J Mol Evol (2010) 71:319–331
DOI 10.1007/s00239-010-9398-z

Automated Removal of Noisy Data in Phylogenomic Analyses

**Vadim V. Goremykin · Svetlana V. Nikiforova ·
Olaf R. P. Bininda-Emonds**

Received: 14 April 2010 / Accepted: 6 October 2010 / Published online: 26 October 2010

Abstract Noisy data, especially in combination with misalignment and model misspecification can have an adverse effect on phylogeny reconstruction; however, effective methods to identify such data are few. One particularly important class of noisy data is saturated positions. To avoid potential errors related to saturation in phylogenomic analyses, we present an automated procedure involving the step-wise removal of the most variable positions in a given data set coupled with a stopping criterion derived from correlation analyses of pairwise ML distances calculated from the deleted (saturated) and the remaining (conserved) subsets of the alignment. Through a comparison with existing methods, we demonstrate both the effectiveness of our proposed procedure for identifying noisy data and the effect of the removal of such data using a well-publicized case study involving placental mammals. At the least, our procedure will identify data sets requiring greater data exploration, and we recommend its use to investigate the effect on phylogenetic analyses of removing subsets of variable positions exhibiting weak or no correlation to the rest of the alignment. However, we would argue that this procedure, by identifying and removing noisy data, facilitates the construction of more accurate phylogenies by, for example, ameliorating potential long-branch attraction artefacts.

Keywords Noise reduction · Saturation · Long-branch attraction · Model testing · Noisy data · Placental mammals · Rodentia

Abbreviations

ML	Maximum likelihood
MP	Maximum parsimony
OTU	Operational taxonomic unit

Introduction

Noise Reduction

It has been long appreciated that multiple substitutions per site (e.g. as derived from an elevated rate of molecular evolution) can impede phylogeny reconstruction because they are indistinguishable from historical signal by phylogeny reconstruction methods (Olsen 1987). The mere presence of noise in the data does not automatically mean that the tree is wrong; however, when the substitution model fails to compensate for the high level of saturation, the quality of topology inference decreases.

Noisy sites are one important source of long branch attraction (LBA) (Felsenstein 1978), a tree-building artefact that causes clustering of the unrelated, fast mutating species on the tree. Though initially thought to affect only maximum parsimony, LBA was later shown to also affect model-based methods under certain conditions (e.g. model misspecification) (e.g. Pol and Siddal 2001). Highly variable sites are especially detrimental when elucidating deeper phylogenetic relationships because they mostly add noise to the data over the generally longer time frames involved (Gribaldo and Philippe 2002).

V. V. Goremykin (✉) · S. V. Nikiforova
IASMA Research and Innovation Center,
Via E. Mach 1, 38010 San Michele all'Adige, TN, Italy
e-mail: Vadim.Goremykin@iasma.it

O. R. P. Bininda-Emonds
AG Systematik und Evolutionsbiologie, IBU – Fakultät V,
Carl von Ossietzky Universität Oldenburg,
26111 Oldenburg, Germany
e-mail: olaf.bininda@uni-oldenburg.de

Several explicit methods for the identification and removal of noisy alignment positions have already been suggested (e.g. Brinkmann and Philippe 1999; Hirt et al. 1999; Lopez et al. 1999; Ruiz-Trillo et al. 1999; Burleigh and Mathews 2004; Pisani 2004; Kostka et al. 2008). From the practical point of view, these methods can be divided into two categories: tree-dependent and tree-independent methods.

Tree-dependent methods, which are the more popular of the two types, were first suggested as a tool in molecular phylogenetics in 1999. The maximum likelihood (ML)-based method by Ruiz-Trillo et al. (1999) and Hirt et al. (1999) involves sorting characters according to their gamma rate assignment by the Tree-Puzzle program, (which uses a neighbour-joining (NJ) tree to assign rates to characters) and the deletion of those sites with the most variable rates. Later, Burleigh and Mathews (2004) used a similar approach, but instead used a maximum parsimony (MP) tree to assign rates to characters. These methods do not require any predefined input tree, but tree-building remains an integral part of the sorting procedure. Thus, their effectiveness depends on the outcome of the tree search through which rates are assigned to positions. All LBA-related artefacts present on the tree can be expected to contribute to the inaccurate estimation of the rates of the sites and ultimately to biasing the results of the subsequent site-stripping. Rodriguez-Espeleta et al. (2007) observed that the tree topology used in the gamma-rate based approach affected the degree to which removal of variability helped to recover the predefined benchmark clades.

Another method, the Slow–Fast approach suggested by Brinkmann and Philippe (1999) and Lopez et al. (1999), does require a pre-determined tree topology to estimate the number of changes for each position within a tree using MP (position-based tree lengths). Thereafter, the position classes with the highest the number of changes are iteratively removed from the data, creating a series of the alignment subsets with the reduced variability. This approach also allows selecting monophyletic clusters (subtrees) on the input tree, thereby escaping the influence of individual branches, which would be suspect in case LBA is present. In this case, the sum of the subtree lengths at every position is used as a proxy for the substitution rate. This taxon set partitioning is implemented in the Slow–Faster program (Kostka et al. 2008) and should, in theory, represent an improvement over the input of a complete tree. Even so, the method remains topology-dependent because omitting some branches does not liberate the results from any potential incorrect order of branching within the chosen subtrees.

By contrast, the rationale underlying development of the compatibility method (Pisani 2004; Pisani et al. 2006) was rooted in the observation that because the length of a character in a MP sense is topology dependent, erroneous tree topologies can lead to character lengths being miscalculated and thus the misidentification of putatively fast evolving sites by the Slow–Fast method. Instead, the method of Pisani (2004) is based on the compatibility principle by Le Quesne (1969), namely, that two characters are compatible if they can be mapped on a tree without homoplasy. Practically, the method involves calculating the number of other positions with which each site in the alignment is incompatible. High incompatibility scores are used to identify and to discard fast-evolving sites. The method is based on the expectation that the correct tree will derive from the largest number of compatible characters (which are all assumed to be uniquely derived) and that noisy characters will be incompatible with this set and can be discarded. (In other words, sites that contain phylogenetic information are expected to show fewer incompatibilities than those containing little or no phylogenetic information.)

Despite the availability of several methods of character-stripping, the removal of noisy data has not yet found routine use in analyses of genome-scale alignments that contain saturated sites (e.g. Springer et al. 2001; Goremykin et al. 2009). This is proximally so because the relative capacity of the character-stripping methods in solving real problems of systematics is still not well-investigated, and ultimately so because the practical problem of finding an objective stopping criterion for character removal remains unsolved. As perhaps first acknowledged by Pisani (2004), “All methods of character selection pose the problem of finding an optimal cut-off value under which characters should not be deleted. How to discriminate characters the deletion of which could improve phylogenetic accuracy, therefore, is the important key. Still, this is the most complex step of any character selection protocol.” In seeking to address this issue, Pisani (2004) listed the significant and systematic deterioration of the results (as indicated by the appearance of obviously nonsensical clusters and/or substantial loss of resolution, support, or decrease in the likelihood of the recovered trees) as the key factor to be considered when deciding to stop the character-stripping process; he also suggested testing for the presence of the clustering signal amongst the deleted characters. Yet no particular guidelines for decision-making (e.g. which concrete stage of the result deterioration should be considered significant? How substantial the loss of support should be?) were suggested in this or the follow-up study (Sperling et al. 2009).

