Available online at www.sciencedirect.com
SCIENCE DIRECT®
Theoretical Population Biology 63 (2003) 17–32
Theoretical Population Biology
http://www.elsevier.com/locate/ytpbi

Statistics for phylogenetic trees

Susan Holmes*

Statistics Department, Stanford University, Mail Code 4065, Sequoia 102 Stanford, CA 94305-4065, USA

Received 7 March 2002; received in revised form 27 August 2002; accepted 28 August 2002

Abstract

This paper poses the problem of estimating and validating phylogenetic trees in statistical terms. The problem is hard enough to warrant several tacks: we reason by analogy to rounding real numbers, and dealing with ranking data. These are both cases where, as in phylogeny the parameters of interest are not real numbers. Then we pose the problem in geometrical terms, using distances and measures on a natural space of trees. We do not solve the problems of inference on tree space, but suggest some coherent ways of tackling them.

Keywords: Phylogenetic trees; Sufficiency; Maximum likelihood; Parsimony; Bootstrap; Bayesian inference

1. Introduction

I will present a statistician's view of some of the difficulties biologists have had to overcome in dealing with the phylogenetic analysis of DNA sequences. In addition to "noise" of various kinds, the foremost difficulty is the number of conflicting signals contained in sequence data, many characters conflict with each other in terms of phylogenetic information. We will see that one possible solution is to let the output of such analysis be distributions on the space of possible rooted binary trees, rather than a unique tree. Distributions on trees are also useful for formulating the Bayesian perspective as well as for constructing confidence regions in a frequentist sense. One barrier to such an approach has been that distributions can be difficult to represent or summarize. Publication standards in the biological journals have dictated the presentation of a unique tree with "Bootstrap values" on the branches to try and give an indication of how sure one is of that particular edge. Statistical procedures have been used in other situations to address such difficulties. In particular, basic ingredients such as distances and probability distributions remain available, even if the classical mean-variance summaries are harder to define.

The statistical approach developed below allows for a seamless extension to consensus and supertree problems that are currently a preoccupation in the phylogenetic literature (Sanderson et al., 1998; Page and Holmes, 2000).

There have been long and painfully contentious polemics in the literature on the "right" method for estimating a phylogenetic tree given a DNA sequence (Kluge and Farris, 1969; Farris, 1983; Felsenstein, 1983). Section 2 shows that the statistical perspective sees the differences between maximum likelihood, maximum parsimony and distance-based methods as much more a matter of "degrees of freedom" allowed in the model than a matter for religious wars.

1.1. The one tree?

It seems that the biologist's dream of a unique "tree of life" for all the genomes and all species is not going to be realized as such. When molecular data started to become available, the hope was that the conflicting picture induced by morphological data would be clarified and there would emerge a unique, correct "species tree." Nothing so simple has occurred. Even when each gene is

*Tel.: +1-650-725-1925; fax: +1-650-725-8977.
E-mail address: susan@stat.stanford.edu (S. Holmes).

S0040-5809/02/$ - see front matter © 2002 Elsevier Science (USA). All rights reserved.
PII: S0040-5809(02)00005-9

taken separately, the data are difficult to coax into a unique tree.[1] Taken together the genes do not currently tell a clear story and definitely "The one tree" dream does not prevail (see Penny and Holmes, 2001, for some actual examples). Variability in this context has often been perceived as noise. In fact, it may be source of information if it is not eliminated too quickly. Trees can still be considered as a most useful unit of summary information and the best coding for each gene.

Statisticians only find this natural: what set of measurements concur in real data? Variability is the rule, not the exception. The study of the variability provides additional information. Statisticians have long elaborated methods for dealing with such variability both with the objective of answering precise questions through hypothesis testing and decision making as well as providing confidence regions instead of unique values.[2] Suppose data is summarized by a set of different trees as suggested by a set of different genes. Analyzing these, the new data are trees themselves. Much can be gleaned by analogy to other statistical problems, however, the very simple statistical algorithm:

1. estimate the parameter,
2. find the sampling distribution of the estimator

needs a sophisticated viewpoint as we venture away from the real line and all that is known in the classical paradigm of real parameters and real-valued vectors.

Phylogenetic trees have both a discreteness and a complexity that justify recourse to quite a battery of tools from statistics and geometry. We will investigate alternatives to simulation studies to access the intricacy of the various challenges raised by phylogenetics. We also emphasize using the correct terminology for various concepts. This helps find statistical resources on the internet or elsewhere where similar issues have already been addressed, instead of having to make up a theory from scratch every time.

1.2. Rounding

There is a tension between discrete combinatorial objects that are the trees as they define the branching order of the taxa and a continuous space of metric trees that come with edge lengths. Combining the combinatorial trees themselves seems to give better answers than using all the data combined to build just one tree. There is a tension in the choice of when the tree building steps should occur. Let us take a similar but simpler situation where this occurs.

Take the example where the parameter of interest is known to be integer $m \in \mathbb{N}$ (number of taxicabs in Barcelona), we seek to find an estimate $\hat{m}$ for m. The data are measurements $\mathscr{X} = \{x_1, x_2, \ldots, x_k\}$, maybe estimates from various authorities, gas stations, taxi companies, each of these observations might be either already be integer, or might have come from some sort of group average (with groups being of different size $n_1, \ldots, n_k$), in which case it need not be an integer but more generally may be real. One can imagine various procedures for estimating m.

- Taking the weighted arithmetic average of various estimates is optimal under a parametric approach[3] but it will not give an integer. A frequentist may just round $\bar{x}$ to the nearest integer.
- A completely nonparametric approach might be to take the median value, few would take the mode, as there are probably few repetitions.
- What will a Bayesian do? With no informative prior, one could put a uniform prior point mass on a basis of a realistic range of possible numbers, say 1000–10,000, then use the data to update the prior, making some distributional assumptions along the way.
- If the data given are the successive taxicab numbers $\mathscr{Y} = \{y_1, y_2, \ldots, y_n\}$ noted by a traveler, taking the maximum y_i will automatically be downward biased, and several authors have suggested schemes for reducing such bias (Lo, 1992).

If the measurements themselves do not come in as integers, but as real numbers, is it better to start by rounding them and then combine them into a mean or is it better to use all the information by directly computing the mean? Some authors have studied the problems involved in such a grouping of data and (Lehmann and Casella, 1998, p. 140) give an account of how to use the fact that one knows that a parameter is a integer with a normally distributed noise, as for as instance in chemistry (Hammersley, 1950). Several papers followed up on Hammersley's study of integer-valued parameters, Hammersley himself showed that the variance of the rounded mean is exponentially small. A discussant of

[1] Maximum likelihood estimation often provides indistinguishably close likelihoods (Steel, 1994), thus making the choice of a unique tree impossible. In parsimony-based methods, finding several most parsimonious trees is quite common.

[2] There is only one value for the speed of light, but there were many measurements, they even had disjoint confidence intervals, Youden (1972).

[3] Averages are justified by the parametric assumption of the bell-shaped curve. They are also nonparametric method of moments estimators, valid for all distributions with a mean.

the paper points out that efficiency[4] may not be the right criteria for comparing estimators. Even in this simple problem, there are several solutions to the statistical estimation problem. One can imagine that the much more complex tree problem is not going to be easy to settle.

The question of how data should be grouped when estimating an integer has also given rise to substantial discussions. This is of interest here because it is analogous to the discussions regarding the "total evidence" controversy in systematics (Page and Holmes, 2000, Chapter 8 or Page, 1996). Should all the data be combined into one long sequence and one tree built or should the information be cut into functional parts, each part then producing a tree and those trees then combined using some averaging or consensus method? In statistics, combining all the data into one sample is known to decrease the bias and make the estimation of the estimator's variance harder. The same tension occurs here with a more intricate bias-variance tradeoff. When doing nonparametric regression for image analysis with a large number of possible features, Amit and Geman (1997) noticed that for training set data, where the response is known, taking subsets of features, making the trees and then averaging them to get a final model, performed better than using all the features at once. This is also the heuristic behind contemporary statistical methods called *bagging* (Breiman, 1996).

1.3. Finding tree summaries for DNA-data

When dealing with integers under very weak assumptions, we can usually give a confidence interval surrounding the $\bar{x}$ and a discrete probability mass to the few integers contained in the interval. This helps mitigate the discreteness of the parameter and the averaging process. The estimate has a different state space than the final parameter. This is unimportant and will also happen for trees.

The essential "rounding" procedure dealt within phylogenetics is the replacement of one functional stretch of DNA (usually a gene) by the gene *tree*. This is obtained by estimation according to a treebuilding mechanism chosen by the biologist depending on personal preferences. After giving some statistical pointers to how to compare these different estimation mechanisms, we look at how to assess the estimate with confidence statements, either frequentist or Bayesian.

