Evolution, 39(3), 1985, pp. 609–622

FACTORS DETERMINING THE ACCURACY OF CLADOGRAM ESTIMATION: EVALUATION USING COMPUTER SIMULATION

KENT L. FIALA AND ROBERT R. SOKAL
Department of Ecology and Evolution, State University of New York, Stony Brook, NY 11794

Abstract.—We developed a simulation model of phylogenesis with which we generated a large number of phylogenies and associated data matrices. We examined the characteristics of these and evaluated the success of three taxonomic methods (Wagner parsimony, character compatibility, and UPGMA clustering) as estimators of phylogeny, paying particular attention to the consequences of changes in certain evolutionary assumptions: relative rate of evolution in three different evolutionary contexts (phyletic, parent lineage, and daughter lineage); relative rate of evolution in different directions (novel forward, convergent forward, or reverse); variation of evolutionary rates; and topology of the phylogenetic tree.

Except for variation of evolutionary rates, all the evolutionary parameters that we controlled had significant effects on accuracy of phylogenetic reconstructions. Unexpectedly, the topology of the phylogeny was the most important single factor affecting accuracy; some phylogenies are more readily estimated than others for simply historical reasons. We conclude that none of the three estimation methods is very accurate, that the differences in accuracy among them are rather small, and that historical effects (the branching pattern of a phylogeny) may outweigh biological effects in determining the accuracy with which a phylogeny can be reconstructed.

Received March 19, 1984. Accepted November 30, 1984

It is desirable to evaluate the accuracy with which tree structures obtained by current numerical taxonomic methods estimate phylogeny (even if, as in phenetics, phylogenetic estimation is not necessarily a goal of the method). Such evaluation can be done only by comparing these tree structures with their respective true phylogenies. Although a few studies (e.g., Baum, 1983; Baum and Estabrook, 1978) have been reported as comparisons of true phylogenies with estimated ones, these are more accurately described as comparisons of better documented estimates with less well documented ones. In general, phylogenies of real organisms are unknown. Rigorous comparisons of estimated and true phylogenies therefore require the use of artificial data. This paper reports the development and use of a simulation model of phylogenesis with which one can obtain a large number of phylogenies and associated data sets for making such comparisons.

In constructing phylogenetic models, evolutionists have made a great variety of assumptions concerning evolutionary change. These assumptions have been the subject of considerable speculation and controversy, in part because macroevolution cannot be studied experimentally and all but the simplest models are analytically intractable. Simulation provides a way of avoiding these difficulties, but one must be concerned with the extent to which simulation results depend upon artificialities of the model rather than upon the nature of evolution. Previous simulation studies (Raup and Gould, 1974; Sokal, 1983*a*, 1983*b*) have shown that simulated data may have the internal structure characteristic of real taxonomic data. This supports the general assumption that evolution may be modeled as a stochastic process, though it does not necessarily validate the specific assumptions of any one model. Rather than basing our results upon the validity of a fixed set of assumptions, our approach is to explore the consequences of changes in certain evolutionary assumptions, and restrict our conclusions to the effects of these changes. The extent

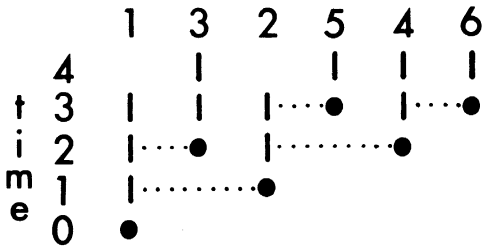

FIG. 1. Illustration of tree generation. The time steps during which a lineage exists are marked with either a filled circle or a vertical bar. A filled circle means the lineage did not exist at the beginning of the time step but originated during the time step, while a vertical bar means a lineage already existed at the beginning of the time step. The filled circle automatically implies that the lineage survived into the next time step, but the vertical bar does not provide any information regarding survival through the time step. The dotted lines connect the origin of a lineage with the parent lineage; the lengths of these lines have no meaning. The four possible combinations of phylogenetic events are illustrated in time step 3. Lineages 1, 2, 3, and 4 all existed at the beginning of time step 3. Lineages 2 and 4 gave rise to new lineages 5 and 6, respectively; lineages 1 and 3 did not give rise to new lineages. Lineages 3 and 4 survived into time step 4; lineages 1 and 2 went extinct during the time step.

to which other aspects of evolution are significant remains a subject for further investigation.

MATERIALS AND METHODS

Simulation Model

Phylogeny.—The phylogenetic branching pattern of a simulated evolutionary tree is generated independently of character evolution, so that replicate character sets can be generated for any topology. Each simulation begins with a single ancestral lineage and proceeds through a sequence of discrete time steps. In each time step, each extant lineage may give rise to a daughter lineage or may not (depending on a branching probability), and then may go extinct or else survive (depending on an extinction probability) (see Fig. 1).

In this study, the branching probability and the extinction probability each had a nominal value of 0.1, but if these probabilities were held constant, the phylogenies tended to radiate rapidly or, more likely, to go extinct. It was therefore necessary to introduce a feedback mechanism for regulating the number of contemporaneous lineages near a specified equilibrium number (Raup et al., 1973). At each time step, the feedback mechanism adjusted either the branching or the extinction probability away from its nominal value so as to tend to stabilize the number of contemporaneous lineages. The simulations were programmed to proceed for at least 100 time steps, then stop at the next time step in which the specified equilibrium number of contemporaneous lineages (20 for this study) was exactly obtained.

