AMB ALGORITHMS FOR MOLECULAR BIOLOGY

REVIEW ARTICLE Open Access

A priori assessment of data quality in molecular phylogenetics

Bernhard Misof[1†], Karen Meusemann[1,2], Björn M von Reumont[1,3], Patrick Kück[1], Sonja J Prohaska[4,5] and Peter F Stadler[5,6,7,8,9,10*†]

Abstract

Sets of sequence data used in phylogenetic analysis are often plagued by both random noise and systematic biases. Since the commonly used methods of phylogenetic reconstruction are designed to produce trees it is an important task to evaluate these trees *a posteriori*. Preferably, however, one would like to assess the suitability of the input data for phylogenetic analysis *a priori* and, if possible, obtain information on how to prune the data sets to improve the quality of phylogenetic reconstruction without introducing unwarranted biases. In the last few years several different approaches, algorithms, and software tools have been proposed for this purpose. Here we provide an overview of the state of the art and briefly discuss the most pressing open problems.

Keywords: Phylogenomics, Tree-likeness, Phylogenetic networks, Multiple sequence alignments, Quartets, Biases

Introduction

Ideally, the evolutionary process generates data that conform to an additive tree structure. This ideal, however, is rarely —if ever— reached in practice. A diversity of natural processes conspire with imperfect models and methods of data analysis to cause sometimes large deviations. An unavoidable confounding factor is noise, introduced by the stochastic nature of sequence evolution itself, leading to a degradation of the phylogenetic signal when divergence times become very large and when data sets are small. Systematic biases are introduced by deviations from tree-like evolution, such as recombination and lateral gene transfer, as well as by violations of the model assumptions on which the data analysis is based, such as parallel evolution.

Nearly all methods of molecular phylogenetics, furthermore, use sequence alignments to obtain estimates of the divergence between taxa. For the purpose of phylogenetic reconstruction, each column of a multiple sequence alignment (MSA) is a character. In other words, the the letters in a column are treated as if they have arisen from a common ancestral state. All the algorithms for computing MSAs, however, explicitly or implicitly optimize cost functions (such as a sum of pair score) that are unaware of the detailed phylogenetic structure of the data set. This optimization problems, furthermore, are NP hard [1,2] and hence can be solved only with (heuristic) approximation algorithms. MSAs, thus, are necessarily only approximations to a perfect assignment of homologous sequence positions. As most alignment methods internally use a guide tree representing a rough estimate of the particular phylogeny to determine the order in which taxa are treated, MSAs incorporate an implicit phylogenetic assumption that can be biased relative to the unknown true phylogenetic tree.

At the current state of the art, these issues are unavoidable at least in the analysis of large data set, although for small examples it may be feasible to employ methods that concurrently estimate alignments and trees directly from unaligned sequence data [3,4]. Even in these cases biases from non-treelike evolution and insufficient knowledge remain. This is in particular true for the mechanisms of in/del formation.

It is good practice in phylogenetic studies, therefore, to estimate the reliability of the phylogenetic reconstructions *a posteriori*. Most commonly, measures such as the

*Correspondence: studla@bioinf.uni-leipzig.de
†Equal contributors
[5]Interdisciplinary Center for Bioinformatics, University Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany
[6]Bioinformatics Group, Dept. of Computer Science, University Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany
Full list of author information is available at the end of the article

bootstrap or jackknife support or parameters such as the *consistency index* or the *retention index*, see e.g. [5] are used. The latter estimate the prevalence of homoplastic characters relative to the reconstructed tree.

An alternative approach, which is less frequently employed in phylogenetic studies, is to investigate the data set for its information content and possible source of problems already *before* even starting to compute trees. In this chapter we briefly review the most promising approaches for *a priori* quality control, focusing on recent developments.

Measures of tree-likeness

Distance-based measures

Tree reconstruction based on uncorrected distances obtained from discrete characters can lead to incorrect trees. Effects such as long branch attraction [6] have long been known and continue to be discussed in the literature, see e.g. [7]. The corresponding distances often deviate from additivity as indicated by conflicting support for alternative trees, and hence indications for misleading signals can be obtained from measures of tree-likeness.

