Molecular Phylogenetics and Evolution
Vol. 17, No. 1, October, pp. 58–75, 2000
doi:10.1006/mpev.2000.0816, available online at http://www.idealibrary.com on IDEAL®

Slicing Hyperdimensional Oranges: The Geometry of Phylogenetic Estimation

Junhyong Kim[1]

Department of Ecology and Evolutionary Biology, Department of Molecular, Cellular, and Developmental Biology, Department of Statistics, Yale University, New Haven, Connecticut 06511

Received September 28, 1999; revised May 15, 2000

A new view of phylogenetic estimation is presented where data sets, tree evolution models, and estimation methods are placed in a common geometric framework. Each of these objects is placed in a vector space where the character patterns are the basis vectors. This viewpoint allows intuitive understanding of various complex properties of the phylogenetic estimation problem structure. This is illustrated with examples discussing data set combinations, mixture models, consistency, and phylogenetic invariants. © 2000 Academic Press

Key Words: **geometry; accuracy; consistency; phylogenetic invariants; mixture models.**

Introduction

As evidenced by the exponential growth of papers containing tree estimates (even outside of evolutionary biology), phylogenetic analysis has become an indispensable part of evolutionary analysis. Yet, the techniques are complex and many problems are unresolved or have complicated relationships with each other. For example, some of the recent topical issues include the effects of combining data sets, the accuracy and consistency of various methods, the effects of weighting, the effects of taxon sampling, and bias due to topology, to name a few (De Queiroz *et al.*, 1995; Hillis, 1995; Huelsenbeck, 1995; Huelsenbeck and Kirkpatrick, 1996; Kim, 1996). These problems and their problem domains are clearly interconnected, but because these connections are rather complicated it is difficult for us to obtain an intuitive understanding of the entire structure of the problems. In this paper, I present a new view of the phylogenetic estimation problem in which data sets, tree models, and estimation methods are placed in a common geometric framework. In particular while the geometry of evolutionary tree models has been previously considered (e.g., Cavender and Felsenstein, 1987; Efron *et al.*, 1996) the construction presented here is new in that geometric representations of estimation methods are also considered. This common framework allows a geometric interpretation of the myriad properties of phylogenetic estimation. The utility of this construction is that it gives us a single picture of the complicated phylogenetic estimation process. It also leads to purely geometric proof techniques that can be useful for difficult-to-analyze problems. In the following, I start with a construction first introduced by James Cavender (1978). The key idea in his paper is to view characters in terms of all possible state assignments at the tips of the trees. This construction allows us to view entire data sets as a point in a large dimensional space. While this is certainly not mechanically different from probability computations given previously (e.g., Felsenstein, 1981), Cavender's paper shifted the viewpoint from algebraic computations to geometric properties. In fact, this geometric perspective was further developed in Cavender and Felsenstein (1987) with its remarkable Fig. 3 (page 69) sketching the geometry of tree models. In this paper, I first review this construction in detail and emphasize its geometric nature. I then describe how tree estimation methods can be seen as a geometrical partitioning of this large dimensional space. Finally, I demonstrate the utility of this view with geometrical interpretations of consistency, accuracy, mixture data sets and mixture models, and phylogenetic invariants. In the following, some of the material is pedantic and appears in many other texts for which I apologize, but I repeat it here to fix ideas.

Data Sets as Points in Large Dimensional Space

With finite state characters there is a fixed number (albeit very large) of possible types of character state assignments for a fixed number of taxa. For example, with four-state characters like the nucleotides of a DNA sequence and, say, five taxa, there are $4^5 = 1024$ possible character state assignments. The assignment

[1] Address for correspondence: Department of Ecology and Evolutionary Biology, P.O. Box 208106, New Haven, CT 06520-8106. Fax: (203) 432-3854. E-mail: junhyong.kim@yale.edu.

1055-7903/00 $35.00

of character states to taxon 1, taxon 2, etc. can be [A, A, A, A, A], [A, A, A, A, C], . . . , [T, T, T, T, T]—up to 1024 different types. I will call a particular assignment of character states to the terminal taxa a *character pattern*—i.e., [A, A, A, A, A] is one character pattern and [A, A, A, A, C] is another pattern, and so on. For example, here is an empirical data set with 10 characters:

CHAR \ TAXA	1	2	3	4	5
1	A	A	C	C	C
2	T	T	A	A	C
3	A	A	C	C	C
4	C	T	C	G	G
5	C	T	C	G	G
6	T	T	A	A	C
7	C	C	C	G	G
8	C	C	C	G	G
9	A	A	C	C	C
10	T	T	A	A	C

This data set consists of 3 character patterns of the type [A, A, C, C, C], 3 of the type [T, T, A, A, C], 2 of the type [C, T, C, G, G], and 2 of the type [C, C, C, G, G]. Suppose also that [A, A, C, C, C] was the 28th character pattern in a numbering scheme, [T, T, A, A, C] was the 878th character pattern, [C, T, C, G, G] was the 234th, and [C, C, C, G, G] was the 212nd. (These numbers are purely made up for illustration purposes. For the numbering scheme, whether [A, A, A, A, A] comes before [A, A, A, A, C] is irrelevant in general. It is relevant if we want to recreate the exact data set from the list of numbers but we can always construct a standard scheme, say, by recursion.) Then, the data set can be represented by the relative frequencies of each character pattern written as an ordered list of numbers,

$$(0, \ldots, 0.3, \ldots, 0.2, \ldots, 0.2, \ldots, 0.3, \ldots, 0, 0, 0).$$

That is, the data set can be represented by a vector in a 1024-dimensional space. This construction gives us a correspondence between a vector and a data set. Geometrically a vector with l elements can be seen as a point in a l-dimensional (Euclidean) space. Therefore, a data set can now be seen as a point in a l-dimensional space, where l is the number of possible character patterns. One restriction is that since these are relative frequencies, the sum of the elements in the vector must equal 1. Therefore, vectors representing data sets are restricted to the simplex contained in the l-dimensional space. (In the following I will refer to this space as the character pattern simplex, or simply the simplex. More formally, this simplex is a subset of $\mathbf{R}^n$ obtained as an isomorphism from the free vector space of character patterns. All geometric properties that I discuss are properties of subsets of $\mathbf{R}^n$ with its usual Euclidean geometry.)

To summarize, given s-state characters and t taxa, the number of possible character patterns is s^t. The relative frequency of each character pattern forms a simplex in the s^t-dimensional space where the axes of this space are the relative frequencies of each kind of character pattern. That is, the coordinates of any point give the relative frequencies of each character pattern.

Model of Character Evolution

Suppose a stochastic model of character evolution for s-state characters and t-taxa were specified. The model would include the tree topology, branch lengths, mode of character change, and possibly many other parameters. But, ultimately, the model would specify the probability of each character pattern—it may take a lot of computations, but eventually a model of character evolution ends up specifying the probability (or a set of probabilities) of each character pattern for all s^t possible patterns. The probability of each kind of character pattern can be listed as a sequence of numbers that sum to 1 (sometimes called the spectra; Lockhart *et al.*, 1994). For example,

$$(0.002, 0.013, 0.009, \ldots, 0.102, \ldots, 0.0, 0.052, 0.011). \quad (1)$$

Therefore, as with the previous construction, a model of character evolution can also be seen as a point on a simplex contained in a s^t-dimensional space.

Given sequence of numbers like (1) as the model of character evolution, finite data sets (samples) from the model can be generated. The sampling distribution of the finite data sets follows a simple multinomial distribution,

$$\begin{aligned} \text{Prob}\{c_1 = f_1, \ldots, c_z = f_z \mid \sum_{i=1}^{z} f_i = n\} \\ = \frac{n!}{f_1! f_2! \cdots f_z!} p_1^{f_1} p_2^{f_2} \cdots p_z^{f_z}, \end{aligned} \quad (2)$$

where c_i is the random variable representing the frequency the ith character pattern, f_i is the observed frequency of the ith character pattern, n is the total number of characters, and p_i is the probability of the ith character pattern as given by the model, e.g., the list (1). The indices run from 1 to $z = s^t$, the total number of possible character patterns. (The sampling distribution of relative frequencies, $r_i = f_i/n$, conditioning on the data sets size, n, is given by the same distribution.) The sampling variance of the relative frequency of the ith character pattern is $f_i(1 - f_i)/n$.

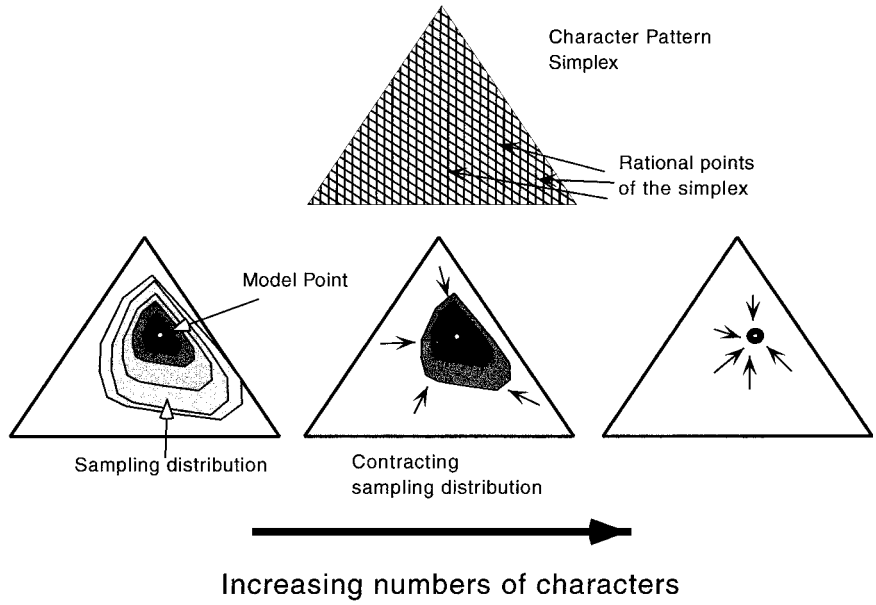

FIG. 1. Schematic picture of data sets seen as points in the character pattern simplex (see text). The triangles represent the simplex space. The rational points of the simplex correspond to finite-sized data sets (top triangle). A tree model of character evolution is a point in simplex and the model point induces a sampling distribution of finite-sized data sets. The sampling distribution contracts around the model point as the size of the data set becomes larger. It eventually contracts to the model point (bottom triangles).

Therefore, when n goes to infinity, the sampling variance goes to zero. This is interpreted as saying that the probability that the relative frequency of the ith pattern, r_i, equals any other value than p_i goes to zero (usually more precisely stated using the law of large numbers, Hogg and Craig, 1978). Or loosely, r_i goes to p_i, justifying, say, the study of consistency properties by "plugging in" the p_i for the relative frequency of each character pattern for infinite-sized data sets (e.g., Felsenstein, 1978; Kim, 1996). In terms of geometry, the model point generates a sampling distribution by inducing a probability mass (the multinomial distribution) on the rational points of the character simplex. This probability mass tends to become increasingly concentrated around the model point as the number of characters increase (Fig. 1).