Characters and Character Evolution.—We regard our simulated characters as discrete morphological characters, however there is no apparent reason to suppose that our results are not equally applicable to other types of data. The starting lineage in a simulation is assigned a character state vector, that is, a set with one character state code for each of 25 characters. A lineage that survives from one time step to the next receives a copy of the character state vector that it had in the previous time step, and a newly generated (daughter) lineage receives as its initial character state vector a copy of the character state vector of the (parent) lineage from which it arose. During a copying step, random changes in character states, or evolution, may occur. The probability of character state change depends on which of three evolutionary contexts the copying occurs in.

A lineage that is about to receive a copy of a preexisting character state vector will be of one of three types: a lineage that has survived from the previous time step and that has not given rise to a daughter lineage during the current time step; a lineage that has survived from the previous time step and that has given rise to a daughter lineage during the current time step; or a newly arisen daughter lineage. We say that any evolutionary change that occurs during the copying of a character state vector to the first type of lin-

eage occurs in the *phyletic context,* to the second type occurs in the *parent lineage context,* and to the third type occurs in the *speciational context.*

The biological justification for this novel terminology is as follows. Evolution occurring in the speciational context corresponds to what Stanley (1979) calls the "speciational component" of evolution. Evolution in the phyletic context is a subset of what is generally called "phyletic" evolution, or what Stanley has called the "phyletic component" of evolution. Assuming that the formal distinction that the computer program makes between parent lineage and daughter lineage is biologically valid, one would also include evolution in the parent lineage context as part of the phyletic component of evolution. However, one might not wish to assume this, but rather to consider both of the lineages branching from a speciation event as being equally distinct from the ancestral lineage. If so, one would regard evolution in the parent lineage context as part of the speciational component. We have defined the three evolutionary contexts in order to retain the flexibility to explore the consequences of either point of view.

Type of Character State Change.—Each time a character state code is copied, it undergoes one of four types of change: *novel* change—to a newly derived state that has never existed before; *convergent* change—to a state that already exists or has existed in some other part of the phylogeny, and that first occurred as a novel change from the same state as the current state; *reverse* change—to the immediately ancestral state; or *null* change—remaining in the current state. The allowed change types are defined in such a way that the set of relations among the states of a character defines a tree, the character state tree. The probabilities of the four change types have been designated *p, q, r,* and *s.* (The mnemonics progressive, quondam, reverse, and static may aid the reader in identifying these symbols.)

A vector of character state change probabilities, [*p q r s*], determines the type of change each time a character state code is copied. A separate vector is defined for each of the three evolutionary contexts, resulting in a 3 × 4 matrix of probabilities. (Three examples of such matrices are shown in Table 1, which will be further explained below.)

Certain constraints require exceptions to the strict application of the state change probabilities. Reversal is not possible from the ancestral state; a null change is substituted. Convergent change is not possible from the tip of a branch of a character state tree; a novel change is substituted. Novel change is not possible if a program limit of 32 states for the character has already been reached; a convergent change is substituted if possible or, if not, a null change. (The 32-state limit was rarely reached.)

Variation of Character State Change Probabilities.—The entries in the matrix of character state change probabilities can be varied to simulate variation in evolutionary rates. To reduce the potential complexity of the variation model, we transform the state change probability matrix as follows. The 3 × 1 vector of probabilities of null change is removed, and the remaining 3 × 3 submatrix of non-null change probabilities is converted into a matrix of conditional probabilities given the occurrence of a non-null change. Determination of change type then becomes a two-step process. First the probability of null change in the current evolutionary context determines whether a null change occurs. If the change is not null, then the specific type is determined by the appropriate entries in the 3 × 3 matrix. Potential variation of evolutionary rate is allowed to occur only in the vector of null change probabilities; the 3 × 3 matrix of conditional probabilities is always fixed for the duration of a simulation.

The simulation allows for three alternative models of variation of evolutionary rates. In each, there is an initial vector of nominal null change probabilities about which any variation occurs. *No*

TABLE 1. Transition probability tables for three context patterns, direction pattern I. To conserve space, the tables for direction patterns II and III are not shown. The table entries can be computed as follows: Novel and convergent change probabilities (p and q) are the same in all three direction patterns. Reversal probabilities (r) are doubled in direction pattern II and tripled in direction pattern III, relative to the direction pattern I values. Null change probabilities (s) are always given by $s = 1 - p - q - r$.

Context pattern	Context	Transition type			
		Novel	Convergent	Reversal	Null
I	phyletic	0.006	0.006	0.006	0.982
	parent lineage	0.006	0.006	0.006	0.982
	speciational	0.006	0.006	0.006	0.982
II	phyletic	0.000	0.000	0.000	1.000
	parent lineage	0.030	0.030	0.030	0.910
	speciational	0.030	0.030	0.030	0.910
III	phyletic	0.000	0.000	0.000	1.000
	parent lineage	0.000	0.000	0.000	1.000
	speciational	0.060	0.060	0.060	0.820

variation is produced by simply leaving the initial vector fixed for the run. *Variation among lineages* is produced by giving each lineage its own vector of null change probabilities, which is inherited and which evolves much like a character state vector. Before each instance in which a lineage receives a copied character state vector, that lineage's null change probability for the relevant evolutionary context is first changed by a random amount. *Variation among characters* is produced by giving each character its own vector of null change probabilities. The character-specific null change vectors are changed from the nominal values by random amounts at the beginning of the simulation, and then left constant during the simulation.

The probability model used to obtain the above-mentioned random amounts of change in the null change probabilities is a random walk of the log of the odds of null change. Negative feedback introduced by setting each new log of the odds equal to the mean of the nominal value and the previous value plus a random deviation keeps the character state change probabilities from quickly drifting to 0 or 1. In instances where the null change probability was set at 1, variation was suppressed.

