

# Combining data in phylogenetic analysis

John P. Huelsenbeck, J.J. Bull and Clifford W. Cunningham

**A**s we learn more about evolutionary processes, it is becoming clear that all data in phylogenetic analysis are not equivalent. For a nucleotide sequence chosen from some organism, individual sites may evolve differently from neighboring sites depending on codon position, functional constraint, or whether or not the site is part of a gene<sup>1-4</sup>. We know, for example, that third codon positions typically evolve at a higher rate than first and second codon positions because of stronger selection against change in the first and second positions. Similarly, and more obviously, pseudogenes (defective copies of genes, which are not transcribed) evolve under different evolutionary processes than their coding paralogs because coding regions experience more constraints than pseudogenes.

Although biologists now recognize that different traits may evolve under different rules, there is still a lack of consensus on how to accommodate this heterogeneity in a phylogenetic analysis. If we can recognize which sets of characters evolve under different rules and can partition the data accordingly, is it best to ignore the partition? Alternatively, should the analysis acknowledge and accommodate the partitions? In short, the question remains whether it is best to (1) always combine partitions, (2) never combine them, or (3) perform a test to decide whether or not to combine them. Box 1 summarizes some of the advantages and disadvantages of the methods (which are described and discussed below) proposed to date.

## Always combine

Kluge proposed that phylogenetic analysis should always be performed using all the evidence (the 'total evidence' approach)<sup>5</sup>. In particular, all of the independent characters available to the systematist should be combined and then analyzed using parsimony. Kluge extended the total evidence argument to include not only character data but also taxa. That is, all of the available taxa (e.g. fossil and living) should be combined in a phylogenetic analysis, as well as all of the available characters (Fig. 1a).

Much of Kluge's justification for total evidence was based on philosophical claims<sup>5</sup>; the total evidence solution is sought because it maximizes the 'informativeness' and 'explanatory power' of the character data used in the analysis. However, many systematists may be more interested in a statistical justification for a method, and the total evidence approach can indeed be justified statistically in many cases. For example, it is often argued that different data (e.g. genes) that are evolving at different rates may interact positively to resolve different levels of a phylogenetic tree<sup>6</sup>. A slowly evolving gene might be useful in resolving older evolutionary splits but be of little use for younger groups, whereas rapidly evolving

**Systematists have access to multiple sources of character information in phylogenetic analysis. For example, it is not unusual to have nucleotide sequences from several different genes, or to have molecular and morphological data. How should diverse data be analyzed in phylogenetic analysis? Several methods have been proposed for the treatment of partitioned data: the total evidence, separate analysis, and conditional combination approaches. Here, we review some of the advantages and disadvantages of the different approaches, with special concentration on which methods help us to discern the evolutionary process and provide the most accurate estimates of phylogeny.**

---

John Huelsenbeck is at the Dept of Integrative Biology, University of California, Berkeley, CA 94720-3140, USA; J.J. Bull is at the Dept of Zoology, University of Texas, Austin, TX 78712, USA; Clifford Cunningham is at the Zoology Dept, Duke University, Durham, NC 27708-0325, USA.

---

genes might be best for accurately resolving recent speciation events. (This argument has often been suggested, but we know of no study that tests its validity.) A more obvious case for total evidence applies when a phylogenetic method is consistent, which means that it converges to the correct phylogenetic tree as more data are included in a phylogenetic analysis. This means that as we add new characters to our analysis (e.g. data from different partitions), the probability of estimating the correct tree converges to 1.0 (Fig. 2). The relationship between number of characters and probability of correctly estimating phylogeny (shown in Fig. 2) corresponds to our intuition about how phylogenetic methods *should* work and justifies a type of total evidence approach when the method is consistent.

## Separate analysis

Miyamoto and Fitch took a position distinct from the total evidence approach<sup>7</sup>. They argued that partitions should not be combined before estimating the phylogenetic tree. Rather, trees should be estimated separately from each partition and the different estimates should be compared using taxonomic congruence (Fig. 1b). Miyamoto and Fitch argued that by performing separate analyses of partitioned data the systematist gains important information that would otherwise be unavailable if the data were combined a priori. The main advantage of the separate analysis approach is that data from different partitions (e.g. genes) are more likely to be independent than characters drawn from the same data partition. The lack of independence among characters increases the sampling variance and thereby increases the chance of erroneous estimates of phylogeny<sup>7,8</sup>. Under the separate analysis approach, each partition represents an independent estimate of the tree, and trees estimated from separate data sets can be used to corroborate certain groupings of taxa. It is often argued that congruence among different data partitions provides some of the strongest evidence that a particular phylogenetic estimate is accurate<sup>9,10</sup>.

Miyamoto and Fitch left open the matter of how to test or define partitions but they proposed several possibilities. For example, one might choose to partition data according to differences in evolutionary rate, linkage, function, or other criteria. We think that a difficulty with this approach is that the trees estimated from each data partition are based on a smaller sample of characters than in the combined analysis and will be more subject to sampling error than combined analyses (whenever the method is consistent for each data partition). Importantly, the separate analysis approach does not distinguish between cases in which sampling error is responsible for the different phylogenetic trees and cases in which the conflicting estimates stem from fundamentally

different evolutionary mechanisms underlying the data. If the different trees result merely from stochastic variation in realized evolutionary changes, then combining the data should actually improve the estimate. An unresolved question remains whether or not the additional information gained from separate analyses outweighs the problem of reduced phylogenetic accuracy.