A Parameterized Tree Model Generates a Model Curve

The kind of character evolution model described above is a "point model" in the sense that the parameter values (i.e., the probabilities of character change) are all fixed and represented by a fixed point in the character simplex. A tree topology denotes a family of character evolution models where the probability of each character pattern is a function of several variable parameters like the branch length and rate matrix, but all sharing the same branching order (= tree topology). The Markov model of character evolution over a tree has been extensively discussed in the literature (e.g., Chang, 1996; Goldman, 1990; Penny *et al.*, 1994) and the common Markov model in the literature is the continuous time Markov model with the conditional probability transition matrix after some time period t written as

$$P(t) = e^{Rt}, \tag{3}$$

where $P(t)$ is the transition matrix and R is the rate matrix, a real valued matrix with rows summing to zero. A discrete Markov evolution model can be obtained from (3) by evaluating (3) on a discrete set of time points (not necessarily with uniform intervals). In the tree models, the branching points of the tree serve as the discrete set of points (vertex-to-vertex). Therefore, this vertex-to-vertex discrete Markov model is a more general model than a continuous time model since any continuous time model can be converted into a discrete vertex-to-vertex evolution model by evaluating Eq. (3) at appropriate time intervals. However, given an arbitrary discrete Markov transition matrix, it cannot be always parameterized with a continuous time parameter unless it is a solution to an equation of the form (3). Therefore, in the following I will use the discrete parameterization and the probabilities of the character patterns will be a polynomial function.

If a tree is fixed and a Markov character evolution model is specified, the probability of any character pattern is easily computed as a function of the Markov transition matrices and the marginal distribution of character states at the root of the tree. This kind of computation is standard in the literature

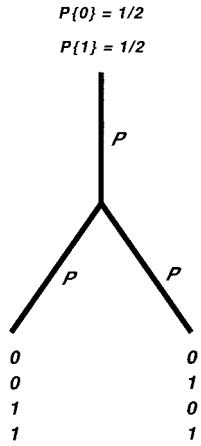

FIG. 2. A simple model of two-state character evolution over a two-taxon tree with a root. Each branch has the same probability, p, of changing from 0 to 1 or 1 to 0. The probability of {0} state or {1} state at the root is $\frac{1}{2}$, respectively. All possible character state assignments for the tree are shown at the tips of the tree.

(e.g., Felsenstein, 1981; Hendy and Penny, 1989) but I repeat a small example to make ideas clear. Suppose we have a two-taxon tree with a third root taxon. Also assume binary state characters and parameters of the model are as given in Fig. 2. That is, the marginal probability of either a 0 or a 1 state at the root is $\frac{1}{2}$ and the transition probability for each of the three branches is the same (denoted by p). With binary states, there are four possible character patterns at the two terminal taxa and the probabilities of each character pattern are,

$$\begin{aligned}
P\{[0,0]\} &= 1/2((1-p)^3 + p^3) \\
&\quad + 1/2(p(1-p)^2 + p^2(1-p)) \\
P\{[0,1]\} &= 1/2(p(1-p)^2 + p^2(1-p)) \\
&\quad + 1/2(p^2(1-p) + p(1-p)^2) \\
P\{[1,0]\} &= 1/2(p(1-p)^2 + p^2(1-p)) \\
&\quad + 1/2(p^2(1-p) + p(1-p)^2) \\
P\{[1,1]\} &= 1/2((1-p)^3 + p^3) \\
&\quad + 1/2(p(1-p)^2 + p^2(1-p)),
\end{aligned} \tag{4}$$

where P{[0, 0]} denotes the probability of the character state 0 at the first taxon and also at the second taxon. (That is, [0, 0] is a character pattern.)

The important part of Eq. (4) is that the probability of a character pattern is a function of the transition matrices and the marginal state distribution at the root. In this simple model, there is only a single variable, p. Equation (4) can be written in the following form to emphasize that the probability of each character pattern is a function of the parameter p:

$$\begin{aligned}
P\{[0,0]\} &= \Phi_1(p) \\
P\{[0,1]\} &= \Phi_2(p) \\
P\{[1,0]\} &= \Phi_3(p) \\
P\{[1,1]\} &= \Phi_4(p).
\end{aligned} \tag{5}$$

In the previous sections, I noted that the probability of each character pattern forms a vector and a point in s^t-dimensional space. The set of equations in (5) forms a vector-valued function in a 4-dimensional simplex parameterized by p. That is, (5) describes a curve in 4-dimensional space. Figure 3a shows the curve drawn by Eq. (4) as the variable p varies from 0 to 0.5 where the three axes are the probability of the character patterns [0, 0], [0, 1], and [1, 0]. (The probability of the pattern [1, 1] was left out.) As can be seen from the figure, in this particularly simple model, the curve drawn by the equation is actually a straight line. This curve (line) then represents the collection of character evolution models generated by the tree topology and the particular parameterization shown in Fig. 2.

The class of models represented by Eq. (4) is the simplest model we can have for this two-taxon tree.

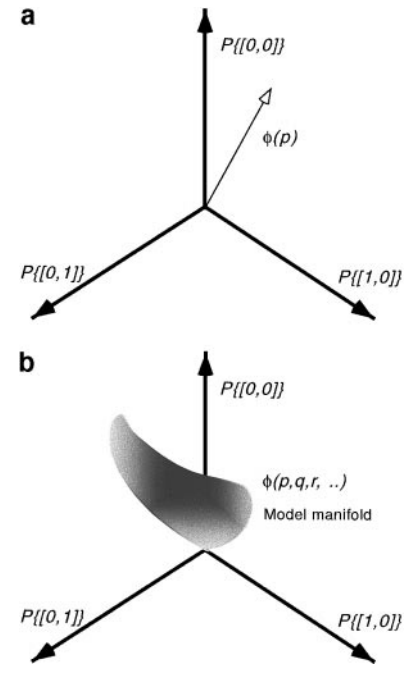

FIG. 3. (a) The tree model manifold (see text) of character pattern probabilities for the model tree shown in Fig. 2. The model manifold is a line parameterized by p, the probability of change per branch. (b) Hypothetical model manifold with more numbers of parameters. The model manifold can have up to as many dimensions as the number of parameters.

More generally, there can be different probabilities for every branch, direction of change (0 to 1 or 1 to 0), and the states at the root. That is,

$$\begin{aligned} &P\{[0,\,0]\} \\ &\quad = \Phi_1(A_0,\, A_1,\, p^1_{00},\, p^2_{00},\, p^3_{00},\, \ldots,\, p^1_{11},\, p^2_{11},\, p^3_{11}) \\ &P\{[0,\,1]\} \\ &\quad = \Phi_2(A_0,\, A_1,\, p^1_{00},\, p^2_{00},\, p^3_{00},\, \ldots,\, p^1_{11},\, p^2_{11},\, p^3_{11}) \\ &P\{[1,\,0]\} \\ &\quad = \Phi_3(A_0,\, A_1,\, p^1_{00},\, p^2_{00},\, p^3_{00},\, \ldots,\, p^1_{11},\, p^2_{11},\, p^3_{11}) \\ &P\{[1,\,1]\} \\ &\quad = \Phi_4(A_0,\, A_1,\, p^1_{00},\, p^2_{00},\, p^3_{00},\, \ldots,\, p^1_{11},\, p^2_{11},\, p^3_{11}). \end{aligned} \tag{6}$$

In this case, since there are more variables the model forms a higher dimensional curve (ignoring rigor, I will call such curves *manifolds;* Fig. 3b). Therefore, a manifold in s^t-dimensional space can be associated with the family of character evolution models of a given tree topology. That is, if a particular set of values for the parameters are chosen, a particular model of character evolution is determined which generates the probability of each character pattern over the given tree topology. This model of character evolution over the tree topology can be seen as a point in a large dimensional space (the vector space of character pattern probabilities). Equations such as (6) with free parameters represent a family of character evolution models—the models that share a tree topology (= branching order) but that vary in other parameters like branch length and rate. This family of character evolution models forms a collection of points in the vector space (of character pattern probabilities). The collection of points forms a geometrical object, the model manifold (e.g., Fig. 3). Thus, a tree topology and the set of character evolution models over the tree topology are represented as a geometrical object.

Geometry of Model Manifolds

Given the characterization of stochastic tree models as manifolds, it is useful to first discuss some general properties of the geometry of the tree model manifolds. The most general tree-evolution model allows each branch to have different transition matrices and within each transition matrix all parameters to vary freely (except the restriction that they are nonnegative and the rows of a Markov transition matrix have to sum to 1.0). A rooted t-taxon binary tree (a tree with no multifurcating vertices) has $2t - 2$ branches. A general Markov model for s-state characters has $s(s - 1)$ parameters. Therefore, the most general Markov model of character evolution over a fixed t-taxon tree has

$$(2t - 2)s(s - 1) + (s - 1)$$

parameters. (The last $(s - 1)$ term is for the marginal probability distribution of the states at the root.) This implies that the character pattern probabilities will be a function of $(2t - 2)s(s - 1) + (s - 1)$ parameters and, geometrically, the manifold will have up to $(2t - 2)s(s - 1) + (s - 1)$ dimensions in the local neighborhood at any point. The quantity $(2t - 2)s(s - 1) + (s - 1)$ is a maximum rather than the exact dimensions because of possible degeneracy. For this class of compact continuously parameterized objects the dimension is given by the rank of the Jacobian of the parametric functions. In fact, degenerate points exist at the boundaries of our models, e.g., zero length branches. It is common to put restrictions on the general model. For example, we often assume symmetry of the character state transitions (e.g., state A to T transition has the same probability as T to A transition). Such symmetry assumptions would reduce the maximum possible dimensions by reducing the number of free parameters (to $(2t - 2)s(s - 1)/2 + (s - 1)$). In the extreme case of simplification, a one-parameter model as in Eq. (5) results in a model curve that is (at most) one dimensional.

Different models of character evolution can be subsets of one another. For example, a model with a single parameter like Eq. (5) is a submodel of a model with more parameters like Eq. (6). Geometrically, a model is a submodel of another model if the model manifold of one model is wholly contained in the model manifold of the other model. (Such relationship has also been called nested models (Yang *et al.*, 1994).) Markov models of character evolution are often classified according to the form of the transition matrices (cf. Hasegawa *et al.*, 1991). For example, the Juke-Cantor model (Jukes and Cantor, 1969) of nucleotide evolution and Kimura two-parameter model (Kimura, 1981) are identical if we set the transition/transversion parameters equal to each other. However, whether this implies that one model is a submodel of another model is not always straightforward since it has to imply one model manifold being the submanifold of another model.

Proposition 1. Given some Markov model of evolution and its induced model manifold, any polynomial relation between the elements of the transition matrix induces a model submanifold (*the submanifold might not be strictly smaller*).