Experimental Design

It is neither practical nor desirable to explore the entire parameter space of the simulation model. Instead, we chose an experimental design that has just four control or treatment variables: relative importance of the different evolutionary contexts; direction of character state change; variation of evolutionary rates; and topology of evolutionary trees.

Relative Importance of Evolutionary Contexts.—One may safely assume that evolutionary rates differ among contexts, but the nature and relative importances of these differences are a subject of controversy. Therefore, we restricted our investigation to just three cases ("context patterns") representing extreme evolutionary models, no one of which is likely to have been realized in actual organisms, but which collectively span a wide range of possibilities. In context pattern I, evolutionary rates are the same in all contexts, so that on the average the amount of evolutionary change between two locations on the phylogenetic tree is proportional to the number of time steps along the evolutionary path between them. Context pattern I is much like the evolutionary models of Astolfi et al. (1981), Tateno et al. (1982), and Nei et

al. (1983). In context pattern II, evolutionary rates are the same in both parent lineage and speciational contexts, but no evolution occurs in the phyletic context. The amount of evolutionary change between two locations is therefore proportional to the number of branch points between them. In context pattern III, evolution occurs only in the speciational context. Raup and Gould (1974), who employed context pattern III in their simulations, have described it as an idealized case of punctuated equilibrium. The three context patterns are shown in Table 1.

Direction of Character State Change.—By manipulating the relative magnitudes of the probabilities of the various directions of character state change, three "direction patterns" were created. In direction pattern I, the probabilities of novel, convergent and reverse change were set equal to each other, that is, $p = q = r$. Thus, changes to relatively derived states (novel and convergent) were collectively more probable than reversals. In direction pattern II, novel and convergent change probabilities were kept equal, but the probability of reversal was doubled, $p = q = r/2$, so that reversals were as likely as derived (novel or convergent) changes. In direction pattern III, the probability of reversal was still greater, $p = q = r/3$, so that reversals were more likely than derived changes. The probabilities for direction pattern I under each of the three context patterns are shown in Table 1.

Variation of Evolutionary Rate.—The third treatment variable, "variation pattern," is simply the evolutionary rate variation rule described above. Variation pattern I is no variation, pattern II is variation among lineages, and pattern III is variation among characters.

Topology of Evolutionary Trees.—The final treatment variable is the topology of the simulated phylogenetic tree. From a large number of different tree topologies obtained by using the identical branching and extinction parameters but different pseudorandom number seeds, we selected two sets of eight topologies each, such that within each set the values of a shape measure, "stemminess," were spaced at equal intervals through the range of values observed in the whole sample. The stemminess of a taxonomic subset (or component [Nelson, 1979]) of a tree is defined as the proportion of the total length of the edges of the subset (including the subtending edge, or "stem") that is accounted for by the length of the subtending edge of the subset (see Fig. 2). The stemminess of a tree is the mean of the values for all subsets. Time was used as the branch length measure.

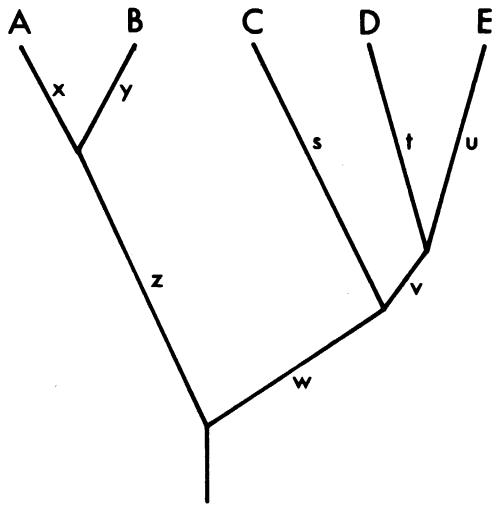

FIG. 2. Illustration of computation of stemminess. For subset AB, stemminess is $z/(x + y + z)$, for subset DE, $v/(t + u + v)$, and for subset CDE, $w/(s + t + u + v + w)$, where lower case letters are internode lengths. The stemminess of the tree is the mean of these values. Note that the length of the root stem is ignored.

The use of two replicate sets of topologies allowed us to test for both a stemminess effect and a topology-within-stemminess level effect.

Table 2 lists the different treatment variables for convenient reference. The two sets of eight trees that we used are illustrated in Figure 3. To summarize the experimental design, we performed 864 different simulations: the main effects

TABLE 2. Summary of experimental treatment variables.

Context patterns
I. Evolution occurs in phyletic, parent lineage, and speciational contexts.
II. Evolution occurs in parent lineage and speciational contexts only.
III. Evolution occurs in speciational context only.
Direction patterns
I. Novel = convergent = reverse.
II. Novel = convergent = reverse/2.
III. Novel = convergent = reverse/3.
Variation patterns
I. No variation in evolutionary rates.
II. Variation in evolutionary rates among lineages over time.
III. Variation in evolutionary rates among characters, constant over time.
Stemminess of phylogenetic tree
Topology assigned one of the following stemminess values: 0.22, 0.26, 0.30, 0.34, 0.38, 0.42, 0.46, 0.50

have 3, 3, 3, and 8 levels respectively; within each of the 216 combinations of these effects, we used two replicate topologies; and within each of these 432 combinations, we simulated two different character sets.

Evolutionary Rates.—Absolute evolutionary rates were not treated directly as independent variables in this study. Taxonomic theory is relatively little concerned with the effect that the overall rate of evolutionary change has on accuracy of cladogram estimation but is more concerned with effects of relative differences in evolutionary rates such as we have just described.