[4] For statisticians, an estimator is efficient if its variance is the smallest possible.

In order to formulate the problem in slightly more precise terms, we will give details about the two essential components in the statistical setting, the data and the parameter space.

1.4. The data

Given an observed data set made on s species and k genes each of n_k characters we may obtain a set of k trees (or networks) and even more (when one gene points to several trees). The data are sequences of DNA or amino acids, one for each species, having been aligned previously.[5] We thus decompose the problem sequentially. For simplicity's sake, the aligned sequences will be considered of equal length thus providing a matrix-block, for which each column is often called a pattern or a character. Below is a little artificial example

```
Homo    G A T A C C T G G
Pan     G A T G C A T G G
Gorilla A C T G C C T G G
Orang   G C T G T A T G G
Lemur   A C T G T A A G G
Rana    A C A G T A A T C
```

Biologists then distill the data into frequencies for each column pattern, often called the spectrum (see Page and Holmes, 2000, Chapter 5) losing the spatial information about the differences between mutation patterns in different zones. For most biologically meaningful models, this constitutes another loss of information. For instance, the second and sixth column are said to be incompatible, because they define incompatible partitions, the second column grouping Homo and Pan together, whereas column 6 separates them.

Unless the data are perfectly treelike, even the case of one gene-sequence will contain incompatibilities that can either be treated as noise or information. For instance if the characters can be considered as a mixture of two different treelike processes, they may conflict, but not necessarily be noisy. Each column defines a partition of the taxa into at most four groups. In most cases, the partition is a dichotomy, and each character can be seen as defining an edge in Fig. 1.

Seen this way, combining characters is the same as combining trees and an associative methodology might be more coherent if the actual sequences are not

[5] Sankoff and Cedergren (1983), Schwikowski and Vingron (1997), Hein (1989) have proposed procedures for making the tree and the alignment at the same time, and this makes a lot of sense (see a complete review in Durbin et al., 1998).

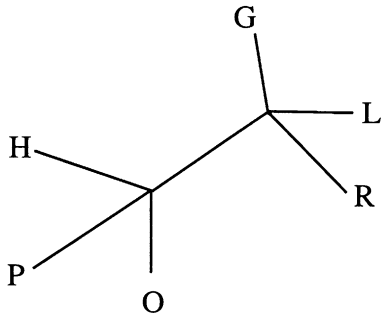

Fig. 1. One edge defined by the first column of the above data.

meaningful entities. If they are, one also has the option of associating not one single tree but a set of trees with weights, called a mixture or distribution of trees.

The reverse operation of coding each edge by a binary character was initially suggested by Farris (1973). Brooks (1981) used it as a way of transforming a tree into a set of binary characters and then combining trees (host and parasite) or combining a tree and other information (biogeographical or morphological). This same coding was used earlier in network addressing by Graham and Winkler (see Van-Lint and Wilson, 1992, Chapter 9 for a nice summary). It is natural to try and apply some of the mathematical development from coding theory to biology.

The inverse problem[6] has been treated in a parametric framework by Steel and Szekely (1999, 2002) who consider this as the inversion of a random function.

1.5. *A first transformation: from data to distances*

Whichever method is used to go from the aligned sequences to the phylogenetic trees, the basic information between sequences is then coded into distances. The changes that are shared by a monophyletic group in the tree, called a clade, are called homologous substitutions. They provide the information that makes tree-building work. If the system were perfect, there would only be just enough substitutions to define the clades and never any 'multiple hits' or changes occurring on the same characters several times. This is exactly what the code of Brooks–Graham–Winkler does. Biology is much messier and there are several difficulties that the estimation mechanisms have to address:

- Some of the changes remain invisible, because a character mutated and then mutated back. This is called a reversal.
- Some changes will be invisible because a second change occurs at the same position, erasing evidence of the first mutation.
- Some changes occur in parallel on different branches of the tree, making nonmonophyletic taxa seem more similar than they actually are. This is called parallelism or convergence.

One way around these problems is to give a Markovian mutation model and use distances that recorrect for the number of changes. The classical mutation models are parametric and vary in the number of degrees of freedom allowed in the 4×4 transition matrix (see Page and Holmes, 2000 or Li, 1997 for details). The sequences are supposed to be in their stationary distribution for the transition matrix. Thus, the characters occur at the root with some probability π. They are then submitted to possible changes along any given branch with probabilities depending on the length of those inner edges or branch lengths.

Given a tree and a mutation matrix, if the probabilities of mutations are all the same in any branch in the tree, the distribution of the sequences will be stationary and all the same. This is rarely observed in practice. Different taxa seem to have different stationary distributions for the nucleotides. Various suggestions for more believable models have been included into different procedures:

- Hidden Markov model for rate variation along the sequences (first introduced by Yang, 1995; Durbin et al., 1998 contains a review of these models in even wider contexts than phylogeny).
- A Gamma distribution of rate variation along sequences is used with much success by Yang (1994).
- Fitch and Markowitz (1970) introduced the concept of concomitantly variable codons (covarions), which asserts that at any one point in time and in any one branch, there are constraints on how many amino acid can be replaced. Covarion models are used for rate changes as in Lockhart et al. (1996, 1998).
- LogDet distances allow for different stationary distributions for different taxa in Lockhart et al. (1994) and Lake (1994).

2. Choosing an estimation method

The data are "rounded" off into trees by an estimation process that chooses trees according to one of several criteria (see Holmes, 1999 for a review):

[6] Going from an observed distribution of characters to a tree.

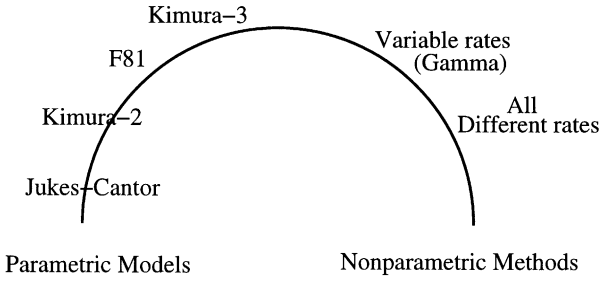

Fig. 2. The meeting of methods.

1. Closest tree (by a spectral analysis developed in Hendy, 1991).
2. Maximum parsimony.
3. Maximum likelihood.
4. A difference between the distances along the tree and the original distances between sequences.[7]

From a statistical perspective, the first two methods are nonparametric—they do not pre-suppose a finite dimensional parametric model, but are based on optimizing potentially infinite dimensional criteria, in the same way nonparametric density or regression methods minimize a smoothness penalty as we will explain in Section 2.1. Likelihood estimation postulates a finite dimensional model, here this is given by specifying the mutation transition matrix (although it seems common practice to both estimate the mutation model parameters and the tree from the same data without further ado).

Distance-based methods are in fact *semiparametric*. They use a parametric model for the estimation of distances between sequences which allows for corrections for multiple hits on a same character using the Jukes-Cantor, Kimura, or F81 models and a nonparametric hierarchical clustering heuristic to build the actual tree.

In Fig. 2, we schematize the three main approaches as a continuum from the right where there are infinitely many parameters to the left where the parametric model is of a small fixed dimension. Then Jukes-Cantor is the leanest and meanest with only one parameter for the mutation model, then the Kimura 2-parameter model. As the mutation model is made more flexible, the number of parameters increases to six in the General Reversible model. We can increase further the number of parameters by allowing different mutation rates in different parts of the tree. The two approaches, parametric and nonparametric, meet where the number of parameters overtakes the amount of information available in the data. This picture has been formally justified in the case of binary characters by Tuffley and Steel (1997)'s observation that parsimony and likelihood estimates coincide when rates are allowed to differ on different edges of the tree for each combination of edge and character.

As a community, statisticians are less polarized on the absolute choice of a method and do not currently have the acrimonious debates present in the biological literature on which method to use for the estimation procedure. The best estimation procedure depends on the data and question at hand. Further, theory often shows many different estimation procedures lead to the same answer, at least to good approximation. Statisticians are pragmatic, there is no unique right answer to which method to use. If the study is in the exploratory stages, with little knowledge of the model, a nonparametric procedure is preferred; it will be robust and informative. If sufficient knowledge of the model becomes available[8] more precise parametric models may be used. These may incorporate commonly accepted mutation rates for those particular genes. Using the same data to estimate the mutation rates and then using them on the tree is a tricky statistical problem often confronting those who practice empirical Bayes (Robbins, 1985).

