurements of differences between units are the same: there is an error; units appear to be more similar than they would be if the complete history were known. Collectively these are called **homoplasy**. There are very clearly documented examples of these in Li (1997), pages 69-70.

Parametric models that take homoplasy into account are the motivation for the 'modified evolutionary distance' computations. Whether they include 1 or 12 parameters they try to retrieve some of the variability lost through homoplasy. Some authors feel that this possibility of error-correction in parametric methods is so essential that it justifies using such models even when they have not been proved to fit the actual phenomenon.

Parsimony methods are sometimes limited to shorter stretches of time to limit the homoplasy; 'long branches' are undesirable in parsimony methods.

6.2. Non-optimality of the solutions. As we saw in section 4, both maximum likelihood and parsimony provide only locally optimum trees, whatever their criteria, because the problems are computationally intractable. The clustering methods used on distance matrices are also only heuristic algorithms, not necessarily providing the global optimum.

So even when the data are perfectly dependable, errors may persist because only a local optimum was obtained, or there may be several optima. Some authors repeat the analysis of the data, interchanging the order of their species, this makes the algorithm choose a different starting point, thus often resulting in a different solution. This appears as option `jumble` in `Phylip` for instance (see the examples in the appendix).

6.3. Many possible trees, little data. When we have boiled down the patterns of different nucleotides, there are often less than 100 of them left, of which usually 80% or more are weak signals because they are singletons. So we have about 100 numbers, the frequencies of each of the patterns, from which to decide about 10^{20} possible trees, a difficult task.

In more detail, although there may be $k = 2000$ characters available, most of these are usually uninformative sites. Even parametric methods that use them (parsimony doesn't) boil them down to 4-5 numbers: the number of columns of each type. For instance, in the `Vitis Vinifera` example of section 1, there were 383 columns of amino-acids of which 187 were all monotypical columns. When one takes out what biologists call **singletons**, (columns with all the same character but for one species) there were only 140 columns of data left. Now if the data are patterned, (the character which appears first is called '1', the next different '2', the third '3' and so on[8]), it can be summarized as a few frequencies, most equal to $\frac{1}{k}$.

7. Evaluating the methods. Some statistical yardsticks such as consistency, efficiency, identifiability, robustness, computational speed, discriminating ability, or versatility may help to compare the methods in an abstract way.

7.1. Accuracy. A first suggestion that comes to mind is *"Which method gives the true tree when we know the answer ?"* Unfortunately, there are few data sets where the truth is known. An example is the tiny organism called bacterophage T7 of which a small phylogeny was generated in laboratory conditions (Hillis, Bull, White, Badgett and Molineux 1992). The programs in the appendix show some of the results obtained on this data for which parsimony seemed to work well, it gave an accurate prediction of the tree. But a sample of size one is no evidence, and as usual the statistician begs the biologists: *Bring us more data!*.

7.2. Consistency. There have been studies of consistency of the estimator $\hat{\mathcal{T}}$ in the classical statistical sense: when the number of characters increases to infinity do the trees provided by the estimators converge to the true tree? Under their own particular assumptions, all methods are consistent. However this is insufficient unless these conditions can be checked. Chang (1996), shows that maximum likelihood is inconsistent when the homogeneity assumption of identical distribution of substitution rates across characters is violated. Parsimony is inconsistent when some branch lengths are long enough to make 'hidden changes' or homoplasies likely.

In fact, the justification of putting this into a classical statistical framework is tricky, because what is being said is that we should consider that characters can be independently sampled from some distribution. Then, as the number of characters increases, we want the estimator to converge to the true tree. However, such increase in the observations is impossible, the genome is finite, and as we sample more and more characters they are less and less independent. (see Sanderson 1995)

Consistency is a quality that should not be considered fundamental. We never have infinite amounts of data, especially as compared to the number of choices that have to be made, on the other hand, it would be

[8]For an example see in the appendix.

most useful to know how long a sequence is necessary to attain a sufficient level of precision in distinguishing between possible trees, this has been named **statistical power** in part of the biological literature. In fact a more precise statistical term would be **efficiency**.

7.3. Efficiency. Historically one could reason backwards and see why biologists have called this **power**, but as no specific testing framework is set up before the analysis, this term seems abusive here. In classical statistical terminology, **efficiency** measures how quickly a method converges to the correct solution as the data size increases, this would be a better term here.

Much theoretical work remains to be done here. For maximum likelihood, classical estimates of efficiency are available. No such information is available for nonparametric estimation methods.

7.4. Robustness. Robustness measures the stability of the method when the data do not fulfill the necessary assumptions. Simulations can be used to test robustness with regards to specific departures from the assumptions. There have been some of these done by biologists. No theory is available, in particular the notion of influence function needs distances to be defined in both tree-space and data-space. Neither have been studied in this context.

7.5. Identifiability. Making the maximum likelihood model more flexible to encompass more biological realism is blocked by the problem of non-identifiability. When both branch lengths and substitution rates are free to vary, the model ceases to be identifiable. This is studied in Chang (1995).Too many parameters and too little data are the plight of phylogeny. This will only get worse as one starts to study real biological data with all the dependency between characters included.

8. Evaluating phylogenies. Various questions that biologists need to answer after building a tree $\hat{\mathcal{T}}$ from their data are:

1. How sure am I that this clade exists?
2. Can I be more confident in clade A than in clade B?
3. If the data were slightly wrong(a bad alignment, a poor reading of the characters), how far off would my tree be?
4. How much support does this clade have from the data?

These questions do not necessarily have to do with a precise stochastic setting as many authors have pointed out. For interesting reflexions on statistics in a non-stochastic setting see Freedman and Lane (1983). For a discussion of the foundation of the use of the bootstrap from the biological point of view see Sanderson (1995).

8.1. The bootstrap. To clarify some of the messy issues here I will try to develop a somewhat geometric analysis of the problem, this is more fully developed in Efron, Halloran and Holmes (1996).