Indeed implementing some of the above criteria is difficult. For instance, relying on a decrease in the support for the tree branches as a criterion requires not only some threshold specification, but also the knowledge of which branches are correct and which are not (since, obviously, decrease of support for LBA artefacts is a good sign). Likewise, at least partial knowledge of the true tree is required to define "nonsensical" clusters. Finally, a decrease in the likelihood scores computed based on the different models and different (shortened) data may be indicative of several factors, not only of the degradation of phylogenetic signal in the data.

We therefore sought to build on the previous methods by designing an automated procedure of noise removal and evaluation of results that also provides a stopping criterion for the removal of the characters. Crucial to our method is that both steps are tree-independent, thereby avoiding the potential errors noted above. We test our method using genome-scale data (which are affected by saturation and long-branch attraction artefacts) for a well-publicized test case: the hypothesis of rodent polyphyly. The issue of the basal phylogeny of placental mammals caused an intensive discussion in the recent past. It is generally held that support for the hypothesis of basal rodent polyphyly derived in part from noisy data and the hypothesis is largely discredited today. The case therefore provides an ideal proof-of-concept for our proposed method. From there we go on to compare the effectiveness of our proposed method with several of the methods introduced above.

The Test Case: Rodent Polyphyly and Relationships Amongst Placental Mammals

The hypothesis of rodent polyphyly, first postulated by Graur et al. (1991) on the basis of MP analyses of nuclear protein sequences is a well-publicized issue of apparent long branch attraction artefacts in phylogenetics. The tree obtained by Graur et al. suggested that the guinea pig (and, by extension, other caviomorph rodents) diverged earlier in the eutherian radiation than did the remaining rodents. This hypothesis received strong criticism at the time from both classical morphologists (e.g. Luckett and Hartenberger 1993) and other molecular phylogeneticists, who suggested the use of either ML-based analyses (e.g. Hasegawa et al. 1992; Cao et al. 1994) or less problematic data (e.g. Allard et al. 1991). Nonetheless, numerous molecular studies supported the hypothesis of rodent polyphyly for nearly a decade, with either Caviomorpha (Li et al. 1992; Graur et al. 1992; D'Erchia et al. 1996) or Myomorpha (mouse, rats, and allies; Ma et al. 1993; Janke et al. 1997; Phillips et al. 2001; Reyes et al. 1998, 2000a) being identified as having diverged earlier than the remaining rodents, and often as one of the earliest diverging branches amongst placental mammals.

Analyses of nuclear genes with increased taxon and character sampling (Madsen et al. 2001; Murphy et al. 2001a, b; Amrine-Madsen et al. 2003; de Jong et al. 2003), however, have laid the foundation for the currently held view on eutherian evolution (see Fig. 1b for a summary; see also Springer et al. 2004) and the virtual abandonment of the rodent polyphyly hypothesis. Here, a monophyletic Rodentia that as sister to Lagomorpha (rabbits and pikas) forms Glires is favoured. Glires together with Euarchonta (containing Primates amongst several other orders) comprise the superorder Euarchontoglires, which, in turn, is sister to the superorder Laurasiatheria (together forming Boreoeutheria). The two remaining superorders, Afrotheria and Xenarthra, are often viewed as more basal lines in eutherian evolution (e.g. Springer et al. 2004), although the placement of the root of the placental tree remains hotly debated (see Prasad et al. 2008 and references therein).

However, analyses of nuclear genes and of complete mitochondrial genomes remain in conflict, even in the face of increased taxon sampling and the use of global GTR-based ML models for the latter. To date, mtDNA-based trees congruent to ones obtained using nuclear (nu) DNA data were achieved only by means of constraining the ML search space (Lin et al. 2002a), by creating a custom substitution model (Gibson et al. 2005), by discarding some mitochondrial genes from the analysis and eliminating synonymous leucine sites (Reyes et al. 2004), or by employing a mix of models to the customary chosen data partitions (Kjer and Honeycutt 2007). By themselves, ML analyses of mitochondrial data employing global models from the GTR family (i.e. with or without correction for invariable sites and rate heterogeneity) tend to support rodents being basal or near-basal and polyphyletic (Reyes et al. 1998, 2000b; Mouchaty et al. 2001; Arnason et al. 2002; Lin et al. 2002b). In investigating the utility of mtDNA data for inferring deep nodes in the mammalian radiation, Springer et al. (2001) made specific reference to saturation in mtDNA data as a factor causing incongruence between mitochondrial and nuclear-based analyses. Saturation in mtDNA data was also noted by several other authors (e.g. Pesole et al. 1999; Phillips and Penny 2003; da Fonseca et al. 2008).

Together, these results indicate that rodent polyphyly is likely an artefact deriving from superimposed mutations in mitochondrial genomic data, thereby providing an ideal test case for our method. If mammalian mtDNA data are indeed more prone to saturation effects than are nuDNA due to their generally higher rate of substitution (see Bininda-Emonds 2007), then the step-wise removal of the most variable characters from an alignment of complete mitochondrial genomes should yield both a monophyletic Rodentia and a topology for placental mammals that is more consistent globally with the currently accepted view.

These data, therefore, provide an important proof of concept for our suggested methodology of character-sorting based on the observed variability. Using sparse taxon sampling characteristic of the data sets from the time of the above discussion and choosing distant outgroups subtended by long branches we re-created the LBA-affected tree (with Rodentia and Lagomorpha appearing at the base of the subtree of placental mammals). Knowledge of the "true tree" of the placentals allowed us to compare the performance of the different character-sorting methods on the basis of the recovery of benchmark clades in this test case.

Fig. 1 Representative changes in tree topology as a result of the ▶ removal of the most divergent characters in multigene alignments for mammals (mtDNA). **a** Topology obtained from the ML analysis of the complete mammalian mtDNA alignment (A_0). **b** Topology obtained immediately after the sharp rise in *r* values in correlation analyses of the conserved (A_n) and each of the variable partitions (B_n and C_n) (see Fig. 2). The tree shown was built from 9549 bp long A_{10} subset. Trees built from A_{11} and A_{12} subsets have the same topology. The bootstrap numbers in the trees were produced by PHYML, using ML start trees (built employing fitted models, GTR + I + G, with the help of PAUP*) and using general specification of GTR + I + G model

Materials and Methods

Description of the Test Procedure

Overview

The complete test procedure we describe here is implemented in the Perl script NoiseReductor.pl. During its execution, the script minimally invokes PAUP* for UNIX (Swofford 2002), and one of two additional Perl scripts depending on the chosen character sorting method. In addition to sorting by the tree-independent criterion of observed variability (OV), or by some other supplied measure of the substitution rate (performed by sorter.pl), a second script (gamma_sorter.pl) can sort the characters according to their gamma rates. Both gamma_sorter.pl and NoiseReductor.pl will invoke ModelTest (Posada and Crandall 1998) as needed to determine the optimal model of evolution. All these Perl scripts are available upon request.

The procedure consists of two main steps. In the first step, the desired measure of character variability is calculated and then the alignment is sorted accordingly. Alternatively, one can put a text file named 'sortme_so' in the current directory, which contains positive numbers, one number per line, that correspond to the desired approximation of the substitution rate for the alignment positions (the first line in the file should have the proxy of the substitution rate for the first alignment position, the second line for the second position and so on). Such files, for instance, (*.plot) are automatically produced by the COMPASS program by Simon Harris (www.ncl.ac.uk/microbial_eukaryotes/downloads.html) in which compatibility-based sorting approach is implemented (Pisani 2004; Sperling et al. 2009). If the file 'sortme_so' s found in the current directory, then the script sorter.pl will skip computation of OV, and will sort positions in the alignment by the measure of substitution rate provided in the file. In the second step, the most variable subsets of the alignment are removed iteratively, with two series of correlation analyses of pairwise distances between the same sets of species pairs being used to determine when character removal should cease. We now describe each step in more detail.