The actual rates used do, however, have an effect on the data generated. The effect is particularly strong with respect to the evolutionary context patterns. Evolutionary opportunities occur twice as often in context pattern II as in context pattern III and roughly 3.5 times as often (empirically determined) in context pattern I as in context pattern II. Therefore we used different absolute rates (see Table 1) in an attempt to make the total number of evolutionary changes similar under each context pattern. No such adjustments were made for the direction pattern or variation pattern treatment variables, for which differences were less extreme. Despite the adjustment, tree length was significantly different among levels of all treatments.

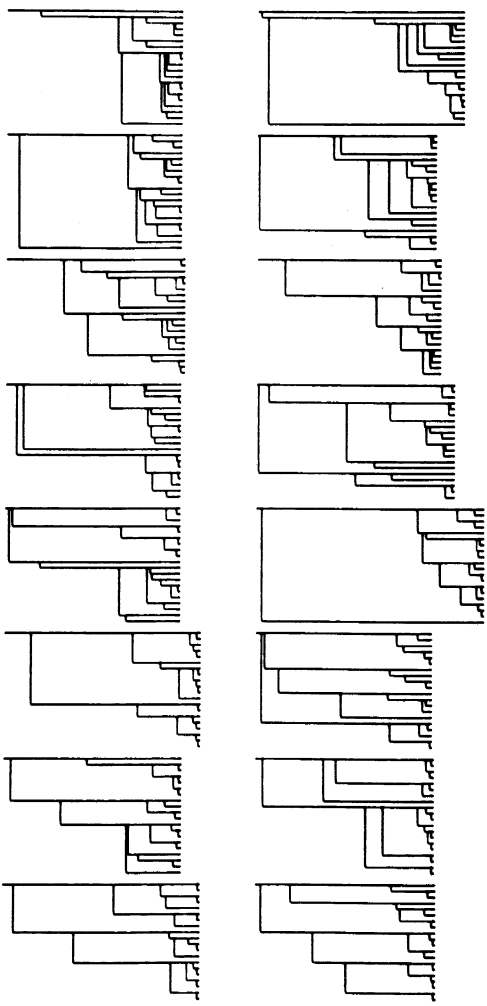

FIG. 3. The simulated tree topologies. Within each column, the topologies are in order from lowest stemminess at the top to highest stemminess at the bottom. Pairs of trees opposite each other have identical stemminess values. Within each tree, the time scale runs from left to right and the vertical scale is arbitrary.

Types of Data Collected

Tree Length Measures.—We computed several measures of length for each tree. Seven of these measures will be referred to as "primary" lengths: N_{true}, C_{true}, and R_{true} are, respectively, the actual numbers of novel, convergent, and reverse steps on the true tree. Analogous measures N_{max}, C_{max}, and R_{max} are computed as the numbers of novel, convergent, and reverse steps that would appear to have occurred if the lineages were mapped onto a tree in which each lineage radiates independently from the root. $L_{min(l)}$ is the length the phylogenetic tree would have if each character state change occurred uniquely (Sokal, 1983*a*); i.e., it is the sum of the number of edges on the character state trees.

From the primary measures, compound measures of tree length and homoplasy defined by Sokal (1983*a*) can be computed as follows: the actual (true) length of the tree is $L_{act} = N_{true} + C_{true} + R_{true}$; the maximum length of the tree is $L_{max(u)} = N_{max} + C_{max} + R_{max}$; the maximum path length of the tree is $L_{max(l)} = L_{max(u)} - 2(R_{max})$; homoplasy is $H = L_{act}/L_{min(l)}$; and the dendritic index is $DI = (L_{max(u)} - L_{act}) / (L_{max(u)} - L_{min(l)})$.

Phylogenetic Reconstruction and Evaluation.—From the data sets resulting from each simulation, dendrograms were obtained by three methods: two cladogram estimation methods (Wagner parsimony and character compatibility) and a phenetic clustering procedure (UPGMA). The data were recoded to additive characters for Wagner analysis, to additive binary characters for compatibility analysis, and to Manhattan distances based on the additive characters for UPGMA clustering. Recoding was required for Wagner analysis because the program used requires additive characters. Recoding for compatibility analysis was not required but was done to increase the degree of resolution of the tree obtained from a primary analysis, because it was impractical to carry out secondary analyses for so many data sets. Although Manhattan distance is not a conventional dissimilarity metric for phenetic analysis (Rohlf and Sokal, 1980), it seemed appropriate here because of the discrete quantal nature of the simulated character state evolution.

After characters invariant among the 20 extant taxa were eliminated, the data sets averaged 24.8 ± 0.01 characters. These characters averaged 18.3 ± 0.07 states, of which 5.5 ± 0.03 were represented among extant taxa. The recoding resulted in an average of 47.0 ± 0.2 additive characters per data set.

The implementation of the Wagner method was that of the program WAGNER 78 by J. S. Farris; compatibility analysis was performed by Fiala's program CLINCH; and UPGMA clustering was performed by subroutines from NTSYS (Rohlf et al., 1980). The Wagner trees were rooted by providing the data for the true ancestor of the whole phylogeny as an outgroup. The compatibility trees were rooted by providing the true primitive state for each character.