Proof. The model manifold is obtained as a variety of the elimination ideal (see section on Phylogenetic Invariants for definition) in the polynomial ring $k[\Phi, \theta]$ to $k[\Phi]$, where Φ is the set of variables for character pattern probabilities (left-hand side of Eq. (6)) and θ is the set of variables representing the transition matrix elements (right-hand side of Eq. (6)). Any polynomial relations in the transition matrix elements can be rep-

resented as a variety of an ideal over $k[\theta]$ which can be extended to $k[\Phi, \theta]$. Therefore, the model manifold induced by polynomial relations is given by the intersection of the two varieties projected to $k[\Phi]$ and is a subset of the points of the original variety.

The relevance of Proposition 1 is that whenever we build submodels using a polynomial relation between the elements of the transition matrix, we can use log-likelihood ratio tests with the standard χ^2 distribution approximation for hypothesis tests between the models. However, this assumes that the maximized likelihood point for either the null hypothesis or the alternative hypothesis is not degenerate. For example, the χ^2 approximation will not hold for degenerate points like zero length internal branches. Symmetry assumptions are special cases of polynomial relations and they can also be seen as having other geometric consequence for the model manifold. For example, with binary states the assumption that 0 to 1 transition has the same probability as 1 to 0 transition implies that Prob{[0, 0, 0, 0]} = Prob{[1, 1, 1, 1]} and so on. That is, the probability of any character pattern and the probability of its binary complement will be identical. This means that the model manifold will be restricted to the linear subspace defined by Prob{[0, 0, 0, 0]} − Prob{[1, 1, 1, 1]} = 0. Similar considerations of this kind lead to the extraction of linear invariants for phylogenetic trees (Lake, 1987a,b; Nguyen and Speed, 1992; Steel *et al.*, 1993, see below). Conversely, the most general model with general transition matrices for each branch does not have any linear invariants, but it may have higher order invariants.

With larger numbers of taxa, we have parametric functions in higher dimensions and there will be one manifold for each possible tree topology. Figure 4 shows an example of model manifolds for binary state four-taxon trees with two free parameters (in fact, the figure shows the set of model trees discussed in Felsenstein, 1978; only 3 out of 16 dimensions are shown). Figure 4 shows the model surfaces for all three tree topologies (labeled as Tree 1, Tree 2, and Tree 3).

The geometrical relationship of the different tree models can be complicated, but an outline of some general features can be sketched. One property can be easily seen from the fact that the tree topologies differ from one another by the internal branches. If all the internal edges are set to zero (or equivalently set the Markov transition matrix to the identity matrix), every tree topology will be equivalent to each other. Therefore, the manifolds corresponding to each tree topology all share a subspace where all the internal edges are zero (up to $ts(s - 1) + (s - 1)$ dimensions). Similarly, if the terminal edges correspond to (converges to) infinite time (or equivalently if the Markov transition matrices have rows with constant and identical elements), regardless of the tree topology, the probability of any

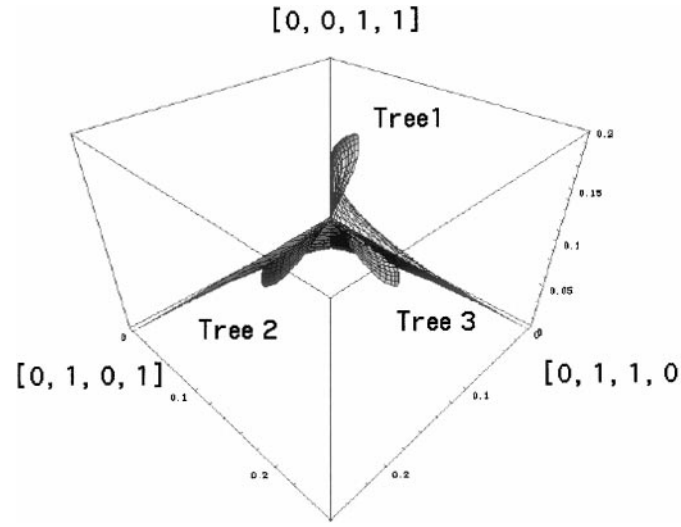

FIG. 4. A low dimensional projection of the actual model manifold for a two-state four-taxon tree parameterized according to Felsenstein (1978). The three different "lobes" correspond to the model manifold for each of the three different possible tree topologies. The axes correspond to informative characters by the parsimony criterion, namely, [0, 0, 1, 1], [0, 1, 0, 1], and [0, 1, 1, 0].

character pattern converges to independent products of the equilibrium distribution of the individual states. (This is assuming that we use the usual nonpathological models that have an equilibrium distribution.) Therefore, the manifolds corresponding to each tree topology all converge toward a single point at or near the center of the simplex. If the marginal equilibrium distribution of the Markov model is uniform then the points at zero length branches and infinite length branches will also meet.

Setting individual branch lengths to zero identifies different topologies. For example, an internal branch separates three different tree topologies that are related to each other by a nearest-neighbor-interchange operation (Swofford *et al.*, 1996). Setting this internal branch length to zero will identify the character pattern probability of the three topologies. Therefore, the manifolds corresponding to three topologies related to each other by a NNI operation share a maximum of $(2t - 3)s(s - 1) + (s - 1)$ dimensional subspace and differ along a maximum of $s(s - 1)$ dimensions. Figure 5 shows a schematic picture of the geometrical relationships of different tree models—we have a high dimensional orange! Now, suppose two models, Tree 1 and Tree 2, differ from one another by a NNI operation around an internal edge and the length of this internal edge is zero. For some set of parameter values for the branches, Tree 1 will generate a particular set of probabilities for the character patterns. Because the length of the internal edge is zero we can find parameter conditions for model Tree 2 that will yield

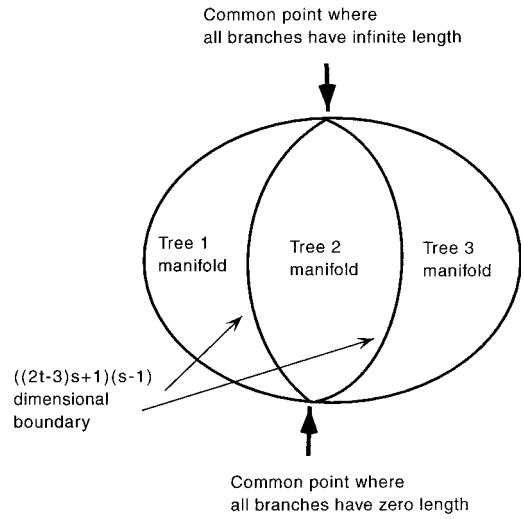

FIG. 5. A cartoon diagram of the spatial arrangement of the model manifolds for three different tree topologies related to each other by nearest-neighbor-interchange. All model manifolds share two points, the point where all branch lengths are zero and the point where all branch lengths are infinite. Two topologies differing by a single internal branch (NNI-related) share a $((2t-3)s+1)(s-1)$ dimensional subspace.

identical probability of the character patterns. This raises the interesting question whether two different tree topologies can be made to yield identical character pattern probabilities if none of the internal edges are allowed to be zero. I discuss this question in a separate section below.

(In addition to these properties the tree model manifold is connected and compact assuming a discrete time parameterization. This is because there is a continuous map from I^w to the model manifold—namely, the polynomial function of character pattern probabilities, where I is the closed interval $[0, 1/s]$—the possible value range for the elements of the transition matrices. Here w is the total number of variables in the character evolution model and s is the number of states. Assuming a continuous time parameterization we have an open point at when time (or branch length) goes to infinity. In this case, we can assume that we are interested in the closure of the model manifold that is connected and compact and is sufficient for the development of our theory.)

Geometrical Interpretation of Tree Estimation Methods

Tree estimation methods take data sets as inputs and produce one or more tree topologies as outputs (and perhaps some other estimates such as branch lengths). As mentioned above, a data set can be seen as a point in a large dimensional space. Then, the action of a tree estimation method can be seen as a labeling of this point; that is, the point is labeled with the identity of some tree topology. If we were to systematically input every possible data set to an estimation method, it would output a tree topology label(s) for each data set, thereby also a label for each point on the s^t-dimensional simplex. Most methods produce a tree for every input data set, therefore, a particular estimation method partitions the character pattern simplex into different regions labeled with different tree topologies (Fig. 6). (Of course, it is not a partition in the strict sense since some data sets can yield more than one tree estimate. This technicality can be resolved by treating such points as the boundary of two or more tree topologies and consider the estimation partitions as open sets.) Now consider the collection of all points labeled with the same tree topology (i.e., a partition). This is the collection of all data sets that reconstruct to the same tree topology. This collection of points also forms a geometrical object that I will call a "method partition." Unless two different estimation methods produce exactly the same tree topology for every data set, different estimation methods form geometrically different method partitions (Fig. 6).

While geometric characterization is not easy for many methods, some methods yield to analysis like

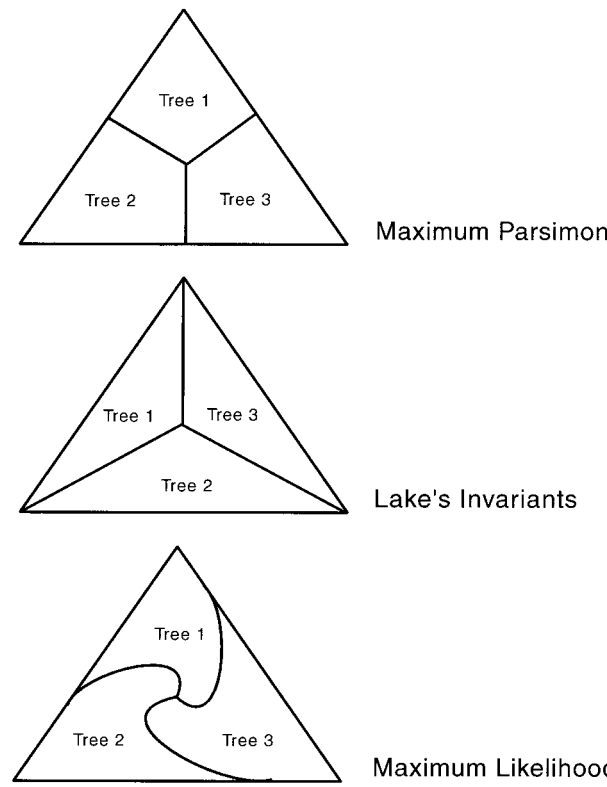

FIG. 6. A schematic diagram of the geometry of estimation methods. A given estimation method can be seen as a partitioning of the character pattern simplex into different tree topology regions. Different methods have geometrically different partition shapes. (The particular shapes shown here are for illustration purposes only and have no real meaning.)

the maximum-parsimony method and phylogenetic invariants.