Character Sorting

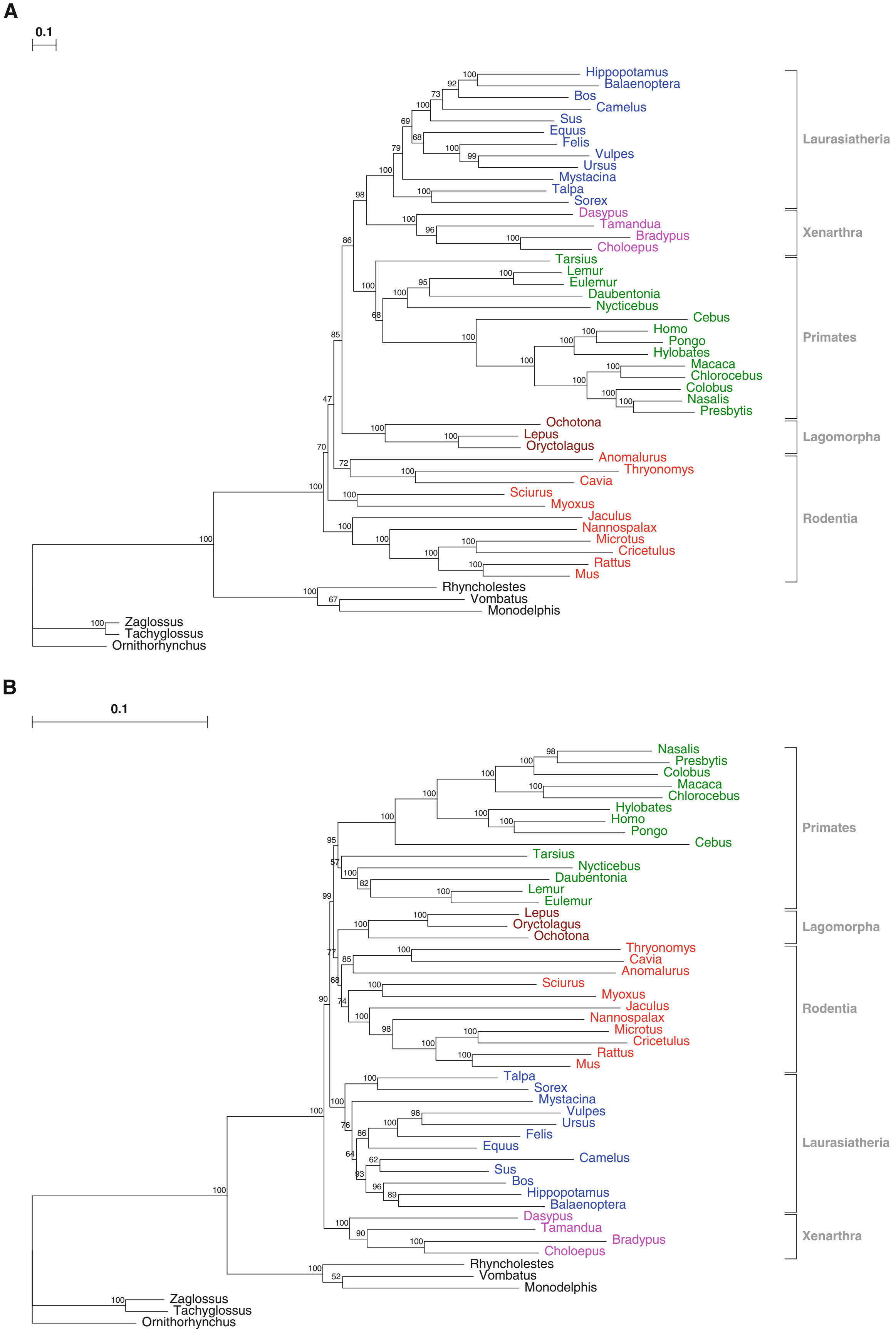
The user can choose to sort positions in ascending order of their variability based on (1) their observed variability independent of any tree topology (2) the assignment of substitution rates provided in the input file 'sortme_so' or (3) their largest contribution to a gamma rate category under the best ML model. Before the analysis, the user should specify a threshold value, either a certain gamma rate class or a certain value of the chosen proxy of substitution rate as needed, at which the shortening process should be stopped to prevent the removal of constant partitions or those of minimal variability.

To calculate our sorting criterion (observed variability, OV), all sequences for a given position are compared in a pair-wise fashion. Mismatches are scored as 1 and matches as 0; the mean value amongst all the comparisons for a given position is used as the measure of character variability in the subsequent data sorting:

$$\mathrm{OV} = \mathrm{sum}(1..k)\{d_{ij}\}/k$$

Here k is the number of pair-wise comparisons made for a given position and d_{ij} is the score of character variability in each pair-wise comparison made (can be either 0 or 1). If n is the number of aligned sequences which do not have a gap at the given alignment position, then $k = (n^2 - n)/2$. The OV measure is actually similar to PI (average pair-wise differences) from population genetics, which is calculated by taking all pairs of individuals in a population and computing the average number of differences between them.

Thus, observed variability (like compatibility scores) is calculated without reference to any specific tree and are free from any systematic bias in the estimation of the substitution rates for all sites in alignment that a wrong input topology might cause. At the same time, an obvious drawback of the method is that it does not incorporate any fitted substitution model, something that has been shown repeatedly to improve the accuracy of phylogenetic inference (e. g. Gadagkar and Kumar 2005; Gaucher and Miyamoto 2005). Indeed, sorting based on the observed variability actually would yield the same results as the sorting based on the native Jukes and


A
0.1
100
92
73
100
69
68
100
99
79
100
100
98
100
96
100
86
100
95
100
68
100
100
100
100
100
100
100
100
85
47
100
100
70
72
100
100
100
100
100
100
100
100
100
67
100
Hippopotamus
Balaenoptera
Bos
Camelus
Sus
Equus
Felis
Vulpes
Ursus
Mystacina
Talpa
Sorex
Dasypus
Tamandua
Bradypus
Choloepus
Tarsius
Lemur
Eulemur
Daubentonia
Nycticebus
Cebus
Homo
Pongo
Hylobates
Macaca
Chlorocebus
Colobus
Nasalis
Presbytis
Ochotona
Lepus
Oryctolagus
Anomalurus
Thryonomys
Cavia
Sciurus
Myoxus
Jaculus
Nannospalax
Microtus
Cricetulus
Rattus
Mus
Rhyncholestes
Vombatus
Monodelphis
Zaglossus
Tachyglossus
Ornithorhynchus
Laurasiatheria
Xenarthra
Primates
Lagomorpha
Rodentia
B
0.1
98
100
100
100
100
100
100
100
95
57
100
82
100
99
100
100
77
85
100
68
74
100
100
98
100
100
90
100
100
76
100
98
86
64
62
93
96
89
100
90
100
100
100
52
100
Nasalis
Presbytis
Colobus
Macaca
Chlorocebus
Hylobates
Homo
Pongo
Cebus
Tarsius
Nycticebus
Daubentonia
Lemur
Eulemur
Lepus
Oryctolagus
Ochotona
Thryonomys
Cavia
Anomalurus
Sciurus
Myoxus
Jaculus
Nannospalax
Microtus
Cricetulus
Rattus
Mus
Talpa
Sorex
Mystacina
Vulpes
Ursus
Felis
Equus
Camelus
Sus
Bos
Hippopotamus
Balaenoptera
Dasypus
Tamandua
Bradypus
Choloepus
Rhyncholestes
Vombatus
Monodelphis
Zaglossus
Tachyglossus
Ornithorhynchus
Primates
Lagomorpha
Rodentia
Laurasiatheria
Xenarthra

Cantor model, applied to a single alignment position (because the latter is proportional to the scope of the observed sequence dissimilarity), which was previously observed to be amongst the worst ML models describing character distribution in non-simulated genome-scale data (Goremykin and Hellwig 2006).