A fundamental theorem of mathematical phylogenetics asserts that a metric d on a finite set X of taxa forms an additive tree if and only if every quartet (set of four taxa) has this property [8,9]. It appears natural, therefore, to use quartets to measure tree-likeness of a data set. Among four taxa $\{A, B, C, D\}$ there are six distances which can be grouped into three pairs: $d(A,B) + d(C,D)$, $d(A,C) + d(B,D)$, and $d(A,D) + d(B,C)$. Ordering these three sums by magnitude, we obtain three parameters $L \geq M \geq S$, from which in turn we derive two split lengths $\alpha = (L-S)/2$ and $\beta = (L-M)/2$, see Figure 1A. The quadruple is a tree if and only if $L = M$, i.e., $\beta = 0$.

The basic idea of *statistical geometry* [11,12] is to consider quadruples first in terms of distances and then in terms of sequence patterns. Three types of quadruple geometries can be distinguished based on distances alone:

1. $L = M = S$. In this case $\alpha = \beta = 0$, the quadruple is an ideal bundle.
2. $L = M > S$. In this case $\alpha > 0$ and $\beta = 0$ so that the quadruple defines a single split.
3. $L > M > S$. In this generic case the data deviate from tree structure.

It can be shown that the ratio β/α approaches 0.5 for random sequences, i.e., complete loss of phylogenetic signal. Averaging the parameters α and β over all quadruples that can be formed from a data set thus provide already a good measure for its tree-likeness.

The *δ-plots* [13] build upon statistical geometry and represent the tree-likeness β/α of quartets in terms of a histogram. For an individual taxon a measure for tree-likeness can be obtained by considering the tree-likeness of all quartets to which it belongs. As suggested e.g. in [13], removing taxa with poor individual tree-likeness can result in increased accuracy of tree estimation.

It is important to note that the assessment of distance measures is not necessarily sufficient. Perfectly tree-like distances can still support an incorrect tree in realistic examples. Instructive cases are discussed in detail in [14]. A possible source of such biases is in particular the use of an incorrect model for the transformation of character differences to distances. A more detailed picture is obtained when the alignment columns for a sequence quadruple are inspected.

Character-based methods

Recording for each column of a given alignment of four sequences only which of the sequences have the same character states we distinguish 16 column types that fall into just five classes k: all equal (1), one triple (4), two pairs (3), one pair (6), and all different (1). For alphabets with less than 4 letters not all of these can be realized. Given a model of evolution it is then possible to compute the expected values of the numbers d_k of occurrences of columns of type k, $0 \leq k \leq 4$, as a function of the divergence (number of substitution events per site). From these value one can then obtain refined parameters for tree- and bundle-likeness, see [11,12] for more details.

An alternative approach is to interpret the 16 column types as support for one of the three possible unrooted trees. Each quadruple, thus, can be associated with a relative support (p_1, p_2, p_3) for the three geometries. Properly normalized these values can be plotted in simplicial coordinates, Figure 1B. This idea underlies the *Quartet Mapping* method [15] and its special case *Likelihood Mapping* [16], where the values of p_i are computed as maximum likelihood estimates for the probabilities of the three tree topologies. Quartet mapping can also be used to resolve the relationships between four groups A, B, C, and D that are *a priori* known to be monophyletic. A special case, in which $A = \{x\}$ is just a single sequence is the problem of assigning individual genes to paralog groups. This technique, implemented in the software tool `quartm`, has been used successfully e.g. to analyze short PCR fragments of homeobox sequences [10,17,18]. Figure 1 deliberately shows an example of extremely information-poor data.

