Syst. Biol. 55(4):652–661, 2006
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150600889617

A Geometric Approach to Tree Shape Statistics

FREDERICK A. MATSEN

Program for Evolutionary Dynamics and the Department of Mathematics, Harvard University, One Brattle Square, 6th Floor, Cambridge, Massachusetts 02138, USA; E-mail: matsen@math.harvard.edu

Abstract.—This article presents a new way to quantify the descriptive ability of tree shape statistics. Where before, tree shape statistics were chosen by their ability to distinguish between macroevolutionary models, the *resolution* presented in this paper quantifies the ability of a statistic to differentiate between similar and different trees. This is termed the *geometric* approach to differentiate it from the model-based approach previously explored. A distinct advantage of this perspective is that it allows evaluation of multiple tree shape statistics describing different aspects of tree shape. After developing the methodology, it is applied here to make specific recommendations for a suite of three statistics that may prove useful in applications. The article ends with an application of the statistics to clarify the impact of taxa omission on tree shape. [Macroevolutionary models; multidimensional scaling; nearest neighbor interchange metric; phylogenetic tree shape; tree estimation bias.]

The analysis of phylogenetic tree shape provides one way of understanding the forces guiding macroevolution and the biases of tree reconstruction methodology. Although it has been a subject of study for many years, a recent editorial in this journal (Simon and Page, 2005) hints that finding the forces guiding tree shape is a long-term challenge which still has not been met. Joe Felsenstein (2004) concludes the chapter on tree shape methodology in his recent book with the simple phrase "[c]learly this literature is in its early days." Indeed, tree shape is still a challenge, and an important one. A complete understanding would help resolve important questions in biology such as the roles of adaptive radiation and environmental change in generating diversity. Tree shape also poses difficult issues of its own, such as the impact of missing or extinct taxa on the understanding of historical biodiversity. Not only are many fundamental questions left unanswered, but the area is ripe for progress: the large number and size of contemporary phylogenies form a fantastic corpus on which macroevolutionary hypotheses can be tested.

In order to use phylogenetic tree shape as a tool, methods are needed to measure and quantify aspects of tree shape. Almost all work to this day has been done with measures of tree "balance," which is the degree to which two sister taxa are of the same or different size. A major vein of research has been to compare the balance of trees created from data to trees produced by one or another null model (Savage, 1983; Guyer and Slowinski, 1991, 1993; Stam, 2002). Kirkpatrick and Slatkin (1993), in one of the early papers in the area, quantified the power of different measures of tree balance in distinguishing between distributions on tree shapes. The two models are extremely simple: one, called the Yule or ERM model, develops a tree by starting with a single species and then choosing uniformly among species to bifurcate. The other, called the PDA model, is simply the distribution on tree shapes induced by the uniform distribution on labeled trees.

Studies have shown that most trees created from data are less balanced than would be expected from the ERM model, yet are more balanced than would be expected from the PDA model (Mooers, 1995; Mooers and Heard, 1997; Purvis and Agapow, 2002). Models of increasing sophistication have appeared, attempting to re-create this observed pattern of tree shape observed in nature. For example, Heard (1996) found that speciation rate variation among lineages can lead to imbalanced trees. Losos and Adler (1995) found that short "refractory periods"—periods before a new species can speciate again—led to more balanced trees, whereas Rogers (1996) found that very long refractory periods led to less balanced trees. Aldous (1995, 2001) was the first to propose a (nonevolutionary) model that interpolated between the ERM and the PDA models. More recently, Steel and McKenzie (2001) and Pinelis (2003) have developed evolutionary models that also interpolate. Another interesting contribution to this area is the "alpha model" of Ford (2005).

With these models, one could presumably arrange parameters to correctly fit the observed pattern of imbalance as reported by a given statistic. But is that really enough? What if other aspects of the tree shape, not measured by the statistic, differ considerably? After all, any single statistic is a one-dimensional summary of a very complex set of data. One might follow the suggestion of Agapow and Purvis (2002) and use two different balance statistics that measure balance in different parts of the tree, but this paper attempts to present a more direct approach.

The only proposal made in the literature that has the potential to encapsulate lots of information about the shape of a tree has been made by Aldous (2001). He suggests first constructing a scatterplot of the interior nodes, where the x coordinate is the size of the subclade subtended by that interior node, and the y coordinate is the size of the smaller daughter clade. The proposal is then to perform nonlinear median regression on the log-log version of this scatter plot and then use the fitted function as a descriptor of tree shape. The log-log scatter plot will be called the "Aldous scatterplot" in this paper.

There are a number of advantages to this approach. It is very natural from a statistical viewpoint relative to the other, more ad hoc, measures of tree balance. The method has the potential to give quite a lot of information about tree shape compared to a single summary statistic.

Finally, it allows comparison of trees of different size by superposition of scatterplots, which is a significant advantage. There is currently no generally accepted method for comparing trees of different size using the standard statistics; this remains a problematic issue (Mooers, 1995; Stam, 2002).

However, there are three disadvantages that may make Aldous' proposal not as practical as might be hoped. The first is that regression works best with many points of data, and thus one can only expect his technique to work with rather large trees. This problem is exacerbated by the fact that isomorphic subtrees are superimposed on one another in the scatterplot, further reducing the number of fittable points. The second is an inherent problem with summarizing a tree as a scatterplot of this sort. Assume that tree T has two nonisomorphic subtrees A and B of the same size. Exchanging A and B in T will not change the scatter plot and thus will not change any regression parameters, although the resulting tree may differ significantly in shape. The third problem is that the resulting output can be hard to interpret. What does, for example, the kth Taylor coefficient of the fitted function actually signify? Despite these issues, this technique seems underutilized and might be the technique of choice when working with large phylogenies.

Overall, it appears that additional methods would be useful for understanding tree shape. This paper attempts to provide some of these new methods.

THE GEOMETRIC APPROACH

The basic philosophy behind the geometric approach is that similar trees should have similar statistics, and that rather different trees should have different statistics. This philosophy is summarized in Figure 1. All of the trees with six tips are evaluated by two hypothetical statistics. The top axis shows what one might consider a good statistic. The maximally balanced trees are on the far left side, and the completely unbalanced tree is on the far right. When a subtree is preserved, the statistic tends not to change too much. The bottom axis shows what might be considered a bad statistic. The extremes of tree balance are now put together, and two similar trees are now on the two extremes of the axis.