Suppose that the number of characters is fixed at k and the number of units or species is N. The data are k characters from an alphabet of length A (maybe $A = 5, \{A, G, C, T, -\}$ or $A = 21$ amino acids and a '-'). The set of all possible columns of N species from an alphabet of size A is $S = A^N$. Under the assumption that the columns of X are exchangeable, a data matrix X can be associated with a unique $\hat{\pi}$, the vector of relative frequencies of each type of possible column. This is not an economic way of coding the data. The vector is of length S and is extremely sparse, but it is conceptually useful here. The tree-estimation process associates to this data vector $\hat{\pi}$ an estimated tree $\hat{\mathcal{T}} = \mathcal{E}(\hat{\pi})$. Of interest are properties of the estimated trees $\hat{\mathcal{T}}^* = \mathcal{E}(\hat{\pi}^*)$ for neighboring $\hat{\pi}^*$.

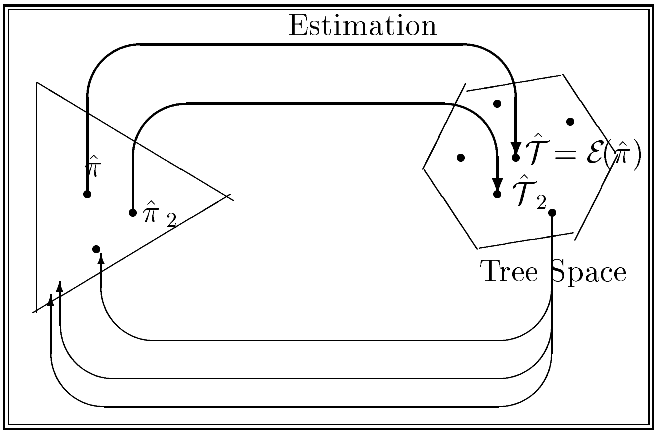

FIG. 9. *Estimation is a function from the data to tree space.*

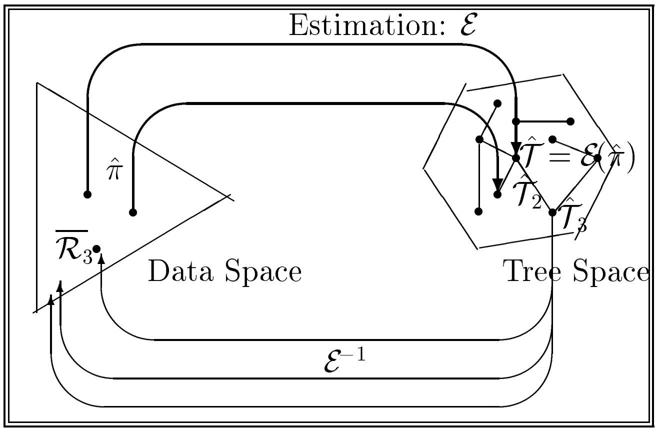

FIG. 10. *Several different data sets could give the same estimated tree.*

Nonparametric bootstrapping is a way of creating a neighborhood of close, plausible frequencies $\hat{\pi}^*$ by redistributing k columns among all the observed columns. Looking at the associated trees provides 'neighboring trees'. Properties of the corresponding neighborhood of $\hat{\mathcal{T}}$ are supposed to represent properties of the neighborhood of the true tree $\mathcal{T}$.

We are interested in trees that are obtained as images of possible frequency vectors $\hat{\pi}^*$. The idea is to find out if for some $\hat{\pi}^*$ "near" $\hat{\pi}$, the tree $\hat{\mathcal{T}}^*$ is different from $\hat{\mathcal{T}}$.

For instance, in the second figure above, region $\mathcal{R}_3$ is the set of all frequencies that would have given the same tree $\hat{\mathcal{T}}_3$.

The above describes nonparametric bootstrapping. Parametric bootstrapping is also a way of studying some aspects of the neighborhood of $\hat{\mathcal{T}}$. This is done by generating new π^* vectors as simulated through the relevant stochastic model taking the tree and the necessary parameters to be those estimated from the data set. `Seq-gen`, Rambaut and Grassly (1997),is one of the software packages available that enables such a study. (See appendix for an example of its use).

Simulating data from a given tree and stochastic model has also been much used to experiment with the nonparametric bootstrap in the absence of necessary theory. (See for instance Chernoff (1997) or Berry and Gascuel (1994)).

Another method for obtaining properties of the neighborhood of the estimated tree is **Bremer support**. This provides a neighborhood directly in tree space by relaxing the optimality criteria somewhat, so that for instance, trees that are up to 10 steps longer are also considered. Or, one could continue to relax optimality until a clade disappears. This was what Bremer (1988)originally suggested. This gives a measure of the diameter of a neighborhood around $\hat{\mathcal{T}}$ defined by contour lines of the function that is optimized.

Once a set of neighboring trees has been generated, there are different ways of using them. Mostly one wants to summarize the properties of these neighborhoods in tree space, again an unsolved statistics problem. One approach is detailed in the following section.

8.2. Summarizing several trees. The **consensus tree** is a notion which is quite useful when several trees have been obtained, either through a perturbation analysis such as bootstrapping, or just because there is not a unique optimal tree but several. One then needs to see how much the various trees concord. Two trees that agree are called **congruent**. Several propositions are available:

- Majority rule consensus.
- Strict consensus.
- Quartets.
- Compatible components.

`Phylip` offers either of the first two. The first creates a tree where the clades are those that have a majority of trees in their favor. In the second, strict consensus, only clades that have unanimous support are shown, others appear as 'unresolved'. An example output from `Phylip` can be found in the appendix.

Current preoccupations of biologists seem concentrated on how to split up the NP-hard problem of finding the optimal tree and recombine various partial solutions. One of the solutions proposed is to divide up the data at random in a cross-validation type procedure, and re-unite the trees with consensus methods. (This is called the `parjack` by Farris, et al. (1996)) Another is to use all the quartets and recombine them. This is called the `quartet puzzling` method and there is available software called `puzzle`, (Strimmer and von Haeseler, (1996)) which provides ways of doing this for the maximum likelihood criterion based trees. Certainly, more study is needed on combining trees.