A generalization to five taxa has been attempted more recently [19]. Finding the best planar visualization of the space representing the 15 possible unrooted trees for a

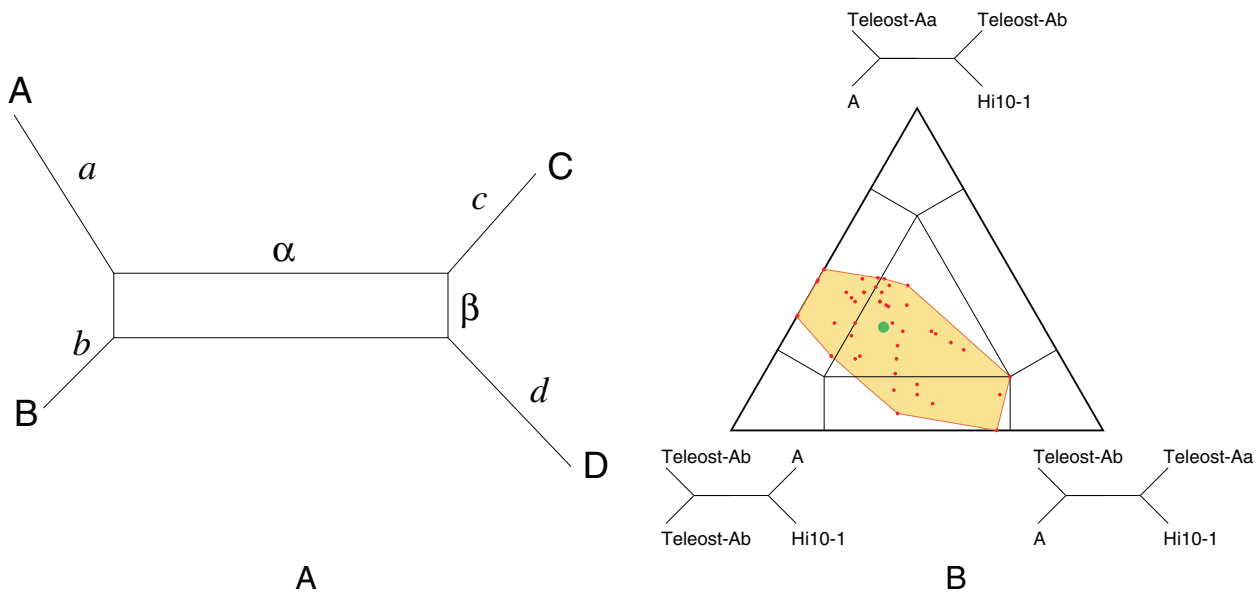

Figure 1 Statistical Geometry and Quartett Mapping. (A) The distances among four taxa $\{A, B, C, D\}$ can always be represented in terms of six parameters: four splits of one taxon compared to the other three, and the two dimensions of central box. For $\beta = 0$ we obtain a tree. **(B)** Comparison of the sequence of a short PCR fragment of the homeobox sequence of a *Hiodon* Hox-A10 gene compared to three sets of homologous Hox-A10 sequences: the HoxA10a and HoxA10b paralogs of crown group teleost fishes and the unduplicated HoxA10 genes of chondrichtyes and sarcopterygians. The colored area is the convex hull of the data points, the big dot indicated the center of mass, from [10].

given set of input data is a rather difficult optimization problem. `PentaPlot` [20] uses a genetic algorithm for this purpose.

On a set X of n taxa there are $2^{n-1} - 1$ distinct splits, i.e., bipartitions of the taxa into exactly two non-empty sets. Weight vectors $\vec{q}$ over the set of splits are related to "pattern probability vectors" $\vec{p}$ assigning probabilities to characters by means of the Hadamard transform $\vec{p} = \mathbf{H}^{-1} \exp[\mathbf{H}\vec{q}]$, where vector exponentiation is interpreted component-wise [21], see [22] for a modern presentation. Intuitively, the Hadamard transformation accounts for multiple state changes of a given character and provides a direct link between observed character states and splits in the underlying data. This connection can be used assess tree-likeness of the data by analyzing the split-spectrum. In addition to a support value q_s for a given split an incompatibility score can be defined as the sum of the supports for all splits s' that cannot occur together with s in the same tree. This information is conveniently summarized in so-called Lento-plots [23]. `Spectronet` provides an implementation [24]. A related method summarizes the Hadamard weight spectrum into three categories: the splits supporting external and internal branches of the optimal tree as well as the splits contradicting this tree. Plotting the relative weights of these three categories in barycentric coordinates produces a "treeness triangle" [25], from which deviation from tree-likeness can be assessed visually.