If one is to apply this sort of intuition on trees, it is necessary to formalize the notion of similar and different for trees. This is done by defining a metric on unlabeled trees.

A Metric for Evolutionary Histories

This section describes a metric on unlabeled trees that can be applied directly to compare tree shapes or can be used to guide the selection of statistics as described below. For this paper the word "tree" signifies a finite strictly bifurcating rooted tree without leaf labels or specified edge lengths. Finite strictly bifurcating rooted trees have been chosen as they correspond most naturally to the output of current macroevolutionary models. This paper concerns itself with tree shape rather than the identity of taxa; thus, leaf labels are ignored. Finally, the intent of this paper is to understand the combinatorial content of the tree; thus, trees are considered without specified edge lengths. The case including edge lengths would be an interesting future extension of this work but would require a significant further development of the methodology.

A metric g is simply a set of distances between pairs of a collection of objects satisfying (i) $g(x, y) = 0$ if and only if $x = y$, (ii) $g(x, y) = g(y, x)$, (iii) the triangle inequality: $g(x, y) + g(y, z) \geq g(x, z)$. One such metric is the nearest neighbor interchange (NNI) metric on unlabeled trees. A single NNI "move" represents a change of branching order of a tree to one of two possible configurations. The two possible moves are depicted in Figure 2. The unlabeled NNI distance from one tree to another is defined to be the minimum number of moves necessary to change one tree to the other. Note that these interchanges have appeared before in Kuhner et al. (1995) as proposal draws for their Metropolis-Hastings approach to estimating population parameters.

Tree space equipped with the NNI metric is shown in Figure 3 for trees on 6 leaves. It is a graph that has connections between any two trees that are a single NNI move apart. Note that the NNI distance is a special case of the shortest-path metric on a graph and thus it satisfies the above conditions to be a metric. Also, although the metric is not explicitly model based, a change of branching order can be thought of as a change of timing of diversification events.

Unsurprisingly, computing this metric is NP-complete, as can be seen by a small modification of a similar proof by DasGupta et al. (2002). Their paper demonstrates that calculating the unrooted NNI distance on unrooted trees is NP-complete. However, the unrooted NNI moves are identical to the moves in Figure 2 when the tree shown in the diagram is chosen

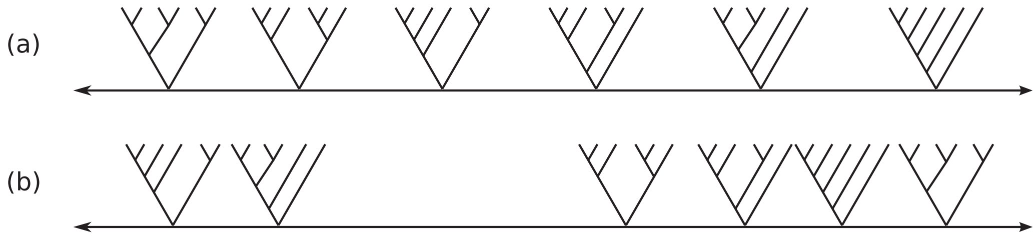

FIGURE 1. Good and bad statistics from the geometric perspective. The horizontal axes represent values of hypothetical statistics. In (a) very different trees are separated, whereas in (b) similar trees are separated and different trees are close together.

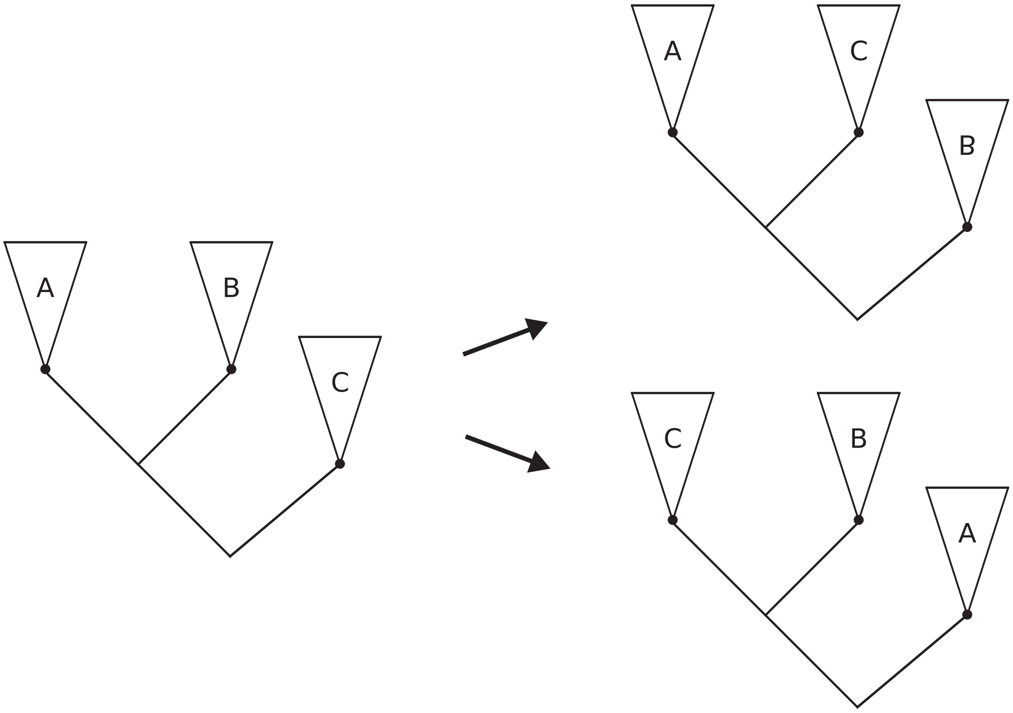

FIGURE 2. A single-rooted NNI move consists of rearranging the tree in either of the two ways shown here. The NNI distance between two trees is the minimum number of moves required to change one tree to another.