Biologists exhibit bootstrap results by drawing a consensus tree. The number of times a given monophyletic group or clade appears on the tree is divided by the total number of trees simulated. This number is written along the branch of the consensus tree as an indication of how 'sure' one could be of the clade. In an abuse of nomenclature, it is called **bootstrap support**.

Interpretation of such a usage is particularly difficult for anyone who has had enough training in probability to use probability trees where the numbers along the branches are conditional probabilities.

This method is so popular with certain schools of biologists that any paper exhibiting a tree without "bootstrap support" numbers on the branches is rejected.

8.2.1. Of the use of p-values?. The bootstrap support is often assimilated to a p-value, the technical discussion of such an interpretation has already been given elsewhere (Efron, Halloran and Holmes, 1996). Although one can ponder whether several p-values associated with a same tree with the same data set one shouldn't worry about the multiple testing aspect, I will only raise a philosophical issue here. Don Ylvisaker pointed out during the workshop that it is becoming customary in court cases to ask statisticians to stand up in court and state their 'p-values' as evidence. This seems to have replaced the notion of an expert. For recognizing fingerprints, the expert says whether the fingerprints were beyond the shadow of a doubt those of a certain person. Although tempting as it may be to quantify the 'beyond the shadow of a doubt' as a number (with eventually several decimal places of precision....) these p values are in fact meaningless. We know that only approximate answers are possible.

8.2.2. Why can the bootstrap run into trouble?. The first time I saw the use of the bootstrap in this context, it seemed that the role of variables and observations had been reversed as compared to the traditional setup. This may create confusion for statisticians accustomed to data matrices with few columns (variables) and many observations (rows).

Here are some other possible sources of error in bootstrapping:

1. Discreteness of the underlying statistic: the tree is a discrete statistic, for which no applicable theory exists for the use of the boot-

strap with reasonable amounts of data. Large deviations as developed by Newton (1996) are unfortunately not applicable. Zharkih and Li (1995) defined the "partial bootstrap" that statisticians will recognize as a m-out-of-n bootstrap that attempts to fix the problems that the bootstrap encounters when the estimated tree $\hat{\mathcal{T}}$ is close to several possible neighboring trees.

2. The statistic $\hat{\mathcal{T}}$ is based on a maximum. It is well-documented that bootstrapping doesn't work for maximums of random variables (Bickel and Freedman, 1982).
3. Overparametrization of the model compared to quantity of data available. In multivariate regression for example, the bootstrap fails completely when the number of variables, and so the number of parameters, becomes of the same order as the number of observations. Work by Freedman and Peters (1984)documents this carefully. Another well explained example can be found in the use of the bootstrap to estimate bias in classification and regression trees (CART), see Breiman and Stone (1984) who explain the magical $1 - \frac{1}{e}$ factor also rediscovered by Harshman (1994).
4. Non-independence of observations: The closer we want the model to adhere to real nucleotide data, the more one sees that the characters are not independent. The codons (triplets of DNA) have to be dependent as there are only certain ones that are possible, those that code for certain amino-acids. There is also well documented **secondary structure** [9] across the sequences which is also evidence against independence. The columns could be considered conditionally independent given the tree, but I have not found any literature explaining this different assumption. It seems that the dependence structure is precisely what one is trying to find in the tree structure.
   Statisticians will recognize here a wonderful field of application for methods of inference more precisely tailored for dependent data, that is, block-bootstrapping, Markov Chains, etc..
5. Non identity of the distribution at different states. Any graphical display such as can be seen in the appendix shows that there are regions where there are many changes and other more stable regions. This spatial dependency should be integrated into the bootstrap.

8.3. Probability distribution on trees. A statistician considering the inferential part of the analysis of trees would characterize how close we believe the estimate to be to the true tree by using sampling theory. This builds on a probability distribution on the space of all trees.

The difficult aspect of this problem is that there are exponentially many possible *trees* that the parameter can take on. The classical non-parametric approach to this would be to put a multinomial probability

[9] The sequences fold and parts react together to perform certain functions.

model on the whole set of trees. This would have dimension $d \sim N^N$, N being the number of species. It could be possible to use a different parametric approach than the substitution model, using prior knowledge on the species' relations or the tree's form. The use of the outgroup strategy explained above is a special case of this. For instance, one could use functions such as the depth of the tree, the number of two-leaved clades, the balance of the tree, etc, and create exponential families through these. The more the parameters, the closer we can come to nonparametric models while keeping a hold on the overall structure of the tree.

Seen this way we can understand that as the maximum likelihood method puts a low dimensional surface through this high dimensional space, its chances of finding a tree 'near' the true tree may be quite low.

The notion of a tree that is *near* the true tree has not been discussed here, nor very much in any of the biological literature. Waterman and Smith define the NNI metric which is at the basis of the 'elementary steps' that Pearl, Doss, Li (1997)use in their Gibbs sampler for generating posterior distributions on tree space. The 'branch swapping' methods used both in PAUP (Swofford 1998) and NONA (Goloboff 1994) also use these elementary steps to explore parts of tree-space searching for optimal trees. We refer the interested reader to Diaconis and Holmes (1998) for reviews of possible distances on trees and random walks on the space of trees that have various desirable properties.

Defining a graph of trees that are nearest neighbors in some sense[10] can be useful. Distances between trees can be defined as the number of edges separting the trees in this graph. Random walk on trees can be seen as random walk on this graph.

Another idea that is possible when a distance between trees has been defined is to look for trees that are suboptimal, but close to **interpretable** trees.

The Bayesian approach advocated and carried through by Doss, Pearl and Li (1996) and Mau, Newton and Larget (1999) defines parametric priors on the space of trees, and then computes the posterior distribution on the same subset of the set of all trees. These enable precise confidence statements in a Bayesian sense.

9. Exploratory indices. For a statistician starting an analysis of aligned molecular data on N species, a first question might be about the relevance of building a tree, or how far the data lie from a 'reasonable tree'. That is basically how tree-like are the data? Of course given complete freedom, we can always build a tree that obeys certain rules and connects the species.