Alignment quality

Large evolutionary distances inevitably entail a large number of homoplastic sites. As most protein-coding genes show dramatic variations in substitution rates that are not uncorrelated across the sequence, this often leads to a patchwork pattern of phylogenetically informative and effectively randomized regions. Alignment errors accumulate in highly variable regions and may produce effectively "homoplastic sites". Both simulation studies [26] and evaluations of real-life data [27] demonstrated that alignment errors can significantly change the outcome of phylogenetic analyses. There is no consensus in the literature, furthermore, how tolerant phylogenetic methods are to multiple substitutions [28-30].

Consequently, one may try to improve the accuracy of tree reconstruction by eliminating all putative homoplastic or otherwise corrupted sites. A simple approach towards this end is to exclude all third-codon positions of protein-coding sequences. Since the quality of tree reconstruction decreases with decreasing sequence length, it is important not to remove too many sites from an alignment, however. For example, while certain first- and second-codon positions may be essentially constant (and therefore phylogenetically useless) or hyper-variable (and hence even misleading), third-codon positions of protein-coding genes can well be informative and thus they should not be discarded outright [31]. Instead, one would like to distinguish clearly homoplastic or otherwise

corrupted sites from putative phylogenetically informative sites so that they — and no others — can be excluded or down-weighted.

The complication with such an endeavor, however, is that, formally, homoplasy is defined relative to a given phylogenetic tree, the very object that molecular phylogenetics is attempting to derive from the alignment. Measures such as the consistency index (the minimum possible number character changes divided by the number of steps observed along the tree) thus cannot be computed prior to estimating the phylogenetic tree itself. Consequently, the *a priori* is a difficult problem since a useful method has to ensure that its approach to homoplasy detection does not implicitly presuppose a phylogenetic tree later to be derived from the same data.

Historically the first tool for removing suspicious parts of alignments was `Gblocks` [32,33], which selects blocks from an input alignment using a set of rules that mimic many researcher's strategy in manually pruning alignments. User-defined parameters set cut-offs so that the retained regions do not contain large segments of contiguous non-conserved positions, are depleted in gap positions, and exhibit high levels of conservation of flanking positions. While intuitively plausible, these rules are not based in some underlying theory. Nevertheless, this approach can lead to better trees, which, surprisingly, often exhibit reduced bootstrap support, indicating that "divergent and problematic alignment regions may lead, when present, to apparently better supported although, in fact, more biased topologies" [33].

EST-based phylogenomic studies are in particular plagued by incomplete sequences and thus by missing data in MSAs. This can introduce surprisingly large biases and substantially compromised phylogenetic accuracy [34,35]. As a remedy, `reap` [34] masks (i) alignment columns containing many gaps and/or highly diverse amino acids and (ii) sequences that either have little overlap with other sequences or appear to be systematically misaligned. The cutoffs used in `reap` were determined empirically to "strike the best compromise between topological accuracy and sequence retention" [34].

Noisy

The `noisy` [36] method is based on the observation that distances derived from pairwise sequence comparisons give rise to fairly robust circular split systems [37]. Circular splits systems can be represented as a circular ordering of the taxa and are consistent with a large number of possible tree topologies [38,39], namely all those that can be inscribed in the circularly ordered taxa without crossings of tree edges. The utility of circular orderings computed e.g. by the `Neighbor-Net` [40] or `Qnet` [41] algorithms for our purposes is that phylogenetically more closely related taxa are preferentially placed closer together in the cyclic ordering. Conversely, similar trees necessarily correspond to similar cyclic orderings. Thus, if a character, i.e., an alignment column, is phylogenetically "useful", its character states will appear "clustered" along the cyclic ordering underlying any tree that is a reasonable approximation of the true phylogeny, independent of the details of the branching order in individual subtrees. In contrast, if a character is completely randomized, we will observe that character states are randomly arranged along the cycle.