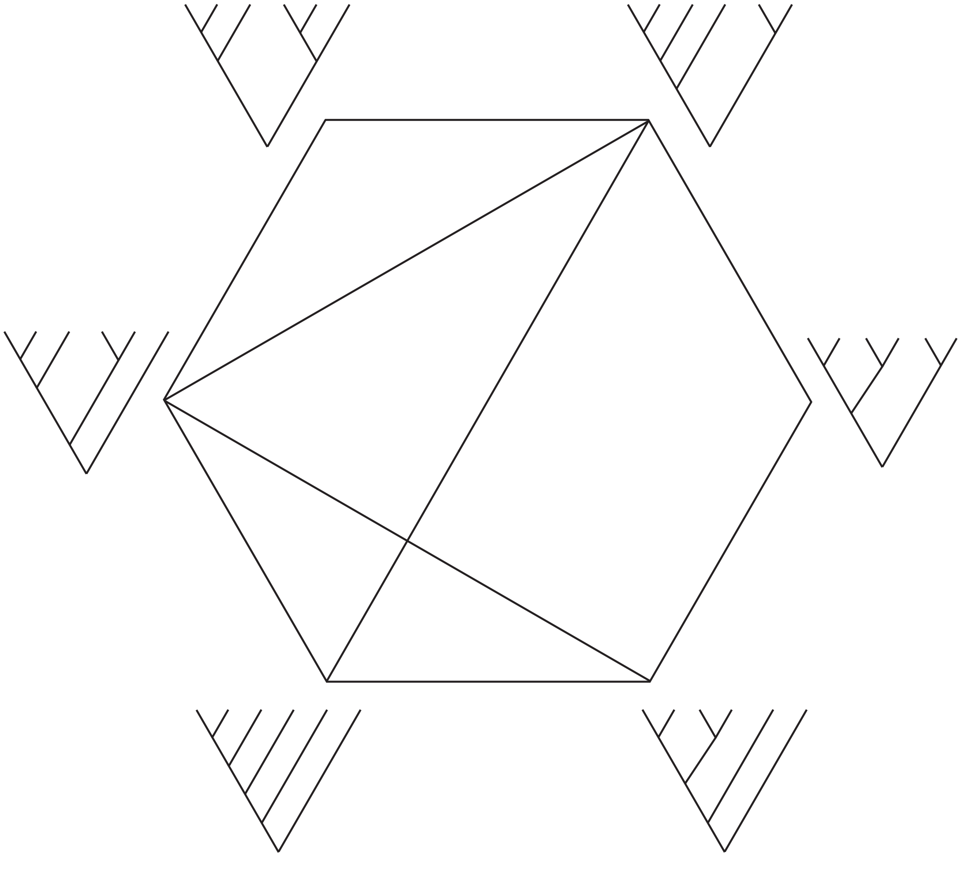

FIGURE 3. Unlabeled tree space equipped with the NNI metric for the trees on six taxa. An edge between two trees means that a single NNI move changes one to the other.

to be anything but the entire tree. Therefore, one can simply root the tree in figure 4 of their paper on the far left side of the main linear tree and the proof proceeds as in their paper.

There are certainly many metrics possible on unlabeled tree space, and NNI is just one choice. For example, an alternative would be the subtree-prune-regraft (SPR) metric, which is similar to the above metric in that it counts the minimal number of moves needed to change one tree to another. However, the SPR moves cut a whole subtree out of the larger tree and then reconnect it in an arbitrary location. Because a single one of these moves can radically alter the shape of a tree, the NNI metric may be more appropriate, which makes smaller changes each step.

For the analysis, tree space was generated for trees of 7 to 15 leaves. The NNI graph was created, and Djikstra's well-known algorithm was used to calculate the shortest paths. There are 10,905 tree shapes of 16 leaves, and though it is certainly possible to perform the analysis below for problems of this size, it was decided to stop with 15. As described later in this paper, the number of tree shapes grows exponentially and the added benefit of another leaf or two did not justify the specialized programming required.

Resolution of Statistics

In this section the notion of the *resolution* of a tree shape statistic with respect to a given metric is defined. The resolution will be the operational definition of performance for tree shape statistics from the geometric perspective.

First fix l, the number of leaves, and enumerate all n trees on l leaves. Let d_{ij} be the distance between trees i and j. Although any metric can be chosen, all of the analysis in this paper will be done with respect to the above NNI metric on tree shapes. Let H be the $n \times n$ "centering matrix"

$$H = I - n^{-1}11'$$

where 1 is the vector with every entry equal to one and $'$ denotes transpose. The application of the centering matrix to a vector subtracts off the average of the entries of the vector from each component. Given a tree shape statistic f, define the vector y_f such that the ith component $(y_f)_i$ is the value of f on the ith tree. Assume that f is not constant on the trees, such that $Hy_f \neq 0$, and define

$$x_f = H\, y_f / \|Hy_f\|.$$

The vector x_f is simply the centered normalized vector of statistics for the n trees. The resolution of the statistic f with respect to a distance matrix $D = (d_{ij})$ is defined as

$$R_D(f) = \frac{1}{2} \sum_{i,j} -d_{ij}^2 (x_f)_i (x_f)_j \qquad (1)$$

This equation formalizes the geometric perspective on tree shape: that a "good" tree shape statistic is one which is similar for similar trees and different for rather different trees. Indeed, an individual term of the sum in (1) will be maximized if $(x_f)_i$ is very negative and if $(x_f)_j$ is very positive or vice versa. The summation and the distances simply combine all of these terms together in a weighted fashion such that ij pairs that are distant carry more weight than ones which are close. Therefore, for a statistic with high resolution, the more distant trees will tend to be farther apart in x-value, and the closer trees will tend to be closer in x-value.

As an example, one can compute the resolution of the tree shape statistics presented in Figure 1. The "good" statistic in Figure 1 has a resolution value of 2.33, whereas the "bad" statistic has a resolution value of −1.23. In this case the upper limit of the resolution is 3.10 and the lower limit is −1.27, which are the eigenvalues of a matrix as described below.

The definition of the resolution is motivated by the statistical method of multidimensional scaling (MDS) (Mardia et al., 1979; Borg and Groenen, 2005). The goal of MDS is to find a set of points $p_1, \ldots, p_n$ in k-dimensional Euclidean space such that the distance between two objects with respect to a metric is well approximated by the Euclidean distance between the corresponding points. Specifically, MDS minimizes the quantity

$$\left[\sum_{i<j} (d_{ij} - |p_i - p_j|)^2 \right]^{1/2}$$

among all collections of n points in k-space.