Each of the different tree-building contexts, parametric, semi parametric and nonparametric can be submitted to such an evaluation in a coherent

[10]they may differ by the transposition of leaves, or by a pruning/reconnection move

way. We follow previous work from Sattvah and Tversky (1977) who study trees as compared to planar multidimensional scaling in the reconstruction of distances for the use of hierarchical clustering for psychological data who suggest the following indices of treelike-ness. For $d(i,j)$ the distance as measured by the distance matrix between units i and j and $d_{\hat{\mathcal{T}}}(i,j)$ the distance as measured along the tree.

$$STRESS = \sum \frac{|d(i,j) - d_{\hat{\mathcal{T}}}(i,j)|}{\sum d(i,j)}.$$

$STRESS$ measures how well the tree reconstructs all the distances.

Measures of how tree-like the data are include the consistency index, a measure used by parsimony-tree builders. It is defined as the ratio of the minimum number of steps a tree with k characters and N species would need, divided by the actual number of steps needed. Thus this index is always between 0 and 1.

Measures of tree-likeness provide size estimates for what could be considered a reasonable neighborhood within which to search for an interpretable tree. Again, there is much to do here.

10. Conclusions. There are suggestions that statistical theory can make to help biologists, here are some examples:

- An evaluation of some of the error could be made by finding how *far* the data are from being tree-like, this could indicate what *size* the neighborhood of possible trees should be. Here I have used notions of distance both in the data space and the tree-space without defining them, this should be a first step for theory (see Diaconis and Holmes, 1998).
- Verification of assumptions and quantification of notions of robustness should go hand in hand.
- There should be a coherence of methods for each part of the analysis. If there *is* an underlying model used in the alignment procedure, then the same model should be employed throughout the tree-building process and its validation. For instance a 2 parameter mutation model used to align sequences implies that a distance based method using Kimura's distance is appropriate. This should be followed by parametric bootstrapping using this same parametric model and software such as `Seq-gen` (Rambaut and Grassly, 1997) [11]. This is not circularity, it is coherence.
  On the other hand, if the method is nonparametric, alignment will probably be better done either by hand in an exploratory fashion or at least without a parametric model, validation methods can include nonparametric bootstrapping, although if independence be-

[11] Using as arguments the estimates of transition/transversion ratio and nucleotide frequencies provided by the data

tween the branches is not assumed, there is no meaning to creating bootstrap numbers along the branches representing averages. It would seem more correct to give the more frequently obtained trees, with their probabilities.

- Monte-Carlo Markov chains for generating sampling distributions on tree space seems like an interesting one (Mau, Newton & Larget, (1999))
- This is a high dimensional problem, *curse of dimensionality* tells us there is NO reason to melt it down to just a planar representation with numbers along the branches without more ado.....

On the other hand there is much to be learnt from the clever algorithms that are being developed by cladists to attack this complex problem, for instance the successive weighting algorithm (Farris 1969) could be transposed into a statistical framework for regression. The procedure reweights the characters after the first tree is found, downweighting those that are discordant with it, and then repeating this until an optimal tree is found. This is like an iterated reweighted least-squares method. Goloboff (1997) has proposed a less computer intensive version of this that creates the weights once only, and assigns an overall cost to each tree taking the weights into account. This is like a downweighting least-squares method.

Those who have run simulation studies on constructing trees have noticed that the bootstrap combined with consensus methods has a propensity to correct bias. Thus Berry and Gascuel (1997) and Erdös, Steel, Szèkely and Warnow (1997) have rediscovered this property of the bootstrap's, already documented in the general case (see Efron and Tibshirani, 1993). Thus combining many bootstrap data sets and then taking a consensus tree has been shown empirically to produce a better estimate than just a one-pass parsimony optimization. However no theory has yet been developed in this case to explain and quantify the improvement.

Serious statistical considerations have also led to many other rediscoveries. For example, Zharkikh and Li (1995) rediscovered the merits of the m-out-of-n bootstrap in the multiple decision context that occurs when there are more than two trees that are plausible given the data.

Certainly I feel that the two fields of evolutionary biology and statistics would gain in more interdisciplinary work.

11. Acknowledgements. Many thanks to those who have wasted their precious time reading and discussing this work; Herman Chernoff, Jerry Davis, Jeff Doyle, Persi Diaconis, Brad Efron, David Freedman, Wen-Hsiung Li and Kevin Nixon. Thanks to Brigitte Charnomordic who wrote the interactive alignment program, and to Tandy Warnow, Lászlo Székely and Joe Chang for sending me copies of their work. And thanks to Betz Halloran for inviting me to this IMA workshop which has been a wonderful opportunity to talk with colleagues.

Note added in proof: I also thank Joe Felsenstein who took the trouble

of sending me many pages of comments that could unfortunately not be accomodated for because of time constraints.

APPENDIX

Examples of data set. I have used two data sets for my examples, the T7 data experimentally generated phylogeny, Hillis et al. (1992) for which the parsimony program will be seen to produce the correct answer. Here is the part of the data set (in `phylip` form) composed of the informative sites:

```
    9   21
R          C C G C C G G C C G G C C A G C G G G G T
J          C C C C G T A C C G G T C A A C G G G G T
K          T C C C G C A C C G A T C A A T G G G G G
L          T C C C G C A C C G A T C A A T G G G G G
M          C T C C G T A C C G G T C A A C G G G G T
N          C C T T A C G T T A G C T G G C A A A A T
O          C T C C G C G C T G G C C G G C A G A A T
P          C C C C A C G C T G G C C G G C A G A A T
Q          C C T T A C G T T A G C T G G C A A A A T
```

If the data set is put into a file called infile it will automatically be processed by any `phylip` program that is called. Otherwise if there is no current infile, `phylip` will ask for a file name, then there is a dialogue menu that allows the user to specify all the options.