For a given cyclic ordering π, the amount of clustering in alignment column i is conveniently quantified as the number $\nu(\pi, i)$ of "break points", i.e., adjacent distinct character states. For constant alignment columns $\nu(\pi, i) = 0$, for non-constant sites we have $\nu(\pi, i) \geq 2$. This number has to be compared with the numbers expected for a random permutation of the letters observed in alignment column i. This background distribution is easily generated by means of shuffling, i.e., by replacing π with a random permutation π' drawn from a uniform distribution. We then measure the fraction $q(\pi, i)$ of sampled random permutations with $\nu(\pi', i) > \nu(\pi, i)$. The value of $q(\pi, i)$ is thus an estimate for the probability that the column i is *not* randomized. The `noisy` program removes all alignment columns with $q < q_{\text{cutoff}}$. It is reassuring to observe that the number of sites that are deemed randomized is minimized by phylogenetically plausible circular orderings π, Figure 2(A).

Two effects have to be considered. On the one hand, columns with small values of q contribute little useful information. On the other hand, a large absolute number of informative sites is necessary to obtain reliable trees. Thus q_{cutoff} must not be too large. The most effective values of q_{cutoff} also depend on the tree topology. As shown in Figure 2(B) caterpillar trees admit larger improvements in bootstrap support than the balanced trees.

The analysis of artificial data sets suggests a set of simple rules that allow the user to decide under which conditions it makes sense to use `noisy` to process MSAs prior to using them for phylogenetic reconstruction:

(1) If the original alignment already yields trees with very high average bootstrap support, there is nothing to be gained.
(2) Data-sets with less than about 10 taxa are unlikely to improve.
(3) The best cutoff value for q depends on the tree topology and in particular on the number of taxa. It pays in general to determine the maximum of the gain in some parameter of tree stability as a function of q and to use the corresponding optimal cutoff value.

The current release [42] of `noisy` can process DNA, RNA, and protein sequences.