The MDS methodology is sketched here in order to relate it to the resolution. Let $D = (d_{ij})$ be the pairwise distance matrix, and let D_s represent the component-wise matrix square of D, such that the ijth component of D_s is d_{ij}^2. Define

$$X_D = -\frac{1}{2} H\, D_s\, H$$

Let $v_1, \ldots, v_k$ be the unit-norm eigenvectors corresponding to the k largest eigenvalues $\lambda_1 \geq \cdots \geq \lambda_k > 0$ of the matrix. The coordinates of the above-described optimal points p_i can be calculated using the formula

$$(p_i)_m = \sqrt{\lambda_m} \cdot (v_m)_i.$$

In short, the best approximation for the distance data in one-dimensional space is the first eigenvector, the best approximation in two dimensions is the pair of the first two eigenvectors, and so on. Therefore, from the multidimensional scaling perspective, the best k tree shape statistics are the first k eigenvectors of X_D. This is the approach taken in most applications of multidimensional scaling.

However, we do not have this luxury. Each dimension of X_D, equal to the number of unlabeled trees, is asymptotically of order $b^l l^{-3/2}$ where l is the number of leaves and $b \approx 2.483$ (Harding, 1971; Semple and Steel, 2003). It is not practical to solve eigenproblems of this size and so the exact eigenvectors are not accessible.

The best one can do is to approximate the eigenvectors by things which are efficiently calculable. The Rayleigh Quotient theorem states that the eigenvector corresponding to the largest eigenvalue of a symmetric matrix M maximizes the quadratic form $q_M(x) = x'Mx$ over all unit-norm vectors x (Ortega, 1987). The current set of tree shape statistics can be calculated efficiently, and so the best choice of statistic from the MDS perspective is the one which maximizes the quadratic form associated with X_D.

This quadratic form associated to X_D is basically (1). In its raw form it is

$$q_{X_D}(x) = -\frac{1}{2}\, x' H\, D_s\, H\, x.$$

First simplify by only considering x which are already centered, i.e., such that $Hx = x$. This is not a loss of generality because any optimal x will certainly be centered; recall that x_f is centered by definition. Therefore, the associated quadratic form on centered vectors of tree shape statistics is

$$R_D(f) \;= -\frac{1}{2} x_f' D_s\, x_f, \qquad (2)$$

which is equivalent to (1). By the Rayleigh Quotient theorem, the largest eigenvalue of $-0.5D_s$ is an upper bound for $R_D(f)$ and the smallest eigenvalue is a lower bound for $R_D(f)$. There is a unique linear transformation transforming the resolution between these two bounds to the interval between zero and one; this will be reported in the tables below and called the *scaled resolution*.

To get all the eigenvectors corresponding to positive eigenvalues of a symmetric matrix M, one can apply the following algorithm: first find the unit-norm v_1 maximizing q_M, which is the first eigenvector. Then project v_1 out of the matrix, resulting in $M_1 = (I - v_1 v_1')\; M\; (I - v_1 v_1')$. Then the second eigenvector will be the v_2 maximizing q_{M_1}; project v_2 out of M_1 to create a matrix M_2, and so on.

As described above, it is difficult to follow this recipe exactly because of the large size of the associated matrices. However, one can approximate the process as follows: first, pick the statistic f_1 that maximizes the resolution, and project out x_{f_1} to create a matrix D_1. Then, pick the statistic f_2 that maximizes R_{D_1}, project again, and so on. Note that it is necessary to take an orthonormal basis for the projecting vectors, as they will not be automatically orthogonal as in the eigenvector case. Exactly these steps will be performed in the following section.

One possible objection to the resolution methodology is that the definition of R_D is implicitly tied to the uniform distribution on tree shapes. That is to say, trees that are never seen in models or from data carry equal weight in the resolution measure as trees which are common. In theory, this could decrease the utility of the resolution measure, especially when considering large trees.

It would be possible to incorporate a distribution on the trees, say p_i, using the following modification to R_D:

$$R_{D,p}(f) \;= \frac{1}{2} \sum_{i,j} -d_{ij}^2\, p_i\, p_j\, (x_f)_i (x_f)_j$$

This extension, although intuitively attractive, can lead to biases in the corresponding high-resolution statistics. For example, assume that one chose for a distribution the Aldous (2001) beta-splitting model with $\beta = -1$. This is a reasonable choice for a tree shape distribution, but performing an analysis analagous to that given below with this weighting scheme could be misleading: statistics chosen with this weighting would have diminished power to reject the beta-splitting model. This is simply because trees which are unlikely under this model will factor less into $R_{D,p}$ and therefore will get a more arbitrary assignment in a high-resolution statistic.

Because of the possibility of bias, the weighted extension will not be followed in this paper. It may make sense to add some weighting to the definition when it is perfectly clear that certain tree shapes will never be seen in models or data; however, this is usually not the case. Furthermore, because of the moderate size of the trees used to evaluate the statistics, this uniform-distribution objection has less impact.

Note that this is not the first application of multidimensional scaling ideas to phylogenetic analysis: Hillis et al. (2005) applied it with interesting results to the space of trees with labeled tips. They used MDS with the Robinson-Foulds distance metric as a tool for visualization and analysis of the output of tree reconstruction software. The intent and methods differ here, as this paper concerns itself with finding near-optimal statistics for understanding unlabeled tree space with the NNI metric.

Results

In this section, the methodology of the previous section is applied to compare the resolution of tree shape statistics. First the standard list of statistics will be evaluated (Kirkpatrick and Slatkin, 1993; Agapow and Purvis, 2002; Felsenstein, 2004) according to the above methodology. Then a best second statistic is searched for given the first, and the best third statistic given a first and second. The chosen criterion for performance is high resolution on the whole unlabeled tree space with the NNI metric as described in the previous section. The tree space was generated and evaluated by an `ocaml` (Chailloux et al., 2000) program whose source is available upon request. All of these tree shape statistics can be calculated for any trees in Newick format using the simple command-line software `simmons` available at http://www.math.canterbury.ac.nz/matsen/simmons/.