A lot of energy has gone into comparing the methods in worst-case scenarios, this is actually quite wasteful, one really wants a method that works best most of the time or on average, and in the same context as the actual data appear. Thus, the enormous discussion on consistency (see Steel and Penny, 2000; pp. 843–846 for a useful review) should have been minimized, since there are never infinite data especially relative to the dimension of the underlying parameter space which is very large. Other measures of performance seem somewhat heavy handed: one such index was the probability $\rho(M, \mathscr{T}, \theta)$ of the method M being right or wrong given the tree $\mathscr{T}$. As a comparison in the taxicab example above, imagine scoring $\hat{m}$ as either right or wrong. Given the enormous size of the state space, one can see that such a clear cut appraisal is not informative, and it seems much more reasonable to consider the average squared distance between the true tree and the estimated one as an analog of the mean square error $\mathrm{MSE} = E(\theta - \hat{\theta})^2$ that statisticians routinely use for real-valued parameters. In our setting this would mean using some distance between trees $d(\theta, \hat{\theta})$ and knowing how to define a probability measure on tree space so as to be able to compute either the expected value, $Ed(\theta, \hat{\theta})^2$ or from a more Bayesian viewpoint $E_p d(\theta, \hat{\theta})^2$ with regards to the posterior p.

[7] The so-called *Distance-based methods*.

[8] Here I am thinking of validation of the molecular clock and independence assumptions.

Note that the analysis of even one data set may give a probability distribution on trees rather than a unique tree. A different way of dealing with disparity from a perfect tree may be to propose confidence regions for the trees or a weighted consensus of the trees compatible with the data.

More general graphs such as the *Splitstree* network championed by Dress et al. (1996); Huson (2000) or networks as developed in Von Haeseler and Churchill (1993) and Strimmer and Moulton (2000) are also interesting alternatives when the data are far from treelike.

Both of these extensions, the probability distribution on trees perspective and the network graph offer the advantage of a painless transition from one gene to k genes.

If we start by considering the combinatorial trees where only the branching order is of importance, the size of the space is the first hurdle. If the data set is forced to fit into only one tree, it will look as if we have an exponential amount of information in the data. Statements a posteriori, such as: we had one chance in the 10^{12} to obtain this tree among all trees are definitely misleading and are of the same caliber as the blade of grass argument.[9]

2.1. Between parametric and nonparametric estimation

Treebuilders have seemed polarized, at least in the past, by the likelihood parsimony debate over which criteria/method is the "right" one. It may seem surprising that a statistician should also see *Maximum Parsimony* as a feasible estimation procedure, however, it can be seen as a nonparametric statistical estimate, in spite of the many assertions to the contrary Felsenstein (1983), Steel and Penny (2000).

From a statistical viewpoint, this is the usual nonparametric/parametric dilemma. When should one be using a restrictive parametric model and estimate the parameters by maximum likelihood and when should a more flexible approach be taken?

Just to give an example of the differences between the two methods, think of a regression context, where a response variable needs to be predicted from some explanatory variable x, and the scatter plot is given in Fig. 3.

If there is some prior information allowing for a parametric model, say: $y = ax^3 + bx + c + \varepsilon$, then a quadratic is fit to the data, or maybe a cubic or other

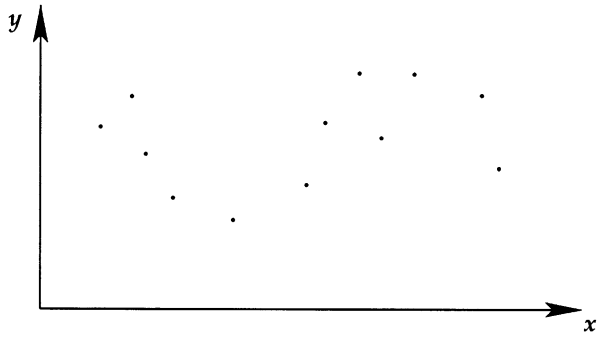

Fig. 3. Scatter plot.

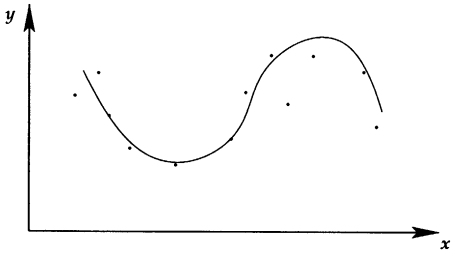

Fig. 4. Cubic polynomial.

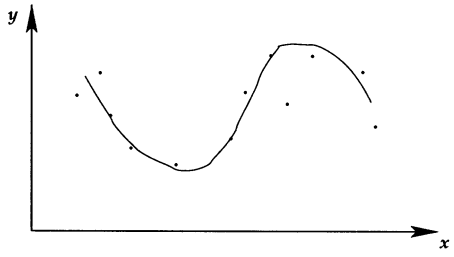

Fig. 5. With smoothness.

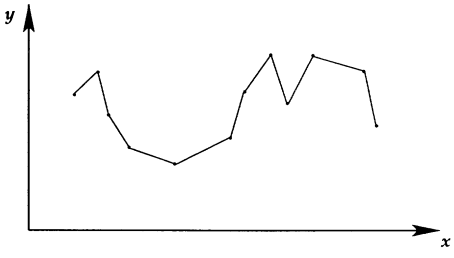

Fig. 6. No smoothness.

polynomial. On the other hand, the data may be fit by a nonparametric smoother defined by minimizing PRESS[10] + non-smoothness penalty. The smoothness penalty is important, without it the final fit interpolates the data perfectly. Of course the problem is that it will not extend well to predict a response from a new x (Figs. 4–6).

[9] Suppose I am sitting on the grass and pick out one blade. I can always ask afterward, what were the chances that I chose that blade? The chances will be very small given the enormous numbers of blades, it will not mean anything about that blade, I had to pick one. It is as if I draw a target around the dart after I threw it on the wall.

[10] Predicted error sum of squares.

Of course as the degree of the polynomial increases, and thus the parametric model becomes more flexible because it has more parameters it will become closer and closer to the nonparametric smooth regression curve.

Many are the discussions relevant to this quandary. Some contributions from theory are surprising: Hodges and Lehmann (1961) showed that nonparametric tests were almost as powerful as the best parametric tests, even if the parametric assumptions were valid. If the parametric assumptions were violated, the nonparametric tests were still valid whereas the parametric tests can be wildly off (Lehmann, 1975). But all statisticians will agree, if the model has already been validated by previous experiments, parametric modeling is more powerful (stricto sensu statisticae). However, if the aim of the study is in any way exploratory, trying to uncover unexpected mutation patterns, irregularities in the molecular clock, interspecies differences in DNA frequencies, a nonparametric approach[11] is in order, this is especially the case when in fact the study is aimed at validating some part of the model, because using the model would lead to a circular argument.

Having spoken a little about the original data and the estimation procedures, I would like to pose some of the problems associated with the very particular structure of the parameter space. The question "What are the chances that this tree is the true evolutionary tree?" comes up soon after an estimated tree has been provided. Unfortunately, this question seems to imply a likelihood-based framework for many biologists, especially traditional cladists (Kluge and Farris, 1969; Farris, 1983). Maximum likelihood estimation with a finite dimensional parametric model is a choice at the estimation level that gives a possible answer to this question. However, it seems to escape many researchers in the field that this question can be posed in any framework, encompassing Bayesian posterior probability computations as well as validation of a nonparametric model. We will see that to answer this question we have to define both a notion of distance and a probability distribution in tree space, but the maximum likelihood framework is not a pre-requisite.

3. What to do when data are trees?

When summarizing conflicting edges into one tree, standard practice discards in some sense the characters that are not compatible with the tree chosen, whereas these characters could in turn be used to construct complementary trees or even networks (as in Bandelt et al., 1995 or Dress et al., 1996) and the data thus would point to several trees at the same time, so that building a unique tree from one data set of conflicting sequences can be seen as a consensus problem. It would seem more satisfactory to decompose the data into successive complementary trees, equivalent to providing a distribution or a mixture of trees as the output instead of a unique tree. In the same way regression provides a decomposition into an estimate and a residual.

3.1. Probability distributions on trees

Building a probability distribution on trees is a complex procedure. We know that choosing optimal trees cannot in general be decomposed into simpler problems. This is the essence of what constitutes computationally intractable problems such as are the maximum likelihood tree and parsimony tree. Most biologists agree that the simplest possible probability distribution on tree space, the uniform distribution is not relevant, a slightly more realistic one is the Yule process, see Aldous (2001) for history and some consequences relevant to balance and depth of phylogenetic trees.