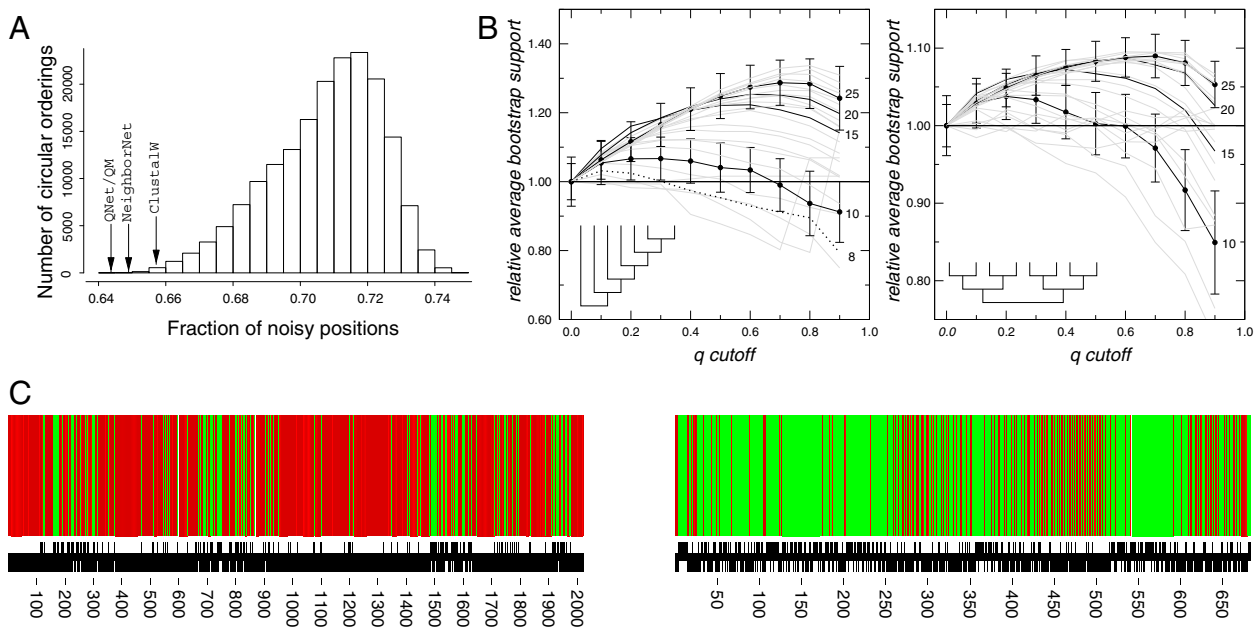

Figure 2 `noisy`. (A) The fraction of sites marked as "randomized" depend on the cyclic ordering. Phylogenetically reasonable ordering such as those computed by `NeighborNet`, `QNet`, or from the guide tree of the alignment program `ClustalW` have a nearly minimal fraction of putative randomized alignment sites. **(B)** The average bootstrap support increases for moderate values of q i.e., as long as not too large a fraction of alignment columns are removed. The effect increases with the size of the data set. **(C)** Distributions of randomized positions can differ substantially between data sets, here 18S RNA of Coleoptera (l.h.s.) and mitochondrial `atp6` gene of squamata (r.h.s.). Red indicates randomized positions, light red singletons, green parsimony informative sites. The bars below indicate included and excluded parts of the alignment, respectively. (Adapted from [36]).

Aliscore

In contrast to `noisy`, `aliscore` has been designed to detect random sequence similarity in MSAs based on pairwise similarity profiling of sequences [43,44]. It is based on the fact that observed sequence motive similarity between a pair of sequences can be distinguished from random similarity by generating a null distribution of random similarity given the motive size and base/aminoacid composition of the sequences. The null distribution is generated by permutations of the original observed sequences generating random similarity. A sliding window is used to generate a profile score of the inferred randomization between pairs of sequences. This can be done with all possible pairwise comparisons within a MSA generating a suite of pairwise profile scores. Finally, these profile scores are used to average over each MSA alignment site in order to generate a consensus profile of sequence similarity within a MSA. This consensus profile informs whether alignment sections contain predominantly random similarity or not, Figure 3. The principle of `aliscore` is thus entirely different to site-focused approaches like `noisy`, `reap` [34] or `gblocks` [32,33]. For a detailed explanation of the algorithm we refer to [43].

The `aliscore` approach has been shown to work well in simulations [43,44], single gene [45-49] and multi-gene approaches [50]. However, as for every masking program arbitrary decisions have to be made as well. For example, the sliding window size has to be set by the user. A larger window size makes the algorithm less sensitive to small sections of randomization. A natural minimal window size is 4, below this window size a distinction between random or non-random similarity is not possible.

A big advantage of the approach is that single splits can be directly evaluated. `aliscore` offers the possibility to define a split in the MSA from which pairwise comparisons are drawn. It thus offers the possibility to generate a consensus profile for just the split under consideration. This tool can become particularly important, if different outgroup taxa are compared with a set of ingroup species. The best outgroup choice is the set of taxa which minimizes the extent of randomization between outgroup and ingroup. The current release of `aliscore` can process DNA, RNA, and protein sequences.

Quality of a data matrix: MARE

A typical feature of phylogenomic data is the frequent occurrence of missing data in concatenated "supermatrices" up to the point where more than 80% of the data are missing [51,52]. The effect of missing data on tree inference is still unclear and it appears that a general rule

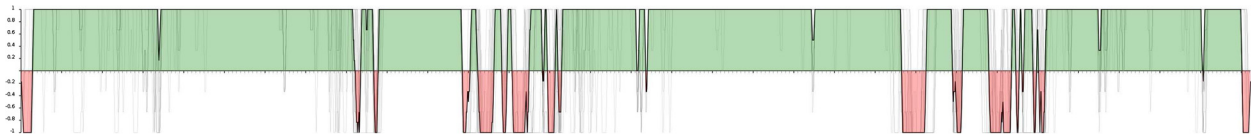

Figure 3 Graphical output of `aliscore` for an alignment of arthropod 18S gene sequences. The consensus profile is colored in green and red. Sections of the consensus profile larger than zero are colored in green, below zero in red. In this particular alignment, several small sections are dominated by strong randomness, indicated in red.