The "classical" list of tree balance statistics was chosen to be $\bar{N}$ and σ_{N}^{2} proposed by Sackin (1972), I_c proposed by Colless (1982), and B_1 and B_2, proposed by Shao and Sokal (1990). To the list was added a rarely used statistic I_2, invented by Mooers and Heard (1997) to provide a measure that weights all nodes equally. The definition of all of these statistics has been included in Appendix 1. Finally, the proposal of Aldous (2001) was implemented to perform median regression as described in the introduction. Median regression was used to fit a quadratic polynomial to the data and the linear and quadratic coefficients were interpreted as descriptive statistics which will be called A_1 and A_2.

Although Aldous' paper did not explicitly specify how to perform the median regression, nonlinear median regression was chosen as described by Koenker and Bassett (1978). This method minimizes the sum of the distances of the estimated median to the data points. Median regression performs better (as a maximum-likelihood estimator) than least-squares regression when errors are non-Gaussian, as in the present case. It can be easily implemented using linear programming; in this case it was implemented in 34 lines of code using an `ocaml` frontend to the GNU linear programming package GLPK.

The results of this analysis are presented in Table 1. First, the scaled resolution of two statistics, I_c and $\bar{N}$, is rather close to one, which is the upper limit. This is quite remarkable, in that two statistics that were designed "by hand" to measure a visible aspect of tree shape end up having almost as much resolution as theoretically possible. The fact that overall tree balance appears as such an important descriptor justifies in a sense the disproportionate amount of attention given to it in the tree shape literature. Another nice fact is that I_c and $\bar{N}$, the two statistics with the highest resolution, are also the two most powerful according to Agapow and Purvis (2002). In this first setting, I_2 does have substantially lower resolution than the other statistics; however, it performs well in other settings. Finally, it appears that the coefficients of the best-fit quadratic polynomial on the Aldous scatterplot should not be used as a first statistic in the simple-minded way presented here on small trees; it is possible that an alternative formulation would yield better results.

TABLE 1. Scaled resolution scores for tree statistics on the NNI distance matrix. For all tables, the maximum scaled resolution is one and the minimum is zero. The number of leaves of the corresponding trees is denoted by l, and I_c, $\bar{N}$, σ_{N}^{2}, I_2, B_1, B_2, A_1, and A_2 denote tree shape statistics as described in the text and Appendix 1.

l	I_c	$\bar{N}$	σ_{N}^{2}	I_2	B_1	B_2	A_1	A_2
7	0.925	0.930	0.902	0.884	0.864	0.925	0.545	0.548
8	0.925	0.912	0.875	0.861	0.832	0.911	0.429	0.436
9	0.918	0.920	0.882	0.853	0.832	0.906	0.316	0.328
10	0.940	0.938	0.898	0.854	0.832	0.908	0.365	0.383
11	0.953	0.951	0.910	0.855	0.837	0.913	0.382	0.397
12	0.953	0.952	0.908	0.850	0.831	0.904	0.383	0.403
13	0.954	0.954	0.907	0.841	0.824	0.899	0.388	0.410
14	0.955	0.954	0.907	0.837	0.819	0.890	0.388	0.412
15	0.954	0.954	0.904	0.829	0.812	0.882	0.398	0.424

So far it is only clear that choosing for maximal resolution gives results that do not seem completely out of the ordinary. However, now something new is possible. Say that I_c is chosen as the first statistic. What is the best second number to know about a tree given that I_c is already known? This question has a mathematical formulation: simply project out the I_c component of the matrix X_D and repeat the above process.

The resolution scores of the previously chosen statistics are listed in Table 2. Note that I_c has low resolution because it has been projected out, and that $\bar{N}$ has rather small resolution, which is to be expected because it is highly correlated with I_c (Blum et al., 2006; Ford, 2005). Comparatively, I_2, A_1, and A_2 now do better.

However, it is possible to improve on existing statistics by explicitly constructing a statistic that measures a different aspect of tree shape than I_c. Plotting the principal components of the X_D matrix suggests that a good second statistic may be the change of balance from the root to the tips. This intuition is implemented here as the "derived statistics" of a given statistic.

The derived statistics attempt to quantify the change of a statistic through the tree. Start by making a two-dimensional scatterplot of the tree, where each subtree is represented by a point with the x axis being the size of the subtree and y being the value of the statistic Y. Now do median regression on this scatterplot and report the slope of the best-fit line or the quadratic coefficient of the best-fit quadratic polynomial. Given an original statistic Y these two derived statistics will be called Y' and Y'' in analogy to the first and second derivatives of calculus. Higher derived statistics and other fittable functions are of course possible but will not be investigated in this paper. In contrast with the Aldous statistics, note that the regression is done on the points directly, rather than on their image under the logarithm.

Also, recall a statistic that has been understood from the theoretical perspective but is not in common usage in the tree shape literature: the number of "cherries" of a tree. A "cherry" is simply a subtree of two leaves. McKenzie and Steel (2000) have shown that the distribution of the number of cherries is asymptotically normal under both the equal rates Markov and the uniform model (see next section) and have derived the mean

TABLE 2. Scaled resolution scores for tree statistics on the NNI distance matrix after projecting out I_c.

l	I_c	$\bar{N}$	σ_{N}^{2}	I_2	B_1	B_2	A_1	A_2
7	0.267	0.283	0.270	0.467	0.362	0.324	0.551	0.557
8	0.252	0.267	0.262	0.484	0.377	0.309	0.511	0.513
9	0.203	0.219	0.212	0.479	0.352	0.298	0.340	0.360
10	0.171	0.185	0.185	0.489	0.355	0.264	0.409	0.424
11	0.154	0.167	0.172	0.503	0.360	0.268	0.425	0.436
12	0.135	0.147	0.157	0.511	0.367	0.252	0.422	0.430
13	0.126	0.136	0.152	0.524	0.377	0.258	0.432	0.441
14	0.117	0.127	0.148	0.535	0.389	0.257	0.427	0.431
15	0.110	0.119	0.145	0.545	0.398	0.262	0.436	0.436

TABLE 3. Scaled resolution scores for tree statistics on the NNI distance matrix after projecting out I_c. A prime (′) or two primes (″) denote the first or second derived statistics, respectively. "Cherries" denotes the cherries statistic as described in the text.

l	Cherries	I'_c	I'_2	B''_1	B''_2	A_1	A_2
7	0.448	0.641	0.576	0.541	0.532	0.551	0.557
8	0.486	0.694	0.683	0.433	0.529	0.511	0.513
9	0.485	0.723	0.720	0.507	0.491	0.340	0.360
10	0.507	0.635	0.636	0.480	0.519	0.409	0.424
11	0.530	0.622	0.618	0.469	0.563	0.425	0.436
12	0.548	0.580	0.571	0.454	0.548	0.422	0.430
13	0.569	0.589	0.587	0.434	0.563	0.432	0.441
14	0.587	0.577	0.562	0.423	0.572	0.427	0.431
15	0.602	0.560	0.544	0.417	0.576	0.436	0.436

and variance for each. The cherries statistic was not included in Table 2 because it does not measure the overall balance of a tree.