Here we are going to give an analog of the nonstandard splits-and-trees data and look at ways statisticians have dealt with these difficulties. The following example involves data that do not belong to the real line $\mathbb{R}$. Commonly called rank data, these come from different observers providing rankings for their preference in chocolates, wine, political candidates, PTA leaders, etc. In the simplest case, we have permutations of the same n objects, say if $n=4$, $x_1=(c_2,c_4,c_1,c_3)$, $x_2=(c_1,c_4,c_2,c_3)$, $\ldots, x_n=(c_4,c_2,c_1,c_3)$, a little more difficult but very related, is the case of partial rankings $y_1=(c_2,c_4,c_3)$, $y_2=(c_1,c_2,c_3)$, $\ldots, y_n=(c_4,c_2,c_1)$ that we would also like to summarize. This data actually occurs in genetics now that gene order data is becoming available (for a review see Pezner, 2000, Chapter 10, an application in Blanchette et al. (1999) or for their use in phylogeny, see Sankoff and Blanchette, 1999).

The ranking data example is a useful analogy because there are several books (Critchlow, 1985; Diaconis, 1988; Marden, 1995; Fligner and Verducci, 1992) on the subject of the statistical analysis of permutation data, with many deep mathematical developments. The simplest example is Mallow's model (Mallows, 1957), defined for rankings obtained from different judges. It assumes that there is a central ranking in the sense of a modal ranking. The mode is the value preferred by the most judges, and suppose that most judges are expected to have ranking around this central one. The relevant probability measures here is centered at the mode and

[11] Or a parametric model with a very large number of parameters.

the probability of a ranking would depend on some distance to this central value. The probability of a given ranking is of the form

$$P(w) = Me^{-\lambda d(w,w_0)},$$

where M is a normalizing constant. This will of course depend on the choice of the center of the distribution w_0, the distance, which can be Spearman's distance, Kendall's or another more refined distance. The parameter λ fixes how tightly the trees are around w_0 and will thus depend on how refined d is. Of course, the simplest possible distribution is the uniform provided by $\lambda = 0$. (Marden, 1995, Chapter 6) is completely dedicated to this type of model—often a first approximation because it supposes so much symmetry in the probability distribution.

This idea can be used to define a probability distribution that carries over to the space of all trees that I will abbreviate $\mathscr{T}$. First, one needs a notion of central tree τ_0 for a configuration of trees. It could be the first tree in the Hadamard decomposition (Hendy and Penny, 1993) or the majority rule consensus tree of many bootstraps of the data (Berry and Gascuel, 1996), or it could be some type of centroid or center of gravity of the trees, from a statistical perspective it should be an estimator of the center of the distribution of trees in $\mathscr{T}$.

Second, one must choose a distance between trees (there are many such distances available see Critchlow et al. (1996) or Dasgupta et al. for reviews), we will also see in Section 5 that a continuous geometry will provide a refined measure of distance that can take into account both branch lengths and branching patterns. The probability distribution proposed will be of the form: for τ in $\mathscr{T}_n$, the space of trees with n leaves, let

$$P(\tau) = Le^{-\lambda d(\tau,\tau_0)}.$$

For fixed distance and τ_0, this distribution is said to belong to an exponential family with λ as a parameter. Exponential families play a special role in statistics because of the notion of sufficiency, Section 3.2. When $\lambda = 0$, we have the uniform distribution, for large λ, the distribution becomes highly concentrated on τ_0. Of course τ_0 and even d can also be treated as parameters.

Many other models and data analytic procedures have been proposed to analyze ranking data; see the references above, these happily coexist without controversy in the statistical community. Many can be carried over to trees.

Before we leave the rank data, I would also like to point out another similarity with the tree data summary problem. The ranking world has been plagued by Arrow's paradox. Arrow (1963) proves that a "sensible" consensus ranking cannot exist under quite a reasonable set of required conditions. His work has given rise to a whole field in economics and political science called social choice theory. One way economic theorists have proposed to resolve this difficulty has been to incorporate a continuum in the space, thus filling in the discrete rankings and then using geometry to solve the paradox (see Chichilnisky, 1980; Baryshnikov, 1997). The same paradox pertains in phylogenetic consensus problems as proved in Steel et al. (2000). We will see that the geometrical representation and metric trees developed below provide just such a filling-in scheme, with the same consequences: we will have clear geometrical notions of consensii and supertree at our fingertips.

Now we will come back to actually trying to summarize a distribution on trees, a Bayesian posterior distribution, or a distribution that could be used to build a frequentist confidence region. In order to do this rigorously, a statistician needs the following baggage, a concept that specifies how to replace a whole distribution by a few summaries, the sufficient statistics.

3.2. Sufficiency

Roughly, sufficiency is a statistical property of a statistic used to estimate a parameter, that ensures that if the problem presents certain symmetries, one can ignore part of the original information. Surveys about sufficiency can be found in Diaconis (1992) or Lehmann and Casella (1998). We present the simplest possible sufficiency model, which most people will recognize. We consider the space of binary n-tuples with probability of having a 1 as the parameter p of the probability distribution on $\mathscr{X} = \{0,1\}^n$. Do we have to keep all the complexity of the observations as $x = (x_1, x_2, \ldots, x_n) = (0,1,0,0,0,1,\ldots,0)$? or can we reduce x to $t = x_1 + x_2 + \cdots x_n = 0 + 1 + 0 + 0 + 0 + 1 \cdots + 0$. In this case we observe one such x and want to estimate θ where $p_\theta(x) = \theta^t(1-\theta)^{n-t}$. Here, the probability at x only depends on t and we can *compress* the information in x down to t.

Formally, $S : \mathscr{X} \to \mathscr{Y}$ is called sufficient for a family of probability distributions defined on $\mathscr{X}$ if the conditional probability

$$\text{If } T(x) \neq t,\ P(x|T(x) = t) = 0 \text{ and if } T(x) = t,$$
$$P(x|T(x) = t) \text{ is the same for all } P \in \mathscr{P}.$$

This provides significant decrease in the complexity of the problem. It can also be seen as an invariance property (see Diaconis, 1992). If the problem is parametric,[12] then the definition of sufficiency just boils

[12] In the sense that the family $\mathbb{P}$ is determined by a finite dimensional parameter $\theta, \mathbb{P} = \{P_\theta, \theta \in \Omega\}$.

down to the fact that

$$P_\theta(x|S(x)=t)$$

is independent of θ. Thus, the parameter θ of the model only makes contact with the data through the function S. This can be shown to be equivalent to the existence of two functions $C(t)$ and $b_\theta(t)$ such that

$$P_\theta(x)=c(x)b_\theta(t)=c(x)b_\theta(S(x)).$$

This is called Fisher's factorization theorem.

In our example of the exponential family model defined in analogy to Mallow's model, the sufficient statistic is

$$\sum_{i=1}^{k} d(\tau_i,\tau_0)=S_k$$

for a collection of k trees and a central tree τ_0. So the computations of an estimate of θ will only depend on the distances between trees. This reduces the data to one number S_k. Of course, without the assumption of the symmetrical distribution sufficient statistics can be more complex. For example, to estimate the mean of a distribution on $[0,1]$ with no parametric assumptions, the sufficient statistic is the full set of ordered observations.

It is to be noted that if we are given a distribution on trees such as the bootstrap distribution obtained either by nonparametric resampling or parametric Monte Carlo, it is often summarized by just the frequencies of edges. These numbers collected in a vector, do not constitute a sufficient statistic for the complete bootstrap distribution. This has been implicitly understood by several authors who work on trees. For instance, Penny and Hendy have proposed a *nearest neighbor* (NN) bootstrap which counts how many times an edge or a *neighboring* split occurs (for an example of its use see Cooper and Penny, 1997). We will see later that the geometrical enhancement of the mathematical picture of tree space make such a NN bootstrap natural.

Different models for the distribution will have different sufficient statistics. Sometimes the exponential model going through a certain number of statistics chosen to be the essential summaries is built. For instance, one could take the average graph distance D_n of a tip to the root and the height of the tree H_n and use those as sufficient statistics, the model would then be:

$$P(\tau)=K\exp(\lambda_1 D_n+\lambda_2 H_n)$$

(See Diaconis, 1989 for many examples in the rankings data problem.) For a k dimensional parameter θ with sufficient statistics $S_1,\ldots,S_k$)

$$P_\theta(\tau)=\exp\Big(\sum\theta_j S_j\Big)$$

Just considering the clade frequencies as a first order approximation can be justified by considering another decomposition of a set of trees $\mathscr{X}$ by a Fourier-type analysis in tree space.[13]

Diaconis and Holmes (1998, 2001) show that the space of all combinatorial trees on n leaves T_n is equivalent to the quotient of the symmetric group on $2(n-1)$ by the subgroup $B_{2(n-1)}$ that leaves the pairs

$$\{(1,2)(3,4)(5,6)\ldots(2n-3,2n-2)\}\quad\text{invariant.}$$

This is called the matching representation of trees where the tree is replaced by all its sibling pairs, including the inner nodes.