### Conditional data combination

Bull *et al.*, Rodrigo *et al.* and de Queiroz presented an alternative approach to the treatment of data partitions that proposes that partitions be subjected to a statistical test of 'homogeneity' (Fig. 1c)<sup>8,11,12</sup>. Heterogeneous data partitions are those that result in significantly different estimates of phylogeny (differences beyond that expected from sampling error) when analyzed separately. If the test result is non-significant, then the data should be combined. This method is similar to the separate analysis method except that the criterion for separating data sets is different. Instead of any difference in evolutionary process between data partitions justifying separate analysis, only those differences in process influencing the estimated trees are used. In essence, the conditional combination approach advocates total evidence except when the data are shown to be incongruent.

Under what circumstances would the systematist want to avoid data combination? Bull *et al.* point out some situations in which combining data can lead to a less accurate estimate of phylogeny than separate analysis of the data<sup>11</sup>. It is now well known that many methods of phylogenetic inference will fail when certain conditions are met; that parsimony, for example, will converge to the wrong tree under many (but specific) conditions (Box 2). What does this inconsistency have to do with combining data? Consider the following example in which two data partitions differ only in their rate of evolution. In this example, the true phylogeny is of the form shown in Fig. 3, and the lengths of the branches are proportional to time. The parsimony method is known to be inconsistent for the tree of Fig. 3 when rates of evolution are high and the characters evolve as a clock. If one of the data partitions is evolving at a *low* rate, then as more data are added to that partition, we would expect to estimate the correct tree using parsimony. If, on the other hand, the characters added to the data partition evolve at a *high* rate, then as more characters are added, we would expect to converge on an incorrect tree. So adding high-rate characters does not help in this case.

With the conditional data combination approach, the data are combined only if trees estimated from different data partitions are not significantly different. In other words, a statistical test is performed on the separate data partitions with the question being 'Could the observed variation in resulting trees be due to stochastic variation in the characters?' Currently, we are developing and evaluating several different statistical tests, but these analyses are not complete.

Several criticisms have been leveled at the conditional combination approach<sup>13-16</sup>. One potential difficulty with this approach is that the appropriate partitions may be unrecognized. A variety of partitions affecting molecular evolution have been identified. For example, different partitions of character data may differ in phylogenetic history (the case of gene trees versus species trees), mode of transmission, coding versus non-coding regions, function, or mutation rate. However, some partitions may not affect phylogeny whereas other partitions that do affect topology may be undetected. There is no solution to this problem, except to keep searching for new biologically relevant partitions of data. However, when valid partitions go undetected, the

### Box 1. Summary of advantages and disadvantages of methods

#### Total evidence

##### Advantages

If data heterogeneity is rare, then the total evidence method is a good rule to adopt, because it will often lead to the best estimate.

##### Disadvantages

If data heterogeneity is common, then combining data may often give an erroneous answer and it further obscures the vagaries of evolution that need to be uncovered in order to improve phylogenetic estimation.

#### Conditional combination

##### Advantages

This method prevents combining heterogeneous data when the partitions have been correctly identified, hence reducing mistakes in phylogeny estimation. Also, conditional combination may result in combining data when appropriate, and otherwise tells us about meaningful conflicts. Hence, we discover something new about the evolutionary process.

##### Disadvantages

If data heterogeneity is very rare, then most cases of heterogeneity that will be detected by statistical tests will be false positives.

#### Separate analysis

##### Advantages

The chance of combining incompatible data sets into one reconstruction (data sets for which the phylogenetic method converges on significantly different trees) is reduced.

##### Disadvantages

Separate analysis does not distinguish between those cases in which combining partitions helps phylogenetic analysis and those cases in which it hinders phylogenetic analysis.

Phylogenies from partitioned data are subject to more sampling variation. This not only means that estimates from partitioned data may be more variable, but also that it may be more difficult to find the same taxonomic grouping in estimates from different partitions.

conditional combination approach errs on the side of total evidence.

There is also the difficult question of how to deal with conflicting phylogeny estimates from different data partitions (this criticism also applies to the separate analysis approach). How does one proceed when different data partitions are in conflict? In the worst case, a systematist would have two (or more) different data partitions, each of which provided a significantly different phylogenetic estimate. How should such a result be interpreted? Several potential resolutions to this problem might prove successful, though none has been evaluated rigorously. One possible resolution is to perform a consensus analysis on trees estimated from both data partitions. In this case, clades that are supported in both data partitions represent the best estimate of phylogeny, with no phylogenetic resolution being available for those clades in conflict. While consensus methods can be used to reconcile conflicting phylogenies<sup>10,17</sup>, they have been strongly criticized (see Refs 13,18). Another possible resolution to the problem is to reanalyze the data using a different (and presumably more realistic) phylogenetic model. The object is to reduce the conflict between the different data partitions<sup>11,19</sup>. For example, a reconstruction model that allows different nucleotide sites to evolve at different rates might lead to improved accommodation of heterogeneity over a method that treats all sites equally. Finally, one could leave the phylogeny of the group unresolved. This is analogous to the situation systematists face routinely with many equally parsimonious trees.

Probably the most serious potential problem with the conditional combination approach involves false positives – the failure to combine data sets that should have been combined. The conditional combination approach advocates the statistical testing of the null hypothesis that the differences among different data partitions could be caused by stochastic variation alone. Any statistical test has an

associated error (Type I and Type II errors). Type I error – the incorrect rejection of the null hypothesis – is typically set to 5%. If the actual incidence of data heterogeneity is much smaller than 5%, then many, if not most, examples of data heterogeneity that will actually be detected will be incorrect. This problem is directly analogous to clinical tests that are designed to detect disease in humans. However, this potential problem can be tested by looking at the distribution of *P* values obtained from tests of homogeneity.

In addition to the advantages of conditional combination in the context of improved phylogenetic estimation, this method also facilitates the discovery of new mechanisms of molecular evolution and their importance to phylogeny reconstruction.