Table 3 presents the results of the resolution method as applied to the distance matrix after I_c has been projected out. The best performance is achieved by the number of cherries, perhaps the simplest possible statistic. Although the performance of the cherry statistic lags behind the above statistics as a first statistic, it has remarkably good performance as a second statistic. Similar performance is achieved by the slightly more complex I'_c. The values of B''_1 and B''_2 were also reported due to their good performance.

Now assume the number of cherries is chosen for the second statistic and we look for a third. As before, project I_c and the number of cherries out of the matrix and compare scores. This time it is B''_2 and I''_c that perform the best.

In the end, what is the best general-purpose suite of statistics to use for tree shape description? For a first statistic, the answer is probably I_c or $\bar{N}$. They have high resolution and are simple to compute. For a second statistic, the number of cherries and I'_c also have good resolution and relatively simple interpretations. For a third statistic, the statistic with the highest resolution is B''_2; however, another recommendation would be the triple (I_c, I'_c, I''_c), which has satisfactory resolution and a somewhat intuitive interpretation.

Example Application

In the introductory section, it was stated that "interpolating" evolutionary models could be used to fit any given pattern of overall imbalance. It was argued that this fact motivates the use of multiple tree shape statistics, as a single statistic may be insufficient to distinguish between trees generated by the original evolutionary model and a fitted one. In this section it is demonstrated that overall balance statistics such as I_c have almost no statistical power to differentiate between the distributions given by two such models, and then it is found which statistics do. The satisfying conclusion is that statistics that have high resolution after projecting out I_c appear to be good at distinguishing between the original and fitted distributions.

The interpolating model chosen for this example application is Aldous' *beta-splitting* model (Aldous, 1995, 2001). It is a simple model with a single parameter, β, which allows interpolation between the maximally imbalanced tree ($\beta = -2$) and the maximally balanced tree ($\beta = \infty$). The "equal rates Markov" or ERM tree (i.e., the coalescent tree distribution) emerges when $\beta = 0$. The "proportional to different arrangements" or PDA tree (i.e., the distribution on tree shapes induced by a uniform distribution on labeled trees) appears when $\beta = -1.5$.

The idea of this model is to recursively split the tips into two subclades using the beta distribution. More precisely, assuming that a clade has l taxa, the probability of the split being between subclades of size i and $l - i$ is

$$q_{l,\beta}(i) = C(l;\beta)\,\frac{\Gamma(\beta+i+1)\Gamma(\beta+l-i+1)}{\Gamma(i+1)\Gamma(l-i+1)}$$

where $C(l;\beta)$ is a normalizing constant. This distribution is equivalent to scattering the taxa on the unit interval and then splitting with the $B(\beta+1, \beta+1)$ distribution (Aldous, 1995).

This model is easily adapted to a maximum-likelihood framework. The likelihood of each tree for a given β is the product of the likelihoods of each split. The likelihood of a collection of trees was chosen to be the product of the likelihoods of each tree. With a trick from Aldous (1995) one can derive a formula for the $C(l;\beta)$ and then find a β that maximizes the log-likelihood of a collection of trees in the standard way.

As an application of the above statistics the effect of missing taxa on phylogenetic tree shape will be investigated using simulation. The pattern of taxon deletion is chosen to model the effect on tree shape of a sequencing strategy which is common in the realm of infectious disease: sequence only those strains which are significantly different from previously sequenced strains. Assume that the original tree emerged from an evolutionary process which has the ERM distribution on trees. Furthermore, assume that the edge lengths are distributed according to a $N(1, .25)$ Gaussian distribution truncated below zero. Given such a tree with l leaves, recursively delete k taxa in the following manner: find the pair of taxa that are closest together in terms of tree distance (including edge length) and randomly delete one of them. After deletion, perform a maximum-likelihood fit as described above on those trees, resulting in a β, and then generate a sample of beta-splitting trees on $l - k$ leaves using this β. Which statistics can distinguish between the original trees and the fitted trees?

This simulation study was performed with a sample size of 500, $l = 100$, and $k = 10$. The β value fitted to the described deletion process was -1.03, corresponding to a decrease in balance from the $\beta = 0$ original tree. Statistics were then compared between 500 of the "fitted" beta-splitting trees and the 500 original trees with deleted taxa. The trees were then evaluated with

TABLE 4. Scaled resolution scores for tree statistics on the NNI distance matrix after projecting out I_c and the number of cherries.

l	B_2''	I_c''	I_2''	A_1	A_2
7	0.747	0.531	0.423	0.662	0.689
8	0.669	0.617	0.536	0.620	0.634
9	0.545	0.525	0.400	0.376	0.396
10	0.556	0.507	0.382	0.449	0.461
11	0.594	0.506	0.392	0.458	0.462
12	0.590	0.533	0.433	0.459	0.457
13	0.602	0.543	0.420	0.450	0.444
14	0.608	0.542	0.419	0.437	0.427
15	0.611	0.550	0.425	0.437	0.424

the two-tailed Wilcoxson rank sum test to find statistical power of each statistic to differentiate between the two distributions. The results of this analysis are in Table 5.

As stated above, the statistical power for this scenario corresponds with the resolution of these statistics when I_c has been projected out. This makes sense because when a tree is fit to the beta-splitting model, the overall balance of the trees would be a primary determining factor. Recall that the three statistics with highest resolution after projection of I_c were the number of cherries, I_c', and I_2. These three statistics are also the most powerful for the example application. The statistics A_1 and A_2 were also included in Table 5 because they performed reasonably well; this corresponds with their good resolution after projecting out I_c as shown in Table 3. It is also not surprising that these statistics perform better on relatively large trees. Finally, as might be expected for a situation in which the overall balance of a tree has been fitted to the model, the statistic I_c has essentially no power to distinguish between the two models.