This representation provides a useful way of enumerating all trees without using a branch-and-bound algorithm. The passage from one tree to the next is done by choosing two sibling pairs and switching one of them. Such a path through all trees is called a Hamiltonian path or a Gray code. In computational applications, it has the advantage that large parts of the tree remain unchanged. However, the distance induced on the number of such moves, although simple to compute is not biologically natural.

The decomposition of functions on tree space is given in Diaconis and Holmes (1998). It is a direct sum decomposition of all functions on tree space into subspaces $S^{2\lambda}$:

$$\mathscr{L}(\mathscr{T}_n)=\bigoplus_{\lambda\vdash(n-1)} S^{2\lambda}$$

Here, the sum is over all partition λ of $n-1$, $2\lambda=(2\lambda_1,2\lambda_2,\ldots,2\lambda_k)$ and $\mathscr{S}^{2\lambda}$ is the associated irreducible representation of the symmetric group $\mathfrak{S}_{2(n-1)}$. The first few terms in the decomposition can be interpreted as follows:

- For $\lambda=(n-1), S^{2\lambda}$ is the space of constant functions. The projection onto $S^{2\lambda}$ counts the number of trees in the data set.
- For $\lambda=(n-2,1)$, the projection onto $S^{2\lambda}$ counts the number of times each particular sibling occurs.
- For $\lambda=(n-3,1,1)$, the projection onto $S^{2\lambda}$ counts the number of times the sibling pair (i,j) occurs at the same time as the pairs (k,l).
- For $\lambda=(n-3,2)$, the projection onto $S^{2\lambda}$ counts the number of times the sibling pair (i,j,k,l) occurs as a clade.

It is in this sense that the sibling pair frequencies are a first order approximation to the complete distribution on trees. It would be useful to be able to say what

[13] This Fourier analysis is not the same as that proposed by Hendy and Penny (1993) and Hendy et al. (1994).

proportion of the information contained in the data set can be reconstructed just by the sibling pair counts. This Fourier-type decomposition follows closely the analysis of ranking data provided by Diaconis (1989). It is most useful for large sets of trees on a small number of species, and provides the basis for the theoretical analysis of times to convergence for the simple random walk on tree space induced by doing transpositions on the matchings (Diaconis and Holmes, 2001).

4. Confidence statements, Bayesian distributions and the need for distances

Trees are high-dimensional parameters, even if they do not lie naturally in an Euclidean space. The best frequentist confidence statements that can be made about them are ones that rely on the notion of a confidence region $\mathscr{R}_\alpha$ defined by statements of the form:

$$\mathbb{P}(\tau \in \mathscr{R}_\alpha) = 1 - \alpha.$$

Much interest in the biological literature has concentrated on "stability" of the estimated tree and it often appears that the question biologists are trying to answer with the bootstrap is one of continuity[14] rather than an inferential one.

In order to discuss the continuity-sensitivity issues, various Bayesian prior and posterior distributions, and the probability measures on tree space that may be relevant to biologists, we have to decide on a satisfactory metric on trees.

If enough prior information is available, it makes sense to use the Bayesian paradigm. This was already suggested by Edwards (1970), then by Wheeler (1991), and more recently implemented by Li et al. (2000), Mau et al. (1999) and Yang and Rannala (1997). They have provided Bayesian algorithms, all set in the parametric framework with posterior probabilities computed using Monte Carlo Markov chains on tree space. Huelsenbeck and Ronquist (2001) provide some software for actually using these ideas.

Finding a unique resulting tree has its own share of difficulties, the mode of the posterior distribution has been suggested Steel and Penny (2000) mainly for its simplicity,[15] however, a more satisfactory result would be a posterior confidence region. This would be a geometric representation of a region containing 95% of the posterior probability. To avoid uniqueness problems[16] usually the smallest region containing 95% is used.

A prior distribution on trees could incorporate the information that an outlier was included to root the tree, thus, we would put a very high probability on the space of trees with an outlier as an outgroup.

A current challenge would be a semiparametric Bayesian method. Such methodology could use Mallow's model extension as in Section 2, with a prior distribution for τ_0 and λ, or could be given by a more general exponential distribution with statistics such as those suggested by Aldous (2001) that could incorporate statistical data[17] collected from large databases of trees such as Sanderson et al. (1994).

More flexible priors can be generated using geometric priors defined using a sensible notion of distance between trees. Let us look at the most useful representation for trees; the geometrical one.

5. Geometrical representation

We have seen that the definition of a distance and a central tree enables a symmetric exponential probability distribution to be defined. In this section, we are going to fill in the set of combinatorial trees, a discrete set, so that various notions of averages can be more refined than in the simple consensus methods. This work is presented in its mathematical technicality Billera et al. (2001) and some of the applications to biological problems are given in Billera et al. (2002).

We start by explaining intuitively what we would like our geometrical representation to provide. For each different tree, a separate region would be contained by a boundary (Fig. 7).

The boundary represents an area of uncertainty of the exact branching order. Two neighboring regions represent neighboring trees. The notion of "neighboring" we have chosen is NN interchange (nni), the rotation distance in $\mathscr{T}_n$. This seems to be the most widely accepted neighborhood relation, although other distances can also be used to define a metric tree space in the same way. The natural way of varying closeness to the boundary or unresolved tree is to make the edge lengths e decrease linearly in the direction of the boundary. This provides a geometrical justification for edge lengths which are not biologically meaningful such as the number of mutations or a "time span" as is the case for trees built under the Markovian models with

[14] In the mathematical sense, continuity of an estimator means that small changes in the data never result in "large jumps" of the estimator.

[15] This is a majority rule type consensus.

[16] Similar to the shortest confidence interval, symmetrical confidence interval, and so on.

[17] Such as measures of tree height, tree balance, etc.

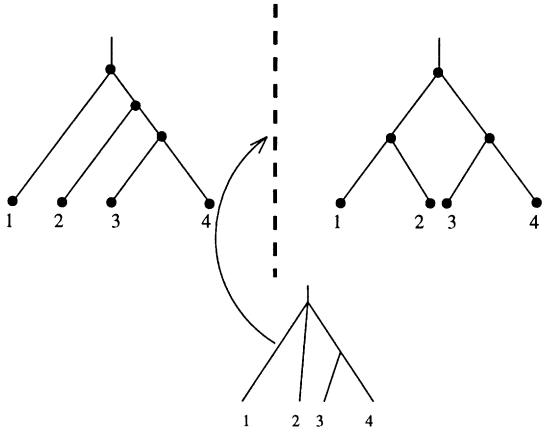

Fig. 7. Two tree regions separated by a boundary—a degenerate tree.

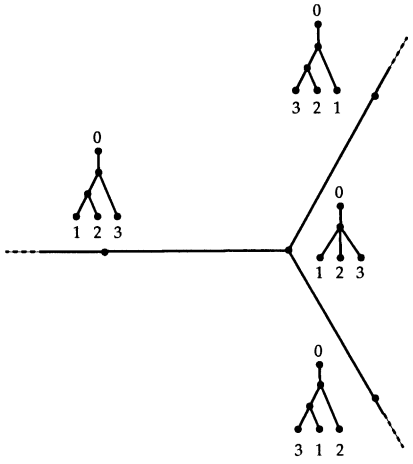

Fig. 8. Trees with three leaves meeting at the star tree point.

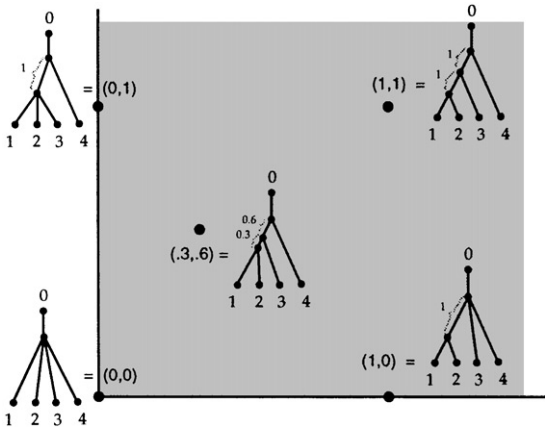

Fig. 9. Trees with four leaves in the same quadrant.

constant mutation rates such as Jukes-Cantor or Kimura models. We will work with rooted trees.