Thus, NoiseReductor.pl also includes a gamma-rate option to increase the range of situations in which the ML-based sorting methods of Ruiz-Trillo et al. (1999) and Hirt et al. (1999) are useful by employing more realistic trees to assign rates to characters. When sorting by gamma category, the settings of the best model are first determined via gamma_sorter.pl using the standard ModelTest procedure (i.e. using a NJ tree built from uncorrected distances to fit the ML model parameters) based on the Akaike information criterion. Optionally, gamma_sorter.pl can further refine the model by having PAUP* compute the ML tree using the settings of the above model, and then using this ML tree as the base tree for a second round of model fitting via ModelTest. Indeed, we suggest to not to use NJ tree for assignment of rates to characters (as is the case in the above studies), but to use it as the foundation for the second round of model fitting, which should make the results of character-sorting less LBA-prone. Using the combination of the best model (chosen either using the NJ or ML tree as the base tree for calculation) and the ML tree obtained using this model, PAUP* is then used to assign sites to up to 100 gamma-rate categories as a prelude to sorting the positions in ascending order of their variability (We have observed that applying 100 rate categories occasionally causes PAUP* to crash because of memory allocation problems. In such cases, gamma_sorter.pl automatically reduces the number of categories by one until a result is achieved.), The user can also optionally perform the model selection procedure under topological constraints as specified in a separate tree file.

Naturally, if the optimal model does not indicate the need for gamma correction, sorting by gamma rate category should not be done. Empirically, however, we have yet to observe a single case where a correction for rate heterogeneity was not recommended by ModelTest for a large genome-scale alignment.

Finally, with regard to the OV-based sorting, it should be noted that sites which support internal splits amongst large groups of species will tend to have higher OV scores than sites that support splits of few versus many species. This will have the effect of reducing internal branch lengths disproportionately compared to external branch lengths. Thus, if OV-based noise removal is applied for doing dating estimates, then it might be advisable to use a two-step procedure where the noise-reduced alignment is used to get the topology and the complete alignment is used to then estimate the branch lengths. Also, it should be noted that, since the part of the total noise in the data causing LBA is associated with concrete taxa, the algorithm can be expected to eliminate the positions leading to LBA from the data more efficiently with increasing proportion of such taxa in alignment.

Correlation Analyses

Regardless of the model-fitting and sorting procedure, the sorted alignment is written to a new file (the master file, A_0) that is subjected to iterative rounds of shortening by repeatedly removing a user-specified number of positions from the most divergent end. At each shortening step (n), three data partitions are written to disc: the shortened alignment (partition A_n), all deleted positions (partition B_n) and only those positions removed in the current shortening step (partition C_n). Note that partition C_n is a subset of partition B_n, and that, after the first shortening step ($n = 1$), partitions B_1 and C_1 are identical.

At each step of the shortening procedure, NoiseReductor.pl then fits the optimal ML model to each of the three partitions according to the same procedure used for the initial model fitting above (i.e. using either the NJ tree or the ML tree). We have observed that execution of some of the models suggested by ModelTest for the most divergent data partitions can occasionally cause PAUP* to crash. In such cases, NoiseReductor.pl automatically modifies the model in an attempt to derive a result: first, for any rate assigned a value of 0.0000 in the substitution rate matrix (such values cause crashes with lset command) numbers from 1 to 9 are added to the fifth decimal place; second, if all rates are higher then 0, then numbers from 1 to 9 are added to the fifth decimal place of the first substitution rate in the matrix (in the case of the GTR model, this would be the A–C rate); and third, if the rates are equal (and absent in the model specification in PAUP format), then numbers from 1 to 9 are added to the fifth decimal position of the first base frequency given in the model specification. If all the above modifications fail to deliver a result, or the model description contains no base frequencies and no rate parameters (which is the case if the base frequencies are in equilibrium and all the rates are equal), then all the above modifications are tried with the optimal model found on the basis of hierarchical likelihood ratio tests. Finally, if this also fails, the script attempts to fit a $GTR + I + G$ model to the data using the lscore command.

For each fitted model, a distance matrix for each data partition (A_n, B_n, and C_n) is determined using PAUP* at each shortening step. (When model fitting is done using the refined two-step procedure above, two matrices are determined, one for each of the models.) These distance matrices form the basis of a pair of correlation analyses between the same taxa in the matrix derived from partition A_n and those in the each of the matrices derived from B_n and C_n to determine when the noisiest sites have been removed from the alignment such that the results based on the remaining data (A_n) are unlikely to be misled significantly by saturation. When model fitting is done using the refined procedure, the script first compares the matrices determined in the first round of model fitting, as outlined above, and then it compares the matrices determined in the second round of model fitting. All correlation analyses use both Pearson product–moment correlation coefficients (*r*) and Spearman's rank correlation coefficients (ρ) based on the absolute distance values and their relative ranks, respectively, within each analogous matrix.

These correlation analyses form the basis of a stopping criterion indicating when the noisy sites have largely been removed from the main data partition. In the ideal case where the model and its settings describe all the data effectively, strong positive correlation will be present between the distances between the same sets of species pairs estimated from each partition because the correction for superimposed mutations works effectively for both data partitions. The correlation would also be relatively linear, and not a curve as for the plot of corrected versus uncorrected distances.

By contrast, the distribution of distance values should lose its linear character when the model fails to describe the data adequately (e.g. at high levels of saturation when the variance in the distance estimation exceeds the distance values themselves). High levels of saturation will lead to a loss of exactness in the prediction of evolutionary distances because of the higher variances in the estimates of those distances. Thus, weak correlations in the relative ranking of distance estimates for the same species pairs between the conserved partition and in each of the variable alignment partitions B_n and C_n would indicate that the chosen model is not compensating adequately for superimposed mutations within the latter partitions (Because weak correlations can also derive from poor model choice, optimal model settings are determined by Noise Reductor.pl via ModelTest each time prior to distance estimation, thereby minimizing the chances of severe model misspecification.). As such, the strength of the correlation should improve with the continued removal of noisy sites. In other words, the absence of strong correlation is interpreted as the failure of the fitted ML model to accommodate multiple substitutions (and the associated non-phylogenetic signals) that remain in A_n and, consequently, as an argument for the discarding and continued removal of the most variable partitions. It is only when all correlation analyses at a given shortening step start to yield strong positive values that the level of saturated data in partition A_n has been reduced to a degree where the tree derived from these data would be largely unaffected by the artefacts from noisy data and could be viewed as robust in the sense of our test.

In addition, NoiseReductor.pl also provides the means to estimate the absolute scope of saturation in the variable partitions B_n, and C_n as further information for the user. These estimates are based on either a correlation analysis (both *r* and ρ values) between the ML and uncorrected *p*-distances estimated from the variable partitions or the deviations between the mean values of each of the ML and *p*-distances calculated from these partitions.

