Syst. Biol. 60(6):833–844, 2011

DOI:10.1093/sysbio/syr064
Advance Access publication on July 29, 2011

A Method for Inferring the Rate of Evolution of Homologous Characters that Can Potentially Improve Phylogenetic Inference, Resolve Deep Divergence and Correct Systematic Biases

CARLA A. CUMMINS AND JAMES O. MCINERNEY*

*Molecular Evolution and Bioinformatics Unit, Department of Biology, National University of Ireland, Maynooth, Co. Kildare, Ireland; *Correspondence to be sent to: Molecular Evolution and Bioinformatics Unit, Department of Biology, National University of Ireland, Maynooth, Co. Kildare, Ireland; E-mail: james.o.mcinerney@nuim.ie.*

Received 3 September 2010; reviews returned 14 February 2011; accepted 2 March 2011
Associate Editor: Michael Charleston

Abstract.—Current phylogenetic methods attempt to account for evolutionary rate variation across characters in a matrix. This is generally achieved by the use of sophisticated evolutionary models, combined with dense sampling of large numbers of characters. However, systematic biases and superimposed substitutions make this task very difficult. Model adequacy can sometimes be achieved at the cost of adding large numbers of free parameters, with each parameter being optimized according to some criterion, resulting in increased computation times and large variances in the model estimates. In this study, we develop a simple approach that estimates the relative evolutionary rate of each homologous character. The method that we describe uses the similarity between characters as a proxy for evolutionary rate. In this article, we work on the premise that if the character-state distribution of a homologous character is similar to many other characters, then this character is likely to be relatively slowly evolving. If the character-state distribution of a homologous character is not similar to many or any of the rest of the characters in a data set, then it is likely to be the result of rapid evolution. We show that in some test cases, at least, the premise can hold and the inferences are robust. Importantly, the method does not use a "starting tree" to make the inference and therefore is tree independent. We demonstrate that this approach can work as well as a maximum likelihood (ML) approach, though the ML method needs to have a known phylogeny, or at least a very good estimate of that phylogeny. We then demonstrate some uses for this method of analysis, including the improvement in phylogeny reconstruction for both deep-level and recent relationships and overcoming systematic biases such as base composition bias. Furthermore, we compare this approach to two well-established methods for reweighting or removing characters. These other methods are tree-based and we show that they can be systematically biased. We feel this method can be useful for phylogeny reconstruction, understanding evolutionary rate variation, and for understanding selection variation on different characters. [Compatibility; maximum likelihood; maximum parsimony; molecular phylogeny reconstruction; site rate variation; site removal; simulation; systematic bias.]

Homologous characters evolve at different rates. Within a given data matrix, some characters might evolve at an appropriate rate to resolve the branching order of the taxa in question (Townsend 2007) whereas others might exhibit high levels of homoplastic noise. Some might be too slowly evolving and therefore mute with respect to phylogenetic statements (Kluge and Farris 1969; Delsuc et al. 2005; Philippe et al. 2005; Townsend 2007). A character could be considered important if it contains useful information about the phylogeny of the group of interest and if it is relatively free of homoplasy for that group. Therefore, for deep phylogenetic relationships, a slowly evolving character might prove useful, whereas for shallower relationships, a more rapidly evolving character could prove to be more useful. Character-state substitution rate (i.e., the rate at which a characters state is transformed into a different state) is an important factor to consider when ranking the informativeness of characters. Knowing a priori the rate of evolution of a character can greatly facilitate the treatment of characters for phylogeny reconstruction.

A number of efforts have been made to evaluate character-specific evolutionary rates. Farris (1969) introduced successive approximations character weighting (SACW) in order to weight characters according to a perceived importance assigned to them. This weighting scheme sought to ensure that characters with a higher degree of correlation with the phylogenetic history were more highly regarded during reconstructions. Farris defined this correlation as the consistency index (CI) for a matrix, or the goodness of the fit of the characters within the matrix to a given tree. The CI for an individual character on a particular tree is derived as the minimum possible character length divided by the observed character length on the considered tree. So, when a character fits on a tree without apparent homoplasy, the CI value is unity. If additional ad hoc hypotheses need to be invoked to explain the evolution of the character on the tree, then the CI value will be less than one (Farris 1969). The CI for a data matrix is obtained by averaging the CI values for all the characters in the matrix. Therefore, a tree must be initially inferred. In his description of the method, Farris preweighted characters according to a weighting system devised by Le Quesne (1969), though he indicated that initial character weights set to unity would also work. As a consequence of the approach, characters that tend to disagree with the initial tree are given a lower weighting in subsequent analyses, in contrast to characters that tend to agree with this initial tree, whose weight remains high.

In the late 1980s, Olsen (1987) noted that among-site rate variation (ASRV) could cause problems in phylogenetic inference, and he attempted to accommodate this variation using a model-based approach that employed

a normal distribution. Using a model to account for rate variation across sites can increase the probability of finding the correct phylogenetic tree topology compared with a method that does not account for rate variation (Yang 1993). By using an evolutionary model that neglects to account for ASRV, sequences will appear to have undergone fewer mutations overall and will appear to be more similar to their relatives compared with an analysis using a model that accounts for ASRV. Therefore, much of the effort to improve phylogeny reconstruction accuracy has focused on methods that deal with accommodating site rate heterogeneous data (Farris 1969; Yang 1996; Brinkmann and Philippe 1999; Hirt et al. 1999; Schmidt et al. 2002).

Yang (1996) modeled ASRV using the gamma distribution. This distribution has some attractive properties, particularly given that its shape can change from being L shaped to being hill shaped, depending on the characteristics of the alignment. Again, this approach tries to incorporate rate variation and it assumes that site rate heterogeneity is well approximated by this model. However, assuming that all sites are free to vary will lead to incorrect estimations when there are sites in the data set that do not or cannot change (Yang 1996). In 1970, Fitch and Markowitz (1970) proposed that for a protein there might be two classes of sites—invariable and variable and they suggested a method of analyzing molecular alignments in order to determine how many positions were invariable and how many were variable. These invariable sites can also confound phylogeny reconstruction and accentuate rate variation across sites. To overcome these issues, some studies have experimented with the removal of sites that violate assumptions of the models that are being used. This has the effect of reducing the range of site-to-site rate variation in the data set.

As an example of a study that effectively reduced site-to-site rate variation, Hirt et al. (1999) not only removed invariant sites, but also removed sites they considered to be fast evolving (Hirt et al. 1999). They identified fast-evolving sites by using two different phylogenetic trees and only removing sites that were considered to be fast evolving on both topologies. In this case, removal of both slow- and fast-evolving sites vastly improved the support values for internal branches on the phylogenetic trees and resulted in a robust placement of the Microsporidia.