**How can we test for data heterogeneity?**

Several possible tests exist for examining partitioned data, which is essential for the conditional combination paradigm. In general, these tests assess the different degrees of support for conflicting phylogenies. One test was described by de Queiroz, who suggested that the bootstrap values for various clades could form a basis for the combination of data partitions<sup>8</sup>. If there is high bootstrap support for conflicting clades, then the data are not combined before estimation. Instead, consensus techniques are used to summarize those portions of the separate estimates that agree. Bull *et al.* suggested that any number of existing methods can be used to test for data homogeneity. Rodrigo *et al.* proposed using the distance between the minimum length trees of each data set as a test statistic. Bootstrapping is used to determine the distribution of this statistic expected from chance<sup>12</sup>.

Another test that can be applied using parsimony employs a simple modification of the Mischevich–Farris index of incongruence among data sets<sup>20</sup>. The test statistic, *I*, is calculated as:

$$I = L_C - \sum_{i=1}^n L_i$$

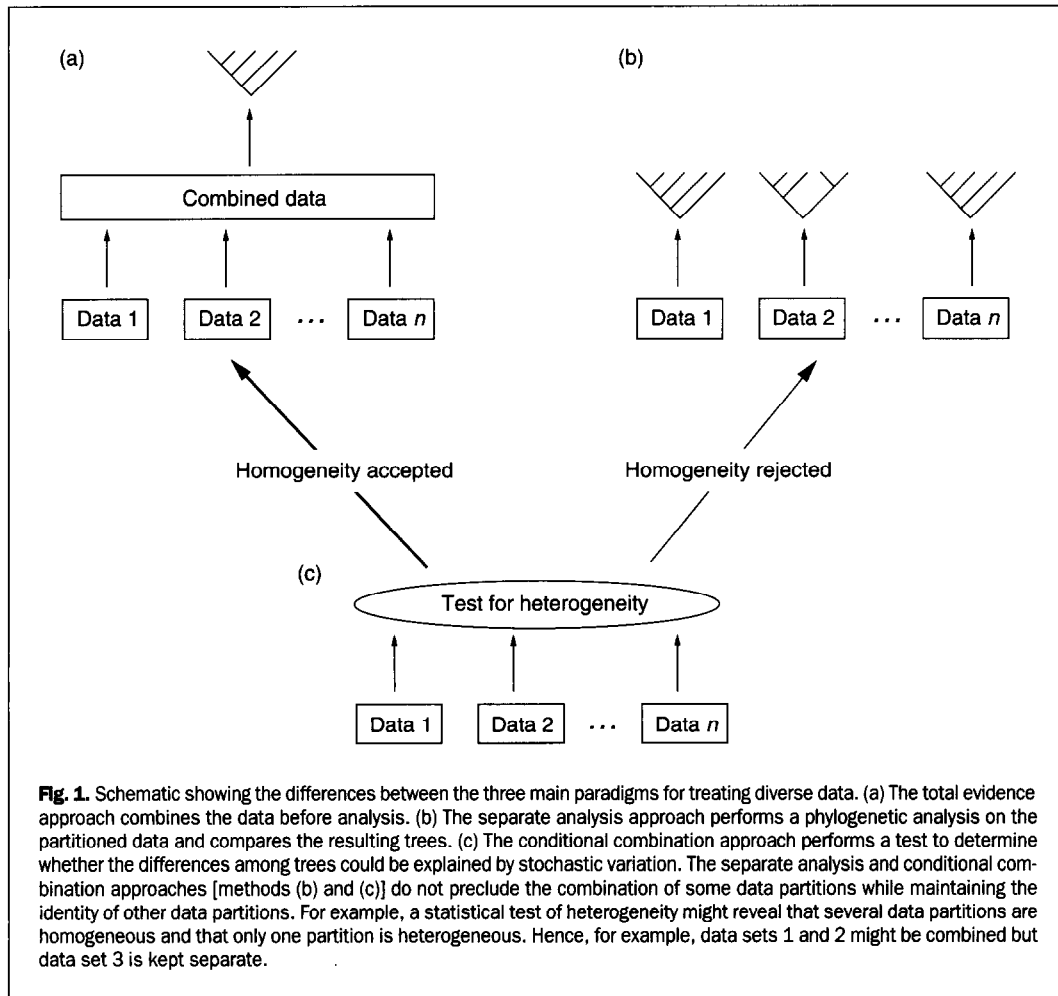
*L<sub>C</sub>* refers to the length of the most parsimonious tree from the combined analysis, whereas *L<sub>i</sub>* is the length of the most parsimonious tree on the *i*-th data set (out of *n* total). Farris *et al.* proposed a sampling method to determine the distribution of *I* values expected from chance alone<sup>21</sup>.

They proposed that the null distribution of *I* be determined by randomly partitioning characters from the combined data into new data partitions of the same size and number as was present originally. *I* is recalculated for each random data partition. If the original *I* is greater than 95% of the *I* values determined through random partitioning, then the null hypothesis is rejected. That is, the test would then indicate that there is more incongruence among the data sets than would be expected from chance alone. One unresolved question addresses the best approach for determining the null distribution of *I*. The approach considered here involves sampling without replacement. However, sampling with replacement (nonparametric bootstrapping) or simulation of the null distribution of *I* (parametric bootstrapping) might also be appropriate. PAUP 4.0 has implemented a version of this test.