Additional Features

Although the Perl scripts are designed primarily to identify noisy positions in a DNA alignment, they can also be used to help facilitate standard ML analysis using PAUP*. For instance, if character sorting via gamma_sorter.pl is not invoked, the two-step model fitting procedure can be used to automate the process of determining the ML tree under the optimal model of evolution. In addition, because all model fitting and tree building analyses for all partitions can be automatically run in parallel by sorter.pl, the scripts will benefit from being run on machines with multiple CPUs or CPU cores. gamma_sorter.pl can also be used to perform a non-parametric bootstrap analysis (Felsenstein 1985) of the original data, with the script starting a user-defined number of PAUP* processes in parallel, each computing a tree for a single bootstrap replicate. The ability of the scripts to automatically make use of multiple processors and cores as far as possible also counteracts the computationally intensive nature of the method, which derives in part from the numerous rounds of model-fitting (i.e. after each shortening step).

Assembly of the Data Set for the Tests

The nucleotide data set was obtained from GenBank in the form of complete mitochondrial genomes for 55 mammalian species (49 eutherians, three marsupials and three monotremes). Alignment was performed initially using ClustalW and subsequently improved by eye, both using the Seaview sequence editor (Galtier et al. 1996). The final alignment length was 11,549 bp after elimination of all regions of insecure alignment as determined by visual inspection.

All model-fitting analyses used the two-step model selection outlined above, and all correlation analyses were based on shortening steps of 200 bp. In testing the efficacy of our procedure, we focused in part on four benchmark clades: a monophyletic Rodentia, Glires, Euarchontoglires (=Glires + Primates here) and Boreoeutheria.

Proxies of the positional substitution rates were also determined used compatibility-based criteria (Pisani 2004; Sperling et al. 2009) implemented in COMPASS program by Simon Harris (www.ncl.ac.uk/microbial_eukaryotes/downloads.html): (1) Le Quesne probability (LQP) values (Le Quesne 1969), and (2) direct compatibility.

Results and Discussion

Failure of the Optimal Global ML Model to Identify the True Tree

ML analyses of the complete mitochondrial data set did not support rodent monophyly, nor did they recover any of the remaining benchmark clades (Fig. 1a). Instead, rodent polyphyly was indicated, with lagomorphs and caviomorph and murid rodents forming successively more distant sister groups to the remaining placentals. Bootstrap support for the relevant nodes was comparatively weak (85%, 47% and 70%, respectively). In addition, the widely accepted monophyly of the benchmark clade Boreoeutheria was contradicted with Xenarthra being strongly supported as sister to Laurasiatheria (bootstrap support = 98%). These general results were true even with the use of the two-step model fitting procedure, which should provide a better choice of the substitution model. We therefore applied our procedure to identify noisy positions and to assess whether their removal led to the recovery of our four benchmark clades. By doing so, we also tested the performance of our procedure with regard to (1) efficiency of character sorting using OV versus assignment to gamma rate categories, Slow–Fast-based sorting and compatibility-based sorting; and (2) determination of a stopping criterion for the removal of noisy data.

Efficiency of Character Sorting

The best character-sorting approach should yield the highest concentration of saturated positions at the most divergent end of the sorted alignment. Estimates of the scope of saturation based on the deviation of the mean values of the corrected ML distances from those of the uncorrected *p*-distances (see above) in the variable partitions B_n revealed that OV-based sorting outperformed gamma-based sorting and compatibility-based sorting in all analyses (Table 1). (These conclusions could not be verified with the Slow–Faster program because it does not create variable data partitions).

OV-based sorting eliminated the basal rodent paraphyly artefact and recovered Euarchontoglires at the seventh shortening step (1400 positions deleted). This was much faster than all other methods tested. Neither the gamma-rate-based sorting nor the two compatibility-based methods tested recovered a single benchmark clade at this or the next four subsequent shortening steps. The Slow–Faster program

Table 1 Relative prowess of character sorting based on observed variability, gamma-rate categories, direct compatibility scores (as implemented in the program COMPASS) and Le Quesne probability (as implemented in the program COMPASS) in concentrating saturated positions towards the most variable end of the sorted alignment

Shortening step	Partition A length	Saturation in B partition, OV sorting	Saturation in B partitions, rate-based sorting	Saturation in B partitions, direct compatibility-based sorting	Saturation in B partitions, LQP-based sorting
1	11,349	11432.00	46.50	219.65	0.57
2	11,149	281.44	56.04	113.08	1.70
3	10,949	205.56	39.98	142.18	2.39
4	10,749	210.84	33.78	65.55	2.38
5	10,549	140.40	53.21	54.32	2.99
6	10,349	158.19	26.45	59.14	3.33
7	10,149	58.98	27.46	101.53	3.34
8	9949	36.36	17.06	116.05	3.27
9	9749	20.29	15.10	92.22	3.38
10	9549	4.90	10.75	79.59	3.83
11	9349	4.63	12.14	72.79	4.07
12	9149	5.26	25.77	47.80	4.05
13	8949	5.74	50.08	11.62	4.33
14	8749	5.82	74.66	9.73	4.38
15	8549	5.60	72.25	7.90	4.27
16	8349	5.55	70.76	7.32	4.24
17	8149	5.46	31.16	6.47	4.11
18	7949	5.07	16.62	5.70	4.10
19	7749	4.80	12.10	5.43	3.93
20	7549	4.34	9.08	4.98	3.76

Saturation in the variable data partitions was estimated using deviations of mean uncorrected distances from the mean evolutionary distances amongst all terminal taxa calculated using optimally-fitted ML models also did not recover a single benchmark clade within this shortening range (up to 2200 deleted positions) even when the input tree (unrooted ML tree shown in Fig. 1b) was correct and Primates, Glires, Laurasiatheria, Xenarthra and the outgroup (the branch subtending monotremes plus marsupials on the above tree) were specified as input taxon partitions.

Similarly, sorting and deletion based on OV also recovered all four benchmark clades faster than all other methods tested (9 shortening steps, 1800 positions removed). Gamma-based sorting required 13 shortening steps (2600 positions deleted) to recover the four clades as did COMPASS with the direct compatibility option turned on. When LQP values were used instead, additional three steps were required (3200 positions removed). The Slow–Faster program with the correct tree divided onto five input subtrees as described above yielded a tree with these clades when 2629 positions were deleted.

It should be mentioned that when we used the (wrong) tree shown in Fig. 1a, as input for the Slow–Faster program, and specified (i) the branch subtending monotremes plus marsupials plus Glires and (ii) the branch subtending Xenarthra plus Laurasiatheria as input subtrees, a tree containing all four benchmark clades was not recovered at any shortening step. By contrast, gamma-rate based sorting with the same wrong input tree (for model specification) still managed to recover the benchmark clades and at about the same point as did COMPASS and Slow–Faster with the correctly specified input branches (see above). These results demonstrate the difficulty in appraising the validity of the Slow–Faster method in situations where the true tree is not known (which unfortunately, is exactly the situation when the method is needed) because the absence of topological changes during the shortening process could be misinterpreted as indicating a robust initial topology. Also, despite the valid arguments regarding importance of topology-independence testing by Pisani (2004), LQP based noise reduction as implemented in COMPASS faces the same problem.

Determination of a Stopping Criterion

Successive removal of the noisiest data from the mammalian data set (see above) yielded a tree that was increasingly congruent with currently accepted views on eutherian evolution: Xenarthra as sister to the remaining placentals and the monophyly of Rodentia, Glires and Euarchontoglires (Fig. 1b). However, the recovery of such benchmark clades as a stopping criterion for character removal is generally unsuitable for two reasons. First, it can often be applied in only a few cases (e.g. when the path of evolution is definitely known from the fossil record). Second, benchmark clades, when available, are often chosen on the basis of their overall robustness, suggesting that their recovery might be more resistant to the negative effects of noisy data.