There are three binary semilabeled trees on three terminal nodes (leaves). We arrange them along three

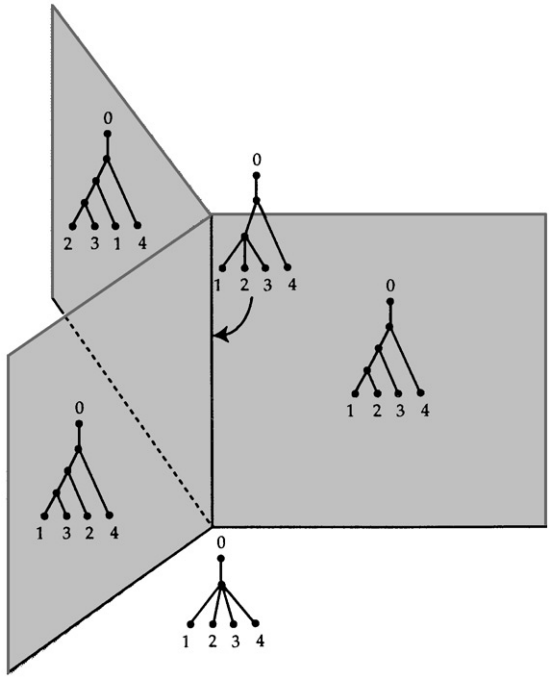

Fig. 10. Three neighboring quadrants.

half lines meeting at the origin which represents the star tree (Fig. 8).

In fact, to make our geometrical space slightly simpler, we restrict ourselves to trees with finite branch lengths, here taken to be in $[0, 1]$. By standardizing the combinatorial trees to all having edge lengths one, we can build the space of trees with n leaves edge lengths as the product of a cube complex $\mathscr{T}_n$ that represents the trees without the pendant edges and the pendant edges $[0, 1]^n$.

For rooted binary trees with four leaves, $\mathscr{T}_4$ is a set of squares pasted together by two edges each. Each square corresponds to a different branching order and the position within the square is determined by the coordinates, each representing one of the two inner edge lengths (Figs. 9 and 10).

Note that the boundary is shared by two other trees. The pendant branch lengths do not appear in this geometrical representation. To obtain a complete coordinate system of binary semilabeled trees, one would have to take the product of $\mathscr{T}_4$ with $[0, 1]^4$.

All quadrants have to have the star tree as one of their corners. So the star tree will have 15 neighboring quadrants. This generalizes to $\mathscr{T}_n$ and explains that at the star tree, the origin of our space, there are exponentially many cubes attached. On the other hand, a tree with only one edge is represented as a segment boundary to three quadrants, thus its neighborhood will contain three "flaps."

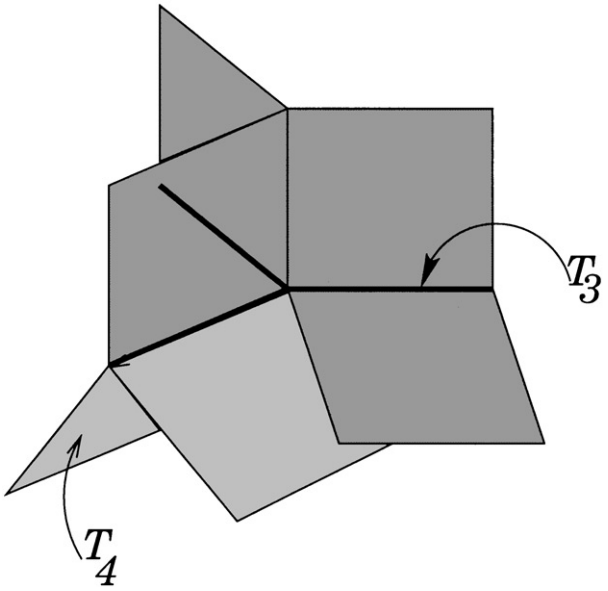

Fig. 11. Embedding of $\mathscr{T}_3$ in $\mathscr{T}_4$.

In fact, if we have a four-leafed tree, but are sure what the outgroup is, the relevant space is the space of rooted trees on three leaves. This embedding is shown geometrically in Fig. 11. Zharkikh and Li (1995) did a simulation study to find how many trees neighbor a given tree. This has consequences for the quality of the bootstrap estimate as is also pointed out also in Efron et al. (1996). We can see that for a tree on four leaves/tips, there can be either no neighbors except trees with the same branching pattern, two neighboring combinatorial trees as in Fig. 13, or 14 neighboring trees (if all the edges are small and we are close to the star tree) (Fig. 12). Of course, for a tree with only two inner edges, this is the only possible way of having these two edges small. This same notion of neighborhood containing 15 different branching orders applies to all trees on as many leaves as necessary, but who have two contiguous "small edges" and all the other inner edges significantly bigger than 0. This picture of tree space frees us from having to use simulations to find out how many different trees are in a neighborhood of a given radius r around a given tree. All we have to do is check how many contiguous edges in the tree are smaller than r, say there is only set of size n_r, then the neighborhood will contain

$$(2n_r - 3)!! = (2n_r - 3) \times (2n_r - 5) \times \cdots 3 \times 1$$

different types of trees. Thus, a point very close to the star tree at the origin will have an exponential number of neighbors. This explosion of the volume of a neighborhood at the origin provides for interesting math problems.

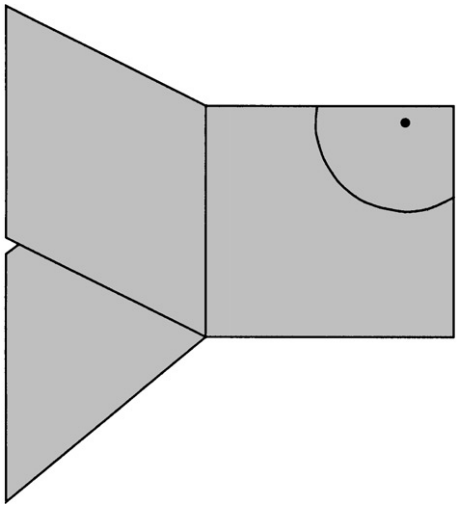

Fig. 12. One tree in the neighborhood.

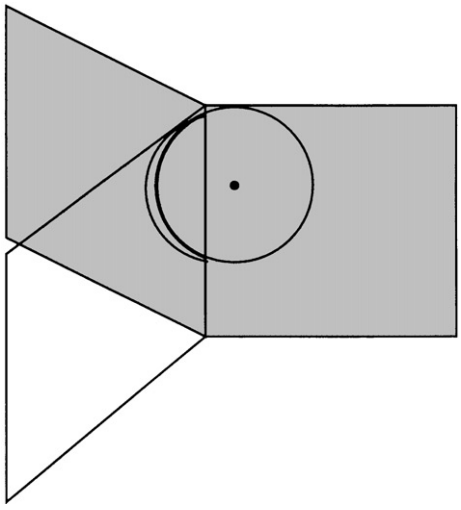

Fig. 13. A tree with two neighbors.

The last element necessary to make a rigorous picture of tree space is the probability measure. We can define such a measure through its probability density or likelihood. A picture of the likelihood contours in the four leaf case is available through the use of the polynomials developed by Chor and Snir (2002), however, the problem of representing more general probability measures on this space is still open.

6. The philosophical question of coherence

A more difficult issue is that of overall coherence in a study, not in the formal statistical sense of de Finnetti, but a more pragmatic one of mixing oranges and apples. Here is an example.

From a statistical viewpoint, the use of the multinomial bootstrap resampling procedure in the maximum likelihood estimation setting is not coherent. Given a parametric model, one should use that model both at the estimation and the validation stages of the analyses. Thus, the parametric bootstrap as implemented in `Seq-Gen` and `Treevolve` by Rambaut and Grassly (1997) is coherent when doing a maximum likelihood estimation. A justification for using the multinomial bootstrap may be for more general model testing, but this leads to confusing conclusions because the alternative is not explicitly defined.