To measure the agreement between each estimated tree and its true counterpart, a strict consensus tree (Sokal and Rohlf, 1981) was constructed, and three consensus indices were computed: CI_C (Colless, 1980; Rohlf, 1982); CI_W ("wrongness"; Fiala, 1983); and d (Robinson and Foulds, 1981). CI_C is a measure of the relative correctness of the estimated tree; it is the number of true taxonomic subsets in the estimated tree divided by the number of subsets in the true tree. It does not take into account the difference between a fully resolved estimated tree that contains some true and some false subsets and an incompletely resolved estimated tree containing the same number of true subsets but fewer incorrect subsets. The strict consensus of either tree with the true tree has the same CI_C, but the tree containing multifurcations might be judged preferable because it omits misleading information. This difference can be quantified by CI_W, the number of incorrect subsets

TABLE 3. Descriptive statistics for length measures. Sample size for each statistic given under treatment label.

		$L_{min(l)}$	L_{act}	H	Wagner length	Total homoplasy	$\hat{H}$	Deviation ratio
Context	I	122.4	236.9	1.957	163.5	1,000	1.338	0.150
pattern	II	133.7	257.1	1.943	188.3	1,439	1.419	0.191
($N = 288$)	III	138.0	266.9	1.951	189.3	1,404	1.386	0.186
Direction	I	143.2	223.3	1.555	189.6	1,123	1.333	0.132
pattern	II	130.9	257.9	1.965	183.0	1,381	1.403	0.186
($N = 288$)	III	119.8	279.7	2.332	168.5	1,339	1.407	0.208
Variation	I	125.1	233.3	1.882	170.7	1,160	1.370	0.168
pattern	II	137.7	268.0	1.968	189.6	1,376	1.386	0.180
($N = 288$)	III	131.1	259.6	2.000	180.8	1,307	1.388	0.179
Stemminess	0.22	133.0	260.2	1.982	194.8	1,397	1.473	0.224
($N = 108$)	0.26	127.6	246.0	1.944	184.1	1,318	1.451	0.199
	0.30	117.0	217.9	1.871	166.4	1,289	1.416	0.188
	0.34	142.3	276.7	1.955	189.6	1,402	1.348	0.177
	0.38	135.6	274.9	2.061	190.4	1,288	1.408	0.173
	0.42	140.4	270.6	1.946	175.3	1,186	1.264	0.145
	0.46	124.2	234.1	1.913	172.4	1,311	1.382	0.167
	0.50	130.4	248.8	1.929	170.4	1,056	1.308	0.132
Overall	Mean	131.3	253.6	1.950	180.4	1,281	1.381	0.176
($N = 864$)	SE	0.088	2.01	0.013	1.15	18.7	0.0048	0.0025

in the estimated tree divided by the number of subsets in the true tree. For the consensus of a fully resolved estimated tree with the true tree $CI_W = 1 - CI_C$; but if the estimated tree is not fully resolved, $CI_W < 1 - CI_C$. In comparisons such as ours, in which at least one of the trees is fully resolved, a simple relationship exists between the d metric of Robinson and Foulds (1981) and CI_C and CI_W, namely, $d = (n - 2)(1 - CI_C + CI_W)$, where n is the number of extant lineages. We therefore present results for only CI_C and d.

RESULTS

Effects of Treatment Variables on Tree Length Measures

Of the numerous measures of tree length computed, we present individual analyses of only $L_{min(l)}$, L_{act}, and H, which are those that correspond most closely to measures that might be estimated in actual taxonomic studies (Table 3). For comparison, Table 3 also includes estimated length measures computed in the Wagner analysis. These include the Wagner tree length, total homoplasy (the sum of the pairwise homoplastic distances among lineages), and deviation ratio (total homoplasy divided by the sum of the pairwise Manhattan distances among lineages) as computed by WAGNER 78, as well as $\hat{H}$, an estimate of H computed as the Wagner tree length divided by $L_{min(l)}$. Note that each Wagner tree length is substantially less than the corresponding L_{act}.

All length measures are significantly dependent on context pattern. But tree length is obviously directly affected by overall evolutionary rates, so differences in tree length among context patterns are completely confounded with our deliberate, though incompletely successful, attempt to choose rates that would equalize L_{act} among patterns. It is interesting, though, that the dependence on absolute rates is loose enough that we were unable to devise a reliable adjustment criterion.

The effects of direction pattern on length are also substantially influenced by absolute rates. L_{act} was greater under those direction patterns in which reversals were more frequent. While $L_{min(l)}$ was less when reversals were more fre-

quent, the effect on L_{act} was more pronounced, with the net result that H increased with increasing frequency of reversals. The decrease in $L_{min(l)}$ with increasing reversals is probably an artifact of the convention that a novel change is substituted when a convergent change from a maximally derived character state is attempted. This situation occurs less frequently when reversals are more frequent, resulting in the generation of fewer states, hence lower $L_{min(l)}$.

Variation pattern II (variation among lineages) produced the highest values of both $L_{min(l)}$ and L_{act}, and variation pattern I (constant rates) the lowest. The difference between pattern I and pattern II can be accounted for by the fact that in pattern II the change probabilities are continually varying and are frequently above the mean because of the positive skew of probabilities under the random walk model used; whereas in pattern I the probabilities are held constant at the mean value. The intermediacy of pattern III is probably due to the fact that the probabilities are randomized once, then are held constant.

There is no suggestion of a trend in the true length measures in relation to stemminess values. This lack of a relation is expected, given the independence of character state change and branching, and is in striking contrast to the inverse relation of stemminess and the estimated length and homoplasy measures from the Wagner analysis (Table 3).

Effects of Tree Length Measures on Accuracy of Phylogenetic Estimates

The results for all six accuracy measures (both CI_C and d for each of three estimation methods) were similar in many respects and, to simplify discussion, will be discussed together wherever possible. Any unqualified reference to "accuracy measures" should be understood to refer to all six indices. Higher accuracy means higher CI_C and lower d.