Thus, we examined if the correlation analyses of pairwise distances might provide an alternative solution to the problem of determining a stopping criterion for character removal. Specifically, we evaluated two approaches: a comparison of the corrected and uncorrected distances calculated from the variable partitions C_n and B_n (referred to hereafter as analysis 1), and a comparison of evolutionary distances calculated from the conserved A_n partition to those from the variable data partitions C_n and B_n (analysis 2). In so doing, we focused solely on OV-based sorting here based on its superior performance in (i) recovering the benchmarks clades; and (ii) concentrating saturated positions towards the most variable end of A_0 compared to all other methods tested (see Table 1 and discussion above).

Intervals exist in both analyses 1 and 2 where *r* values rise sharply (ends marked with orange lines in Fig. 2). These 'break intervals' are short, never exceeding two shortening steps, and are observed at the same or immediately adjacent shortening step in analyses 1 and 2 (step 9 or 10, when 2000 bps ((C_n), results not shown) or 2200 bps ((B_n), Fig. 2) have been removed, respectively). Inspection of the variable alignment partitions (B_n and C_n) of removed positions from before the break intervals revealed that they contained positions that are absolutely dissimilar amongst certain OTUs, with pairwise ML distances being unreliable and unrealistically high (from 10 to $\sim$20,000 substitutions per site). At such high values, small changes in base composition can cause disproportionally large differences in ML distance values (results not shown). Examination of all distance matrices produced during the shortening process revealed that the break intervals occur after all such outliers are finally removed, in part because of the sensitivity of Pearson correlation coefficients to outliers. Thus, *r* values estimated in comparisons involving the most divergent alignment partitions behave erratically on the left side of the graphs and show greater similarity to the Spearman's rank correlation coefficients on the right side after the removal of the outliers.

Empirically, all mtDNA-based trees calculated from the conserved alignment partition after the end of the break intervals but before the shortening step leading to the point of information loss (defined by the start of branch collapse leading to unresolved clusters on the tree; marked by red lines in Fig. 2) recover all benchmark clades, including a monophyletic Rodentia. Indeed, all benchmark clades were recovered also before the break intervals (see above), indicating that the data set appeared to contain more residual noise than just the substitutional saturation associated with the rodents and surrounding clades. Examination of the first

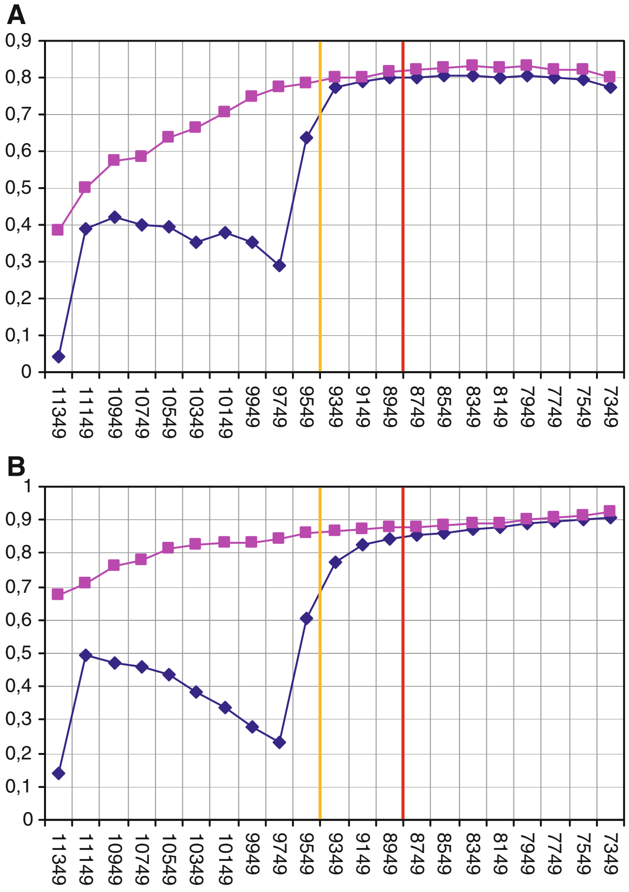


Fig. 2 Results of the correlation analyses involving pairwise comparisons between the conserved (A_n) and variable partitions (B_n). *Pink* and *blue* dotted lines on the graphs represent ρ and r values, respectively. Digits on the x-axis depict shortening steps (i.e. lengths of the remaining A_n partitions). *Orange* lines on the graphs denote the ends of the break intervals (see main text), whereas *red* lines denote points beyond which removal of variable positions caused the collapse of some tree branches. Graph A presents the results of the analysis 2 (see main text) and graph B presents the results of the analysis 1

tree obtained after the break interval at 10th shortening step (Fig. 1b) reveals higher bootstrap support for all nodes along the backbone of the tree than were obtained using the entire alignment. Support for Glires (bootstrap = 77%) and Rodentia (bootstrap = 68%) is admittedly relatively low, but still generally higher than for the conflicting nodes in Fig. 1a. Virtually all the trees obtained after the break interval but before the point of information loss are congruent to the topology shown in Fig. 1b. The exception is the last tree obtained at the thirteenth shortening step, where Tarsius was placed as the sister group to the remaining primates and lagomorphs were sister to the primates. This tree, however, might indicate the first signs of information loss in the matrix.

In general, our correlation-based stopping criterion only operates effectively when the sorting method is good and efficiently concentrates noisy data to the right-hand side of the sorted alignment. For example, in applying LQP-based sorting, which was the slowest in recovering the four benchmark clades (see above), we observed a steady growth of the correlation values without any breaking intervals (result not shown). This was because the sorting algorithm failed to efficiently concentrate saturated positions in the variable partitions, such that ML distance matrices based on these partitions did not include any outliers. This failure, in turn, would have led to the erroneous conclusion that the tree built from A_0 partition could have been accepted in the sense of our correlation test based on the absence of a breaking interval in variability reduction process.

By contrast, OV-based character sorting does appear able to effectively identify and concentrate the most highly variable sites from genome-scale alignments, showing good performance compared to other sorting methods (see Table 1). Thus, we can expect the appearance of some extremely high distances in the variable matrices, with the consequence that failure of the optimal ML model to correct for superimposed substitutions in divergent data partitions will become apparent in the correlation analyses as break intervals (Fig. 2). We therefore suggest that the removal of variable positions from large-scale alignments should be continued at least until the very end of sharp rise in r values in any series (A or B) of correlation analyses, as these most variable alignment subsets contain the most misleading positions in the data. By contrast, lack of any break intervals during the shortening process with the OV method would indicate that the tree obtained from the original unsorted alignment data is not strongly biased by noise, and can be accepted in the sense of our test. This does not mean that the tree is correct; it means that the tree structure is not likely to be affected by one factor which might pose obstacles to correct phylogeny reconstruction—superimposed mutations. Other factors, related, for example, to the compositional heterogeneity, to the assumptions about time-reversibility of the substitution process implemented in the currently available substitution models, to the regional variation in rates of substitutions amongst different taxa, to concerted gene evolution, etc. might cause errors too and, thus, additional data exploration is always advisable.

Conclusions

To date, recommendations in the literature to improve the accuracy of phylogenetic analyses have focused largely on increased taxon and character sampling (including the reduction of missing data) and the application of increasingly more complex ML models. Although such steps are undoubtedly beneficial, they still might fail to counter the negative impact of non-phylogenetic signals, which, as Jeffroy et al. (2006) and Rodriguez-Espeleta et al. (2007) have pointed out, are associated with saturated positions. Our study highlights in part the potential impact of saturated positions in phylogenomic analyses. This is not a trivial issue given the growing use of unfiltered genomic-scale data to infer phylogenetic relationships. Phylogenetic analyses of angiosperms represent a prime example of this trend (e.g. Stefanovich et al. 2004; Leebens-Mack et al. 2005; Qiu et al. 2005; Moore et al. 2007; Jansen et al. 2007). In fact, Stefanovich et al. (2004) and Leebens-Mack et al. (2005) explicitly advocate the retention of all characters in their studies.