Maximum-Parsimony Method

The maximum parsimony method defines a parsimony length function of the form, $l(c, T)$, where c is a character pattern and T is a tree topology. For example, with a 4-taxon tree, and the character pattern [0, 0, 1, 1], the length function value for the topology {{1, 2}, {3, 4}} is 1, while for the topology {{1, 3}, {2, 4}}, the value is 2. The length function is easily computed for any character pattern and tree topology combination (Maddison, 1989; Swofford and Maddison, 1992). The parsimony length of a data set is obtained by summing the parsimony function over individual characters. In practice, we do not need to compute the length function for every character, but only for every different character pattern. The computed value is then multiplied by the frequency of the character pattern. Therefore, the parsimony length of a data set, D, and topology, T, is given by

$$\begin{aligned} L(D, T) &= f_1 \cdot l(c_1, T) + f_2 \cdot l(c_2, T) + \cdots \\ &\quad + f_z \cdot l(c_z, T) \\ &= \sum_{i=1}^{z} f_i \cdot l(c_i, T), \end{aligned} \tag{7}$$

where c_i is the ith character pattern, f_i is the frequency of that pattern in the data set and the index runs from 1 to $z = s^t$, the total number of possible character patterns for t taxa. Because the length function is summed over the character in this additive manner,

$$L(D_1 + D_2, T) = L(D_1, T) + L(D_2, T), \tag{8}$$

where $D_1 + D_2$ denotes the combining of two data sets by summing the frequency of the character patterns (equivalent to concatenating the two data sets). For relative frequencies,

$$L(D_1 + D_2, T) = \alpha L(D_1, T) + (1 - \alpha)L(D_2, T), \tag{8'}$$

where α is the relative size of data set D_1 compared to D_2. Therefore, the parsimony length function is additive with respect to concatenating data sets.

The parsimony method estimates the tree topology by choosing the topology that minimizes the data set length [Eq. (7)]. Suppose, $\overset{*}{T}$ is the maximum-parsimony tree for some data set D_1. Also, suppose that the same topology, $\overset{*}{T}$, is the maximum-parsimony tree for another data set D_2. Then it is seen by the additivity characteristic of (8) that $\overset{*}{T}$ is also the maximum parsimony tree for the combined data set $D_1 + D_2$.

Proof. Suppose T′ were a shorter tree for the combined data set. Then $L(D_1 + D_2, T') = L(D_1, T') + L(D_2, T') > L(D_1 + D_2, \overset{*}{T}) = L(D_1, \overset{*}{T}) + L(D_2, \overset{*}{T})$ which contradicts the statement that T' is a shorter tree than $\overset{*}{T}$. The first and the last equality is by (8) and the inequality is because $\overset{*}{T}$ is the maximum-parsimony tree for the individual data sets and L is a nonnegative function.

This shows that the set of data sets that is labeled as the same tree by the maximum-parsimony method is a convex set. That is, if $\overset{*}{T}$ is the maximum-parsimony tree for one data set and it is also the maximum-parsimony tree for another data set, then it is the maximum-parsimony tree for any convex combination (in terms of the frequency of each character pattern) of the two data sets. This means that the method partition of the maximum-parsimony method is a convex partition. I will call such methods convex methods and say

Proposition 2. The maximum parsimony estimation method is a convex method.

Proof. This immediately follows from the additive objective function as proven above.

In fact, it is a linear subspace of the character pattern simplex (because the parsimony method chooses a tree according to a set of linear inequalities). Convex methods have the nice property that they are robust to mixture models—in a particular sense that I will discuss below.

Phylogenetic Invariants

Phylogenetic invariants were previously introduced as a new approach to phylogenetic inference (Cavender, 1978; Cavender and Felsenstein, 1987; Fu and Li, 1992; Lake, 1987b; Navidi *et al.*, 1991; Nguyen and Speed, 1992). The invariants are functions of character pattern probabilities consisting of a set of polynomial equations. The equations denote a set of values that are "invariant" (constant) only on particular trees. Different trees have different sets of invariant equations. Phylogenetic estimation is done by estimating the character pattern probabilities with the observed frequencies, substituting into the invariant equations for each tree, and asking if the values deviate significantly from the expected value (usually zero).

Initial derivations of the invariants involved deductions from branch lengths of a tree (e.g., Cavender, 1978), algebra (e.g., Nguyen and Speed, 1992), or heuristic algorithms (e.g., Sankoff, 1990) which tended to be algebraically involved. From the geometry of phylogenetic estimation, we can obtain a more intuitive view of phylogenetic invariants. As mentioned above, a tree and a model of stochastic evolution yield a set of equations that represent the probability of each character pattern, for example, of this form,

$$\begin{aligned} p_1 &= \phi_1(\alpha_1 \cdots \alpha_t) \\ p_2 &= \phi_2(\alpha_1 \cdots \alpha_t) \\ &\vdots \\ p_n &= \phi_n(\alpha_1 \cdots \alpha_t), \end{aligned} \tag{9}$$

where each p_i denotes the probability of ith character pattern and α's are the parameters of the character evolution model. Equation (9) then is a parametric description of the model manifold embedded in the space of character pattern probabilities. For example, a parametric description of a unit circle in the plane is given by

$$\begin{aligned} x &= \cos(\theta) \\ y &= \sin(\theta). \end{aligned} \tag{*}$$

Equation (*) gives, for every value of θ, a value of the x coordinate and a value of the y coordinate. Another representation of a unit circle is given by the equation $x^2 + y^2 = 1$, or equivalently,

$$x^2 + y^2 - 1 = 0. \tag{**}$$

The parametric description, equation (*), specifies the circle by giving the x, y coordinates as a function of a parameter, θ. The equation (**) specifies the circle by prescribing a constraint condition on the set of points that belong to the unit circle—namely, if a point with the coordinates (p, q) is part of the unit circle, it has to satisfy the condition (**). That is, the sum of the squared coordinates minus one must always satisfy the invariant quantity zero. Given a random point with the coordinates (p, q), we can ask whether it belongs to the collection of points making up a unit circle by computing $p^2 + q^2 - 1$ and asking whether this quantity is zero.

Many geometric objects of interest (or, at least the connected compact objects) can be described by a set of constraint equations like (**). In which case, the object is described by the set of "roots" of the equation. The collection of such "root points" of an algebraic equation is called an *algebraic variety* (Cox *et al.*, 1992). The same geometric object can be also described by a parametric function like (*). The equations like (**) are the implicit function (or equation) forms of the parametric functions of the form (*). The two forms are (almost) equivalent descriptions of the same object. ("Almost," because the geometric object described by the implicit form can be a larger object that strictly contains the geometric object described by the parametric form.) Therefore,

Proposition 3. Phylogenetic invariants are the implicit function form of the parametric tree model functions of the form (9).

For example, consider Eq. (4) from above. From (4) we obtain,

$$\begin{aligned} P\{[0,0]\} &= 1/2(1 - 2p + 2p^2) \\ P\{[0,1]\} &= p - p^2 \\ P\{[1,0]\} &= p - p^2 \\ P\{[1,1]\} &= 1/2(1 - 2p + 2p^2). \end{aligned} \tag{10}$$

Immediately the following implicit equations are found in terms of the coordinates:

$$\begin{aligned} P\{[0,0]\} - P\{[1,1]\} &= 0 \\ P\{[0,1]\} - P\{[1,0]\} &= 0 \\ P\{[0,0]\} + P\{[0,1]\} + P\{[1,0]\} & \\ + P\{[1,1]\} - 1 &= 0. \end{aligned} \tag{11}$$

Each of the three equations in (11) is an equation of a hyperplane. The geometric object is described as the simultaneous roots of all three equations, that is, the intersection of three hyperplanes in four dimensions and therefore a line as we have seen above. For more complex models with more taxa, the parametric function describing the model is more complex and it is difficult to find its implicit form. There are standard computational algebraic geometry techniques to derive implicit forms from the parametric model functions called Groebner basis (also found in popular packages such as Mathematica). The basic idea is to rewrite equations of the form (9) as

$$\begin{aligned} p_1 - \phi_1(\alpha_1 \cdots \alpha_t) &= 0 \\ p_2 - \phi_2(\alpha_1 \cdots \alpha_t) &= 0 \\ &\vdots \\ p_n - \phi_n(\alpha_1 \cdots t\alpha_t) &= 0. \end{aligned} \tag{12}$$

That is, as a set of equations in $n + t$ variables. If we denote all possible polynomials in $n + t$ variables (over some field k) as $k[p_1 \ldots p_n, \alpha_{n+1} \ldots \alpha_{n+t-1}]$, Eq. (12) is a subset of such polynomials whose zero points (variety) correspond to the model manifold when projected to $p_1 \ldots p_n$. The problem is to find the set of polynomials in $k[p_1 \ldots p_n]$ whose variety contains the variety of the polynomials in $k[p_1 \ldots p_n, \alpha_{n+1} \ldots \alpha_{n+t-1}]$ projected to $p_1 \ldots p_n$. Such a set of polynomials is called the *elimination ideal* (Cox, 1992) and can be systematically found using Groebner basis algorithms with lexical ordering of the variables. Using Mathematica, the four-taxon two-state symmetric character invariants are found as

$$\begin{aligned} h_1 = {} & p_3p_5^2 + p_7p_5^2 + p_7^2p_5 + p_3p_6p_5 + p_3p_7p_5 + p_4p_7p_5 \\ & + p_6p_7p_5 + p_7p_8p_5 + p_4p_7^2 - p_4p_5p_6 - p_4p_6^2 \\ & - p_5p_7 + p_6p_8 + p_4p_7p_8 - p_3p_6p_8 - p_4p_6p_8 \\ & - p_5p_6p_8 - p_6^2p_8 - p_3p_7p_8 - p_6p_7p_8 - p_3p_8^2 \\ & - p_6p_8^2 \end{aligned}$$

$$
\begin{aligned}
h_2 &= p_7^2 + p_2p_7 + p_3p_7 + p_5p_7 + p_6p_7 + p_8p_7 - p_7 \\
&\quad + p_3p_5 - p_4p_6 + p_2p_8 \\
h_3 &= p_6^2 + p_2p_6 + p_3p_6 + p_4p_6 + p_5p_6 + p_7p_6 + p_8p_6 \\
&\quad - p_6 + p_2p_5 - p_4p_7 + p_3p_8 \\
h_4 &= p_1 + p_2 + p_3 + p_4 + p_5 + p_6 + p_7 + p_8 - 1, \qquad (13)
\end{aligned}
$$

where p_I's denote the probability of each character pattern (eight total given the symmetric model). These equations are more complex than Cavender and Felsenstein's (1987) original elegant set,

$$
\begin{aligned}
K &= (p_5 - p_8)(p_3 - p_2) - (p_7 - p_6)(p_1 - p_4) = 0 \\
T &= (p_2 + p_3)(p_5 + p_8) - (p_1 + p_4)(p_6 + p_7) = 0.
\end{aligned} \qquad (14)
$$

But by division of polynomials

$$
\begin{aligned}
K &= h_2 - h_3 + (p_6 - p_7)h_4 \\
T &= h_2 + h_3 - (p_6 + p_7)h_4,
\end{aligned} \qquad (15)
$$

so Cavender and Felsenstein's invariants describe a subset of the model manifold. This raises the question of how many equations are required to implicitly describe the model manifold. Felsenstein (1991) examined this question but left it incompletely answered. In general, the total number of simultaneous equations that define a geometric object is indefinite. This is because if the roots of $g(x) = 0$ describe a geometric object, the equation $f(x)g(x) = 0$ also has the same geometric object as a subset of its roots. However, a bound on the number of equations in the minimal set is found as the dimensions of the character pattern simplex minus the number of free variables in the tree model (the codimension) as postulated by Felsenstein (1991). For differentiable functions like the character pattern probabilities, the rank of the Jacobian matrix, **J**, gives us the local dimension of the model manifold. In the appendix, I give a proof of the following statement.