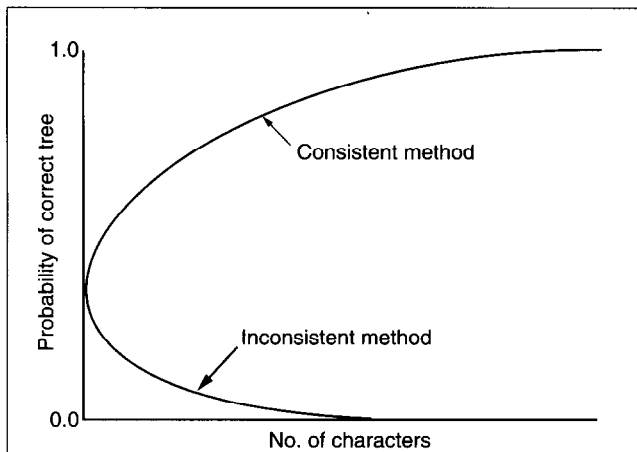
Huelsenbeck and Bull proposed a likelihood ratio test for data heterogeneity (the likelihood heterogeneity test)<sup>22</sup>. In essence, this test compares the likelihood obtained under the constraint that the same phylogeny underlies all of the data sets with the likelihood

obtained when this constraint is relaxed. The test statistic,  $\delta$ , is calculated as  $2(\ln L_1 - \ln L_0)$ . *L<sub>0</sub>* is the likelihood of the tree when the same tree is assumed to underlie all data partitions, whereas *L<sub>1</sub>* is the likelihood of the tree when different trees can underlie each data partition. The null distribution of  $\delta$  is calculated using simulation; new data partitions of the same size as the original are simulated under the null hypothesis. Parameters for the simulation are obtained from the maximum likelihood estimates of the parameters.

Although many other tests could potentially be applied to the partition problem, the properties of such tests are unknown. Ideally, tests would indicate combining data when combining data actually increases the accuracy of phylogenetic estimation. Conversely, the tests should not indicate data combination when in actuality it is worse to combine the data. The behavior of these tests could be examined using simulation or well-supported phylogenies<sup>23</sup>.



**Fig. 1.** Schematic showing the differences between the three main paradigms for treating diverse data. (a) The total evidence approach combines the data before analysis. (b) The separate analysis approach performs a phylogenetic analysis on the partitioned data and compares the resulting trees. (c) The conditional combination approach performs a test to determine whether the differences among trees could be explained by stochastic variation. The separate analysis and conditional combination approaches [methods (b) and (c)] do not preclude the combination of some data partitions while maintaining the identity of other data partitions. For example, a statistical test of heterogeneity might reveal that several data partitions are homogeneous and that only one partition is heterogeneous. Hence, for example, data sets 1 and 2 might be combined but data set 3 is kept separate.



**Fig. 2.** The relationship between number of characters and the probability of correctly estimating phylogeny for a consistent method and for an inconsistent method. The consistent method converges to the correct phylogeny as more data are analyzed, whereas the inconsistent method converges to an incorrect phylogeny.

**An example of data heterogeneity**

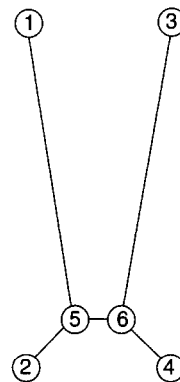
We applied the likelihood heterogeneity test, described above, to a controversial phylogenetic problem – the phylogeny of amniotes (mammals, birds, crocodylians and lizards). Over the past decade, phylogenetic analyses of morphological data and of at least 16 genes have provided different estimates of this phylogeny<sup>14,24–29</sup>. Usually, one of two trees is estimated (Trees 1 and 3 in Fig. 4). One is consistent with a bird–crocodylian relationship (birds and crocodylians are each other’s closest relatives), whereas the other is consistent with a bird–mammal relationship. Phylogenetic analyses that include fossils support the traditional hypothesis that birds and crocodylians form a monophyletic group to the exclusion of mammals<sup>25</sup>. Most of the debate has centered on which nucleotide data are the most reliable and on how to combine the different sources of data to provide an accurate estimate of phylogeny. What has not been addressed is whether the differences in the phylogenetic trees can be explained simply by stochastic variation.

The sequences of five genes [12S, 16S, 18S and 28S rRNA (ribosomal RNA), and tRNA (transfer RNA) (Valine) sequence] from four amniote taxa were examined: lizard (*Sceloporus undulatus*), alligator (*Alligator mississippiensis*), bird (*Gallus gallus*) and mammal (*Mus musculus*). The alignment of Hedges<sup>26</sup> was used with the exception that all sites with gaps or ambiguous states were removed<sup>26,29</sup>. The likelihood of each tree was calculated assuming two models of DNA substitution – the Jukes–Cantor, and Hasegawa, Kishino, Yano (HKY) models. The Jukes–Cantor model assumes a very simple process of DNA substitution in which each possible nucleotide transition has the same probability of change in any small time interval<sup>30</sup>. The HKY model is much more complicated (parameter rich) in that it allows for differences in equilibrium nucleotide frequencies as well as for a transition:transversion bias<sup>31</sup>. Furthermore, the implementation of the HKY model used here assumes rate heterogeneity among sites<sup>32</sup>. Each gene was treated as a separate data set. Table 1 shows the log likelihoods of the best trees for each gene under the two models of DNA substitution. Different genes lead to different estimates of phylogeny, with some genes providing a bird–mammal estimate and other genes providing a bird–crocodylian estimate. Both models of substitution recover the traditional phylogeny uniting birds and crocodylians for four out of six genes, although for

**Box 2. The inconsistency problem**

One of the more distressing properties a phylogenetic method can have is inconsistency. An inconsistent method will converge to an incorrect estimate as more data are added. What is more, as data are added, the confidence that will be placed in that incorrect grouping (whether assessed using bootstrapping or a randomization technique) will increase. How is it possible that a method, such as parsimony, can converge on an incorrect phylogenetic estimate as more characters are included?

Felsenstein first showed that parsimony can converge to an incorrect phylogeny when rates of evolution do not follow a molecular clock for the four-taxon case<sup>33</sup>. Consider the unrooted four-taxon tree shown below.