Many different methods exist for the identification of sites with a high substitution rate (Farris 1969; Kuhner and Felsenstein 1994; Brinkmann and Philippe 1999; Hansmann and Martin 2000; Schmidt et al. 2002; Pisani 2004). The majority, though not all, of these methods are tree based. Tree based methods identify rapidly evolving sites based on a tree either provided by the user or inferred by the method before site identification. For instance, TREE-PUZZLE (Schmidt et al. 2002) and DNArates (Maidak et al. 1996; Olsen et al. 1998) estimate evolutionary rates for each character based on a given tree and process of character-state substitution. TREE-PUZZLE can employ a discrete gamma distribution to estimate site rates, with sites allocated to a different category based on their likelihood score on the tree. The DNArates program has been used in conjunction with the fastDNAml program (Olsen et al. 1994) in order to partition alignments of homologous characters into rate categories (Fischer and Palmer 2005). Fischer and Palmer (2005) used a procedure that is not unlike the SACW approach in order to reweight characters for subsequent analyses. For a data set that was aimed at settling the placement of Microsporidia, they found that early unweighted data sets resulted in a variety of placements of the taxon, whereas successive rounds of character reweighting tended to result in fewer tree topologies and finally the authors settled on a placement of the microsporidia with the fungi that was best supported by the successively reweighted data.

Brinkmann and Philippe (1999) developed a method known as "slow-fast" where an alignment is split into groups (Brinkmann and Philippe 1999; Kostka et al. 2008). These groups are generally user-defined taxonomic groups. The evolutionary rate at a given site is calculated as the sum of the number of changes at the same position in all the groups individually. Although groups are, technically, user defined, any prior knowledge of the group will be based on previous tree inferences and, therefore, the slow-fast method is, by proxy, a tree-based method. In addition, due to the nature of this method, it is not suitable for small data sets.

The problem with tree-based methods is that the true tree is rarely known with certainty. Therefore use of an incorrect initial tree can result in incorrect assignation of an evolutionary rate to a character. Each character is compared with the given tree topology, whether correct or incorrect. A character is considered rapidly evolving if it conflicts with the initial tree or has a high level of homoplasy when mapped onto the tree. By assuming a topology prior to site rate identification, a slowly evolving site could potentially appear to be rapidly evolving, simply because the tree onto which it is mapped is incorrect. This initial error can become a source for systematic biases. Therefore, it may be preferable to have a method of determining evolutionary rate for a character that is independent of any a priori tree estimation procedure.

Tree-independent approaches to differentially weighting characters for phylogeny reconstruction include the Le Quesne (1969) test of character compatibility, which provided a "coefficient of character-state randomness" that could be used, if desired, to exclude characters from subsequent analysis. Essentially, this test evaluates two characters and if they can be mapped onto the same tree without homoplasy, then they are compatible, otherwise they are incompatible. Characters that have the highest amounts of incompatibilities with the other characters might be considered candidates for removal prior to subsequent phylogenetic analysis. Le Quesne (1989) later introduced the notion of compatibility within data being indicative of the level of phylogenetic information. This work was further extended by Meacham (1994), who developed his "Frequency of Compatibility

Attainment" statistic. Wilkinson (1998) highlighted the advantages of creating split patterns for sites when detecting conflict. Conflict, as defined by Le Quesne, becomes much easier to identify and rank when using a universal coding system for sites. Pisani (2004) utilized this idea to identify fast-evolving sites. According to the method of Pisani (2004), each site in the alignment receives an Le Quesne Probability (LQP) score, which is "[. . .] the probability of a random character having as low or lower incompatibility with the rest of the data than does the original character.". Pisani used this probability measure to explore arthropod relationships using different strategies for removal of characters with differing LQP values.

Hansmann and Martin (2000), in contrast with the compatibility strategies, proposed a very simplistic non–tree-based method for identifying rapidly evolving characters. They used the number of different character states in an alignment column as a proxy for evolutionary rate (Hansmann and Martin 2000). They cite the intuitiveness of the relationship between higher numbers of polymorphisms at a site and speed of evolution at that site. The set of most polymorphic characters would, therefore, be enriched in homoplastic sites (Hansmann and Martin 2000). However, each site is treated as a separate entity and consequently, this approach does not include information that may be contained in the data set as a whole, apart from ranking the sites from least to most polymorphic. In this paper, we present our method, TIGER (Tree Independent Generation of Evolutionary Rates), which is based on a similar concept to Le Quesne (1989), Wilkinson (1998) and Pisani (2004). TIGER analyzes similarity within characters (Wilkinson 1998). We expect that fast-evolving characters have lost some, most, or all of their phylogenetic signal and therefore should demonstrate reduced similarity with other sites that are more slowly evolving. Rather than comparing sites and only allowing them to be compatible or incompatible, our method allows sites to be scored according to varying degrees of similarity. This approach should provide a more fine-grained or nuanced result than the one that scores sites as being either compatible or incompatible.

In this report, we analyze synthetic data sets in order to explore the behavior of our approach and then, to demonstrate the utility of the method, we analyze two well-known problematic data sets. Additionally, we show that tree-based site removal approaches have significant problems, particularly when the data set contains a systematic bias (e.g., convergent base compositional bias), whereas our tree-independent approach can overcome these biases.

METHODS

Set Partitions

Our method is based on the analysis of set partitions at each position in a matrix. This matrix could be any type of data, including alignments of DNA or protein sequences or a matrix of homologous morphological characters.

A partition of a set X is a set of nonempty subsets of X such that every element x in X is in exactly one of these subsets. We treat each character in the matrix as a set and partition this set based on character states. A set partition is denoted, for example, as {{1}, {2, 3}, {4}, {5}} or 1/2,3/4/5. The partition 1/2,3/4/5 shows that for this character, taxa 2 and 3 have the same character state which is different from all the others, taxon 1, taxon 4 and taxon 5 each have unique character states—both different from each other and different from taxa 2 and 3. In this way, each character's partition is determined in order to enable pairwise comparisons with the rest of the characters in the data set. For example, in a nucleotide alignment of six taxa, character J = AAGGGC and character K = TTCCCA (assuming the order of the taxa is the same for both characters in this example). The partition set for both J and K is 1,2/3,4,5/6, despite having different character states.

Using this kind of data transformation, we can measure the degree of similarity between characters based on the similarity of their set partitions. We find that a character with a set partition that is similar to many other characters in the data matrix can usually, though not always, be a more slowly evolving character than a character with a set partition that is less similar to the rest of the characters in the matrix. Therefore, we can use the average similarity of a character's set partition to the rest of the matrix as a proxy for evolutionary rate.

The rate r_i for the character at position i is defined as:

$$r_i = \frac{\sum_{j \neq i} \mathrm{pa}(i,j)}{n-1} \quad (1)$$

where n is the total number of characters in the matrix and $\mathrm{pa}(i, j)$ is the partition agreement score. This is defined as

$$\mathrm{pa}(i,j) = \frac{\sum_{x \in P(j)} a(x, P(i))}{|P(j)|}, \quad (2)$$

where $|P(j)|$ is the number of groups in the partition of the jth character and $a(x, P(i))$ equals 1 if $x \subseteq A$ for some $A \in P(i)$. $P(i)$ may be defined as a partition in character i.

It is important to note that, given two sets A and B, if $A \subseteq B$, it is not necessarily commutative and, often, $B \not\subset A$. In this case, $\mathrm{pa}(i, j) \neq \mathrm{pa}(j, i)$. Also, because the rate is based on averaging of combinations of 1 or 0 values, it will always have a range between 0 and 1. A constant site, that is, a site with only one character state, will have $r = 1$ given that the pa, will be one for every comparison.