Proposition 4. The rank of the Jacobian matrix of the parametric equation equals the number of independent parameters in the Markov transition matrices attached to each branch.

Proof. See appendix.

Therefore the codimension, $n - r$, is the minimal number of equations that we need. There must be at least this many equations to describe the model manifold. However, possibly more are needed for a "nice" representation. This is seen the four-taxon two-state symmetric case of Eq. (14). The number of parameters in this model is 5 for the probability of change in each branch. The dimension of the character pattern space is 8 (because of the binary symmetries). Therefore, the codimension is 3 and we expect to need at least three invariants; however, it turns out that four as given in equation (13) are the best polynomial descriptors of the model manifold because they divide all other polynomial descriptors.

Finally, invariants derived using elimination ideals are guaranteed to be the smallest variety that contains the parametric model, though it may be larger. That is, the set of zero points given by (13) may be larger than the model manifold. This question can be answered by examining the entire sequence of elimination ideals (36 equations with hundreds of terms for the four-taxon two-state model) and asking whether the solutions to (13) extend to the entire sequence. An examination of the 36 equations shows that, in fact, the solutions to (13) do extend to the full ideal over $k[p_1 \ldots p_n, \alpha_{n+1} \ldots \alpha_{n+t-1}]$ (not repeated here, the equations are 20 pages long). Therefore, Eq. (13) describes the model manifold for the four-taxon two-state model exactly. It is difficult to examine this in general terms and each model must be examined individually. As well much of the theorems are given over the set of complex numbers whereas only the real numbers are relevant for phylogeny problems (but see Hagedorn, 1999, for a general exposition and Hagedorn and Landweber, 1999, for a survey).

The Combined Geometry of Data, Model, and Estimation Method

In the sections above, I discussed the geometry of data sets, tree models, samples of the tree models, and tree estimation methods. The description of each of these objects involved the space of character patterns—that is, the objects were described as geometrical objects in a large dimensional space where the axes of the space denote the frequency or probability of each kind of possible character patterns. All of these geometric objects can be embedded in a single picture. Figure 7 shows a schematic abstraction of such a picture. The triangle in the picture symbolizes that the geometrical objects are contained within a simplex. Every rational point in this simplex represents an empirical data set or a sampled data set from a model. A tree topology with a model of character evolution draws a manifold in this space (what I called a model manifold). If a point is fixed on the tree model manifold, this is equivalent to choosing particular values for the parameters of character evolution. Given such a point, a sampling distribution of the model is generated as probability values on the rational points of the simplex. Sampling distribution of the entire family of models for a given tree topology can be generated integrating the conditional sampling distribution over the set of points in the tree manifold with respect to a suitable probability measure for the points. The tree estimation

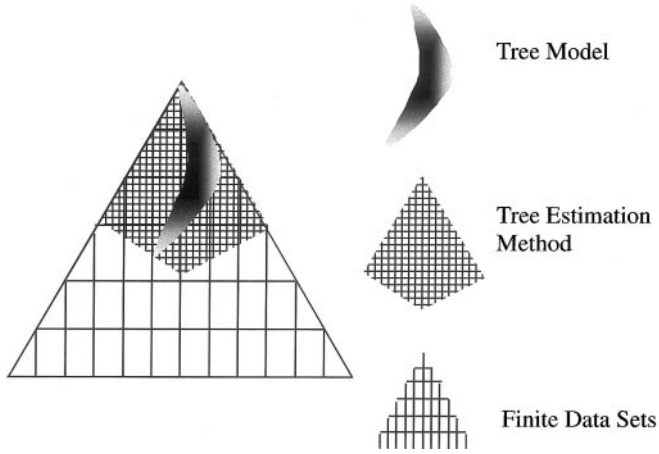

FIG. 7. A schematic diagram of the geometry of phylogenetic estimation. The triangle represents the character pattern simplex. The tree model is a subspace of this simplex (shown as the gray object). A tree estimation method is a partition of the simplex into tree topologies (shown as the speckled object). Finite-sized data sets are rational points of the simplex (shown as a grid).

methods now can be drawn as a partitioning of the simplex into labeled points (each label being the identity of the tree topology). In the following sections, I use this geometric setup to give a pictorial description of a variety of concepts and phenomena that arise in phylogenetic estimation.

Applications

Combining Data Sets and Mixture Models

Combining data sets has been extensively discussed in the literature (see De Queiroz *et al.,* 1995). Given the view of data sets as points in the character pattern simplex, we can obtain a very simple geometrical picture of a combined data set. Associate with data set X a point x in the simplex, and with data set Y a point y in the simplex. Then the combination of the two data sets lies on a line connecting the two data sets because it is a convex linear combination of the frequencies of character patterns in each data set. The exact position on the line will depend on the relative sizes of the data sets such that the ratio of the distance from the combined data set to the two original data sets will be proportional to the relative sizes of the data sets (Fig. 8a).

The same picture can be drawn for mixture models where some of the characters are generated from one kind of model (say a model with fast rates of evolution) and the other characters are generated from a different kind of model (say a model with slow rates of evolution). Again, the mixture model lies on some point within the line drawn between the two points representing the original models. The position of the point representing the mixture model depends now on the model of the mixture. For example, we might have a mixture model consisting of $\frac{1}{2}$ probability of drawing from model X (with the corresponding point x) and $\frac{1}{2}$ probability of drawing from model Y (with the corresponding point y). Then the mixture model will be represented by a point halfway on the line connecting x and y. More generally, there might be a mixture model of drawing from model X with probability α and drawing from model Y with probability $1 - \alpha$. Then the point representing the mixture model is α distance away from point y (assuming that the line connecting x and y has unit distance; Fig. 8a). More complicated models can be generated with the assumption that α has some prior distribution like the beta distribution. Then, the unconditional mixture model is obtained by integrating over the prior distribution (in a suitable sense) and the mixture model point will be $a/(a + b)$ distance from point y, where a and b are the parameters of the beta distribution.

This discussion can be extended to the commonly used model of gamma distributed rate mixture models

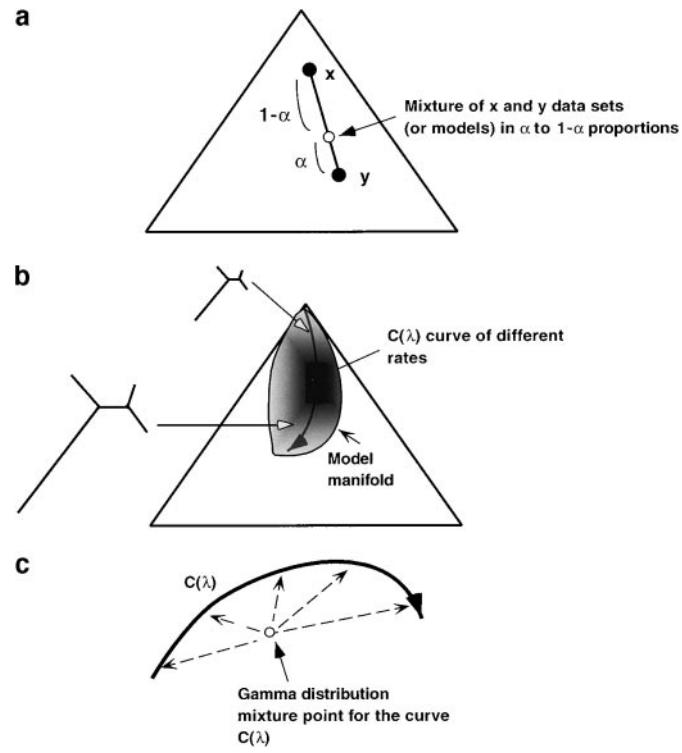

FIG. 8. (a) When two models or two data sets are combined, it corresponds to drawing a line between the two model points in the simplex (shown labeled as x and y) and selecting a point on the line. The particular position of the mixture (combination) point is proportional to the mixture proportions. (b) A model of rate variation across sites can be seen as a model curve that is a subset of model manifold. The (vector-valued) model curve, $\mathbf{c}(\lambda)$, is parameterized by the rate variable, λ. Points along this curve correspond to taking a particular tree shape and expanding the shape or contracting the shape (shown by the trees on the left). (c) A gamma distribution model of rate variation across sites is equivalent to integrating the model curve, $\mathbf{c}(\lambda)$. The result of the integration is a point inside the convex hull of the model curve.

(Yang, 1994). In this class of mixture models, there is a rate parameter that is a scalar quantity, say λ, applied as a multiplier to the rate matrix R in Eq. (3). The model then assumes that λ is drawn from a gamma distribution (usually with only one varying parameter) and the characters are drawn from a tree model conditioning on the value of λ. To geometrically understand this mixture model, we need to first determine the ensemble of tree models that are being "mixed." First note that a single rate parameter is being applied over the whole tree (of course, the expected amount of change on any given branch can be different since it is proportional to the length of the branch). Therefore, this is equivalent to picking some tree shape and tracing out all the uniform "contractions" or "expansions" of the tree (Fig. 8b). In the tree model manifold, this is equivalent to picking a point and drawing an one-dimensional (vector-valued) curve, $\mathbf{c}(\lambda)$, that passes through the point. The curve is one-dimensional because it varies according to the single parameter λ. In other words, an one-dimensional curve parameterized by the variable λ represents the family of possible "uniform" rate variations for a given tree shape. This curve is, of course, a subset of the tree model manifold. The mixture model consists of mixing the points of this curve together in proportions according to a gamma distribution of the λ parameter. The particular point represented by the gamma mixture model can be found by integrating over the (vector-valued) curve in a suitable sense (i.e., integrate the vector $\mathbf{c}(\lambda)$ with respect to the push-forward measure of λ; Fig. 8c). The important point is that the gamma distributed variable sites model is a single point in the character simplex rather than multiple points or a family of points for fixed values of λ. Therefore, the characters drawn according to this model are identically and independently distributed (*iid*). In fact, geometrically,

Proposition 5. A non-iid model of character evolution is given by a set of points in the character pattern simplex, not a single point.

Proof. This follows simply from a geometric interpretation of the definitions.

For example, if we have fixed rate categories (e.g., we assume that the third position evolves at a different rate than the first and second position of a codon), the resulting data set is not drawn from an *iid* model and the probability of the character patterns is given by three different points in the character pattern simplex. Certainly more complicated mixture models can be generated. For example, different rate distributions might be applied for every branch of the tree, etc. The convex hull of a geometrical object represents the all possible convex linear combination of points of the geometrical object. Therefore,

Proposition 6. The family of all possible mixture models is contained in the convex hull of the tree model manifold.

Proof. The convex hull of the models is a compact convex set. A compact convex set is the closure of convex combinations and therefore contains the integral sum with respect to a normalized measure.

In the previous discussion, I mentioned convex methods as methods whose method partition is convex. Therefore, if a convex method estimates some tree T for data set A and the same tree for data B, then it will estimate the same tree for any convex combination of the two data sets. Therefore,

Proposition 7. Let a convex method be a consistent estimator over some restricted parameter set, $\boldsymbol{S}$, then it is a consistent estimator for any mixture models of the parameter set $\boldsymbol{S}$.