The internal branch and two opposing peripheral branches of this tree are very short, whereas the remaining two peripheral branches are very long. It is informative to imagine what character patterns would be expected at the tips of the tree as the three short branches tend to length zero.

Very few substitutions would be expected to occur along a very short branch, so the probability of actually observing a change along all three short branches is very small. Imagine that at point 2 of the tree, we have the nucleotide ‘A’. Because the three branches separating points 2 and 4 are very small (tending toward zero), this means that we can expect to observe an ‘A’ at points 4, 5 and 6 also. What are the possible nucleotides that we could observe at points 1 and 3? Remember that the branches leading to points 1 and 3 are long enough so that we can expect to see changes along these branches every so often. The possible states that could be observed at points 1 to 4 are as follows:

Pattern	Point 1	Point 2	Point 3	Point 4
1	A	A	A	A
2–4	A	A	C, G, or T	A
5–7	C, G, or T	A	A	A
8	C	A	G	A
9	G	A	C	A
10	A	A	T	A
11	T	A	A	A
12	C	A	T	A
13	T	A	C	A
14	C	A	C	A
15	G	A	G	A
16	T	A	T	A

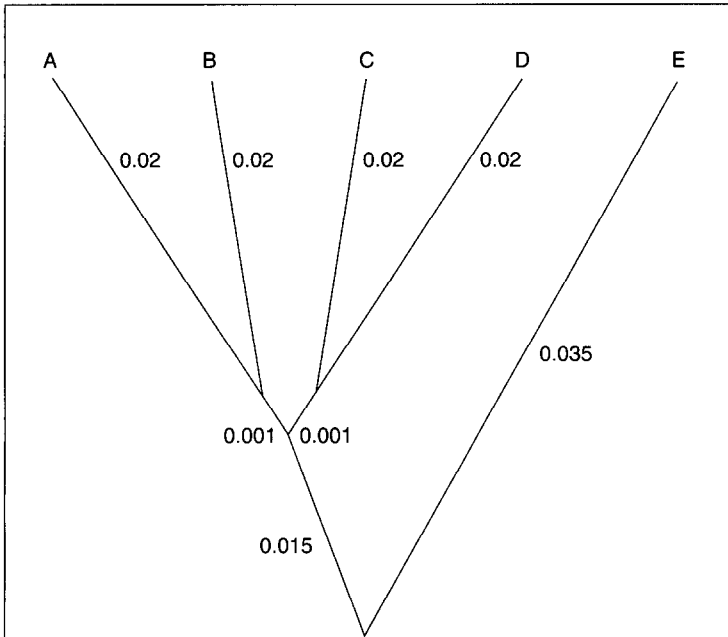
Pattern 1 corresponds to the case where no change is observed along the long branches leading to points 1 and 3. Patterns 2 to 7 are cases in which a change is observed along one long branch and patterns 8 to 16 are cases in which a change is observed along both long branches. Interestingly, the only patterns that parsimony considers informative (makes use of when deciding among possible phylogenies) are patterns 14 to 16. Unfortunately, these patterns are informative for the incorrect tree. In the worst-case scenario, every informative character is for the incorrect phylogeny!

The set of parameter conditions (branch lengths, transition:transversion bias, and so on) for which a method becomes inconsistent has been termed the ‘Felsenstein Zone’<sup>44</sup>. This phenomenon has also been described as ‘long-branch attraction’ because the observation in cases in which parsimony is inconsistent is that the long branches of the phylogeny are incorrectly joined together. Although for the four-taxon case, the parsimony method becomes inconsistent only when evolution does not follow a molecular clock, Hendy and Penny have shown that the method can become inconsistent even under clock-like conditions for more than four taxa<sup>45</sup>.

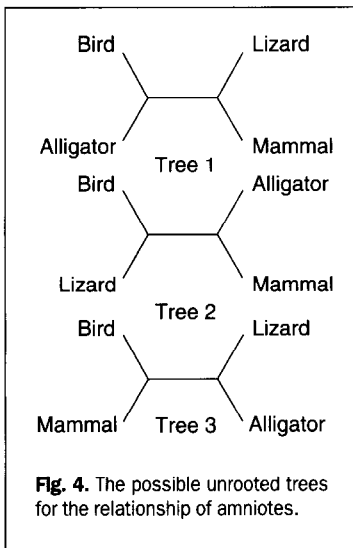
The implication of the Felsenstein Zone for data combination is that phylogenetic methods can fail when the method’s assumptions are not satisfied. It is quite possible that a phylogenetic method will be compatible with the evolutionary processes of one data partition but not with another. Combining the data from both partitions could cause an incorrect phylogeny to be estimated.

each substitution model the set of four genes was not identical. Only the 18S rRNA gene united birds and mammals under both substitution models.

How would the different methods for treating partitioned data handle this example? With the total evidence approach, the best phylogenetic estimate is obtained by first combining all of the genes into one data set and then performing the phylogenetic estimate. The total evidence approach yields support for the bird–crocodylian tree (Tree 1 from Fig. 4),



**Fig. 3.** A tree of five taxa (A–E) for which the parsimony method can converge to an incorrect estimate of phylogeny. The rates are in terms of number of substitutions per site. When the rates are as labeled on the tree, the parsimony method will converge to a correct phylogenetic estimate. However, when the rates of evolution are increased by a factor of five, the parsimony method will converge to an incorrect phylogenetic estimate.