Building on previously suggested methods of noise-reduction, and on the novel method of noise reduction based on the observed sequence variability, we present here an automated procedure of noise removal and provide guidelines for its optimal use. At a minimum, the method provides means of data exploration for the end user with respect to the potential impact of noisy, saturated positions on the inferred phylogenetic hypotheses. Rate variability amongst nucleotide positions in the alignment can be estimated using any published measure. However, our results indicate that our suggested proxy of the substitution rate, observed variability, appear to be more effective at identifying the most saturated positions compared to other noise-reduction methods tested.

Importantly, our results point to the existence of an apparently objective stopping criterion, indicating the point at which the noisiest and most misleading positions have been removed from the alignment. The stopping criterion is designed to estimate the ability of the fitted ML model to compensate for multiple substitutions per site in light of the observation that the mere presence of saturation does not automatically impact negatively on likelihood-based phylogenetic analysis because ML can account for non-observed, superimposed substitutions. Instead, the potential for negative effects depends on whether or not the model applied can accommodate the given level of saturation. This can be observed by comparing the pairwise ML distances calculated from two partitions of the same alignment, one containing conserved characters and the other the variable characters. Absence of correlation between ML distances based on conserved and variable partitions indicates that correction for superimposed substitution does not work well in the variable partition. Strong positive correlation indicates that correction works well, and character removal process may be stopped. Such analysis of evolutionary distances in alignment partitions have been used to measure and compare the degree of correlation between synonymous and non-synonymous substitutions in coding sequences (e.g. Goremykin et al. 1996; Jabbari et al. 2003). However, we are not aware of its application to guide the identification and potential removal of saturated positions.

This is not to say, however, that all noisy data are removed at the point where the stopping criterion is applied. Visual inspection of the scatterplots of the evolutionary distances versus non-corrected distances in B_n partitions after application of the stopping criterion in our two examples revealed that the distribution of the dots is still curve-like, which is indicative of at least some degree of saturation in the data. However, because variability in DNA sequence data is a continuum (from hypervariable to invariant sites), there is no method to define noisy sites per se. Indeed, our criterion might be held to be too liberal in that some noisy sites are retained. However, it does seem to point to a natural dichotomy in the data, where r values increase sharply indicating improved performance of ML model in compensating for multiple substitutions per site. Whether character removal beyond this point is justified cannot be answered on the basis of our tests. As variability in the data will often be increasingly reduced with subsequent character removal, correlation values might continue to increase to the point where the drawbacks from the loss of phylogenetic information outweigh any additional gain from the removal of saturated positions. Because the correlation analysis that we suggest here provides estimates only for the removal of noise and not for the retention of phylogenetic signal, it can only be used to identify and eliminate that part of the alignment causing ML models to yield extremely unreliable estimates of the substitution process.

We believe that the scripts that we present here can help one to identify better supported hypotheses of phylogenetic relationships of species, and therefore, can enrich the means available to assess the reliability of phylogenetic analyses. As our analyses of the mammalian mtDNA data set showed, our method can arguably improve the net result of phylogeny reconstruction, especially for deeper nodes, where the inference of phylogeny is especially obscured by multiple substitutions and the resulting long-branch attraction. In many such cases, including amongst the basal angiosperms, the usual suggestion of breaking of long branches cannot be applied (e.g. for isolated taxa with extinct relatives, taxa that are not infrequently found at the base of the tree). Instead, coupled with more traditional approaches of adding more characters and species to the alignment, the assessment and potential removal of saturated positions should be a necessary procedure in helping to improve building of phylogenetic trees.

Acknowledgments O.B.E. received funding support in part granted by the Deutsche Forschungsgemeinschaft (BI 825/3-2).

References

Allard MW, Miyamoto MM, Honeycutt RL (1991) Test for rodent polyphyly. Nature 353:610–611

Amrine-Madsen H, Koepfli KP, Wayne RK, Springer MS (2003) A new phylogenetic marker, apolipoprotein B, provides compelling evidence for eutherian relationships. Mol Phylogenet Evol 28:225–240

Arnason U, Adegoke JA, Bodin K, Born EW, Esa YB, Gullberg A, Nilsson M, Short RV, Xu X, Janke A (2002) Mammalian mitogenomic relationships and the root of the eutherian tree. Proc Natl Acad Sci USA 99:8151–8156

Bininda-Emonds ORP (2007) Fast genes and slow clades: comparative rates of molecular evolution in mammals. Evol Bioinf 2007:3:59–85

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

Burleigh JC, Mathews S (2004) Phylogenetic signal in nucleotide data from seed plants: implications for resolving the seed plant tree of life. Am J Bot 91:1599–1613

Cao Y, Adachi J, Yano T, Hasegawa M (1994) Phylogenetic place of guinea pigs: no support of the rodent-polyphyly hypothesis from maximum-likelihood analyses of multiple protein sequences. Mol Biol Evol 11:593–604

D'Erchia AM, Gissi C, Pesole G, Saccone C, Arnason U (1996) The guinea-pig is not a rodent. Nature 381:597–600

da Fonseca RR, Johnson WE, O'Brien SJ, Ramos MJ, Antunes A (2008) The adaptive evolution of the mammalian mitochondrial genome. BMC Genomics 9:119

de Jong WW, van Dijk MAM, Poux C, Kappé G, van Rheede T, Madsen O (2003) Indels in protein-coding sequences of Euarchontoglires constrain the rooting of the eutherian tree. Mol Phylogenet Evol 28:328–340

Felsenstein J (1978) Cases in which parsimony or compatibility methods can be positively misleading. Syst Zool 27:401–410

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791

Gadagkar SR, Kumar S (2005) Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous. Mol Biol Evol 22:2139–2141

Galtier N, Gouy M, Gauthier C (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput Appl Biosci 12:543–548

Gaucher EA, Miyamoto MM (2005) A call for likelihood phylogenetics even when the process of sequence evolution is heterogeneous. Mol Phylogenet Evol 37:928–931

Gibson AV, Gowri-Shankar P, Higgs G, Rattray MA (2005) Comprehensive analysis of mammalian mitochondrial genome base composition and improved methods. Mol Biol Evol 22:251–264

Goremykin VV, Hellwig FH (2006) A new test of phylogenetic model fitness addresses the issue of the basal angiosperm phylogeny. Gene 381:81–91

Goremykin VV, Bobrova V, Pahnke J, Troitsky A, Antonov A, Martin W (1996) Non-coding sequences from the slowly evolving chloroplast inverted repeat in addition to *rbc*L data do not support gnetalean affinities of angiosperms. Mol Biol Evol 13:383–396

Goremykin V, Viola R, Hellwig F (2009) Removal of the noisy characters from the chloroplast genome-scale data suggests a revision of phylogenetic placements of *Amborella* and *Ceratophyllum*. J Mol Evol 68:197–204

Graur D, Hide WA, Li WH (1991) Is the guinea-pig a rodent? Nature 351:649–652

Graur D, Hide WA, Zharkikh A, Li W-H (1992) The biochemical phylogeny of guinea-pigs and gundis, and the paraphyly of the order Rodentia. Comp Biochem Phys B 101:495–498