Proof. Immediately follows from proposition 6.

For example, maximum parsimony is not a consistent estimator. But, suppose we restrict our attention to the parameter space where it is a consistent estimator, then it is a consistent estimator for any mixture models of the restricted parameter space.

Finally, a mixture model also has geometrical structure. One thing that is obvious is that since we are integrating over the rate parameter we loose one dimension, while gaining the dimensions of the parameters of the mixture distribution. In fact, if the parameters of the mixture distribution are fixed (say from a separate estimation process) then the resulting mixture models is a smaller model (in the number of dimensions) than the original model—that is, in some sense the mixture model is a simpler model. For some mixture models, phylogenetic invariants of the mixture model can be derived. (That is, the entire set of invariants not just the linear set which obviously contains the mixture model.) However, since the parametric equations include an integral there is no guarantee that we will end up with an algebraic variety. That is, the parametric equation may be an analytic function and the model manifold may not be exactly expressible with a set of algebraic equations (over $\mathbf{R}^n$).

Consistency and Separation of Tree Topologies

A phylogenetic estimator is consistent if the estimate converges to the model parameter (Felsenstein, 1978; Kim, 1998). The consistency property also has a very simple geometrical picture. In a previous section, I noted that the probability mass of a sampling distribution of data sets generated by a tree model (with fixed character evolution parameters) converges to a point on the character pattern simplex as the size of the

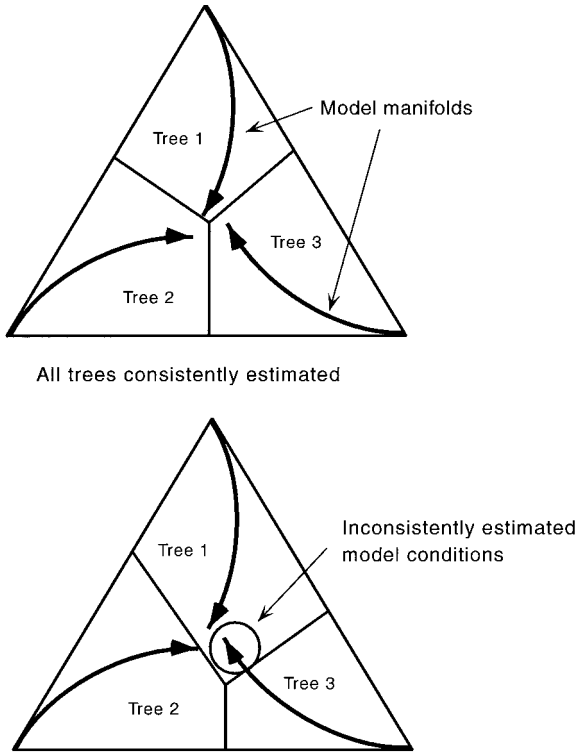

FIG. 9. A tree model is consistently estimated if its model point is completely contained within the estimation method partitions. If the entire manifold is completely contained inside the partitions, the tree topology is consistently estimated (top figure). If a part of the model manifold extrudes outside the estimation method partition like Tree 3 in the lower figure, a wrong tree is estimated (with infinite data) for those model conditions (Tree 3 is estimated as Tree 1 in this figure).

sample (the number of characters) increases. Therefore, if we have infinite sized data sets, we can identify the collection of such infinite-sized data sets for a given tree topology with the model manifold itself. Given the view of tree estimation methods as partitions of the simplex, a tree estimation method is consistent if every tree model manifold is completely contained in the tree estimation partition of that particular tree (Fig. 9, top). Otherwise, the parts of the tree model manifold that "stick out" of the partition are model conditions that result in the wrong estimate (Fig. 9, bottom). (The part that is sticking out is labeled as a different tree since it is in a different partition.)

From this geometrical picture, it is easy to gain intuition about various different scenarios. Immediately from the geometry it can be seen that consideration of the "borders" of the model manifolds is critical for the consistency property. That is, the method partition must "cleanly cut" through the borders of the method manifolds to separate the models. This means that we need to examine what happens to these borders, the model conditions at which two tree topologies meet—namely, the tree topologies with zero length internal branches or certain topologies where some of the branches have infinite length (see previous section on model manifolds). For example, it has been noted that a maximum-likelihood tree with the "wrong" model can be inconsistent (e.g., Kim, 1998). An easy geometrical proof can be given. Suppose that the data are generated under model X and estimated with a different model Y. By geometrical reasoning, the maximum likelihood estimate will be consistent if and only if the model manifold for the borders given by model Y is either identical to or completely contained in the borders given by model Y. This will only happen if the model manifold for X is a subset of the model manifold for Y. Otherwise, the maximum-likelihood estimate using the wrong model will be an inconsistent estimate.

For another example, suppose that the estimation partition was not a convex partition (e.g., the standard maximum-likelihood method). The tree manifold for the usual character evolution model is also not a convex collection of points. Suppose there is a method partition that completely contains a tree manifold (and therefore consistent), but not convex. It is easy to conjecture that a mixture model consisting of some combination of models from different points in the tree model manifold might not be consistently estimated because it can be made to stick out (Fig. 10). Of course, nonconvex geometrical objects (the method partition) can completely contain convex geometrical objects (the collection of mixture models) and more exact proofs and examples are given in Steel *et al.* (1994) and Chang (1996b). I noted above that general mixture models are contained in the convex hull of the model manifold. Since the convex hull is a much larger object than the original object it is natural to conjecture that general mixture models will not be consistently estimated, which in fact is the case because different tree models can not be separated (Steel *et al.*, 1994), something I discuss next.

The consideration of the consistency property led to a direct examination of the geometry of the tree model

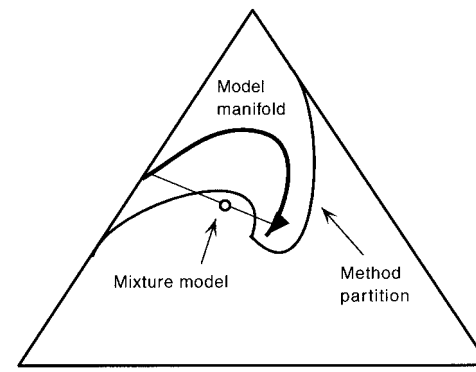

FIG. 10. If the estimation method partition is nonconvex, a mixture model can be inconsistently estimated (see text) even when the pure models are consistently estimated.

manifold. Drawing pictures of the tree model manifold seems to raise a question that is more fundamental than the questions concerning consistency: namely, can two different tree topologies yield exactly the same probability model on the character patterns? Geometrically, we are asking whether two tree model manifolds intersect one another in the character simplex in a nontrivial manner (there is always a trivial intersection since two tree topologies can be made exactly the same with appropriate zero length branches). Since the points of the character simplex define the probability of each kind of character pattern, the points of intersection of two tree model manifolds represent model conditions where the two trees define exactly the same probability and, therefore, cannot be distinguished by any estimation method. For obvious reasons it would be desirable if such intersections do not exist. When such intersections can be ruled out the tree topologies are separable and estimable by some reconstruction method. Chang (1996) gives comprehensive treatment of such a separable class of models. On the other hand, from the discussion of the convex hull of model manifolds it can be imagined that it is not so easy to separate the convex hull of two (or more) tree models. In fact, Steel *et al.* (1994) has shown that for a certain general class of mixture models, all tree topologies intersect in a nontrivial manner. Many of the interesting questions concerning estimable models and consistency can be reduced to questions about intersections of the geometry of tree models.

Accuracy, Power, and Complexity of the Model

The accuracy of an estimation method for a particular tree model can also be given a simple geometric interpretation. Fix a tree model by choosing a tree topology and the various character evolution parameters. Therefore, as discussed previously, this is a point in the character simplex. This point induces a sampling distribution on the (rational) points of the simplex. Now, if an estimation method is chosen, the accuracy of the estimation method for this particular tree model is determined by how much of the sampling distribution's probability mass is contained within the method partition of this estimation method for this tree (Fig. 11). That is, let this particular tree topology be called T and let the method partition for T by the estimation method X be called $P_X(T)$. Then, $P_X(T)$ represents the set of data sets that X will estimate as T. Therefore, the accuracy of the method X for this tree model is related to how much of its sampling distribution is in $P_X(T)$. (I use the word "related" rather than "determined" because notions of accuracy can be varied; cf. Kim, 1998.)

The above discussion of accuracy assumed a particular fixed set of parameter values for the character evolution over the tree topology. Normally it would be desirable to discuss these concepts for the entire family

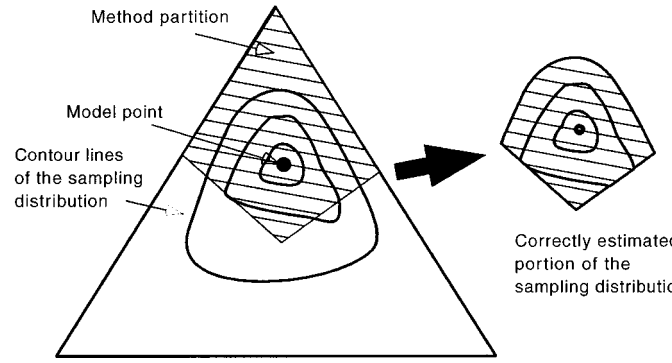

FIG. 11. The accuracy of an estimation method for a particular model (represented by a point) is proportional to the probability mass of the sampling distribution that falls inside of the estimation method partition for the tree topology of the model.

of character evolution models over the tree topology and speak of, say, the accuracy for the entire tree topology. It is obvious this cannot be done (at least using a probabilistic language) unless a distribution assumption over the family of character evolution models is imposed. The difficulty with choosing appropriate distributions has been discussed extensively in the literature (Hillis, 1995; Huelsenbeck and Hillis, 1993; Kim, 1998; Strimmer and von Haeseler, 1996), especially, with respect to how it can bias the assessment of the performance of a given estimation method. However, it may still be desirable to roughly answer questions like how the complexity of the character evolution model relates to the performance of the estimation methods. (In the following, I use loose language for an intuitive picture.)