We computed regressions of accuracy measures on several classes of tree length measurements. These were: the primary length measures; the compound length measures L_{act}, $L_{max(l)}$, and $L_{max(u)}$; homoplasy measurements H and DI; and estimated length measurements computed in the Wagner analysis.

Results were so unimpressive that we simply note that even when all the variables in any one class are included in the regression model, the obtained R^2 values are very low. For the primary measures they range from 0.08 to 0.12, for the compound measures from 0.06 to 0.10, and for H and DI from 0.05 to 0.14.

Unexpectedly, the Wagner length measures (length, total homoplasy, deviation ratio, and $\hat{H}$) proved to be better (though still poor) predictors of accuracy, not just for the Wagner results, but for compatibility and UPGMA as well. The R^2's for the regressions of CI_C and d on Wagner length measures were, respectively, 0.24 and 0.28 for Wagner estimates, 0.29 and 0.25 for compatibility estimates, and 0.17 and 0.16 for UPGMA estimates. Accuracy was correlated negatively with deviation ratio, Wagner length, and $\hat{H}$, and positively with Wagner total homoplasy. Note that these results do not mean that for two trees estimated from the same data by different methods, the one with better Wagner statistics is likely to be more accurate. Rather, they imply that for two different data sets, the one for which better Wagner statistics are obtained will be more accurately analyzed regardless of method.

Effects of Treatment Variables on Accuracy of Phylogenetic Estimates

Mean consensus indices for the experiment are shown in Table 4, broken down by experimental treatments. The accuracy measures were analyzed using a repeated measures analysis of variance model (Hull and Nie, 1981) in which context pattern, direction pattern, variation pattern and stemminess rank were included as main effects, with estimation method as the "within subjects" factor, and replicates 1 and 2 of 8 stemminess

TABLE 4. Descriptive statistics for CI_C and d between estimated trees computed by each of three methods, and the corresponding true tree. Sample size for each statistic given under treatment label.

		CI_C			d		
		WAGNER 78	CLINCH	UPGMA	WAGNER 78	CLINCH	UPGMA
Context	I	0.717	0.689	0.778	8.7	8.1	7.8
pattern	II	0.778	0.744	0.778	7.5	7.1	7.9
($N = 288$)	III	0.694	0.661	0.644	9.9	9.4	12.8
Direction	I	0.750	0.728	0.750	7.6	7.2	9.0
pattern	II	0.739	0.706	0.739	8.5	8.2	9.3
($N = 288$)	III	0.694	0.661	0.717	10.0	9.3	10.2
Variation	I	0.733	0.706	0.744	8.5	8.0	9.2
pattern	II	0.722	0.694	0.717	8.8	8.4	10.1
($N = 288$)	III	0.728	0.694	0.739	8.7	8.2	9.3
Stemminess	0.22	0.572	0.522	0.567	13.8	12.4	15.6
($N = 108$)	0.26	0.667	0.606	0.650	10.6	10.4	12.4
	0.30	0.711	0.678	0.767	9.1	8.7	8.3
	0.34	0.744	0.728	0.744	8.4	7.8	9.2
	0.38	0.744	0.711	0.744	8.4	8.1	9.1
	0.42	0.800	0.778	0.778	5.9	6.1	7.8
	0.46	0.778	0.756	0.789	7.1	6.8	7.6
	0.50	0.811	0.811	0.833	6.1	5.5	6.0
Overall	Mean	0.728	0.700	0.733	8.7	8.2	9.5
($N = 864$)	SE	0.005	0.005	0.005	0.15	0.13	0.17

values were nested within stemminess rank. The results for CI_C are summarized in Tables 5 and 6. The ANOVA results for d parallel those for CI_C and are not shown.

We first discuss the multivariate effects on the accuracy measures of all three estimation methods, then the differences among methods within cells of the analysis.

For both accuracy measures, the effects of context pattern, direction pattern, stemminess rank, and topology within stemminess rank were highly significant. The linear and quadratic components of the sum of squares for stemminess rank were also significant, using orthogonal polynomials. Variation pattern was not significant, although it approached significance for d. The only significant interaction term was that between context pattern and stemminess rank, for both indices.

The context pattern effect is not readily interpretable because it may be confounded with effects of absolute rates of evolution, as we have noted above in regard to the effect of context pattern on tree length. The effect of direction pattern is readily interpretable: as the relative importance of reversal as a type of character state change increases, accuracy of estimation drops. The direction pattern effect could well be described as a homoplasy effect because of the strong relationship between direction pattern and homoplasy (see Table 3).

The most intriguing effects are those of tree shape. Stemminess rank is the single most important component of variation. Yet there is an additional highly significant effect due to topology within stemminess rank (i.e., the two replicate trees for each stemminess rank tended to have different effects on accuracy). Thus, while our stemminess measure is a good predictor of the accuracy of the estimate, it does not utilize every relevant aspect of the structure of the phylogeny.

The above discussion has concerned the multivariate effects on accuracy for all three tree estimation measures. There is, in addition, significant heterogeneity among methods within cells of the experiment (Table 6). As ranked by CI_C, Wagner parsimony and UPGMA are

TABLE 5. Multivariate analysis of variance of arcsin-transformed CI_C between true trees and estimates from WAGNER 78, CLINCH, and UPGMA. Tests of significance for among treatments.[1] Each "topology within stemminess" term is tested against "within cells." Each other term is tested against the corresponding "topology within stemminess" term. Significance levels: * = $P < 0.05$; ** = $P < 0.01$; *** = $P < 0.001$.