For example, consider two sites A = CTTAA and B = AGGGG with partition sets 1/2,3/4,5 and 1/2,3,4,5, respectively. $\mathrm{pa}(A, B) = 0.5$ because, out of two partitions in B ({1} and {2,3,4,5}), only {1} $\subseteq P(A)$. Given that {2,3,4,5} is not a subset of any partition in A, a({1}, $P(A)$) = 1 and a({2,3,4,5}, $P(A)$) = 0 $\therefore$ $\mathrm{pa}(A, B) = 0.5$. As mentioned, this calculation is not commutative, so $\mathrm{pa}(B, A) \neq 0.5$. $\mathrm{pa}(B, A) = 1$ because all partitions in $A \subseteq P(B)$.

This approach is designed to measure how much a particular character tends to agree with the other characters in the data. If a character shares partitions with many other characters, then it is likely that they hold similar information. This may be viewed as a signal in the data. Conversely, a character whose set partition greatly differs from the other signals in the data may be thought of as noise. To put it another way, a rapidly evolving character is likely to have sustained multiple substitutions, some or all of whom might be superimposed on earlier substitutions, therefore, this character is more likely to have a set partition that agrees less with more slowly evolving characters.

It is reasonable to suggest that a character that shares few partitions with the majority of other characters could be considered rapidly evolving. On the other hand, a slowly evolving character is more likely to share partitions with, or at least have fewer that conflict with, many other characters. The first assumption might not hold true in a situation where all or most characters in a matrix are rapidly evolving. It is most likely to hold true when evolutionary rates are moderate and when there is a gradient of evolutionary rates from slow to fast. Note that the rate of evolution that is assigned to a particular character is measured in arbitrary units and will vary with the data matrix being used. It is not a measure of substitutions per unit of time and indeed there are no units associated with the rate. This method can be used to analyze DNA, protein, morphological, or other arbitrary homologous characters.

It should be noted that for the current analyses, we did not attempt to deal with missing data. Missing data can be a feature of both molecular and morphological data sets, usually because a particular gene or morphological character has not been sampled or found. Missing data can be accommodated by an appropriate pruning of the characters so that only character states that have been observed are being compared.

Binning

It is often useful or convenient to group sites with similar evolutionary rates together and in our implementation of this method a range of rates can be divided into a user specified number of partitions, or bins. Sites are placed into bins depending on their rate value. The slowest rate and the fastest rate are determined and bins are constructed by splitting the rates into equal partitions. In this paper, we have used a variety of binning schemes, from 8 bins to 20 bins. In theory, any number of bins can be constructed, as long as the number is less than or equal to the number of characters in the matrix.

Data Simulations

In order to test the features of the method, we generated a number of artificial nucleotide data sets, using a phylogenetic tree and a prespecified model of nucleotide substitution. In the first instance, we simply wanted to know if data sets with different patterns of ASRV would return different patterns when analyzed using TIGER. Second, we wanted to see if removing characters had a beneficial effect on the fit of the data matrix to all possible trees or produced the desirable effect of improving the fit of the data to "good" trees while worsening the fit of the data to bad" trees. Our third simulation experiment involved the evaluation of whether or not the TIGER approach to character removal would improve the likelihood of resolving deep relationships.

In this report, we have used nucleotide data for reasons of ease of interpretation and also because of the ready availability of excellent computer software (Rambaut and Grassly 1997) to generate the data; however, in principle we could have used protein, morphological, or any kind of multistate character matrices.

Varying gamma shapes.—Using Seq-Gen (Rambaut and Grassly 1997), we simulated two data sets over the same 49-taxon tree (Maddison 2004) (the tree is available in Supplementary Material, available from http://www.sysbio.oxfordjournals.org/) and we employed a model that used a discrete approximation to the gamma distribution, with four categories of sites. In order to assess whether or not the TIGER algorithm could detect different patterns of ASRV, two different α values were used in simulations—0.5 and 20.0 reflecting two different distribution shapesthe first is L shaped and the second is hill shaped. Both alignments were 999 bp in length and simulated under the JC model (Jukes and Cantor 1969). We experimented with other models of sequence evolution and different tree shapes and numbers of taxa and the results are essentially the same as presented here, so we only present the results of the JC simulations on this data set.

Changing fit of the data to all trees in treespace.—Removal of homoplastic characters in a matrix should have the effect of improving the fit of the data to the true tree whereas worsening the fit of the matrix to trees that are very different from the true tree. However, given that it is possible to edit any tree to change its topology into any other tree, if we perform any data modification it will most likely influence the goodness-of-fit of the data to all trees in some way. Some trees are very similar to the true tree and some are very dissimilar, consequently, whereas incrementally removing larger numbers of characters (grouped into bins), we investigated the change in fit of the data to all possible phylogenetic trees for an eight-taxon data set. In our experiments, we measured the change in the CI for all trees as bins were sequentially removed, starting with the bin containing the most rapidly evolving characters (a total of 10 bins were used in this experiment). In effect, for the set of all trees, T, we computed the CI for the original data set on tree t ($t \in T$) and compared this value with the CI value for the data set with Bin10 removed. We then

plotted this value against the "nodal" distance (Puigbo et al. 2007) between the true tree and tree *t* (when *t* is not the true tree). For the true tree, the nodal distance is always zero. We carried out the same procedure when we removed Bin9+Bin10, Bin8+Bin9+Bin10, and Bin7+Bin8+Bin9+Bin10.

TIGER rates versus likelihood scores.—Using the correct tree and the correct model, site-specific likelihood scores can give a very good estimate of character evolutionary rate. We wished to test how well the TIGER approach could identify these characters without any knowledge of a tree. We used 100 different seven-taxon trees chosen at random from treespace (which contains 945 unrooted trees). A nucleotide alignment of 999 positions was generated under the JC model for each of these 100 trees. We generated site-specific likelihood scores in PAUP* (Wilgenbusch and Swofford 2003) for all 945 trees for each data set and we measured the ranking of sites on each tree to TIGER rankings. That is to say, the site(s) with the highest likelihood value are ranked as #1 and the site(s) with the lowest value as #999 and likewise for TIGER rates. The Euclidian distance between all likelihood rankings and TIGER rankings was calculated. This is a very simple measure of the average difference in rank for a character in the two lists.

Deep branching tree.—Rapid evolution can obfuscate deep relationships on a tree, often leading to unwanted polytomies. This situation is particularly problematic when long unbroken branches subtend a series of rapid cladogenetic events. To test whether the TIGER approach could help resolve deep relationships where there is very little phylogenetic signal, we used the JC model of sequence evolution to produce 100 simulated 999 bp nucleotide data sets across the eight taxon tree shown in Figure 2. The short deep branches combined with long terminal branches presents a difficult problem for phylogenetic analysis, mostly due to the confounding effects of rapidly evolving characters. To ensure that the data generated displayed poor phylogenetic resolution, we built a majority-rule consensus tree from maximum likelihood (ML) trees constructed from each of the data sets prior to any site removal. This was repeated after removal of sites dictated by TIGER and to test the performance of a tree based method in this scenario, we also repeated the analysis after removal of rapidly evolving sites identified by ML. The ML tree was estimated using PAUP* and the sites were categorized on this tree using TREE-PUZZLE.