Gribaldo S, Philippe H (2002) Ancient phylogenetic relationships. Theor Popul Biol 61:391–408

Hasegawa M, Cao Y, Adachi J, Yano T (1992) Rodent polyphyly? Nature 355:595–595

Hirt RP, Logsdon JM Jr, Healy B, Dorey MW, Doolittle WF, Embley TM (1999) Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins. Proc Natl Acad Sci USA 96:580–585

Jabbari K, Rayko E, Bernardi G (2003) The major shifts of human duplicated genes. Gene 317:203–208

Janke A, Xu X, Arnason U (1997) The complete mitochondrial genome of the wallaroo (*Macropus robustus*) and the phylogenetic relationship among Monotremata, Marsupialia, and Eutheria. Proc Natl Acad Sci USA 94:1276–1281

Jansen RK, Cai Z, Raubeson LA, Daniell H, Depamphilis CW, Leebens-Mack J, Muller KF, Guisinger-Bellian M, Haberle RC, Hansen AK et al (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci USA 104:19369–19374

Jeffroy O, Brinkmann H, Delsuc F, Philippe H (2006) Phylogenomics: the beginning of incongruence? Trends Genet 22:225–231

Kjer KM, Honeycutt RL (2007) Site specific rates of mitochondrial genomes and the phylogeny of eutheria. BMC Evol Biol 7:8

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 Bioinf 9:34

Le Quesne WJ (1969) A method of selection of characters in numerical taxonomy. Syst Zool 18:201–205

Leebens-Mack J, Raubeson LA, Cui LY, Kuehl JV, Fourcade MH, Chumley TW, Boore JL, Jansen RK, de Pamphilis CW (2005) Identifying the basal angiosperm node in chloroplast genome phylogenies: sampling one's way out of the Felsenstein zone. Mol Biol Evol 22:1948–1963

Li W-H, Hide WA, Zharkikh A, Ma D-P, Graur D (1992) The molecular taxonomy and evolution of the guinea pig. J Hered 83:174–181

Lin Y, Waddell P, Penny D (2002a) Pika and vole mitochondrial genomes increase support for both rodent monophyly and glires. Gene 294:119–129

Lin YH, McLenachan PA, Gore AR, Phillips MJ, Ota R, Hendy MD, Penny D (2002b) Four new mitochondrial genomes and the increased stability of evolutionary trees of mammals from improved taxon sampling. Mol Biol Evol 19:2060–2070

Lopez P, Forterre P, Philippe H (1999) The root of the tree of life in the light of the covarion model. J Mol Evol 49:496–508

Luckett WP, Hartenberger J-L (1993) Monophyly or polyphyly of the order Rodentia: possible conflict between morphological and molecular interpretations. J Mamm Evol 1:127–147

Ma D-P, Zharkikh A, Graur D, VandeBerg JL, Li WH (1993) Structure and evolution of opposum, guinea pig, and porcupine cytochrome b genes. J Mol Evol 36:327–334

Madsen O, Scally M, Douady CJ, Kao DJ, DeBry RW, Adkins R, Amrine HM, Stanhope MJ, de Jong WW, Springer MS (2001) Parallel adaptive radiations in two major clades of placental mammals. Nature 409:610–614

Moore MJ, Bell CD, Soltis PS, Soltis DE (2007) Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proc Natl Acad Sci USA 104:19363–19368

Mouchaty SK, Catzeflis F, Janke A, Arnason U (2001) Molecular evidence of an African phiomorpha-south america caviomorpha clade and support for hystricognathi based on the complete mitochondrial genome of the cane rat (*Thryonomys swinderianus*). Mol Phylogenet Evol 18:127–135

Murphy WJ, Eizirik E, Johnson WE, Zhang YP, Ryder OA, O'Brien SJ (2001a) Molecular phylogenetics and the origin of placental mammals. Nature 409:614–618

Murphy WJ, Eizirik E, O'Brien SJ, Madsen O, Scally M, Douady CJ, Teeling E, Ryder OA, Stanhope MJ, de Jong WW et al (2001b) Resolution of the early placental mammal radiation using Bayesian inference. Science 294:2348–2351

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

Pesole G, Gissi C, de Chirico A, Saccone C (1999) Nucleotide substitution rate of mammalian mitochondrial genomes. J Mol Evol 48:427–434

Phillips MJ, Penny D (2003) The root of the mammalian tree inferred from whole mitochondrial genomes. Mol Phylogenet Evol 28:171–185

Phillips MJ, Lin Y-H, Harrison GL, Penny D (2001) Complete mitochondrial sequences for two marsupials, a bandicoot and a brushtail possum. Proc R Soc Lond Ser B 268:533–1538

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

Pisani D, Mohun MM, Harris S, McIterney JO, Wilkinson M (2006) Molecular evidence for dim-light vision in the last common ancestor of the vertebrates. Curr Biol 16:318–319

Pol D, Siddal ME (2001) Biases in maximum likelihood and parsimony: a simulation approach to a 10-taxon case. Cladistics 17:266–281

Posada D, Crandall KA (1998) ModelTest: testing the model of DNA substitution. Bioinformatics 14:817–818

Prasad AB, Allard MW, Green ED (2008) Confirming the phylogeny of mammals by use of large comparative sequence data sets. Mol Biol Evol 25:1795–1808

Qiu YL, Dombrovska O, Lee J, Li L, Whitlock BA, Bernasconi-Quadroni F, Rest JS, Davis CC, Borsch T, Hilu KW et al (2005) Phylogenetic analyses of basal angiosperms based on nine plastid, mitochondrial, and nuclear genes. Int J Plant Sci 166:815–842

Reyes A, Pesole G, Saccone C (1998) Complete mitochondrial DNA sequence of the fat dormouse, Glis glis: further evidence of rodent paraphyly. Mol Biol Evol 15:499–505

Reyes A, Gissi C, Pesole G, Catzeflis FM, Saccone C (2000a) Where do rodents fit? Evidence from the complete mitochondrial genome of Sciurus vulgaris. Mol Biol Evol 17:979–983

Reyes A, Pesole G, Saccone C (2000b) Long-branch attraction phenomenon and the impact of among-site rate variation on rodent phylogeny. Gene 259:177–187

Reyes A, Gissi C, Catzeflis F, Nevo E, Pesole G, Saccone C (2004) Congruent mammalian trees from mitochondrial and nuclear genes using Bayesian methods. Mol Biol Evol 21:397–403

Rodriguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H (2007) Detecting and overcoming systematic errors in genome-scale phylogenies. Syst Biol 56:389–399

Ruiz-Trillo I, Riutort M, Littlewood DT, Herniou EA, Baguna J (1999) Acoel flatworms: earliest extant bilaterian Metazoans, not members of Platyhelminthes. Science 283:1919–1923

Sperling EA, Peterson KJ, Pisani D (2009) Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of eumetazoa. Mol Biol Evol 26:2261–2274

Springer MS, Debry RW, Douady C, Amrine HM, Madsen O, de Jong WW, Stanhope MJ (2001) Mitochondrial versus nuclear gene sequences in deep-level mammalian phylogeny phylogeny reconstruction. Mol Biol Evol 18:132–143

Springer MS, Stanhope MJ, Madsen O, de Jong WW (2004) Molecules consolidate the placental mammal tree. Trends Ecol Evol 19:430–438

Stefanovic S, Rice DW, Palmer JD (2004) Long branch attraction, taxon sampling, and the earliest angiosperms: Amborella or monocots? BMC Evol Biol 4:35

Swofford DL (2002) PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4.0b10. Sinauer Associates, Sunderland, MA