This simple simulation exercise further demonstrates that the resolution measure can help guide the selection of good general-purpose tree shape statistics. Although these statistics were chosen on purely geometric grounds, they were also the most powerful for this somewhat arbitrary model.

EXTENSIONS

There are a number of limitations to this methodology that point the way for future development. The first is that this application of the MDS technique was to a specific model of tree space, namely that with the unlabeled NNI distance. It is possible that this is not a good choice. However, the general framework presented here is not tied to the NNI metric, thus other models may be used in the future if desired. Another angle on this issue is the fact that the resolution is implicitly tied to the uniform distribution on tree shapes. As mentioned above, a nonuniform distribution could be accomodated but may lead to undesirable biases. Nevertheless, a careful examination of resolution in the nonuniform case could lead to interesting results.

TABLE 5. Comparison of the scores for various statistics when applied to trees from two different models. "NM" signifies the median of the statistic when applied to a sample of ERM trees of size 90; "DM" signifies the median when applied to a sample of beta-splitting trees with leaves deleted as described in the text. The last line shows the P-value for the two-sided Wilcoxson rank-sum test.

Distance	I_c	Cherries	I_2	I_c'	A_1	A_2
NM	0.077	30	0.47	0.015	0.62	0.056
DM	0.076	29	0.49	0.019	0.51	0.089
P	0.16	7.6×10^{-32}	5.1×10^{-13}	4.6×10^{-7}	4.4×10^{-6}	1.1×10^{-6}

Second, this methodology offers nothing to the debate of how to compare the shape of trees of different size. This is a very fundamental problem, which may be more philosophical than technical: what does it actually mean to say that a tree of one size has a similar shape to one of a different size? A common response in the literature (Mooers, 1995; Stam, 2002) is to compare in one way or another the shape of a given tree to a sample of trees from a fixed distribution; knowing the distribution of the statistic as for the number of cherries (McKenzie and Steel, 2000) makes this an attractive option for some statistics. However, if descriptive theory independent of perhaps over-simple models is desired, some other method will have to be found. This is clearly an interesting avenue for future research.

Third, because as mentioned above the number of unlabeled binary trees is exponential in the number of leaves, this analysis is limited to moderately small trees. This may skew the analysis in that statistics that perform poorly for small trees may perform quite well for large trees; an example case might be Aldous' descriptors of tree shape. One response to this objection is that the results show a certain level of stability as l increases: statistics that are good for smaller l appear to be good for larger l as well. Furthermore, although increasingly large trees are now available, the analysis of trees of intermediate size is still a challenge and at worst the above methodology is applicable to that case. However, this is a problem for future research, and new methods may solve this problem.

Fourth, edge length information is conspicuously absent in tree shape analysis. Typically, information about timing of speciation (or other branching) events is analyzed in a completely different manner, as a lineages-through-time plot, which is then used to estimate speciation and extinction rates with maximum likelihood (Nee et al., 1994). Any analysis of this sort eliminates topological information that may aid in choosing an evolutionary model. The tree shape literature has already shown that the standard birth-death process where each leaf is equally likely to split or be eliminated does not construct trees that seem to reflect the imbalance seen in nature; nevertheless, this assumption is implicit in lineages-through-time analysis. More work is needed to integrate the tree shape and timing literature.

Fifth, the statistics that are examined here are for the most part ad hoc. With the exception of Aldous' statistics, they are designed to describe a visible aspect of tree shape. A less ad hoc algebraic approach has been examined in Matsen and Evans (2006). Furthermore, it is possible to optimize over algebraic expressions that give tree shape statistics in a natural way. The author has implemented a genetic algorithm to do so; this work will be described in a forthcoming paper (Matsen, 2006). There is certainly room for improvement: with the notable exception of the first tree shape statistic, the statistics described in this paper have substantially less resolution than the upper bound (see Tables 2 and 4).

Finally, there is a limitation that is fundamental to any discussion of trees: with very few exceptions, trees are not actual data. They are almost certainly flawed reconstructions of historical events. A common response to this problem by coalescent theorists trying to estimate evolutionary parameters is to simply "integrate out" the history by performing MCMC iteration over possible histories (Kuhner et al., 1995). However, there seems to be a signal in tree shape that stands out from the noise and that can guide the selection of evolutionary models. Tree shape also has a role in understanding potential problems and biases of tree reconstruction methods.

In summary, a new method is developed here for evaluating tree shape statistics, which is called the *resolution* of a statistic. This method formalizes the intuition that a good statistic takes on similar values for similar trees and different values for rather different trees. It has the advantage that it can help choose a kth statistic given that $k-1$ other statistics are already known; this opens up the possibility of finding a useful suite of statistics to describe a tree. The method is applied to make specific recommendations for such a suite of three statistics. Finally, the results of the geometric analysis are compared to two model-based tree distributions and find that statistics with good resolution were also the ones that had high power to distinguish the two distributions. This paper represents a small step in an area that may continue to pose interesting questions for years to come.

Acknowledgments

The author would like to thank Akira Sasaki for asking him the question "What is a good way to numerically describe the shape of a tree?" 2 years ago, as well as David Aldous, Steve Evans, Joseph Felsenstein, Arne Mooers, Montgomery Slatkin, Mike Steel, and John Wakeley for stimulating discussion and valuable comments. Susan Holmes supplied the original idea of applying multidimensional scaling in this setting. F.A.M. was supported by a Graduate Research Fellowship from the National Science Foundation.

References

Agapow, P., and A. Purvis. 2002. Power of eight tree shape statistics to detect nonrandom diversification: A comparison by simulation of two models of cladogenesis. Syst. Biol. 51:866–872.

Aldous, D. 1995. Probability distributions on cladograms. Pages 1–18 *in* Random discrete structures (D. Aldous and R. Pemantle, eds.). Springer, Berlin.

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

Blum, M. G. B., O. François, and S. Jauson. 2006. The mean, variance and joint distribution of two statistics sensitive to phylogenetic tree balance. Ann. Appl. Prob., in press.