I first note that regardless of the estimation method, estimation would be easier if the sampling distributions of the data sets have "minimal overlap" under alternative tree models. That is, suppose there were only two possible tree models, say A and B and two possible data sets X and Y. Suppose under A, $\text{Prob}_A(X) = 1$ and $\text{Prob}_A(Y) = 0$; similarly under B, $\text{Prob}_B(X) = 0$ and $\text{Prob}_B(Y) = 1$. Under this scenario we can imagine being able to construct a well-performing estimation method, namely, the one that estimates A for the first of the data sets and estimates B for the second of the data sets. On the other hand, if all data sets are equally probable by either of the tree models, no method could be expected to perform well. Now fix a tree topology and suppose that we have two character evolution models, "Simple" and "Complex," where "Simple" is a geometric subset of "Complex." Then the sampling distribution of the data sets for this tree topology is obtained by integrating the conditional sampling distribution over the possible parameter set for each model. Except for pathological cases, the sampling distribution of a more complex model will have a higher variance than that of a simpler model. This

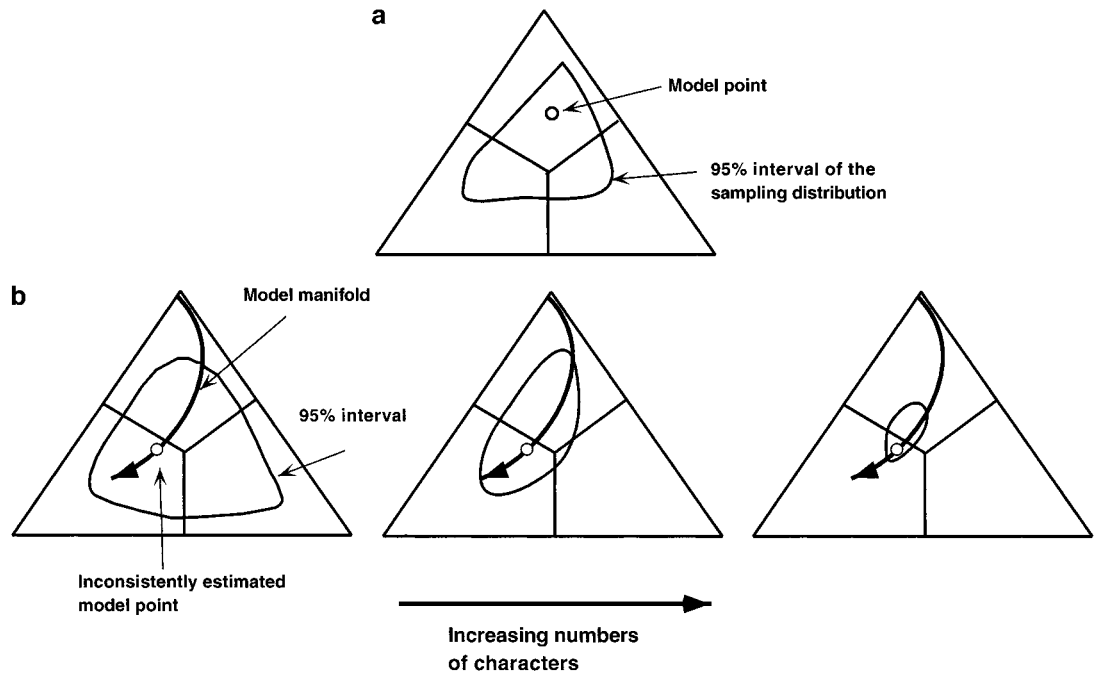

FIG. 12. A hypothetical scenario for the behavior of an estimation method with increasing amounts of data. For a given data set size, the expected accuracy of the estimation method can be visualized with a confidence interval for the sampling distribution from the model point (e.g., 95% interval). The confidence interval is drawn as a contour around the model point (a). As the data set size increases, the contour line contracts around the model point (b). Even if the model point is such that it will be inconsistently estimated, accuracy can increase for a while because the relative amount of the contour "captured" by the method partition can increase for a while depending on the "shape" of the contour as in contracts.

gives a geometrical interpretation of why one class of models can be harder to estimate than another class of models; if one model class is more "space filling" than another class, it can "smear" the sampling distribution more severely. This intuition holds well when one model is a subset of another model. It is harder to make similar statements precisely when we do not have such subset relationships.

Consistency and efficiency are large-sample properties statistical estimators (Kim, 1998). More often the behavior of estimators at finite sample sizes is the more interesting property. The geometric insight can be also used to explore questions of how the estimators behave with respect to increasing amounts of characters within a finite range. As usual, extreme cases are easier to explore and I use one pathological case to demonstrate geometric reasoning. Again, recall that a tree model of character evolution induces a sampling distribution over the simplex. As the sample size (numbers of characters) increases, the sampling distribution becomes increasingly "dense" around the model point, eventually becoming a point mass at the model point. Suppose we were to draw a 95% confidence interval of the sampling distribution around the model point (in a suitable manner, e.g., using the minimum volume). This can be visualized as a kind of an amoeba-like outline surrounding the model point (Fig. 12a). As the sample size increases, this amoeba will contract eventually to a point (see Fig. 1). The manner of this contraction will determine the finite sample behavior of the accuracy of the estimators. Suppose we have a tree model that is inconsistently estimated by an estimation method, say maximum parsimony. This implies that if we have infinite amounts of data, we will converge on the wrong tree; i.e., the probability of estimating the correct tree will go to zero. But, does this mean that it will monotonically go to zero? Doodling with pictures, it can be imagined that if the "amoeba" contracts in the manner shown in Fig. 12b, the corresponding probability of the correct tree can increase for a finite range of values before decreasing. Figure 13 shows an example of a six-taxon tree. By algebraic calculations this tree is definitely in the "Felsenstein's zone" and the estimate will converge to the wrong tree (shown on the right). However, for finite numbers of characters, the probability of obtaining the correct estimate goes up with increasing numbers of character for a finite period (it will eventually start decreasing toward 0). (The figure shows two plots with the probability of correct estimate computed including multiple most-parsimonious tree as correct and computed as incorrect.) It can also be imagined that the converse situation might happen. A consistently estimated tree

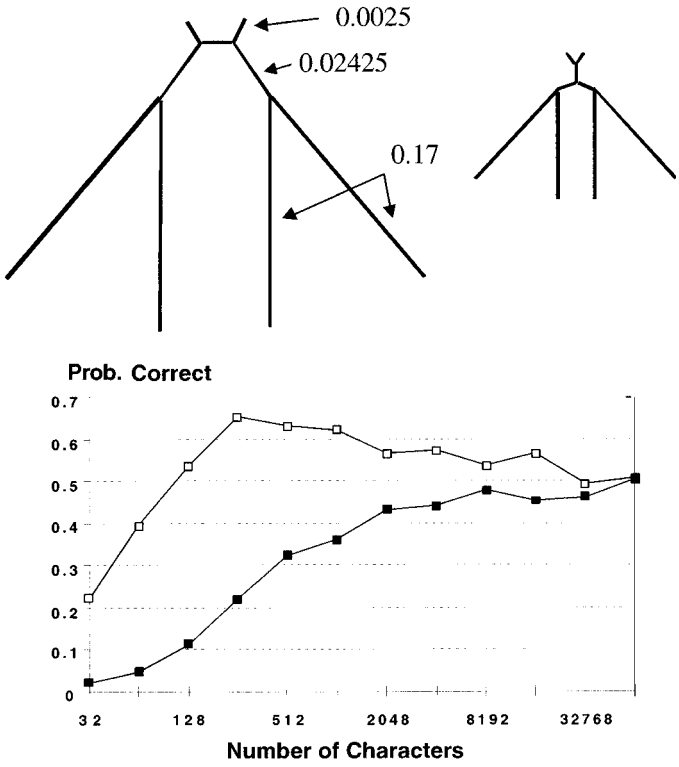

FIG. 13. A pathological example where accuracy increases for a while with increasing numbers of characters even for a model tree that is inconsistently estimated with the maximum-parsimony method. The model tree is shown on the top left. With infinite amounts of characters, the tree shown on the top right is a shorter tree. The graph on the bottom shows the probability of obtaining the correct estimate as a function of the numbers of characters. The open boxes represent the results when we allow multiple most-parsimonious trees to count as correct. The closed boxes represent the results when only a single most-parsimonious tree is considered correct. These lines will eventually decrease down to zero but in this pathological example accuracy is still increasing for ~64,000 characters.

model might be estimated with decreasing accuracy for a finite range of sample sizes.

Lack of Power in Phylogenetic Invariants

As the final example of geometric reasoning, I will examine the problem of power in using phylogenetic invariants for tree estimation. Cavender (1978) first proposed phylogenetic invariants as a method for establishing confidence intervals for tree estimates. Lake (1987b) derived a set of linear invariants for the Kimura two-parameter model (no linear invariants with phylogenetic information exists for a Jukes-Cantor type of model) and suggested estimating trees by asking whether the expected invariant quantities are indeed invariant when computed with the observed data. Subsequent studies have noted that Lake's linear invariants result in consistent but very weak estimates (Huelsenbeck and Hillis, 1993). An easy geometrical reason can be found as to why, in fact, little power should be expected with the linear invariants.

I noted above that phylogenetic invariants are the implicit function form of the parametric tree model functions. The set of points comprising the roots of the phylogenetic invariants contains the model manifold. Often, describing a geometrical object as the root set of equations requires more than one equation. For example, suppose we were to describe the two points $\{1/\sqrt{2}, 1/\sqrt{2}\}$ and $\{-1/\sqrt{2}, -1/\sqrt{2}\}$ as the roots of equations in two dimensions. Then one possibility is to write this as the simultaneous roots of

$$\begin{cases} x^2 + y^2 - 1 = 0 \\ x - y = 0 \end{cases}. \qquad (***)$$

That is, the two points are described as the intersections of a unit circle and the line $y = x$. Suppose that we are given a point, say $\{p, q\}$, and we want to know whether this point is one of the above two points. We can insert the test point, $\{p, q\}$, first into the unit circle equation and ask whether $p^2 + q^2 - 1$ equals zero. If it does, then we can insert it into the next equation and ask whether $p - q$ equals zero. Suppose that $p^2 + q^2 - 1$ equals zero but $p - q$ is not zero. Then we know that the point $\{p, q\}$ lies somewhere in the unit circle but is not one of the original two points we were interested in. The purpose of this example is to show that when we are interested in a geometric object described as the simultaneous roots of many equations, satisfaction of one (or a subset) of the equations is insufficient to "nail it down" to the actual object. This is the situation in using Lake's linear invariants. The actual tree model manifold is contained in the intersection of the roots of several different equations. Lake's invariants are only one subset of such equations. Therefore, just like this toy example (***), satisfying the linear invariants is like knowing that the unit circle equation has been satisfied but not knowing where in the circle the actual points lie. To use phylogenetic invariants as an estimation method, we need to have on hand all the invariants, not just a subset of the invariants. Otherwise, an incomplete description of the model tree results and we can well expect the lack of power in our estimates.

Finally, consider the following two different set of equations:

$$\begin{cases} f(x) = 0 \\ g(x) = 0 \end{cases} \qquad (16)$$

$$f(x)g(x) = 0. \qquad (17)$$

Equation (16) describes a geometrical object that is the intersection of the set of points that make $f(x)$ go to zero and the set of points that make $g(x)$ go to zero. On the other hand, Eq. (17) goes to zero if either $f(x)$ goes to zero or if $g(x)$ goes to zero. That is, the geometrical

object described by (17) is the union of the geometrical objects represented by $f(x) = 0$ and $g(x) = 0$. Suppose we have a four-taxon tree with three possible tree topologies. Also suppose that $f(x) = 0$, $g(x) = 0$, and $h(x) = 0$ are the phylogenetic invariants for each of the trees respectively. (Of course, several equations are required for each tree topology but I am simplifying here for notational convenience.) Then the equation $f(x)g(x)h(x) = 0$ describes the model manifolds for all three tree topologies simultaneously. In principle this construction can be extended to any number of tree topologies. While this still does not allow smooth parameterization of the tree topologies (because of the singular points at the boundaries) it shows that we do not necessarily have to treat different tree topologies separately as discrete objects unlike the treatment given in Yang *et al.* (Yang *et al.*, 1995). (For example, the equation form $f(x)g(x)h(x) = 0$ can be used in a Lagrange equation approach to maximizing the maximum-likelihood function with the assumption that the solution does not lie inside the singular points. This assumption can also be relaxed if we settle for an approximate (but very close) maximum.)