**Fig. 4.** The possible unrooted trees for the relationship of amniotes.

which is the phylogenetic tree that is also obtained if fossils are included in the analysis<sup>14,25</sup>. Does inclusion of the conflicting phylogenetic signal from the 18S rRNA gene hurt phylogenetic reconstruction? In fact, bootstrap support for this tree is about ten percent lower than for the data without the 18S gene. Also, the total evidence approach does not provide any information regarding the extent of incompatibility among data sets. The separate analysis approach might treat each gene separately. There are several ways in which the trees estimated from each gene can be compared, with consen-

sus techniques being one of them. In the amniote case, the separate analysis approach would indicate conflict in the trees estimated using different genes. One possible interpretation of conflict is that we cannot have much confidence in the relationships of amniotes. Finally, with the conditional combination approach, a test is performed to determine whether the differences in phylogenies observed among the different genes is consistent with sampling variation. The likelihood heterogeneity values, in this example, were significant at  $P = 0.03$  for both models of DNA substitution, indicating the presence of conflicting phylogenetic signal among the genes at a level greater than expected purely from sampling error. When the 18S rRNA gene was excluded, the null hypothesis of no heterogeneity among genes is tentatively accepted ( $\delta = 1.09394$ ,  $P = 0.15$  for Jukes–Cantor model;  $\delta = 1.93994$ ,  $P = 0.24$  for HKY- $\Gamma$  model). This test suggests, therefore, that the 18S rRNA gene has been subject to different processes of molecular evolution than the other genes analyzed here, and that the differences in process lead to a different estimate of phylogeny. The conditional combination approach would combine the data from the 12S, 16S, 28S rRNA and tRNA genes because they do not conflict. We are left with two different estimates and some indication that the phylogenetic methods are failing for at least one of the data partitions (18S or the combination of the other genes).

What could be different about the 18S rRNA gene? Several mechanisms, including long branch attraction, horizontal gene transfer, ancestral polymorphism, paralogy, and convergence in nucleotide content between different lineages, have been suggested as possible causes of conflicting phylogenetic signal<sup>11,33–41</sup>. However, several of these possibilities are implausible for the amniote example. Horizontal gene transfer does not seem likely because the mechanisms known for transferring essential genes between vertebrates (e.g. hybridization) are confined to closely related organisms. Ancestral polymorphism is likewise implausible because gene conversion is known to homogenize rRNA sequences within populations<sup>42</sup>. Both of these explanations are further discounted because the 18S rRNA gene is part of a cluster with other nuclear rRNA genes (the 28S gene, which was included in this analysis). Long branch attraction and shifts in equilibrium nucleotide frequencies are possible explanations for the heterogeneity, though we note that phylogenetic analysis of the 18S rRNA gene with LogDet<sup>43</sup> distances – which corrects for shifts in equilibrium nucleotide frequency – still produces a bird–mammal estimate with this gene.

This analysis offers definitive evidence of different genes providing significantly different estimates of phylogeny in

higher organisms. This result is unique in that it is not likely to be explained by horizontal gene transfer or by ancestral polymorphism. Heterogeneity of this sort has a profound impact on the larger realm of phylogenetic analysis (and fields that use phylogenies such as epidemiology) because it suggests that the models of DNA substitution used are making mistakes by failing to capture the relevant information about molecular evolution. Identifying heterogeneity is thus an important step in improving these

**Table 1. Log likelihoods of trees under the Jukes–Cantor and HKY- $\Gamma$  models of DNA substitution for five genes**

Model	Tree	Gene				
		12S rRNA	16S rRNA	18S rRNA	28S rRNA	tRNA (Val.)
JC	1	<b>-2451.37</b>	<b>-3603.93</b>	-2089.59	<b>-447.30</b>	-223.98
	2	-2458.43	-3623.90	-2091.62	-454.53	-224.27
	3	-2453.23	-3628.92	<b>-2072.38</b>	-454.53	<b>-223.43</b>
$\delta_1 = 2 [(-8798.43) - (-8816.19)] = 35.52 (P = 0.03)$						
HKY- $\Gamma$	1	-2357.30	<b>-3487.60</b>	-2058.73	<b>-432.02</b>	<b>-205.38</b>
	2	-2358.10	-3497.91	-2058.81	-434.70	-205.46
	3	<b>-2356.33</b>	-3498.20	<b>-2054.67</b>	-434.70	-205.43
$\delta_2 = 2 [(-8536.03) - (-8541.06)] = 10.06 (P = 0.03)$						

models. If significant heterogeneity in tree estimates is widespread, systematists need to reconsider their methods of analysis as well as the accuracy of their trees.

### Conclusion

Although it is widely understood that data from different partitions often evolve differently, the influence that evolutionary process has on phylogenetic reconstruction is not generally appreciated. The fit between the assumptions made in a phylogenetic analysis and the evolutionary processes that generated the character data are centrally important in phylogenetic analysis. When the assumptions we make are unacceptably false, many phylogenetic methods fail (become inconsistent). Theoretically and in practice, phylogenetic analysis of data from different genes can provide different and well-corroborated estimates. That is, analysis of one data set provides one highly corroborated phylogeny whereas analysis of another data set provides a different highly corroborated phylogeny.

Three different approaches can be taken to resolve differences in phylogenetic estimates from different data partitions – the total evidence, separate analysis, and conditional combination approaches. Interestingly, one can view these methods as adopting different rejection criteria for the null hypothesis that the differences among phylogenies estimated from different partitions is due to stochastic variation. With the total evidence approach, no amount of evidence will reject the null hypothesis, with the separate analysis approach, no amount of evidence will accept the null hypothesis, whereas with the conditional combination approach, some predetermined level of heterogeneity among different data sets will reject the null hypothesis. Hence, the conditional combination approach is a hybrid between the total evidence and separate analysis methods.