can not be derived from several simulation and empirical studies [51,53,54]. The take-home message of these studies is that data masking, which can increase data saturation, seems advisable. In its simplest form, we are given a bipartite graph $G = (X \cup Y, E)$ describing by its edges $\{x, y\} \in E$ which gene $x \in X$ is present in which species $y \in Y$. An ideal data set is a maximal biclique, i.e., a maximal complete subgraph of G [55,56]. Since this would lead to the removal of too many genes and taxa, in a relaxed version, one seeks a *quasi-biclique* [57], requiring that each gene is present at least in a prescribed fraction of taxa, and each taxon is represented by a minimum fraction of genes. It is worth noting that the same problem appears in the analysis of protein-protein interaction networks and has received considerable attention in this context [58]. Although the maximum vertex biclique problem is solvable in polynomial time [59], many of its variants [60] and in particular the more relevant quasi-biclique problems are NP-complete [58,61]. Thus exact algorithms are applicable only to to relatively small data. In addition, earlier methods do not consider differences in the information content of taxa and genes, which might be a major drawback.

In simulation studies we were able to show that the likelihood of reconstructing a correct tree dramatically decreases if data saturation is below 30%. Selection of a data subset of less genes and taxa but with higher data saturation can potentially alleviate the problem. However, it seems advisable that during the process of data selection, potential phylogenetic signal of each single gene and taxon should be considered in order not to only maximize data saturation but also information content of the data set. The proposed algorithm implemented in the software package `mare` does exactly this, Figure 4. It is designed in a way that (1) the potential information content of genes and taxa is evaluated using geometry mapping [15] and (2) this information is used in combination with information on missing data to select an optimal data subset. The selection of the optimal data subset is based on a simple optimization algorithm in which the reduction of the total data matrix is penalized and the increase in total information content of the matrix favored. The selected optimal data subset corresponds to a quasi-biclique with high information content. Simulations show that the chance to reconstruct the correct tree increases tremendously when the raw data are processed in this manner [62].

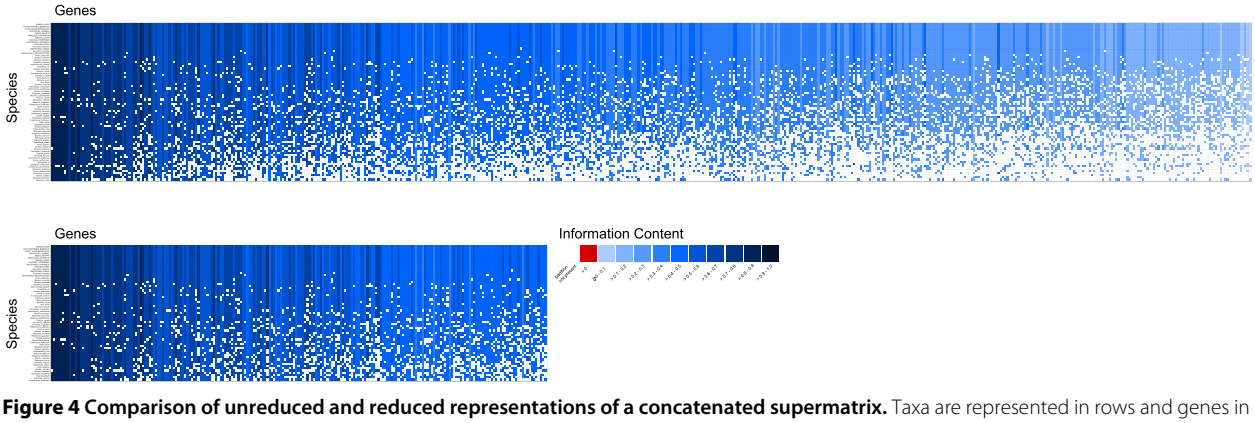

Figure 4 Comparison of unreduced and reduced representations of a concatenated supermatrix. Taxa are represented in rows and genes in columns. If a gene has not been identified or sequenced in a taxon, this entry is left white in the matrix, blue entries indicate the presence of gene sequences for that taxon. Shades of blue correspond to information content of the specific gene. Dark blue represents high information content and light blue low information content. The representation of the original supermatrix is placed in the upper panel. Columns are sorted according to their information content. The reduced supermatrix in the lower panel was generated with the software `mare` and represents an optimal selection of taxa and genes from the original supermatrix according to the criteria developed in this `mare` approach.