Parsimony tree. This is part the output from the `phylip` command `dnapars`:

```
One most parsimonious tree found:

                 +--------O
     +-----------6
     !           !  +-----P
     !           +--7
     !              !  +--Q
     !              +--8
  +--5                 +--N
  !  !
  !  !                 +--L
  !  !           +-----3
  !  !           !     +--K
--1  +-----------2
  !              !     +--M
  !              +-----4
  !                    +--J
  !
  +-----------------------R

  remember: this is an unrooted tree!

requires a total of     25.000

 steps in each site:

         0   1   2   3   4   5   6   7   8   9
     *-----------------------------------------
    0!       1   2   2   1   2   2   1   1   1
   10!   1   1   1   1   1   1   1   1   1   1
   20!   1   1
```

The output file called `treefile` contains the following line (the tree in parentheses format):

```
(((O,(P,(Q,N))),((L,K),(M,J))),R);
```

Maximum likelihood trees: Output from `phylip` program `dnaml`:

```
Nucleic acid sequence Max. Likelihood, vers. 3.572c
Empirical Base Frequencies:
   A       0.27778    G 0.22685
   C       0.22325  T(U)0.27212
Transition/transversion ratio =   2.000000
(Transition/transversion parameter =   1.519971)
  +J
  !
  !        +R
  !     +--1
  !     !  !  +N
  !     !  +--4
  !     !     !  +O
  !  +--5     +--3
  !  !  !        !  +P
  !  !  !        +--2
--7--6  !           +Q
  !  !  !
  !  !  +L
  !  !
  !  +M
  !
  +K
Ln Likelihood =  -344.10331
Examined   95 trees
 Between        And       Length      Approx. Confidence Limits
 -------        ---       ------      ------- ---------- ------
   7          J          0.00006     (     zero,    infinity)
   7             6       0.00003     (     zero,    infinity)
   6             5       0.00006     (     zero,    infinity)
   5             1       0.00936     (     zero,     0.02236) **
   1          R          0.00466     (     zero,     0.01384) **
   1             4       0.00469     (     zero,     0.01389) **
   4          N          0.00462     (     zero,     0.01369) **
   4             3       0.00003     (     zero,    infinity)
   3          O          0.00462     (     zero,     0.01369) **
   3             2       0.00003     (     zero,    infinity)
   2          P          0.00462     (     zero,     0.01369) **
   2          Q          0.00003     (     zero,    infinity)
   5          L          0.00006     (     zero,    infinity)
   6          M          0.00003     (     zero,    infinity)
   7          K          0.00003     (     zero,    infinity)
   *  = significantly positive, P < 0.05
  ** = significantly positive, P < 0.01
```

How tree-like were the data?. Here is the distance as computed by Jukes-Cantor distance between the bacteriophage species:

```
R    0    4    8    8    5   11    6    6   11
J    4    0    3    3    0   12    6    6   12
K    8    3    0    0    4   15   10   10   15
L    8    3    0    0    4   15   10   10   15
M    5    0    4    4    0   13    6    7   13
N   11   12   15   15   13    0    5    3    0
O    6    6   10   10    6    5    0    1    5
P    6    6   10   10    7    3    1    0    3
Q   11   12   15   15   13    0    5    3    0
```

as compared to the distances along the branches of the 'best' distance based tree:

```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
[1,]    0    7   10   10    8   15    9    9   15
[2,]    7    0    5    5    1   16   10   10   16
[3,]   10    5    0    0    5   19   13   13   19
[4,]   10    5    0    0    5   19   13   13   19
[5,]    8    1    5    5    0   16   11   11   16
[6,]   15   16   19   19   16    0    8    6    0
[7,]    9   10   13   13   11    8    0    2    8
[8,]    9   10   13   13   11    6    2    0    6
[9,]   15   16   19   19   16    0    8    6    0
```

For which the stress was:

```
> sqrt( sum((d7a-d72f)^2)/sum(d72f^2))
[1] 0.1205607
```

Parametric bootstrap generation of sequences. Suppose we had the `treefile` from a previous `phylip` output, the generation of sequences is done using `Seq-gen` (Rambaut and Grassly, 1997) by :

```
seq-gen -mHKY -t3.0 -l27 -n100 < treefile > example.T7
```

For which the output looks like:

```
Sequence Generator - seq-gen, Version 1.04
(c) Copyright, 1996 Andrew Rambaut and Nick Grassly
Department of Zoology, University of Oxford
South Parks Road, Oxford OX1 3PS, U.K.
Simulating 11 taxa, 27 bases
  for 1 tree(s) with 100 dataset(s) per tree
Branch lengths assumed to be number of substitutions
per site
Rate homogeneity of sites.
Model=HKY
  transition/transversion ratio = 3 (kappa=6)
  frequencies = A:0.25 C:0.25 G:0.25 T:0.25
0%|____________________|100%
  [....................]
Time taken: 0.12 seconds
```

The data file example.T7 generated looks like this:

```
 11 27
Pfa4      CCGACCTCCAAGATTCGCTATGACAAT
Pvi10     CCGACCTCCAAGATTCGCTATGACAAT
Pcy9      CCGACCTCCAAGATTCGCTATGACAAT
Pkn8      CCGACCTCCAAGATTCGCTATGACAAT
Pfr7      CCGACCTCCAAGATTCGCTATGACAAT
Pbe5      CCGACCTCCAAGATTCGCTATGACAAT
Pma3      CCGACCTCCAAGATTCGCTATGACAAT
Pga11     CCGACCTCCAAGATTCGCTATGACAAT
Plo6      CCGACCTCCAAGATTCGCTATGACAAT
Pme2      CCGACCTCCAAGATTCGCTATGACAAT
Pre1      CCGACCTCCAAGATTCGCTATGACAAT
 11 27
Pfa4      ATGGTAGCGGATAACTGACTTCATCGA
Pvi10     ATGGTAGCGGATAACTGACTTCATCGA
Pcy9      ATGGTAGCGGATAACTGACTTCATCGA
Pkn8      ATGGTAGCGGATAACTGACTTCATCGA
Pfr7      ATGGTAGCGGATAACTGACTTCATCGA
Pma3      ATGGTAGCGGATAA.........etc
```