Several questions about the treatment of diverse data remain unresolved. For example, the performance of the heterogeneity tests has not been thoroughly examined. We have little idea how powerful these tests are at detecting heterogeneity. Also, these tests have not been applied on a wide scale, so we have little idea how prevalent significant differences among data partitions are in nature. Addressing these two questions should go a long way toward resolving which method of data analysis is generally the best. If data heterogeneity is very rare, then total evidence would be a good rule of thumb to use in phylogenetic analysis. If, on the other hand, significant heterogeneity among data partitions is very common, then the separate analysis or conditional combination approaches would be the best methods to use.

### References

- Harvey, P.H. and Pagel, M.D. (1991) *The Comparative Method in Evolutionary Biology*, Oxford University Press
- Wolfe, K.H. *et al.* (1988) Mutation rates differ among regions of the mammalian genome, *Nature* 337, 283–285
- Luo, C.-C. *et al.* (1989) Structure and expression of dog apolipoprotein A-I, E, and C-I mRNAs: Implications for the evolution and functional constraints of apolipoprotein structure, *J. Lipid Res.* 30, 1735–1746
- Stewart, C.-B. and Wilson, A.C. (1987) Sequence convergence and functional adaptation of stomach lysozymes from foregut fermenters, *Cold Spring Harbor Symp. Quant. Biol.* 52, 891–899
- Kluge, A.G. (1989) A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes), *Syst. Zool.* 38, 7–25
- Hillis, D.M. (1987) Molecular versus morphological approaches to systematics, *Annu. Rev. Ecol. Syst.* 18, 23–42
- Miyamoto, M.M. and Fitch, W.M. (1995) Testing species phylogenies and phylogenetic methods with congruence, *Syst. Biol.* 44, 64–76
- de Queiroz, A. (1993) For consensus (sometimes), *Syst. Biol.* 42, 368–372
- Penny, D. and Hendy, M.D. (1986) Estimating the reliability of evolutionary trees, *Mol. Biol. Evol.* 3, 403–417
- Swofford, D.L. (1991) When are phylogeny estimates from molecular and morphological data incongruent? in *Phylogenetic Analysis of DNA Sequences* (Miyamoto, M.M. and Cracraft, J., eds), pp. 295–333, Oxford University Press
- Bull, J.J. *et al.* (1993) Partitioning and combining data in phylogenetic analysis, *Syst. Biol.* 42, 384–397
- Rodrigo, A.G. *et al.* (1993) A randomization test of the null hypothesis that two cladograms are sample estimates of a parametric phylogenetic tree, *N. Z. J. Bot.* 31, 257–268
- Chippindale, P.T. and Wiens, J.J. (1994) Weighting, partitioning, and combining characters in phylogenetic analysis, *Syst. Biol.* 43, 278–287
- Eernisse, D.J. and Kluge, A.G. (1993) Taxonomic congruence versus total evidence, and the phylogeny of amniotes inferred from fossils, molecules and morphology, *Mol. Biol. Evol.* 10, 1170–1195
- Jones, T.R. *et al.* (1993) When theories and methodologies clash: a phylogenetic reanalysis of the North American ambystomatid salamanders (Caudata: Ambystomatidae), *Syst. Biol.* 42, 92–102
- Kluge, A.G. and Wolf, A.J. (1993) Cladistics: what's in a word? *Cladistics* 9, 183–199
- Lanyon, S.M. (1993) Phylogenetic frameworks: towards a firmer foundation for the comparative approach, *Biol. J. Linn. Soc.* 49, 45–61
- Barrett, M. *et al.* (1991) Against consensus, *Syst. Biol.* 40, 486–493
- Miyamoto, M.M. *et al.* (1994) A congruence test of reliability using linked mitochondrial DNA sequences, *Syst. Biol.* 43, 236–249
- Michevich, M.F. and Farris, J.S. (1981) The implications of congruence in *Menidia*, *Syst. Zool.* 30, 351–370
- Farris, J.S. *et al.* (1995) Testing significance of incongruencies, *Cladistics* 10, 315–319
- Huelsenbeck, J.P. and Bull, J.J. (1996) A likelihood ratio test for detection of conflicting phylogenetic signal, *Syst. Biol.* 45, 92–98
- Graybeal, A. (1994) Evaluating the phylogenetic utility of genes: a search for genes informative about deep divergences among vertebrates, *Syst. Biol.* 43, 174–193
- Gardiner, B.G. (1982) Tetrapod classification, *Zool. J. Linn. Soc.* 74, 207–232
- Gauthier, J. *et al.* (1988) Amniote phylogeny and the importance of fossils, *Cladistics* 4, 105–209
- Hedges, S.B. *et al.* (1990) Tetrapod phylogeny inferred from 18S and 28S ribosomal RNA sequences and a review of the evidence for amniote relationships, *Mol. Biol. Evol.* 7, 607–633
- Løvtrup, S. (1985) On the classification of the taxon Tetrapoda, *Syst. Zool.* 34, 463–470
- Hedges, S.B. and Maxson, L.R. (1991) Pancreatic polypeptide and the sister group of birds, *Mol. Biol. Evol.* 8, 888–891
- Hedges, S.B. (1994) Molecular evidence for the origin of birds, *Proc. Natl Acad. Sci. USA* 91, 2621–2624
- Jukes, T.H. and Cantor, C.R. (1969) Evolution of protein molecules, in *Mammalian Protein Metabolism* (Munro, H., ed.), pp. 21–132, Academic Press
- Hasegawa, M. *et al.* (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA, *J. Mol. Evol.* 22, 160–174
- Yang, Z. (1993) Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites, *Mol. Biol. Evol.* 10, 1396–1401
- Felsenstein, J. (1978) Cases in which parsimony or compatibility methods will be positively misleading, *Syst. Zool.* 27, 401–410
- Dykhuizen, D.E. and Green, L. (1991) Recombination in *Escherichia coli* and the definition of biological species, *J. Bacteriol.* 173, 7257–7268
- Maynard Smith, J. *et al.* (1991) Localized sex in bacteria, *Nature* 349, 29–31
- Médigue, C. *et al.* (1991) Evidence for horizontal gene transfer in *Escherichia coli* speciation, *J. Mol. Biol.* 222, 851–856
- Souza, V.T. *et al.* (1992) Hierarchical analysis of linkage disequilibrium in *Rhizobium* populations: Evidence for sex? *Proc. Natl Acad. Sci. USA* 89, 8389–8393
- Valdez, A.M. and Piñero, D. (1992) Phylogenetic estimation of plasmid exchange in bacteria, *Evolution* 46, 641–656
- Wilson, A.C. *et al.* (1977) Biochemical evolution, *Annu. Rev. Biochem.* 46, 473–639