Summary

In this paper, I showed how variously different properties and concepts found in phylogenetic estimation can be put into a common geometric framework. This geometric viewpoint allows a better understanding of various complex relations that arise in phylogenetic estimation. In addition, it gives precise form to one estimation method, the phylogenetic invariants, such that we can adopt existing techniques from mathematical geometry to further develop this method. I gave several examples of how properties like consistency, accuracy, power, and data mixtures can be given a geometric interpretation. Many other cases can be stated in these kinds of geometric pictures. One important problem, which relates to the current questions about large-scale phylogenies (see Kim, 1998), is how the geometry of phylogenetic estimation for one number of species relates to another number of species. It is clear that the number of dimensions increases as a product, say from $2^4 = 16$ for a two-state four-taxon trees to $2^5 = 32$ for a two-state five-taxon trees, suggesting a tensor product space structure (this can also be seen from the basis vectors). However, many details must be worked out and it will be pursued in the future.

Appendix

Assume that we have a binary unrooted tree with n terminal vertices (leaves of the tree) and $n - 2$ internal vertices. Also assume that we have a k by k transition matrix (for k-state characters) attached to each branch (edge) of the tree and we label the elements of the transition matrices as α_{jk}^{i} for the ith branch and the transition between jth and kth state. Let $\mathbf{p} = \boldsymbol{\phi}(\boldsymbol{\alpha})$ be the vector valued function of character pattern probabilities and $\boldsymbol{\alpha}$ is the vector of parameters which are the elements of the transition matrices. For notational convenience relabel the elements of the transition matrix from 1 to z ($= k(k-1)(2n-3)$) If we look at a row of $\mathbf{p}$, it has the form,

$$p_i = \underbrace{\underbrace{\alpha_1\alpha_2\cdots\alpha_k}_{2n-3 \text{ terms}} + \alpha_{k+1}\alpha_{k+2}\cdots\alpha_{k+l}\cdots + \alpha_m\alpha_{m+1}\cdots\alpha_z}_{k^{n-2} \text{ terms}}.$$

There are $2n - 3$ terms within a product corresponding to a particular state assignment of the vertices. I will call a particular set of state assignments to the vertices a Markov path. There are k^{n-2} terms in the summation corresponding to the number of possible state assignments at the internal vertices that lead to the particular character pattern at the terminal vertices. That is, there are k^{n-2} possible Markov paths that lead to a particular character pattern.

The Jacobian matrix of $\mathbf{p} = \boldsymbol{\phi}(\boldsymbol{\alpha})$ is then

$$\mathbf{J} = \left[\frac{\partial}{\partial\alpha_1}\vec{\phi}, \ldots, \frac{\partial}{\partial\alpha_z}\vec{\phi}\right].$$

I want to now show that the rank of $\mathbf{J}$ is z, the number of parameters of the transition matrices. The rows of $\mathbf{p}$ are all sums of monomials of the form $\alpha_1\alpha_2\ldots\alpha_m$ and each row has a unique combination of k^{n-2} terms with $(n-3)k(k-1) + n(k-1)$ different. The column vectors of $\mathbf{J}$ are partial derivatives with respect to each α_i from 1 to z. Take the first row of $\mathbf{p}$. Then if the monomials in the first row of $\mathbf{p}$ contain α_i, the ith column of $\mathbf{J}$ is again a sum of monomials, otherwise it is zero. Suppose that rank of $\mathbf{J}$ is less than z. Then we have

$$c_1\mathbf{j}_1 + c_2\mathbf{j}_2 + \cdots + c_z\mathbf{j}_z = 0 \text{ (where } \mathbf{j}_I \text{ are the column vectors of } \mathbf{J}),$$

for some set of c_I not all zero. Without loss of generality suppose $\{\mathbf{j}_1, \mathbf{j}_2 \ldots \mathbf{j}_p\}$ are the subset of column vectors of $\mathbf{J}$ with a monomial sum in the first row. Since the sum of different monomials are linearly independent the coefficients c_I in front of these vectors must be zero and only those columns of $\mathbf{J}$ with zero in the first row can have non-zero coefficients. But, the columns with zeros in the first row must also have monomial sums in some of the rows and by same reasoning the coefficients for those columns must also be zero. Therefore, $c_I = 0$ for all i, and the rank of $\mathbf{J}$ is z.

ACKNOWLEDGMENTS

I thank Michael Sanderson, Alan de Queiroz, and Gunter Wagner for their continued support and criticism of this work. Two anonymous reviewers helped make this a better paper. This work was initiated 7 years ago under Margaret Kidwell at University of Arizona inspired by discussions in the Phylogenetic Discussion Group. This work was supported in part by NSF Grant DEB-9806570.

REFERENCES

Cavender, J. A. (1978). Taxonomy with confidence. *Math. Biosci.* **40:** 271–280.

Cavender, J. A., and Felsenstein, J. (1987). Invariants of phylogenies in a simple case with discrete states. *J. Classif.* **4:** 57–71.

Chang, J. T. (1996). Full reconstruction of markov models on evolutionary trees: Identifiability and consistency. *Math. Biosci.* **137:** 51–73.

Cox, D., Little, J., and O'Shea, D. (1992). "Ideals, Varieties, and Algorithms," Springer-Verlag, New York.

De Queiroz, A., Donoghue, M. J., and Kim, J. (1995). Separate versus combined analysis of phylogenetic evidence. *Annu. Rev. Ecol. Syst.* **26:** 657–681.

Efron, B., Halloran, E., and Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci.* **93:** 7085–7090.

Felsenstein, J. (1978). Cases in which parsimony and compatibility methods will be positively misleading. *Syst. Zool.* **27:** 401–410.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17:** 368–376.

Felsenstein, J. (1991). Counting phylogenetic invariants in some simple cases. *J. Theor. Biol.* **152:** 357–376.

Fu, Y.-X., and Li, W. H. (1992). Necessary and sufficient conditions for the existence of linear invariants in phylogenetic inference. *Math. Biosci.* **108:** 203–218.

Goldman, N. (1990). Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process model of DNA substitution and to parsimony analyses. *Syst. Zool.* **39:** 345–361.

Hagedorn, T. (1999). Determining the number and structure of phylogenetic invariants. *Adv. Appl. Math.*, in press.

Hagedorn, T., and Landweber, L. F. (1999). Phylogenetic invariants and geometry. *J. Theor. Biol.*, in press.

Hasegawa, M., Kishino, H., and Saitou, N. (1991). On the maximum likelihood method in molecular phylogenetics. *J. Mol. Evol.* **32:** 443–445.

Hendy, M. D., and Penny, D. (1989). A framework for the quantitative study of evolutionary trees. *Syst. Zool.* **38.**

Hillis, D. M. (1995). Approaches for assessing phylogenetic accuracy. *Syst. Biol.* **44:** 3–16.

Hogg, R. V., and Craig, A. T. (1978). "Introduction to Mathematical Statistics," Macmillan, New York.

Huelsenbeck, J. P. (1995). Performance of phylogenetic methods in simulation. *Syst. Biol.* **44:** 17–48.

Huelsenbeck, J. P., and Hillis, D. M. (1993). Success of phylogenetic methods in the four-taxon case. *Syst. Biol.* **42:** 247–264.

Huelsenbeck, J. P., and Kirkpatrick, M. (1996). Do phylogenetic methods produce trees with biased shapes? *Evolution* **50:** 1418–1424.

Jukes, T. H., and Cantor, C. R. (1969). Evolution of protein molecules. In "Mammalian Protein Metabolism" (H. N. Munro, Ed.), Vol. 3, pp. 21–132, Academic Press, New York.

Kim, J. (1996). General inconsistency conditions for maximum parsimony: Effects of branch lengths and increasing numbers of taxa. *Syst. Biol.* **45:** 363–374.

Kim, J. (1998). Large scale phylogenies and measuring the performance of phylogenetic estimators. *Syst. Biol.* **47:** 43–60.

Kimura, M. (1981). Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. USA* **78:** 454–458.

Lake, J. A. (1987a). Determining evolutionary distances from highly diverged nucleic acid sequences: Operator metrics. *J. Mol. Evol.* **26:** 59–73.

Lake, J. A. (1987b). A rate-independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. *Mol. Biol. Evol.* **4:** 167–191.

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

Maddison, W. P. (1989). Reconstructing character evolution on polytomous cladograms. *Cladistics* **5:** 365–377.

Navidi, W. C., Churchill, G. A., and Haeseler, A. V. (1991). Methods for inferring phylogenies from nucleic acid sequence data using maximum likelihood and linear invariants. *Mol. Biol. Evol.* **8:** 128–143.

Nguyen, T., and Speed, T. (1992). A derivation of all linear invariants for a nonbalanced transversion model. *J. Mol. Evol.* **35:** 60–76.

Penny, D., Lockhart, P. J., Steel, M. A., and Hendy, M. D. (1994). The role of models in reconstructing evolutionary trees. In "Models in Phylogeny" (D. Siebert, Ed.), Oxford Univ. Press, London.

Sankoff, D. (1990). Designer invariants for large phylogenies. *Mol. Biol. Evol.* **7:** 255–269.

Steel, M. A., Szkely, L. A., Erdös, P., and Waddell, P. J. (1993). A complete family of phylogenetic invariants for any number of taxa. *N. Zeal. J. Bot.* **31:** 289–296.

Steel, M. A., Szekely, L. A., and Hendy, M. D. (1994). Reconstructing trees when sequence sites evolve at variable rates. *J. Comp. Biol.* **1:** 153–163.

Strimmer, K., and von Haeseler, A. (1996). Accuracy of neighbor joining for n-taxon trees. *Syst. Biol.* **45:** 516–532.

Swofford, D. L., and Maddison, W. P. (1992). Parsimony, character-state reconstructions, and evolutionary inferences. *In* "Systematics, Historical Ecology, and North American Freshwater Fishes" (R. L. Mayden, Ed.), pp. 186–223, Stanford Univ. Press, Stanford, CA.

Swofford, D. L., Olsen, G. J., Waddell, P. J., and Hillis, D. M. (1996). Phylogenetic inference. *In* "Molecular Systematics" (D. M. Hillis, C. Moritz, and B. K. Mable, Eds.), pp. 407–514, Sinauer, Sunderland, MA.

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., Goldman, N., and Friday, A. (1994). Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol.* **11:** 316–324.

Yang, Z., Goldman, N., and Friday, A. (1995). Maximum likelihood trees from DNA sequences: A peculiar statistical estimation problem. *Syst. Biol.* **44:** 384–399.