This file example .T7 was then submitted to the `phylip` program `dnapars` with the option multiple data sets indicating that there were 100 data sets to analyze, the first part of the output from this looked like this:

```
((R,(((((M,K),L),N),Q),(J,P))),O)[0.0100];
((R,(((((M,K),L),N),(J,Q)),P)),O)[0.0100];
((R,(((((M,K),L),(J,N)),Q),P)),O)[0.0100];
((R,(((((M,K),(J,L)),N),Q),P)),O)[0.0100];
((R,(((((M,(J,K)),L),N),Q),P)),O)[0.0100];
((((((((J,M),(R,K)),L),N),Q),P),O)[0.0100];
((((((J,(R,M)),K),L),N),Q),P),O)[0.0100];
(((((((R,J),M),K),L),N),Q),P),O)[0.0100];
((R,((((((J,M),K),L),N),Q),P)),O)[0.0100];
((((((R,(J,M)),K),L),N),Q),P),O)[0.0100];
(((R,J),(((((M,K),L),N),Q),P)),O)[0.0100];
((J,(R,(((((M,K),L),N),Q),P))),O)[0.0100];
((R,(J,(((((M,K),L),N),Q),P))),O)[0.0100];
((R,((J,((((M,K),L),N),Q)),P)),O)[0.0100];
((R,(((J,(((M,K),L),N)),Q),P)),O)[0.0100];
((R,((((J,((M,K),L)),N),Q),P)),O)[0.0100];
((R,(((((J,(M,K)),L),N),Q),P)),O)[0.0100];
(((J,(R,M)),((((K,L),N),Q),P)),O)[0.0100];
((((R,J),M),((((K,L),N),Q),P)),O)[0.0100];
(((R,(J,M)),((((K,L),N),Q),P)),O)[0.0100];
((M,((R,J),((((K,L),N),Q),P))),O)[0.0100];
(((R,J),(M,((((K,L),N),Q),P))),O)[0.0100];
(((R,J),((M,(((K,L),N),Q)),P)),O)[0.0100];
(((R,J),(((M,((K,L),N)),Q),P)),O)[0.0100];
(((R,J),((((M,(K,L)),N),Q),P)),O)[0.0100];
(((R,(M,(J,K))),(((L,N),Q),P)),O)[0.0100];
((((J,M),(R,K)),(((L,N),Q),P)),O)[0.0100];
(((R,((J,M),K)),(((L,N),Q),P)),O)[0.0100];
```

Notice at the end of each tree is associated a weight.

Putting trees together: consensus.
These trees are usually summarized by programs like `phylip`'s `consense`. This has an output tree that looks like this:

```
Majority-rule and strict consensus tree program,
                                 version 3.572c
CONSENSUS TREE:
the numbers at the forks indicate the number
of times the group consisting of the species
which are to the right of that fork occurred
among the trees, out of 100.00 trees
                      +----L
                 +-22.0
            +-22.0    +----K
            !    !
       +-22.0    +---------M
       !    !
       !    +--------------N
       !
  +100.0              +----O
  !    !         +-22.0
  !    !    +-22.0    +----R
  !    !    !    !
  !    +-22.0    +---------J
  !         !
  !         +--------------P
  !
  +------------------------Q

   remember: this is an unrooted tree!
```

Proof of sparsity of the data. Here are all the possible "boiled down" patterns in the `Vitis Vinifera` data:

```
11111111111111111111 11111111111111111121 11111111111111112111 11111111111111121111
                 187                    2                   17                    4
11111111111111211111 11111111111111213111 11111111111112111112 11111111111112112112
                   4                    1                    1                    1
11111111111231111113 11111111112111111211 11111111112111113211 11111111121111111131
                   1                    3                    1                    1
11111111121211111111 11111111121211111131 11111111122111121222 11111111211111111111
                   7                    1                    1                    5
11111111211111112111 11111111211111113111 11111111211111133111 11111111211111211111
                   1                    1                    1                    1
11111111211111331111 11111111221111111111 11111111221211322111 11111111221221112112
                   1                    1                    1                    1
11111111223121314352 11111111231311241111 11111111234311141441 11111112111111111111
                   1                    1                    1                    6
11111112111111112111 11111112111111113111 11111112112111113221 11111112345141141514
                   1                    1                    1                    1
11111121111111111111 11111121111111112111 11111121131311111111 11111123415141411114
                   3                    1                    1                    1
11111211111111113111 11111231112111111111 11112111111111111111 11111211311111111111
                   1                    1                    1                    1
11111211321111114111 11111221133132111132 11112111111111111111 11112111111111211111
                   1                    1                   10                    2
11112111111211111211 11112111112321131212 11112111113111111311 11112111131111111111
                   1                    1                    1                    1
11112111211111111111 11112111211111311111 11112113311111141111 11112113212121121121
                   1                    1                    1                    1
11112211111131113113 11112211111121111112 11112211112111113211 11112211311121211111
                   1                    1                    1                    1
11112211333321145362 11112212121211111111 11112212131321111111 11112311111111114111
                   1                    1                    1                    1
11121111111112111111 11121111113131111333 11121113423212444311 11121112133321113332
                   2                    1                    1                    1
11121112322121122111 11121221341111114115 11122111111121111112 11211111111111111111
                   1                    1                    1                    2
11211111112111113221 11211111112121111113 11211111122211111221 11211111131111111111
                   1                    1                    1                    1
11211113222121121211 11211211211111111111 11211221211211112221 11211321211111114111
                   1                    1                    1                    1
11212221211111111111 11212221211111112111 11212221211312112332 11212221213131111111
                   3                    1                    1                    1
11212221222122121212 11213111222221112231 11213331211331121321 11222222111111111111
                   1                    1                    1                    1
11222222221222212112 12111111111112111111 12111111111121112112 12113222122222111242
                   1                    1                    1                    1
12121111111112131112 12121112111111111111 12121112111112111111 12121112122121121121
                   1                    1                    1                    1
12121112122212211211 12121112122222222222 12121112312131211111 12121113111112114111
                   1                    1                    1                    1
12121332121411212411 12122222111212212222 12123112222222222222 12221111111112111111
                   1                    1                    1                    1
12221112122222222222 12221112222222222222 12221222211112111111 12221223212222222222
                   1                    1                    1                    1
12222112212221231221 12222222211111311111 12222222211122113112 12222222112112121211
                   1                    1                    1                    1
12222222211311211131 12222222212121212211 12222222212222222222 12222222222212112222
                   1                    1                    1                    1
12222222222212121221 12222222222222222212 12222222222222222222 12222222222222223234
                   1                    1                    1                    1
12222222223222233332 12222222345463455546 12222223442522162222 12222223454242426424
                   1                    1                    1                    1
12222322421242442144 12223222222222222222 12223222322332333223 12223332212121214111
                   1                    1                    1                    1
12223332241122141112 12223442325252533556 12232222111112124111 12232222111122111112
                   1                    1                    1                    1
12321112144141242111 12321112322222232222 12323332311411255441 12323323342432526223
                   1                    1                    1                    1
12324222312112112231 12324444444412144441 12324552464432782443 12333322111412112111
                   1                    1                    1                    1
12333334355251324326 12334333155553155565 12343335312522226562 12344443533635273661
                   1                    1                    1                    1
12345222121112252211
                   1
```