Borg, I., and P. J. F. Groenen. 2005. Modern multidimensional scaling. Springer series in statistics, 2nd edition. Springer, New York.

Chailloux, E., P. Manoury, and B. Pagano. 2000. Développement d'applications avec Objective CAML. O'Reilly, Sebastopol, CA english translation available at http://caml.inria.fr/pub/docs/oreilly-book/.

Colless, D. 1982. Phylogenetics: The theory and practice of phylogenetic systematics. Syst. Zool. 31:100–104.

DasGupta, B., B. Filler, and B. Filler. 2000. On computing the nearest neighbor interchange distance. Pages 125–143 *in* Proceedings of the DIMACS Workshop on Discrete Problems with Medical Applications (D. Z. Du, P. M. Pardalos, and J. Wang, eds.) DIMACS series in discrete mathematics and theoretical computer science, Volume 55. American Mathematical Society Providence, RI.

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Press, Sunderland, Massachusetts.

Ford, D. J. 2005. Probabilities on cladograms: Introduction to the alpha model. http://arxiv.org/abs/math/0511246.

Guyer, C., and J. Slowinski. 1991. Comparisons of observed phylogenetic topologies with null expectations among 3 monophyletic lineages. Evolution 45:340–350.

Guyer, C., and J. Slowinski. 1993. Adaptive radiation and the topology of large phylogenies. Evolution 47:253–263.

Harding, E. F. 1971. The probabilities of rooted tree-shapes generated by random bifurcation. Adv. Appl. Prob. 3:44–77.

Heard, S. 1996. Patterns in phylogenetic tree balance with variable and evolving speciation rates. Evolution 50:2141–2148.

Hillis, D., T. Heath, and K. S. John. 2005. Analysis and visualization of tree space. Syst. Biol. 54:471–482.

Kirkpatrick, M., and M. Slatkin. 1993. Searching for evolutionary patterns in the shape of a phylogenetic tree. Evolution 47:1171–1181.

Koenker, R., and G. Bassett. 1978. Regression quantiles. Econometrica 46:33–50.

Kuhner, M., J. Yamato, and J. Felsenstein. 1995. Estimating effective population-size and mutation-rate from sequence data using metropolis-hastings sampling. Genetics 140:1421–1430.

Losos, J., and F. Adler. 1995. Stumped by trees—A generalized null model for patterns of organismal diversity. Am. Nat. 145:329–342.

Mardia, K., J. Kent, and J. Bibby. 1979. Multivariate analysis. Academic Press, New York.

Matsen, F. 2006. Optimization over a class of tree shape statistics. Accepted to IEEE/ACM Trans. Comput. Biol. Bioinform. http://arxiv.org/abs/q-bio.PE/0605034

Matsen, F., and S. Evans. 2006. Ubiquity of synonymity: Almost all large binary trees are not uniquely identified by their spectra or their immanantal polynomials. http://arxiv.org/abs/q-bio/0512010.

McKenzie, A., and M. Steel. 2000. Distributions of cherries for two models of trees. Math. Biosci. 164:81–92.

Mooers, A. 1995. Tree balance and tree completeness. Evolution 49:379–384.

Mooers, A., and S. Heard. 1997. Evolutionary process from phylogenetic tree shape. Q. Rev. Biol. 72:31–54.

Nee, S., R. M. May, and P. H. Harvey. 1994. The reconstructed evolutionary process. Philos. Trans. R. Soc. Lond. B Biol. Sci. 344:305–11.

Ortega, J. M. 1987. Matrix theory: A second course. Plenum Press, New York.

Pinelis, I. 2003. Evolutionary models of phylogenetic trees. Proc. Roy. Soc. B 270:1425–1431.

Purvis, A., and P. Agapow. 2002. Phylogeny imbalance: Taxonomic level matters. Syst. Biol. 51:844–854.

Rogers, J. 1996. Central moments and probability distributions of three measures of phylogenetic tree imbalance. Syst. Biol. 45:99–110.

Sackin, M. 1972. Good and bad phenograms. Syst. Zool. 21:225–226.

Savage, H. 1983. The shape of evolution—Systematic tree topology. Biol. J. Linn. Soc. 20:225–244.

Semple, C., and M. Steel. 2003. Phylogenetics. Oxford University Press, New York.

Shao, K., and R. Sokal. 1990. Tree balance. Syst. Zool. 39:266–276.

Simon, C., and R. Page. 2005. The past and future of systematic biology. Syst. Biol. 54:1–3.

Stam, E. 2002. Does imbalance in phylogenies reflect only bias? Evolution 56:1292–1295.

Steel, M., and A. McKenzie. 2001. Properties of phylogenetic trees generated by Yule-type speciation models. Math. Biosci. 170:91–112.

First submitted 17 December 2005; reviews returned 8 February 2006; final acceptance 9 March 2006
Associate Editor: Mike Steel

APPENDIX 1

Statistics previously defined in the literature are presented here for the convenience of the reader. Assume a tree has been chosen with l leaves. Let N_i represent the number of internal nodes between leaf i and the root (inclusive). Let r_j and s_j be the number of leaves of the two subtrees above the internal node j. For an internal node j and leaf i let d_{ji} be the number of edges on the path connecting j to i. Let M_j be the maximum of d_{ji} over leaves i subtended by j. Let $\mathcal{L}$ be the leaves of the tree and $\mathcal{I}$ the internal nodes except for the root. The root is denoted r.

$$I_c = \frac{2}{(n-1)(n-2)} \sum_{j \in \mathcal{I} \cup \{r\}} |r_j - s_j|$$

$$I_2 = \frac{1}{n-2} \sum_{\substack{j \in \mathcal{I} \cup \{r\} \\ r_j + s_j > 2}} |r_j - s_j| / |r_j + s_j - 2|$$

$$\bar{N} = \frac{1}{n} \sum_{i \in \mathcal{L}} N_i$$

$$\sigma_N^2 = \frac{1}{n} \sum_{i \in \mathcal{L}} (\bar{N} - N_i)^2$$

$$B_1 = \sum_{j \in \mathcal{I}} M_j^{-1}$$

$$B_2 = \sum_{i \in \mathcal{L}} N_i / 2^{N_i}$$