- 40 Pesole, G. *et al.* (1991) **The branching order of mammals: Phylogenetic trees inferred from nuclear and mitochondrial molecular data**, *J. Mol. Evol.* 33, 537–542
- 41 Doyle, J.J. (1992) **Gene trees and species trees: Molecular systematics as one-character taxonomy**, *Syst. Bot.* 17, 144–163
- 42 Hillis, D.M. *et al.* (1991) **Evidence for biased gene conversion in concerted evolution of ribosomal DNA**, *Science* 251, 308–310
- 43 Lockhart, P.J. *et al.* (1994) **Recovering evolutionary trees under a more realistic model of sequence evolution**, *Mol. Biol. Evol.* 11, 605–612
- 44 Huelsenbeck, J.P. and Hillis, D.M. (1993) **Success of phylogenetic methods in the four-taxon case**, *Syst. Biol.* 42, 247–264
- 45 Hendy, M. and Penny, D. (1989) **A framework for the quantitative study of evolutionary trees**, *Syst. Zool.* 38, 297–309

# The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance

John Wakeley

**T**he four bases of DNA are classified by their structures into purines and pyrimidines (see Fig. 1). Nucleotide substitutions within each structural class (transitions) occur with greater frequency than changes between structural classes (transversions). This phenomenon is called transition bias and was discovered when the first comparisons of molecular sequences were made<sup>1,2</sup>. It is now recognized as a general property of DNA-sequence evolution, having been observed in nuclear, mitochondrial and chloroplast DNA, and in prokaryotes, eukaryotes and viruses. Transition bias has been found in the DNA sequences of pseudogenes and functional genes<sup>3,4</sup>, in transfer RNA (tRNA) (Ref. 5), in ribosomal RNA (rRNA) (Ref. 6) and in the functional but non-coding mitochondrial control region<sup>7</sup>. While always present, transition bias appears more pronounced in animal mitochondrial<sup>8</sup> than in nuclear<sup>3,4</sup> or chloroplast<sup>9</sup> DNA.

Studies of transition bias are valuable for two reasons. First, estimates of the pattern of nucleotide substitution are important to our understanding of DNA-sequence evolution. Knowledge of transition bias facilitates inferences about mutational patterns and about the type and strength of natural selection. Second, reliable estimates of transition bias are important to evolutionary-distance correction methods (see Goldstein and Pollock<sup>10</sup> for a recent example). Changes in estimates of transition bias can substantially alter distance corrections<sup>11</sup>. Corrected distances are used as input into phylogenetic-tree-building algorithms and to estimate divergence times.

## Models of nucleotide substitution

One cannot rightly discuss transition bias without invoking models of nucleotide substitution, and many of these have

**Estimates of transition bias provide insight into the process of nucleotide substitution, and are required in some commonly used phylogenetic methods. Transitions are favored over transversions among spontaneous mutations, and the direction and strength of selection on proteins and RNA appears to depend on mutation type. As the complexity of the nucleotide-substitution process has become apparent, problems with classical methods of estimating transition bias have been recognized. These problems arise because there is a fundamental difference between ratios of numbers of differences among sequences and ratios of rates, and because classical methods are not easily generalized. Several new methods are now available.**

John Wakeley is at the Dept of Biological Sciences, Nelson Biological Labs, PO Box 1059, Busch Campus, Rutgers University, Piscataway, NJ 08855-1059, USA.

been proposed. The definition of transition bias adopted here is the one used by most workers and is easily applied to all substitution models. Transition bias is measured by the ratio of the overall rate of transitions ( $T$ ) to the overall rate of transversions ( $V$ ). This quantity is referred to here as the  $T:V$  rate ratio.

Jukes and Cantor<sup>12</sup> introduced the first and simplest substitution model. Since all changes are considered equally likely, this model contains no transition bias. However, the  $T:V$  rate ratio is 1:2, simply because there are twice as many transversions as transitions (see Fig. 1). Transition bias is indicated when the  $T:V$  rate ratio is greater than 1:2, and many substitution models incorporate this possibility explicitly. Kimura's<sup>13</sup> two-parameter model is a model of only transition bias. Transitions happen at rate  $\alpha$  and transversions at rate  $\beta$ , so the  $T:V$  rate ratio is

$\alpha:(2\beta)$ . When  $\alpha = \beta$ , this model reduces to that of Jukes and Cantor, and when  $\alpha > \beta$ , there is transition bias.

As the complexities of DNA-sequence change were elucidated, new models were introduced in an attempt to capture something more of the reality of nucleotide substitution. Almost all of these models include transition bias. Box 1 shows three currently used models and gives expressions for the  $T:V$  rate ratio for each.

## The causes of transition bias

In the same year as their proposal of the structure of DNA, Watson and Crick<sup>14</sup> suggested a mechanism of point mutation that immediately favors transitions. Their idea was that while the maintenance of the double helix demanded that purines always paired with pyrimidines and vice versa, mutations could occur if disfavored tautomeric forms of the