How some regions are stable where others are are variable.

The regions of low variability have mostly dots in them, high variability is shown by the letters, (except for complete columns that indicate just a difference with the first taxa).

These patterns can be made even more visible by the use of color, (see Charnordic and Holmes, 1997).

A consensus bootstrap tree for the vitis vinifera data.

```
                                   +----Tre
                              +-100.0
                         +-48.3    +----Psa
                         !    !
                    +-37.6    +---------Ath
                    !    !
               +-32.1    +--------------Pde
               !    !
          +-19.0    +-------------------Phy
          !    !
          !    !                   +----Pta
     +-49.7    +----------------61.7
     !    !                        +----Fan
     !    !
     !    +-----------------------------VVi
     !
     !                   +--------------Osa2
     !                   !
+-100.0             +-54.2         +----Hvu3
!    !              !    !      +-82.7
!    !              !    +-84.3    +----Tae
!    !         +-100.0        !
!    !         !    !         +---------Hvu2
!    !         !    !
!    !         !    +-------------------Zma2
!    +------100.0
!              !                   +----Pgl
!              !              +-45.4
!              !         +-75.4    +----Hvu1
!              !         !    !
!              +------95.5    +---------Zma1
!                        !
!                        +--------------Osa1
!
+---------------------------------------Stu


remember: this is an unrooted tree!
```

REFERENCES

[1] ALDOUS D. A., *Probability Distributions on Cladograms*, in Random Discrete Structures, IMA series, **vol. 76**, (1996), pp. 1–18, Springer Verlag, NY.

[2] ANDERSON T. W. AND RUBIN H., *Statistical inference in factor analysis*, Berkeley Symposium on Math. Stat. and Probab., (Third), Ed. J. Neyman, **vol. 5**, (1956), pp. 111–150.

[3] BERRY AND GASCUEL O., *Strict Consensus Parsimony*, COCOON, 1997.

[4] BICKEL P. AND FREEDMAN D., *Some asymptotic theory for the Bootstrap*, Annalls Statistics, **9**, pp. 1196–1217.

[5] BREMER K., *The limits of amino-acid sequence data in angio-sperm phylogenetic reconstruction*, Evolution, **42**, (1988), pp. 795–803

[6] CHANG J., *Inconsistency of Evolutionary Tree Topology Reconstruction Methods when Substitution Rates Vary across Characters*, Mathematical Biosciences, **134**, (1996), pp. 189–215.

[7] CHANG J., *Full reconstruction of Markov Models on Evolutionary Trees: Identifiability and Consistency*, Mathematical Biosciences, **137**, (1996), pp. 51–73.

[8] CHARNOMORDIC B. AND HOLMES B., `Dnaview`, an interactive viewer for alignment and tree building, (1997), Unpublished manuscript and software.

[9] CHERNOFF H., *Problems with Bootstrapping Phylogenies*, IMA conference on Statistics and Genetetics, (1997), unpublished communication.

[10] DIACONIS P. AND EFRON B., *Computer intensive methods in statistics*, Scientific American, **248**, (1983), pp. 116–130

[11] DIACONIS P. AND HOLMES S., *Random walks on phylogenetic trees*, Techical report, Biometrics Unit, Cornell, (1997).

[12] DOYLE J.J., *Gene trees and species trees: Molecular systematics as one-character taxonomy*, Syst. Bot., **17**, (1992), pp. 144–163.

[13] EDWARDS, A. W. F. AND L. L. CAVALLI-SFORZA, *Reconstruction of evolutionary trees*, pp. 67–76, in Phenetic and Phylogenetic Classification, ed. V. H. Heywood and J. McNeill., Systematics Association **vol. 6**, Systematics Association, London, (1964).

[14] EFRON B., HALLORAN E. AND HOLMES S., *Bootstrap confidence levels for phylogenetic trees*, Proc. National Academy Sciences, **vol. 93**,(1996), pp. 13429–34.

[15] EFRON B. AND TIBSHIRANI R., *An Introduction to the Bootstrap*, Chapman and Hall, (1993), London.

[16] ERDÖS P. L., STEEL M. A., SZÈKELY L., AND WARNOW T. J, *Inferring big trees from short sequences*, to appear in Proceedings of ICALP, (1997).

[17] ERDÖS P. L., STEEL M. A., SZÈKELY L., AND WARNOW T. J, *A few logs suffice to build (almost) all trees*, (I) and (II) Tech reports, U. Penn. Computer Science Dept, (1997).

[18] FARRIS J.S., *A successive approximations approach to character weighting*, Syst. Zool, **18**, (1969), pp. 374–385.

[19] FARRIS J. S., *Methods for computing Wagner trees*, Syst. Zool., **219**, (1970), pp. 83–92.

[20] FARRIS J. S., *The logical basis of phylogenetic analysis*, in Advances in cladistics, **vol. 2**, (N. Platnick and V. Funk, eds.) (1983), pp. 7–36.