Empirical Testing

Thermus data set.—In order to further understand TIGER's functionality, two empirical data sets were used. A 1273-column alignment of bacterial 16S ribosomal RNA genes known as the *Thermus* data set is well studied (Embley et al. 1993; Mooers and Holmes 2000), and we used this data set to examine whether the TIGER approach is useful for accounting for base compositional biases. This data set contains three thermophiles, *Aquifex aeolicus*, *Thermatoga maritima*, and *Thermus aquaticus* whose sequences are enriched in G and C nucleotides and two mesophiles, *Bacillus subtilis* and *Deinococcus radiodurans* whose nucleotide composition is more balanced. A combination of compositional bias and distant relationships can mean that when there is only a weak phylogenetic signal, it can be overcome by the similarity in base composition of the most rapidly evolving positions in the alignment. In general, many methods of phylogenetic analysis will group the thermophiles together in this data set, despite the fact that there is strong evidence that *T. aquaticus* and *D. radiodurans* are sister taxa (Embley et al. 1993). We refer to a tree displaying the mesophiles as a monophyletic group to the exclusion of the thermophiles as the ATTRACT tree and this is the tree recovered by most tree inference methods using the whole sequence alignment. We refer to a phylogenetic tree that places *T. aquaticus* and *D. radiodurans* together as the TRUE tree. Due to this well-characterized strong compositional attraction, we wished to investigate whether site removal using the TIGER approach could influence recovery of the correct tree. However, to demonstrate the different effects of site removal in a tree-independent fashion compared with the traditional ML approaches, we also compared the topology inferred after removal of rapidly evolving sites identified by TIGER with the topology recovered after removal of rapidly evolving sites according to TREE-PUZZLE (Schmidt et al. 2002) and SACW (Farris 1969). We did not use TREE-PUZZLE to infer the tree, we simply used the method implemented by TREE-PUZZLE to assign evolutionary rates to sites, based on a tree that we supplied to the software.

Primate data set.—It has generally been accepted that humans share a close relationship with orangutans, gorillas, and chimpanzees (Hayasaka et al. 1988; Begun 1992; Adachi and Hasegawa 1995; Shoshani et al. 1996; Ruvolo 1997; Satta et al. 2000; Ebersberger et al. 2007). From this group, it is generally agreed that orangutans are the least closely related to humans and that humans, chimps, and gorillas form a monophyletic group, though there are some conflicting opinions (Schwartz 1984; Grehan and Schwartz 2009).

The relationships of interest, therefore, concern the human, chimpanzee, and gorilla lineages (Satta et al. 2000). The separation of these three lineages is thought to have occurred in quick succession (Hayasaka et al. 1988; Adachi and Hasegawa 1995), and this makes the phylogeny difficult to resolve and the two alternative hypotheses—human, chimp together (HC hypothesis) or chimp, gorilla together (CG hypothesis)—receive almost equal support from this data set. Because of the controversy surrounding this topology, the second empirical data set we use is a well-known primate mitochondrial data set (see Supplementary material)

consisting of 12 sequences and 898 aligned nucleotide positions (Hayasaka et al. 1988).

In a parsimony analysis of the data set, with all characters being equally weighted, both the HC and the CG hypotheses are equally good, with 1153 steps required to explain the data. We used the tree-based methods of assigning character evolutionary rates and use alternatively the HC and the CG trees in order to carry out the inferences. We compared and contrasted the results from tree-based analysis with the tree-independent method described here.

Results and Discussion

Varying Gamma Shapes

Our first analysis of the behavior of the TIGER method focused on the analysis of simulated data sets for 49 taxa with different patterns of rate variation across sites. We chose the 49-taxon data set that is distributed with the MACCLADE software (Maddison 2004) because it contains a reasonable range of branch lengths and has a moderately large number of taxa. We simulated two separate data sets that differed by the ASRV model used to generate the data. In the first case, we used a gamma distribution with an α parameter of 20 and in the second the α parameter was set to 0.5, reflecting very different evolutionary scenarios. We then used the TIGER approach to place sites into 20 bins sorted by their rate of evolution (Fig. 1a,b).

There are two interesting points to be made about Figure 1. First of all, the two graphs are not the same and furthermore Figure 1b, which is generated from the data set with an α parameter of 0.5, is more L shaped than Figure 1a, which was generated from the data with an α parameter of 20. This indicates that the TIGER approach is detecting the different ASRV patterns. What is of further interest is that within each graph there is a clear multimodality. There are four clusters of bars on the histograms (indicated by the alternative shading and clear zones on the diagrams). When the seq-gen software generates data, it uses an approximation to the gamma distribution and in these cases an approximation was employed that used four categories of sites. The TIGER approach has identified these subtle patterns and has placed the different sites into clusters.

True Tree versus Incorrect Trees

If the removal of rapidly evolving characters really is a good idea for improving the chances of recovering the correct phylogenetic tree, then we expect that removal of these characters would improve the goodness-of-fit of the data to the true tree while worsening the goodness-of-fit of the data to other trees. In order to test this hypothesis, we generated a simulated data set containing eight taxa and using the JC model, according to the protocols previously described. We progressively removed the fastest evolving sites, as judged by the TIGER approach, until we had removed the four fastest categories of sites. We then examined the goodness-of-fit of the data to the correct tree (the tree used to simulate the data) and also the goodness-of-fit of the data to all the other possible trees. We plotted the goodness-of-fit measure (CI) against the nodal distance (as measured by the TOPD/FMTS software, Puigbo et al. 2007) for the unstripped data set for each possible tree topology and we plotted the change in CI (ΔCI) against nodal distance for each of the data sets where sites were stripped. The results of these experiments are seen in Figure 2. In total, there were 10,395 trees examined for each treatment of the data.

With all sites included in the alignment, the CI for the correct tree was 0.825. The worst CI value in the data set was 0.612 and the tree with the largest nodal distance from the true tree had a distance of 2.44949 and a CI value of 0.616. In general, there is a negative correlation between CI and nodal distance from the true tree.

When we stripped out the Bin10 category of sites, we saw the CI values increased for some trees and decreased for others. The CI value with the largest increase for any of the 10,395 trees was the CI value for the true

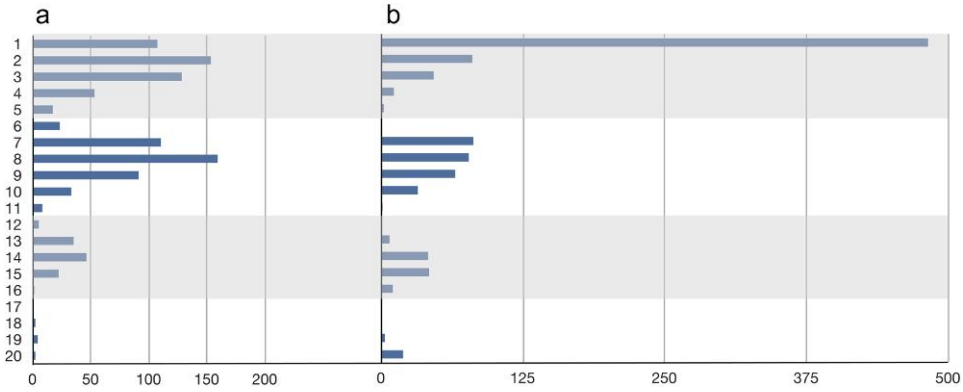

FIGURE 1. Histograms of binning results for two different data sets with different ASRV. a) A 999-bp, 49-taxon data set generated using the tree in S1 and ASRV modeled using a gamma distribution with a shape parameter of 0.5, and (b) data set of the same size and topology but with ASRV modeled using a gamma shape parameter of 20.0. The alternating shaded and clear areas indicate the four categories of sites that approximate the gamma distribution. This figure is available in black and white in print and in color at *Systematic Biology* online.