Can we allow ourselves to switch paradigms as we deem fit? Starting with a parametric model for evolution, such as Jukes-Cantor, does it make sense, after a tree has been estimated, to switch to a different paradigm at the validation stage and use a nonparametric bootstrap to compute confidence levels on the tree? This is an open problem in statistics as well as in phylogenetics. Some statisticians often switch paradigms in the middle of their studies, from parametric to nonparametric, usually what can be actually proved is that if the parametric model is correct there is a loss in power (sensu statisticae strictu) when switching to a nonparametric procedure. Other statisticians abhor this paradigm switching; see the discussion to Efron (1986)'s *Why isn't everyone a Bayesian?*

As to the mixture Bayesian-frequentist, empirical Bayes (Robbins, 1985, 1980, 1983) is a typical example of loose boundaries that exist when choosing different paradigms at different stages of an analysis. The current *resolution* within statistics of the problems posed by Empirical Bayes goes as follows; frequentists treat it as just another method to be evaluated by its long-term frequency. Bayesians react with amusement or horror.[18]

6.1. Validation procedures and their implementation

There has been much interest in the stability of an analysis, and much of the use of the bootstrap in phylogeny is aimed at answering the question: *If my data were slightly perturbed, how much would the estimated tree have changed*? This poses the question of the continuity of the estimator used rather than the broader question of inference to a larger possible population of characters.

Little progress has been made on the robustness of the results to the underlying assumptions implicit in various methods, however, there are many simulation tools available that can be combined to provide useful approximate statements.

[18] Dennis Lindley wrote "there is no less Bayesian thing to do!"

6.2. Bootstrapping phylogenies

One of the simulation tools most commonly used has been the bootstrap, introduced to the field by Felsenstein (1985).

In the statistical literature, the theorems that justify the use of the bootstrap usually state that the distribution of the distance between the true parameter and the estimate can be well approximated by the distribution between the estimate and the bootstrap resampled estimate, something that can be summarized as

$$Distribution(d(\mathscr{T}, \hat{\mathscr{T}})) \approx Distribution(d(\hat{\mathscr{T}}, \hat{\mathscr{T}}^*)).$$

However, most of the theoretical work involves an assumption of independent, identically distributed variables and some strong assumptions on the properties of the distance. No actual theory in the phylogenetic context exists at present, although referring to theoretical arguments in other cases does provide useful insight into the most sensible simulations to undertake.

As pointed out in Efron et al. (1996), the bootstrap performs better when the number of neighboring trees is not too large, the calculations in the previous section give the indication that this will be linked both to the number of continguous small branches and the probability measure on the tree space.

Care should be taken to simulate problems of a comparable size as the real data, as conclusions from a study on four-leaved trees may not be generalized to trees with hundreds of leaves pointed out by Hillis' study of the effects of long branch attraction have shown (Hillis, 1996). We conclude with three key points.

- The bootstrap procedure as used by statisticians supposes that the observations are independent. That means that the columns should be independent of each other. Hidden Markov models secondary/tertiary structure models seem more believable. An alternative to the independent bootstrap should thus be preferred, PHYLIP Felsenstein (1993) and Seqgen Rambaut and Grassly (1997) are tools that cater to these extended dependent models.
- Covarion models lean towards models where the observations (columns of the sequence matrix) are not identically distributed either but depend on a covariable. In their studies, a binary one, that could be either on or off. This, is also modeled by "hotspots" along the sequences, see Tang and Lewontin (1999).
- For a practical implementation of Bayesian procedures, providing posterior probabilities using Monte Carlo Markov chains, Huelsenbeck and Ronquist (2001) is a useful tool, that needs to be extended to more general models.

7. Summary

Staying within a coherent framework runs counter some biologists' method of multitasking their assumptions in parallel. Many like to keep all eventualities concurrently in mind. This does not allow for a linear analysis, and unless more biologists decide to carry forth their analyses within a Bayesian paradigm, giving each contingent assumption a prior probability and then looking at the posterior probabilities, this leads to very confusing conclusions.

Perhaps the biggest change in statistics over the last 20 years has been the decrease in the use of p-values. Tukey and his co-workers in Exploratory Data Analysis (see Tukey, 1975) have shown us the importance of keeping as much of the data in mind as possible. A picture is worth a thousand words, and geometry has much to offer in this complex multidimensional analysis of DNA sequences through trees, graphs and their assorted confidence regions. As remarked by Felsenstein (2001), there are far more interesting problems than statistical consistency to address at this point.

Some open problems for my mathematically inclined colleagues would include

- Giving an indication of the order of magnitude of the number of simulations necessary in the MCMC-type methods to ensure that the stationary distribution is attained.
- Study of robustness of the conclusions of a phylogenetic analysis to the assumptions would help calibrate the level of detail within which one should stay, given that uncertainty is involved at every level of such a complex analysis.
- Can we extend some of the work on trees to networks? This would be useful, both for analysing regulatory networks, but also would enable one to test whether data are actually treelike.[19]

Acknowledgments

I would like to thank all the participants of the Doom 2001 New Zealand phylogenetics meeting for their stimulating discussions and Sam Karlin for his patience. Hua Tang, Persi Diaconis, Joe Felsenstein and the anonymous referees carefully read previous versions of this piece and provided invaluable references and corrections.

[19] Theoretical exploration of treelikeness has been undertaken by Dress et al. (2001).

References

Aldous, D., 2001. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. Statist. Sci. 16, 23–34.

Amit, Y., Geman, D., 1997. Quantization and recognition with randomized trees. Neural Comput. 9, 1545–1588.

Arrow, K., 1963. Social Choice and Individual Values. Wiley, NY.

Bandelt, H.J., Foster, P., Sykes, B.C., Richards, M.B., 1995. Mitochondrial portraits of human populations using median networks. Genetics 141, 743–753.

Baryshnikov, Y.M., 1997. Topological and discrete social choice: in a search of a theory. Soc. Choice Welf. 14, 199–209.

Berry, V., Gascuel, O., 1996. Interpretation of bootstrap trees: threshold of clade selection and induced gain. Mol. Biol. Evol. 13, 999–1011.

Billera, L., Holmes, S., Vogtmann, K., 2001. The geometry of tree space. Adv. Appl. Math. 27, 771–801.

Billera, L., Holmes, S., Vogtmann, K., 2002. A geometrical perspective on the phylogenetic tree problem. Technical Report xx, Statistics, Sequoia Hall, Stanford, CA 94305.

Blanchette, M., Kunisawa, T., Sankoff, D., 1999. Gene order breakpoint evidence in animal mitochondrial phylogeny. J. Mol. Evol. 49, 193–203.

Breiman, L., 1996. Bagging predictors. Mach. Learning 24, 123–140.

Brooks, D.R., 1981. Hennig's parasitological method: a proposed solution. Syst. Zool. 30, 229–249.

Chichilnisky, G., 1980. Social choice and the topology of spaces of preferences. Adv. Math. 37, 165–176.

Chor, B., Snir, S., 2002. Four taxon ML fork under molecular clock: analytic solutions. Whitianga-New Zealand Phylogenetics Meeting, Whitianga, New Zealand.

Cooper, A., Penny, D., 1997. Mass survival of birds across the cretaceous-tertiary boundary: molecular evidence. Science 275, 1109–1113.

Critchlow, D.E., 1985. Metric Methods for Analyzing Partially Ranked Data. Springer-Verlag, Berlin.

Critchlow, D.E., Pearl, D.K., Qian, C., 1996. The triples distance for rooted bifurcating phylogenetic trees. Syst. Biol. 45, 323–334.

Dasgupta, B., He, X., Jiang, T., Li, M., Tromp, J., Wang, L., Zhang, L., Computing distances between evolutionary trees. Proceedings of the Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, 5–7 January 1997, New Orleans, LA.

Diaconis, P., 1988. Group Representations in Probability and Statistics. Institute of Mathematical Statistics.

Diaconis, P., 1989. A generalization of spectral analysis with application to ranked data. Ann. Statist. 17, 949–979.

Diaconis, P., 1992. Sufficiency as statistical symmetry. In: American Mathematical Society Centennial Publications, Vol. II (Providence, RI, 1988), American Mathematical Society, F. Browder (Ed.), Providence. RI, pp. 15–26.

Diaconis, P., Holmes, S., 1998. Matchings and phylogenetic trees. Proc. Natl. Acad. Sci. USA 95, 14600–14602 (electronic).

Diaconis, P., Holmes, S., 2001. Random walks on trees and matchings. Technical Report, Statistics Department, Stanford, CA 94305.

Dress, A., Huson, D., Moulton, V., 1996. Analysing and visualizing sequence and distance data using splitstree. Appl. Math. 71, 95–109.

Dress, A., Holland, B., Huber, K., Koolen, J., Moulton, V., Weyer-Menkhoff, J., 2001. Delta additive and Delta ultra-additive maps, Gromov's trees, and the Farris transform. Discrete Appl. Math., submitted for publication.

Durbin, R., Eddy, S., Krogh, A., Mitchison, G., 1998. Biological Sequence Analysis. Cambridge University Press, Cambridge, UK.

Edwards, A., 1970. Estimation of the branch points of a branching diffusion process (with discussion). J. R. Statist. Soc. B 32, 155–174.