Source of variation	Sum of squares	*d.f.*	Mean square	*F*	Significance of *F*
Within cells	9.53	432	0.022		
Topology within stemminess	1.75	8	0.219	9.93	***
Stemminess	21.77	7	3.111	14.19	***
Context by topology within stemminess	0.41	16	0.026	1.16	ns
Context	6.17	2	3.084	120.33	***
Context by stemminess	1.83	14	0.130	5.09	**
Direction by topology within stemminess	0.44	16	0.027	1.24	ns
Direction	1.80	2	0.902	32.97	***
Direction by stemminess	0.15	14	0.011	0.39	ns
Variation by topology within stemminess	0.47	16	0.030	1.34	ns
Variation	0.17	2	0.085	2.88	ns
Variation by stemminess	0.40	14	0.029	0.98	ns
Context by direction by topology within stemminess	0.75	32	0.024	1.07	ns
Context by direction	0.22	4	0.055	2.34	ns
Context by direction by stemminess	0.57	28	0.020	0.86	ns
Context by variation by topology within stemminess	0.71	32	0.022	1.01	ns
Context by variation	0.24	4	0.059	2.66	ns
Context by variation by stemminess	0.68	28	0.024	1.08	ns
Direction by variation by topology within stemminess	0.56	32	0.018	0.79	ns
Direction by variation	0.05	4	0.014	0.78	ns
Direction by variation by stemminess	0.38	28	0.014	0.78	ns
Context by direction by variation by topology within stemminess	1.12	64	0.018	0.79	ns
Context by direction by variation	0.07	8	0.008	0.47	ns
Context by direction by variation by stemminess	0.83	56	0.015	0.85	ns

[1] Treatment labels: "context" = context pattern; "direction" = direction pattern; "variation" = variation pattern, "stemminess" = stemminess rank.

more accurate overall than compatibility; by contrast, as ranked by *d*, compatibility is the most accurate method, then Wagner parsimony, and then UPGMA (Table 4). The difference between the results for the two measures of accuracy reflects the previously noted differences in the concept of consensus. Wagner parsimony and compatibility frequently produced estimated cladograms containing multifurcations, while UPGMA dendrograms were nearly always fully resolved. (The multifurcations that did occur in UPGMA are due to the occurrence of identical OTUs.)

The only highly significant interaction of tree estimation method with a main treatment is that with context pattern (Table 6). In context pattern I, UPGMA is the most accurate method, while in context pattern III it is the least accurate method. In context pattern II, UPGMA is equally as accurate as Wagner and superior to compatibility as measured by CI_C, but is the least accurate method as measured by *d*.

Stemminess itself does not interact significantly with method, but there are significant interactions of topology within stemminess and method, and of context pattern by topology within stemminess and method. Thus, the unidentified as-

TABLE 6. Multivariate analysis of variance of arcsin-transformed CI_C between true trees and estimates from WAGNER 78, CLINCH, and UPGMA. Tests of significance for within treatments.[1]

Source of variation	Averaged F	Significance of F
Method	52.01	***
Topology within stemminess and method	3.10	***
Stemminess and method	1.36	ns
Context by topology within stemminess and method	3.18	***
Context and method	15.00	***
Context by stemminess and method	0.94	ns

[1] "Method" = method of tree construction (Wagner parsimony, compatibility, or UPGMA). Other labels as in Table 5. Results for direction and variation patterns and for interactions of context, direction and variation patterns are non-significant or marginally significant and are not shown.

pect of tree structure that is not measured by stemminess seems to affect different methods differently, whereas stemminess exerts an overall influence that is equal among methods.

DISCUSSION

Three results of this study are particularly striking: none of the three estimation methods is especially good at reconstructing phylogenies accurately; the differences among the methods are rather small, at least small enough to be completely overshadowed by the common deficiencies; and the historical pattern of branching of a phylogeny plays a far more substantial role in determining the accuracy with which the phylogeny can be reconstructed than do such biological factors as the nature and rate of evolutionary change. Such results are not causes for optimism by those who wish to estimate phylogenies.

Relatively low accuracy of phylogenetic estimation has also been noted in the simulation studies of Astolfi et al. (1981), Tateno et al. (1982), and Nei et al. (1983). In the single simulation analyzed extensively by Sokal (1983*a*, 1983*b*), no taxonomic method perfectly reconstructed the phylogeny. Thus this result appears to have substantial generality.

The issue of phylogenetic topology requires further discussion. Several authors (Colless, 1970; Felsenstein, 1978; Sokal, 1983*b*) have considered the effect that the distribution of edge lengths (as measured by number of character state changes) within the true phylogeny may have on the accuracy with which various phylogenetic methods estimate the phylogeny. Our demonstration of the importance of stemminess shows that the distribution of edge lengths is in fact important, but that the historical pattern of speciation events and extinctions within a particular phylogeny is the major determinant of the distribution. This is because of the obvious fact that the length of an edge as measured in time units has a major influence on its length in character state changes. Stemminess is in effect a measure of the relative amount of time available for divergence to occur, as opposed to time available for convergence, and phylogenies will be more accurately estimated when evolutionary change that sets taxa apart from sister groups predominates over convergent change within taxa (Colless, 1970; Felsenstein, 1978; Sokal, 1983*b*).

A branch-length effect related to stemminess was also noted by Tateno et al. (1982). Of the two extreme topologies they illustrate, the one described as being more subject to errors has the lower stemminess.