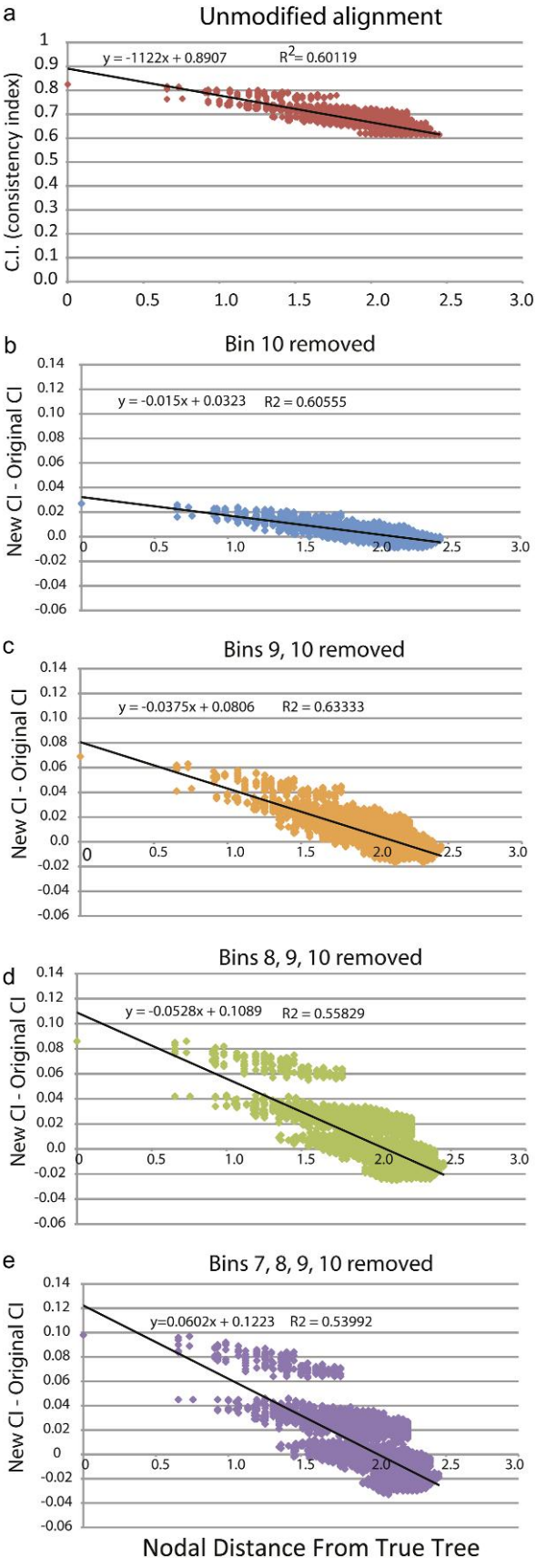

FIGURE 2.

tree—an increase to 0.852. In contrast, the tree with the largest nodal distance from the true tree experienced a decrease in CI value and its new value was 0.612. Overall, a total of 5364 trees (51.6% of the total) saw an increase in CI value, whereas 5031 trees experienced a decrease in CI value.

Continued site stripping resulted in a progressive increase in CI value for the true tree and a progressive decrease in CI value for the tree with the largest nodal distance from the true tree. When Bin categories 9 and 10 were removed, the values changed to 0.894 and 0.609, respectively, with 5403 (51.9%) of the trees now experiencing an increase in CI value. When Bin categories 8, 9 and 10 were removed, the values changed to 0.911 for the true tree and 0.601 for the worst tree with 3811 of the trees having an increased CI value. Finally, when we removed Bin categories 7, 8, 9, and 10, the values changed to 0.923 and 0.597, respectively, with 3257 of the trees experiencing an increase in CI value (31.3%), whereas 7138 had a decreased CI value (68.6%).

Therefore, we can see for this data set that site stripping has resulted in a bias in the fit of the data to different trees. In general, those tree topologies that are close to the true tree will begin to fit the data better, whereas those trees that are least similar in topology to the true will begin to fit the data worse. The tree that is most positively affected by site stripping is the true tree. It must be remembered that the TIGER approach is not tree based and at no time was the TIGER software aware of the topology of the true tree.

TIGER Rates versus Likelihood Scores

To see how well TIGER can approximate site-specific rates we compared it with likelihood scores for each site on every possible seven-taxon unrooted tree. The Euclidian distance from TIGER ranking to the likelihood rankings on all trees were recorded for all data sets, with particular emphasis on where the distance between TIGER rankings and the likelihood rankings on the known true tree fell with respect to the other trees. In 100% of data sets, this distance fell within the top 0.3% of all scores. In 95% of all cases, the distance from TIGER rankings to the likelihood rankings on the true tree was the smallest distance recorded to any tree in the data set.

This shows that the TIGER approach will produce an ordering of the evolutionary rates of the sites that is usually closer to the ranking of sites according to the true tree than to other incorrect trees.

Deep Branching Tree

In order to see whether it is possible for our method to improve the resolution of deep relationships where

←

FIGURE 2. Change in CI with increasing site removal. On the abscissa is the nodal distance of a tree from the correct tree and on the ordinate is either the CI or the difference in CI value between the unstripped alignment and the stripped alignment (ΔCI).

phylogenetic signal is weak, we simulated 100 different DNA alignments based upon a single phylogenetic tree with long external branches and very short internal branches (Fig. 3a). This alignment was designed to represent a difficult problem of phylogenetic inference and was simulated using the JC model of sequence evolution. ML trees for each of the data sets were inferred under the JC model. As expected, prior to removal of rapidly evolving sites, the majority-rule consensus analysis using the JC model produced a tree with polytomies and poor resolution (Fig. 3b), and the only branch that is resolved has a bipartition frequency (BF) of 55% was for a split that separates taxa C and D from the rest. We used the TIGER approach to identify the rapidly evolving characters in the matrices and place all characters into 10 bins with increasing evolutionary rate. Removal of the most rapid category of sites, Bin10, which contained between 183 and 502 sites with an average of 424 between the 100 data sets, entirely resolved all except the basal polytomy (Fig. 3c), with BF ranging from 67% to 99%. We wished to test our method against a tree-based method. We used TREE-PUZZLE (Schmidt et al. 2002) on the same simulated data. Removing the most rapidly evolving category of sites using the TREE-PUZZLE approach (ranging from 269 to 481 sites, mean of 334 sites removed) the tree remained equally unresolved as prior to any site removal, with the BF of the split separating C and D rising to 61 (Fig. 3b).