Efron, B., 1986. Why isn't everyone a Bayesian? (c/r: P5-11; p 330–331). Am. Statist. 40, 1–5.

Efron, B., Halloran, E., Holmes, S., 1996. Bootstrap confidence levels for phylogenetic trees. Proc. Natl. Acad. Sci. USA 93, 13429–13434.

Farris, J., 1973. On comparing the shapes of taxonomic trees. Syst. Zoo. 2, 50–54.

Farris, J.S., 1983. The logical basis of phylogenetic analysis. In: Platnick, N., Funk, V. (Eds.), Advances in Cladistics. Columbia University Press, New York, pp. 7–36.

Felsenstein, J., 1983. Statistical inference of phylogenies (with discussion). J. R. Statist. Soc. A 146, 246–272.

Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783–791.

Felsenstein, J., 1993. PHYLIP, (Phylogeny Inference Package) version 3.5c. Department of Genetics, University of Washington, Seattle, version 3.5c. edition.

Felsenstein, J., 2001. The troubled growth of statistical phylogenetics. Syst. Biol. 50, 465–467.

Fitch, W.M., Markowitz, E., 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genetics 4, 579–593.

Fligner, M.A., Verducci, J.S.E., 1992. Probability Models and Statistical Analyses for Ranking Data. Springer-Verlag, Berlin.

Hammersley, J., 1950. On estimating restricted parameters (with discussion). J. R. Statist. Soc. Ser. B 12, 192–240.

Hein, J., 1989. A new method that simultaneously aligns and reconstructs ancestral sequences for any number of homologous sequences, when the phylogeny is given. Mol. Biol. Evol. 6, 649–668.

Hendy, M.D., 1991. A combinatorial description of the closest tree algorithm for finding evolutionary trees. Discrete Math. 96, 51–58.

Hendy, M.D., Penny, D., 1993. Spectral analysis of phylogenetic data. J. Classification 10, 5–23.

Hendy, M.D., Penny, D., Steel, M.A., 1994. A discrete fourier analysis for evolutionary trees. Proc. Natl. Acad. Sci. 91, 3339–3343.

Hillis, D.M., 1996. Inferring complex phylogenies. Nature 383, 130.

Hodges, J.L.J., Lehmann, E.L., 1961. Comparison of the normal scores and Wilcoxon tests. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, Vol. 1, pp. 307–317.

Holmes, S., 1999. Phylogenies: an overview. In: Halloran, E., Geisser, S. (Eds.), Statistics and Genetics, IMA, Vol. 81. Springer-Verlag, NY.

Huelsenbeck, J., Ronquist, F., 2001. Mr Bayes. Bayesian inference of phylogenetic trees. Bioinformatics 17, 754–755.

Huson, D., 2000. Splitstree. Web address. URL:http://www.mathematik.unibielefeld.de/~huson/phylogenetics/splitstree.html.

Kluge, A.C., Farris, J.S., 1969. Quantitative phylogenetics and the evolution of Anurans. Syst. Zool. 18, 1–32.

Lake, J.A., 1994. Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc. Natl. Acad. Sci. 91, 1455–1459.

Lehmann, E.L., 1975. Nonparametrics: Statistical Methods Based on Ranks. Holden-Day, San Francisco, CA.

Lehmann, E.L., Casella, G., 1998. Theory of Point Estimation, 2nd Edition. Springer, Berlin.

Li, S., Pearl, D.K., Doss, H., 2000. Phylogenetic tree construction using mcmc. J. Am. Statist. Assoc. 95, 493–503.

Li, W.H., 1997. Molecular Evolution. Sinauer, Boston.

Lo, S.-H., 1992. From the species problem to a general coverage problem via a new interpretation. Ann. Statist. 20, 1094–1109.

Lockhart, P.J., Steel, M.A., Hendy, M., Penny, D., 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. 11, 605–612.

Lockhart, P.J., Larkum, A.W.D., Steel, M.A., Waddell, P.J., Penny, D., 1996. Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. Proc. Natl. Acad. Sci. 93, 1930–1934.

Lockhart, P.J., Steel, M.A., Barbrook, A.C., Huson, D.H., Howe, C.J., 1998. A covariotide model describes the evolution of oxygenic photosynthesis. Mol. Biol. Evol. 15, 1183–1188.

Mallows, C.L., 1957. Non-null ranking models. I. Biometrika 44, 114–130.

Marden, J.I., 1995. Analyzing and Modeling Rank Data. Chapman & Hall, London.

Mau, B., Newton, M.A., Larget, B., 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics 55, 1–12.

Page, R., 1996. On consensus, confidence and total evidence. Cladistics 12, 83–92.

Page, R., Holmes, E., 2000. Molecular Evolution, A Phylogenetic Approach. Blackwell Science, Oxford.

Penny, D., Holmes, S., 2001. Doom01: biological mathematics in evolutionary processes. Trends Ecol. Evol. 16, 275–276.

Pezner, P., 2000. Computational Molecular Biology, an Algorithmic Approach. MIT press, Cambridge, MA.

Rambaut, A., Grassly, N.C., 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13, 235–238.

Robbins, H., 1980. An empirical Bayes estimation problem. Proc. Natl. Acad. Sci. 77, 6988–6989.

Robbins, H., 1983. Some thoughts on empirical Bayes estimation. Ann. Statist. 11, 713–723.

Robbins, H., 1985. Linear empirical Bayes estimation of means and variances. Proc. Natl. Acad. Sci. 82, 1571–1574.

Sanderson, M.J., Donoghue, M.J., Piel, W., Eriksson, T., 1994. Treebase: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. Am. J. Bot. 81, 183.

Sanderson, M., Purvis, A., Henze, C., 1998. Phylogenetic supertrees: assembling the trees of life. Trends Ecol. Evol. 13, 105–109.

Sankoff, D., Blanchette, M., 1999. Phylogenetic invariants for genome rearrangements. J. Comput. Biol. 6, 431–445.

Sankoff, D., Cedergren, R., 1983. Simultaneous comparison of three or more sequences related by a tree. In: Time Warps, String Edits, and Macromolecules: the Theory and Practice of Sequence Comparison. Addison-Wesley, New York, pp. 253–264.

Schwikowski, B., Vingron, M., 1997. The deferred path heuristic for the generalized tree alignment problem. J. Comput. Biol. 4, 415–431.

Steel, M.A., 1994. The maximum likelihood point for a phylogenetic tree is not unique. Syst. Biol. 43, 560–564.

Steel, M., Penny, D., 2000. Parsimony, likelihood and the role of models in molecular phylogenetics. Mol. Biol. Evol. 17, 839–850.

Steel, M.A., Szekely, L.A., 1999. Inverting random functions. Ann. Combin. 3, 103–113.

Steel, M.A., Szekely, L.A., 2002. Inverting random functions (II): explicit bounds for discrete maximum likelihood estimation, with applications. SIAM J. Discrete Math. 15 (4), 562–575.

Steel, M., Dress, A., Bockner, S., 2000. Some simple but fundamental limits for supertree and consensus tree methods. Syst. Biol. 42, 363–368.

Strimmer, K., Moulton, V., 2000. Likelihood analysis of phylogenetic networks using directed graphical models. Mol. Biol. Evol. 17, 875–881.

Tang, H., Lewontin, R., 1999. Locating regions of differential variability in DNA and protein sequences. Genetics 153, 485–495.

Tuffley, C., Steel, M., 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bull. Math. Biol. 59, 581–607.

Tukey, J., 1975. Mathematics and the picturing of data. In: Proceedings of the International Congress of Mathematicians, Vol. 2, Vancouver, pp. 523–531.

Van-Lint, J., Wilson, R., 1992. A Course in Combinatorics. Cambridge University Press, Cambridge.

Von Haeseler, A., Churchill, G., 1993. Network models for sequence evolution. J. Mol. Evol. 37, 77–85.

Wheeler, W., 1991. Congruence among data sets: a Bayesian approach in phylogenetic analysis of DNA sequences. Miyamoto M.M., Cracraft, J. (Eds.). Phylogenic Analysis of DNA Sequences. Wiley, NY.

Yang, Z., 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39, 306–314.

Yang, Z., 1995. A space–time process model for the evolution of DNA sequences. Genetics 139, 993–1005.

Yang, Z., Rannala, B., 1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. Mol. Biol. Evol. 14, 717–724.

Youden, W.J., 1972. Enduring Values. Technometrics. 14, 1–11.

Zharkikh, A., Li, W.H., 1995. Estimation of confidence in phylogeny: the complete and partial bootstrap technique. Mol. Phylogenet. Evol. 4, 44–63.