[21] FARRIS J. S., *The information content of the phylogenetic system*, Syst. Zool., **28**, (1979), pp. 483–519.

[22] FARRIS, J. S., ALBERT, V. A., KÄLLERSJO, M., LIPSCOMB,D. AND KLUGE A.G., *Parsimony jackknifing outperforms neighbor-joining*, Cladistics, **12**, (1996), pp. 99–124.

[23] FELSENSTEIN, J., *Statistical inference of phylogenies (with discussion)*, Journ. Royal Stat. Soc. A, **146**, (1983), pp. 246–272.

[24] FELSENSTEIN, J., `PHYLIP`, *(Phylogeny Inference Package) version 3.5c.*, Distributed by the author. Department of Genetics, University of Washington, Seattle, (1993).

http://evolution.genetics.washington.edu/phylip.html
[25] Foulds L. R. and Graham R. L., (1982) *The Steiner tree problem in phylogeny is NP-complete*, Adv. Appl. Math., **3**, (1982), pp. 43–49.
[26] Freedman D. and Lane D., *Significance testing in a non stochastic setting*, Festschrift for Eric Lehmann, (1983), pp. 185–208.
[27] Freedman D. and Peters S. C., *Some notes on the bootstrap in regression problems*, Journ. Bus. Ec. St., **2**, (1984), pp. 406–409.
[28] Friedman J. H. and Rafsky L., *Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests*, Annals Statistics, **7**, (1979), pp. 697–717.
[29] Gardner M., *The Last Recreations*, Copernicus-Springer Verlag, NY, (1997).
[30] Goloboff P., *Nona*, available from J. Carpenter, Entomology Dept, American Museum of NAtural History , 79th st.,New York, NY 10024-5192 (1995).
[31] Goloboff P., *Self-weighted optimization: Tree searches and character state reconstructions under implied transformation costs*, Cladistics, **12**, (1997), pp. 225–246.
[32] Harshman J., *The effect of irrelevant characters on bootstrap values*, Syst. Biol., **43**, (1994), pp. 419–424.
[33] Harvey P. H. and Pagel M. D., *The comparative method in Evolutionary Biology*, Oxford University Press, Oxford, (1991).
[34] Hillis D. M., Bull J. J. White M. E., Badgett M. R. and Molineux I. J., *Experimental Phylogenies: generation of a known phylogeny*, Science 255, (1992), pp. 589–592.
[35] Hillis D., Moritz C., and Mable B.,, *Molecular Systematics*, Sinauer, (1996), Boston.
[36] Kluge, A. C. and Farris J. S., *Quantitative phylogenetics and the evolution of anurans*, Syst. Zool., **18**, (1969), pp.1–32.
[37] Lavit, C., Escoufier, Y., Sabatier, R., and Traissac, P., *The ACT (STATIS method)*, Comput. Statist. Data Analysis, **18**, (1994), pp. 97–119.
[38] Li W. H., *Molecular Evolution*, Sinauer, Boston, (1997).
[39] Li S., Pearl D. K., Doss H., *Phylogenetic Tree Construction using MCMC*, Technical report no 583. Ohio Statistics Dept., (1996), submitted to Journ. American Statistical Association.
[40] Mau, B., Newton, M. A., and Larget B., *Bayesian phylogenetic inference via Markov Chain Monte Carlo Methods*, (1999) to appear Biometrics, vol.55.
[41] Newton, M. A., *Bootstrapping Phylogenies: Large deviations and dispersion effects*, Biometrika, **83**, (1996), pp. 315–328.
[42] Page R. D. and Charleston M., *From gene to Organismal Phylogeny: Reconciled Trees and the Gene Tree/Species Tree Problem*, (1997), Tech rep. Univ.Glascow., http://taxonomy.zoology.gla.ac.uk/rod/pubs.html
[43] Rambaut, A. and Grassly, N. C., *Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees*, Comput. Appl. Biosci., **13**, (1997), pp. 235–238.
[44] Rice K., Steel M., Warnow T. and Yooseph S., *Getting better topology estimates of difficult evolutionary trees*, U. Penn. Computer Science, Tech. Report, (1997).
[45] Sanderson M., *Objections to bootstrapping phylogenies: a critique*, Syst. Biol., **44**, (1995), pp. 299–320.
[46] Sarni-Manchado, P., Verriès C. and Tesnière C., *Molecular characterization and structural analysis of one dehydrogenase gene (GV-adh 1) expressed during ripening of grapevine (Vitis vinifera L.) berry*, Plant Science, **125**, (1997), pp. 177–187.
[47] Sattah S. and Tversky A., *Additive similarity trees*, Psychometrika **vol 42 no 3**,(1977), pp. 319–345.
[48] Schröder E., *Vir Combinatorische Probleme*, Zeit. Pur. Math. Phys., **vol 15**, (1870), pp. 361–376.

[49] Stanley R., *Enumerative Combinatorics*, **vol I**, 2nd edition (1996).
[50] Strimmer, K. and von Haeseler, A., *Quartet Puzzling: a quartet maximum likelihood method for reconstructing tree topologies*, Mol. Biol. Evol., **13**, pp. 964–969.
[51] Swofford, PAUP 4.0, (1998), Available from Sinauer,Boston.
[52] Tuffley and Steel M., (1997) *Links between Maximum Likelihood and Maximum Parsimony under a simple model of substitution*, Technical Report.
[53] Waterman M. S. and Smith T. F., *On the similarity of dendograms*, Jour. Theoret. Biology, **73**, (1978), pp. 789–800.
[54] Wheeler W., *MALIGN*, Dept. of invert., Am. Museum of Natural History, NY.
[55] Yokoyama S. and Harry D. E., *Molecular Phylogeny and evolutionary rates of alcohol dehydrogenases in verterbrates and plants*, Mol. Biol. Evol., **10**, (1993), pp. 1215–1226.
[56] Zharkikh, A. and Li W. H., *Estimation of confidence in phylogeny: The complete and partial bootstrap technique*, Mol. Phylogenet. Evol., **4**, (1995), pp. 44–63.