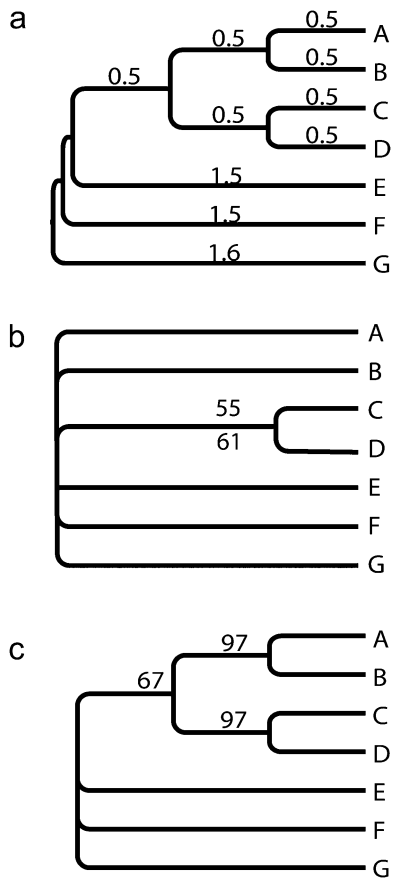

FIGURE 3. Effect of site removal on deep closely spaced cladogenetic events. a) The topology of the tree used to generate the simulated data (see text for details of simulation). b) Majority-rule consensus ML tree after before site removal and also after site removal using ML. The bootstrap support value for the unstripped alignments is above the line and the value after site removal using likelihood is below the line. c) Majority-rule consensus ML tree after removal of Bin10, the fastest evolving sites, according to the TIGER method.

This shows both the pitfall of the tree-based method and the advantage of our tree-independent method. The sites identified as most rapidly evolving by TREE-PUZZLE are those that do not agree with the initial tree inferred by ML. For this reason, removal of these sites does not clarify signals in the data, rather it merely strengthens the signal for the initial groupings. The tree-independent method, however, does not need any initial tree, therefore it is not biased toward any single tree and, instead, it picks out genuine signals in the data.

Thermus Data Set

The *Thermus* data set consists of 1273 aligned nucleotide positions from the 16S rRNA gene and is available as Supplementary Material. Using ML phylogenetic reconstruction implemented in PAUP4.0b10, we examined the differences in tree topology when removing characters judged to be rapidly evolving according to TIGER versus characters judged to be rapidly evolving according to TREE-PUZZLE (with a user-supplied tree, constructed using ML). In addition, we used the *reweight* command in PAUP to apply SACW (Farris 1969) and evaluate the effect that this approach had on the chances of recovering the correct tree. Using the original alignment of 1273 aligned positions (see Supplementary Material) and a GTR+I+G model of sequence evolution, we produced the phylogenetic tree in Figure 4a. Using the TREE-PUZZLE software, we categorized sites according to the GTR+I+G model using a discrete approximation to the gamma distribution to model ASRV, with a total of eight categories of sites. The category of sites with the fastest rate of evolution was removed from the alignment (a total of 186 sites) and the analysis was re-run using this newer shorter data set (consisting of 1087 sites). In this case, the same ATTRACT tree was recovered. The most significant difference between the two bootstrap analyses was that the bootstrap support values for the data set with the sites removed were much higher and each of the internal nodes was recovered in 100% of the bootstrap pseudoreplicates (Fig. 4b). It must be remembered that the rates of evolution of the sites had been determined using the ATTRACT tree, which is the tree that is obtained in the analysis of the unstripped data set.

In order to investigate the SACW method, we first inferred the most parsimonious phylogenetic tree with all sites equally weighted and using an exhaustive search of tree space and the parsimony optimality criterion. Support for this tree was assessed using 1000 rounds of bootstrap resampling, with the results summarized by a majority-rule consensus procedure. The most

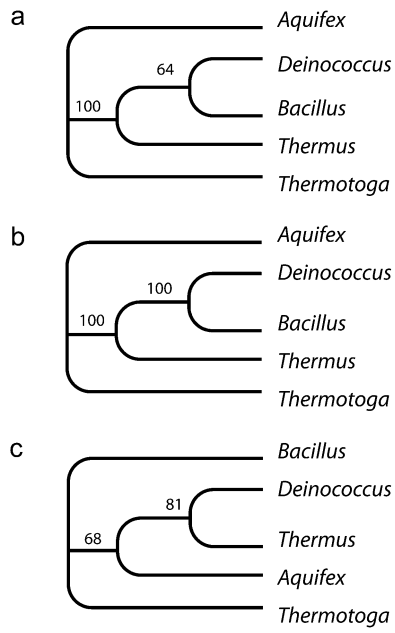

FIGURE 4. Analysis of the *Thermus* data set. a) Topology and support prior to site removal. b) The tree recovered after removal of sites identified by PUZZLE and using SACW. c) The resulting tree after removal of sites identified by TIGER.

parsimonious tree was once again the ATTRACT tree, with bootstrap support values of 92% for the grouping of *D. radiodurans* and *B. subtilis* and 96% for a clan containing *A. aeolicus* and *T. maritima*. Using the *reweight* command in the PAUP software, we weighted the characters according to their CI value on this tree. We then carried out another bootstrap resampling analysis to assess support for groups on the tree. This time the ATTRACT tree was once again recovered, but the support for all internal edges was at 100%.

We used the TIGER approach to identify rapidly evolving sites in the rRNA data set. We placed all sites from the alignment into one of eight bins according to how rapidly they evolve. The most rapidly evolving category of sites contained 108 sites and these were removed for subsequent ML analysis. Using the GTR+I+G model of sequence evolution on the remaining 1165 sites, we recovered the TRUE phylogenetic tree. After 1000 bootstrap replicates, we observed that the grouping of *D. radiodurans* and *T. aquaticus* in 81% of the replicates and the grouping of *T. maritima* and *B. subtilis* was observed in 68% of the replicates. The ATTRACT topology that groups *D. radiodurans* and *B. subtilis* together was seen in 19% of the replicates.

We carried out an additional analysis of the sites that are identified as being rapidly evolving. In all cases, we analyzed the most rapidly evolving sites on their own to see if there was any strong phylogenetic signal in those sites. As these sites are saturated for change, we do not expect to see a single phylogenetic signal, rather a number of incongruent signals. In our analyses, only the sites in Category 8 of the ML analysis contained any congruent phylogenetic signal. There was 80% bootstrap support for the TRUE tree in these sites. This result demonstrates that not only does such an ML approach result in strong support for the incorrect topology but also the characters that it discards contain more true phylogenetic signal than the characters that it retains. This needs to be viewed as a systematic error.

Primate Data Set

Our last analysis involves an 898 bp data set of 12 primate mitochondrial sequences (Hayasaka et al. 1988). Two equally most parsimonious trees, requiring 1153 steps can be obtained by analysis of these sequences. One of these trees places the human and chimpanzee (*Pan troglodytes*) together as sister taxa, whereas the other tree groups the Chimpanzee with the Gorilla. We wanted to investigate two things with this data set. First, in this case, where two phylogenetic hypotheses are strongly competing and where there is no greater support for one topology over the other, whether the TIGER approach would recover the accepted tree (human and chimp together) with confidence. Second, whether the tree-dependent method would be influenced strongly by the tree that is used to determine the evolutionary rate of the characters, or whether it would work well irrespective of the tree that it used initially for character reweighting. More specifically, we wished to see if using a particular tree in order to generate evolutionary rates would tilt the balance in favor of this topology in a bootstrap analysis. In other words, we wanted to explore whether character removal, based on an incorrect tree, could override the (albeit small) amount of extra support for the true tree and subsequently provide strong support for the incorrect tree.

When the tree that places *Homo* and *Pan* together was used in SACW in order to reweight characters according to the CI, then this same tree was recovered in the majority-rule consensus tree following bootstrapping. The bootstrap support value for this relationship was 79%, compared with a 51% value for the equally weighted data set (10,000 bootstrap replicates). We then used the other equally parsimonious tree in order to carry out character weighting for SACW. Using character reweighting according to the CI, we obtained a bootstrap support value of 77% for the grouping of *Pan* and *Gorilla* together. This shows that the initial tree that is used for character weighting can override small phylogenetic signals and because characters that tend not to agree with this initial tree are down weighted, this has a huge affect on which tree is supported in subsequent analyses.