Two aspects of our model may lead some to have reservations about its generality. First, we have tacitly assumed that evolutionary rate is roughly uniform over the phylogeny. Our variation patterns introduced some nonuniformity of rates, but primarily as uniform "noise" rather than major localization of high or low rates. An alternative interpretation of our model is possible, however. If we regard our time scale as an arbitrary construct of the simulation model, rather than as a proper time scale, and measure branch length in units of expected amount of

evolution, rather than as time, stemminess becomes a measure of the nonuniformity of the distribution of the amount of evolution over the phylogeny. Our conclusions about the effect of stemminess might then, to some extent, be taken as conclusions about the effect of variability of evolutionary rates. A possibly serious shortcoming of this view as an evolutionary model, however, is that it implies that the terminal edges of the tree are constrained to have just that amount of evolution that will equalize the average total amount of evolution leading to each terminal lineage.

Second, our model probably simulates convergence less effectively than it simulates divergence. Like other models to date (Raup and Gould, 1974; Astolfi et al., 1981; Tateno et al., 1982; Nei et al., 1983), ours assumes complete independence of character evolution, and thus provides little insight into the effect of convergence in large suites of characters, such as might be produced by directional selection for functional adaptation (Raup and Gould, 1974). The convergence generated by our simulation is essentially just uniform noise due to independent multiple origins of character states, whereas adaptive selection should produce correlated multiple origins of character states.

It would be interesting to attempt to incorporate solutions to these problems into the simulation. Their addition might well have the effect of increasing the differences in accuracy among methods, perhaps in favor of cladistic methods. For example, we speculate that the lack of correlated convergence in the model underlies the general similarity of accuracy among all three phylogenetic inference methods used. Presumably, all make roughly equally good use of divergence, but phenetic clustering should be more easily misled by convergent similarity. However, we expect that any effects of improving the realism of the model would take the form of differences in deterioration of performance rather than differences in improvement of performance, and, given the generally poor accuracy of the estimated phylogenies obtained under the model as it is, it seems that such differences would be of little practical import.

Acknowledgments

We thank J. G. Felsenstein, G. Hart, F. J. Rohlf, and B. A. Thomson for their comments on the manuscript, and N. Oden for suggesting the probability model for variation of rate of character change.

Contribution No. 519 in Ecology and Evolution from the State University of New York at Stony Brook. Research supported by grant no. BSR 8306004 from the National Science Foundation to R.R.S.

Literature Cited

Astolfi, P., K. K. Kidd, and L. L. Cavalli-Sforza. 1981. A comparison of methods for reconstructing evolutionary trees. Syst. Zool. 30:156–169.

Baum, B. R. 1983. Relationships between transformation series and some numerical cladistic methods at the infraspecific level, when genealogies are known, pp. 340–345. *In* J. Felsenstein (ed.), Numerical Taxonomy: Proceedings of a NATO Advanced Study Institute. Springer-Verlag, Berlin.

Baum, B. R., and G. F. Estabrook. 1978. Application of compatibility analysis in numerical cladistics at the infraspecific level. Can. J. Bot. 56:1130–1135.

Colless, D. H. 1970. The phenogram as an estimate of phylogeny. Syst. Zool. 19:352–362.

———. 1980. Congruence between morphometric and allozyme data for *Menidia* species: a reappraisal. Syst. Zool. 29:288–299.

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

Fiala, K. L. 1983. A simulation model for comparing numerical taxonomic methods, pp. 87–91. *In* J. Felsenstein (ed.), Numerical Taxonomy: Proceedings of a NATO Advanced Study Institute. Springer-Verlag, Berlin.

Hull, C. H., and N. H. Nie. 1981. SPSS Update 7–9. McGraw-Hill, N.Y.

Nei, M., F. Tajima, and Y. Tateno. 1983. Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. J. Molec. Evol. 19:153–170.

Nelson, G. 1979. Cladistic analysis and synthesis: Principles and definitions, with a historical

note on Adanson's *Familles des Plantes* (1763–1764). Syst. Zool. 28:1–21.

RAUP, D. M., AND S. J. GOULD. 1974. Stochastic simulation and evolution of morphology—towards a nomothetic paleontology. Syst. Zool. 23:305–322.

RAUP, D. M., S. J. GOULD, T. J. M. SCHOPF, AND D. S. SIMBERLOFF. 1973. Stochastic models of phylogeny and the evolution of diversity. J. Geol. 81:525–542.

ROBINSON, D. F., AND L. R. FOULDS. 1981. Comparison of phylogenetic trees. Math. Biosci. 53: 131–147.

ROHLF, F. J. 1982. Consensus indices for comparing classifications. Math. Biosci. 59:131–144.

ROHLF, F. J., J. KISHPAUGH, AND D. KIRK. 1980. NTSYS. Numerical taxonomy system of multivariate statistical programs. Tech. Rep. State Univ. of New York, Stony Brook.

ROHLF, F. J., AND R. R. SOKAL. 1980. Comments on taxonomic congruence. Syst. Zool. 29:97–101.

SOKAL, R. R. 1983*a*. A phylogenetic analysis of the Caminalcules. I. The data base. Syst. Zool. 32:159–184.

———. 1983*b*. A phylogenetic analysis of the Caminalcules. II. Estimating the true cladogram. Syst. Zool. 32:185–201.

SOKAL, R. R., AND F. J. ROHLF. 1981. Comparing numerical taxonomic studies. Syst. Zool. 30:459–490.

STANLEY, S. M. 1979. Macroevolution: Pattern and Process. Freeman, San Francisco.

TATENO, Y., M. NEI, AND F. TAJIMA. 1982. Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. J. Molec. Evol. 18:387–404.

Corresponding Editor: W. R. Atchley