It should be noted that in this particular case, the ML approach to site stripping was not as sensitive as the SACW approach and indeed was quite insensitive to the initial tree that was used for site classification. When the HC hypothesis tree was used, and the

TREE-PUZZLE software was asked to put sites into a total of 10 categories, then a total of 114 sites were put into the fastest category. When the CG hypothesis tree was used, then a total of 121 sites were put into the fastest category. Irrespective of the tree that was used to categorize sites, when category 10 was removed, we always recovered strong support for the HC hypothesis. We should note, however, that when the HC hypothesis was used to categorize sites, the resulting bootstrap support value was 99%, whereas when the CG hypothesis tree was used to categorize sites, then support for the HC hypothesis after site stripping was somewhat lower at 81%.

We used the TIGER approach to categorize characters in a tree-independent manner and to place them into a total of 10 bins according to their average split similarity with the other characters in the matrix. We removed the fastest category of sites, Bin10, which contained a total of 192 characters. We then used maximum parsimony bootstrapping to evaluate support for groups in the phylogeny. We recovered a grouping of *Homo* and *Pan*, with 87% support after 10,000 bootstrap replicates. The alternative hypothesis, grouping *Pan* and *Gorilla* together received 8.8% bootstrap support. Using ML, the HC hypothesis received 90% bootstrap support, whereas the CG hypothesis received 6% support.

Conclusion

In this article, we report the development of an algorithm, based on those of Le Quesne (1989), Wilkinson (1998) and Pisani (2004) that uses similarity in the pattern of character-state distributions between characters as a proxy for speed of evolution in a data matrix of homologous characters. We expect that rapidly evolving characters are likely to lose some, most, or all of their phylogenetic information and will tend to have a character-state distribution that is closer to random than the distribution expected from a more slowly evolving character. A character is assumed to be rapidly evolving if it has a character-state distribution pattern that, on average, is not very similar to the patterns observed in other characters. This assumption is only likely to hold in some (though probably very many) situations. Specifically, in a data matrix where each character is effectively randomized, due to a very rapid rate of evolution or a long evolutionary timespan, we do not expect that this kind of approach will work well. Notwithstanding this caveat (which is a situation that would confound most, if not all, phylogenetic methods), we have observed some very interesting and desirable properties of this approach that make it a useful addition to the phylogenetic arsenal.

The TIGER approach identified differing patterns of ASRV, distinguishing alignments that had extreme variation in among-site evolutionary rates from those alignments that had a more even distribution of rates. Additionally, it was able to identify subtleties in the data such as the four clusters of rates in each alignment—a by-product of the simulation process.

The TIGER approach helped improve the fit of the data to the correct tree in our simulations. Removing sites that TIGER identified as being rapidly evolving resulted in a better fit of the data to good trees and worse fit of the data to bad trees, with the true tree being affected most positively. Additionally, using the TIGER approach, we could improve the resolution of deep lineages where rapid cladogenesis resulted in very difficult-to-resolve branches. Worryingly, the likelihood approach to removing rapidly evolving sites proved to be problematic—the sites that were removed were those that did not agree with the initial tree, resulting in a situation where, out of 100 simulations, there was little improvement in the recovery of the deep diverging rapid cladogenesis tree.

For the ribosomal RNA data set, we observed a number of issues. First, the TIGER approach seems to have some merit as an approach to removing sites that interfere with phylogeny reconstruction. Additionally, two other tree-dependent methods—site identification using a ML model of ASRV and site identification using the fit of the data to an initially constructed phylogenetic tree—are systematically biased toward favoring the first phylogenetic tree they construct. We, therefore, feel it is important to be cautious when using tree-based methods of assigning evolutionary rates to sites, unless the evolutionary history is known with certainty. We note, however, that a sophisticated compositionally heterogeneous model of sequence evolution is capable of identifying the correct topology for this data set, without the necessity of deleting or reweighting characters (Foster 2004).

The point concerning tree-based attribution of evolutionary rate is quite clearly exemplified by the primate mitochondrial data set and maximum parsimony analysis. Here two hypotheses are equally good when using the parsimony criterion. Character reweighting based on one of the two equally most parsimonious trees will skew subsequent analyses toward supporting this particular topology, whereas the same is true for the alternative topology. Ultimately, the TIGER analysis, which does not use a tree, recovers the correct phylogenetic hypothesis (which has been confirmed by numerous other studies) while not using an a priori determined phylogenetic tree in order to do so. We find that support for the grouping of *Pan* and *Gorilla*, to the exclusion of *Homo* is an artifact that is due to the most rapidly evolving sites. This also shows that site stripping can be beneficial for resolution of recent relationships, not just ancient relationships. We should also state here that ML analysis of this data set produces the correct tree, using the Tamura–Nei model, with bootstrap support for (*Homo, Pan*) at 94%.

Ultimately, TIGER is an interesting device for identifying characters that do not agree with the majority of the data. We argue here that in many cases this disagreement can be diagnostic of rapid evolution. At the very least, the converse is likely to be true—rapid character evolution is likely to produce a pattern that is not very similar to other characters. Removal of these kinds

of characters can greatly improve the accuracy of successive phylogenetic analysis by removing conflicting signals.

There are surely limits to what site removal can accomplish and with certainty site removal is a poor alternative to precise model definition. However, precise model definition comes with a cost. Models that adequately describe the evolution of a set of DNA or protein sequences might, of necessity, be very parameter rich (e.g., using a combination of Dirichlet processes for both site rate identification, Huelsenbeck and Suchard 2007, and site-specific profiling, Lartillot and Philippe 2004, as implemented in the CAT model) and require a large amount of sequence before they become statistically consistent. The most commonly used models of sequence evolution are often inadequate to describe the evolution of the sequences being studied. Model selection approaches often "max-out," where the most parameter-rich method of analysis is the one that is selected by a likelihood ratio test, Akaike information criterion or Bayesian information criterion (Keane et al. 2006), indicating that perhaps there are not enough parameters available. Therefore, it might not be an option to use a precisely described model. In the case of the rRNA sequences being analysed in this study, the raw alignment exhibited significant compositional heterogeneity and none of the standard, compositionally homogeneous time-reversible models of sequence evolution can adequately account for this heterogeneity. By identifying and removing the most rapidly evolving characters, the models are better able to account for the evolution of the sequences.

We have no good theoretical framework for knowing precisely how many sites to remove from an alignment. It is likely that in many cases there is no need to strip out any sites. At the moment, we only have an ad hoc approach to site stripping and this must be considered a major problem. Ideally, we wish to remove sites that only contribute noise and do not contribute any phylogenetic signal. Our recommendation is the testing of congruence across a progressively larger number of the fastest evolving characters using methods such as the permutation-tail-probability test (Faith and Cranston 1991) or likelihood mapping (Strimmer and von Haeseler 1997). This would result in the removal of sites that show very little consistency with the rest of the data and very little consistency with one another. However, this is also ad hoc and should be used as nothing more than a rule-of-thumb.

We also note that bootstrap support values or Bayesian clade probability values are probably meaningless when there is a directed attempt to remove sites that disagree with the rest of the data. It is likely that the support values will tend to increase when incongruent data are removed. When we use bootstrap support values, we wish to show that the data have been strongly influenced by the character removal; we do not wish to imply that bootstrapping should follow character removal, as, in most cases, the resulting bootstrap scores are likely to be higher.

Given that there are limits to what can be achieved by character removal, we conclude by advising that this method should be used as one part of an overall experimental programme of data exploration. We expect that additional tree-independent methods of analyzing evolutionary rate variation can be developed.

Supplementary Material

Supplementary material can be found at http://www.sysbio.oxfordjournals.org/.

Funding

C.A.C. is funded by a Science Foundation Ireland Research Frontiers Programme award [07/RFP/EEEOBF654] to J.O.Mc.I.

Acknowledgments

The work was supported by the NUIM High-performance computing resource and the Irish Centre for High End Computing.

References

Adachi J., Hasegawa M. 1995. Improved dating of the human/chimpanzee separation in the mitochondrial DNA tree: heterogeneity among amino acid sites. J. Mol. Evol. 40:622–628.

Begun D. 1992. Miocene fossil hominids and the chimp human clade. Science. 257:1929–1933.

Brinkmann H., Philippe H. 1999. Archaea sister group of bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol. Biol. Evol. 16:817–825.

Delsuc F., Brinkmann H., Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. 6:361–375.

Ebersberger I., Galgoczy P., Taudien S., Taenzer S., Platzer M., von Haeseler A. 2007. Mapping human genetic ancestry. Mol. Biol. Evol. 24:2266–2276.

Embley T., Thomas R., Williams R. 1993. Reduced thermophilic bias in the 16S rDNA sequence from *Thermus ruber* provides further support for a relationship between Thermus and Deinococcus. Syst. Appl. Microbiol. 16:25–29.

Faith D., Cranston P. 1991. Could a cladogram this short have arisen by chance alone?: on permutation tests for cladistic structure. Cladistics. 7:1–28.

Farris J. 1969. Successive approximations approach to character weighting. Syst. Zool. 18:374–385.

Fischer W. M., Palmer J. D. 2005. Evidence from small-subunit ribosomal RNA sequences for a fungal origin of Microsporidia. Mol. Phylogenet. Evol. 36:606–622.

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

Foster P. G. 2004. Modeling compositional heterogeneity. Syst. Biol. 53:485–495.

Grehan J., Schwartz J. 2009. Evolution of the second orangutan: phylogeny and biogeography of hominid origins. J. Biogeogr. 36: 1823–1844.

Hansmann S., Martin W. 2000. Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: influence of excluding poorly alignable sites from analysis. Int. J. Syst. Evol. Microbiol. 50:1655–1663.

Hayasaka K., Gojobori T., Horai S. 1988. Molecular phylogeny and evolution of primate mitochondrial DNA. Mol. Biol. Evol. 5: 626–644.

Hirt R., Logsdon J., Healy B., Dorey M., Doolittle W., and Embley T. 1999. Microsporidia are related to fungi: evidence from the largest subunit of RNA polymerase II and other proteins. Proc. Natl. Acad. Sci. U.S.A. 96:580–585.

Huelsenbeck J., Suchard M. 2007. A nonparametric method for accommodating and testing across-site rate variation. Syst. Biol. 56: 975–987.

Jukes T., Cantor C. 1969. Evolution of protein molecules. In: Munro, editor. Mammalian protein metabolism. New York: Academic Press. p. 240–253.

Keane T., Creevey C., Pentony M., Naughton T., McInerney J. 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. BMC Evol. Biol. 6.

Kluge A. G., Farris J. S. 1969. Quantitative phyletics and the evolution of anurans. Syst. Zool. 18:1–32.

Kostka M., Uzlikova M., Cepicka I., Flegr J. 2008. SlowFaster, a user-friendly program for slow-fast analysis and its application on phylogeny of Blastocystis. BMC Bioinformatics. 9:341.

Kuhner M. K., Felsenstein J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11:459–468.

Lartillot N., Philippe H. 2004. A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21:1095–1109.

Le Quesne W. 1969. A method of selection of characters in numerical taxonomy. Syst. Biol. 18:201–205.

Le Quesne W. 1989. The normal deviate test of phylogenetic value of a data matrix. Syst. Zool. 38:51–54.

Maddison D. R. 2004. Testing monophyly of a group of beetles. Study 1 in Mesquite: A Modular System for Evolutionary Analysis. Version 1.04. Available from: http://mesquiteproject.org.

Maidak B. L., Olsen G. J., Larsen N., Overbeek R., McCaughey M. J., Woese C. R. 1996. The ribosomal database project (RDP). Nucleic Acids Res. 24:82–85.

Meacham C. A. 1994. Phylogenetic relationships at the basal radiation of angiosperms: further study by probability of character compatibility. Syst. Bot. 19:506–522.

Mooers A., Holmes E. 2000. The evolution of base composition and phylogenetic inference. Trends Ecol. Evol. 15: 365–369.

Olsen G. 1987. Earliest phylogenetic branchings: comparing rRNA-based evolutionary trees inferred with various techniques. Cold Spring Harbor. Symp. Quant. Biol. 52:825–837.

Olsen G. J., Matsuda H., Hagstrom R., Overbeek R. 1994. fastDNAmL: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Comput. Appl. Biosci. 10:41–48.

Olsen G., Pracht S., Overbeek R. 1998. DNArates. Version 1.1.

Philippe H., Zhou Y., Brinkmann H., Rodrigue N., Delsuc F. 2005. Heterotachy and long-branch attraction in phylogenetics. BMC Evol. Biol. 5:50.

Pisani D. 2004. Identifying and removing fast-evolving sites using compatibility analysis: an example from the Arthropoda. Syst. Biol. 53:978–989.

Puigbo P., Garcia-Vallve S., McInerney J. O. 2007. TOPD/FMTS: a new software to compare phylogenetic trees. Bioinformatics. 23: 1556–1558.

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

Ruvolo M. 1997. Molecular phylogeny of the hominoids: inferences from multiple independent DNA sequence data sets. Mol. Biol. Evol. 14:248–265.

Satta Y., Klein J., Takahata N. 2000. DNA archives and out nearest relative: the trichotomy problem revisited. Mol. Phylogenet. Evol. 14:259–275.

Schmidt H., Strimmer K., Vingron M., von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics. 18:502–504.

Schwartz J. 1984. The evolutionary relationships of man and orangutans. Nature. 308:501–505.

Shoshani J., Groves C., Simons E., Gunnell G. 1996. Primate phylogeny: morphological vs molecular results. Mol. Phylogenet. Evol. 5:102–154.

Strimmer K., von Haeseler A. 1997. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. Proc. Natl. Acad. Sci. U.S.A. 94:6815–6819.

Townsend J. P. 2007. Profiling phylogenetic informativeness. Syst. Biol. 56:222–231.

Wilgenbusch J. C., Swofford D. 2003. Inferring evolutionary trees with PAUP*. Curr. Protoc. Bioinformatics. Chapter 6: Unit 6.4.

Wilkinson M. 1998. Split support and split conflict randomization tests in phylogenetic inference. Syst. Biol. 47:673.

Yang Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10:1396–1401.

Yang Z. 1996. Among-site rate variation and its impact in phylogenetic analyses. Trends Ecol. Evol. 11:367–372.