The current implementation of `mare` handles protein sequences only.

Concluding remarks

Although many studies have been directed at a better understanding of artifacts in phylogeny reconstruction such as long branch attraction or homoplasy [7,63,64], we still lack a comprehensive understanding of how biases can be recognized in data sets prior to the estimation of a phylogenetic tree. Instead, often time extensive computational resources are expended to reconstruct phylogenies with disappointing results that can be identified only *a posteriori* as artifacts. It is then problematic at best to distinguish artefactual input data from issues such as inadequate models of evolution.

In this minireview we have briefly discussed first attempts at an *a priori* assessment of different aspects of data quality that aim at the identification of potentially problematic taxa or characters. It is of utmost importance to ensure that such methods do not make any assumptions on phylogenetic relationships, because such implicit information may then inadvertently be enforced in the "data cleaning" step, and thus transmitted to the phylogenetic reconstruction methods.

Despite very encouraging results obtained with tools such as `noisy`, `aliscore`, and `mare`, much additional research focused on dissecting confounding signal will be necessary for a comprehensive understanding of analyses artifacts. `Noisy` and `aliscore` address the decay of phylogenetic signal induced by multiple saturation, the `aliscore` algorithm can deal with heterogeneous composition of nucleotide sequences, and `mare` indirectly scores the influence of missing data as well. However, all of these approaches do not directly dissect the separate influence of these confounding factors on tree reconstructions. This must be a focus of future work, because substitutional saturation, heterogeneous sequence composition, non-stationary substitution processes, and the non-random distribution of missing data can constitute strong confounding factors, in particular in phylogenomic analyses. It surprising, therefore, that a standard canon of tools to study these effects *a priori* to tree reconstructions is still missing.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

BM and PFS wrote the first draft of this review. All authors read and approved the final manuscript.

Author details

[1]Zoologisches Forschungsmuseum Alexander Koenig, Adenauerallee 160, D-53113 Bonn, Germany. [2]CSIRO Ecosystem Sciences, Australian National Insect Collection, Clunies Ross Street, AU-2601 Acton, Canberra, Australia. [3]The Natural History Museum London, Dept. of Life Sciences, Cromwell Road, GB-SW7 5BD Acton, London, UK. [4]Computational EvoDevo Group, Dept. of Computer Science, University Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany. [5]Interdisciplinary Center for Bioinformatics, University Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany. [6]Bioinformatics Group, Dept. of Computer Science, University Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany. [7]Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany. [8]Fraunhofer Institut for Cell Therapy and Immunology, Perlickstraße 1, D-04103 Leipzig, Germany. [9]Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Vienna, Austria. [10]Santa Fe Institute, 1399 Hyde Park Rd., NM87501 Santa Fe, USA.

Received: 21 May 2014 Accepted: 24 July 2014
Published online: 12 September 2014

References

1. Just W: **Computational complexity of multiple sequence alignment with SP-score.** *J Comput Biol* 2001, **8:**615–623.
2. Wang L, Jiang T: **On the complexity of multiple sequence alignment.** *J Comput Biol* 1994, **1:**337–348.
3. Lunter G, Miklós I, Drummond A, Jensen JL, Hein J: **Bayesian coestimation of phylogeny and sequence alignment.** *BMC Bioinformatics* 2005, **6:**83.
4. Redelings BD, Suchard MA: **Joint bayesian estimation of alignment and phylogeny.** *Syst Biol* 2005, **54:**401–418.
5. Farris JS: **The retention index and the rescaled consistency index.** *Cladistics* 1989, **5:**417–419.
6. Felsenstein J: **Cases in which parsimony or compatibility methods will be positively misleading.** *Syst Zool* 1978, **27:**401–410.
7. Telford MJ, Copley RR: **Animal phylogeny: fatal attraction.** *Curr Biol* 2005, **15:**296–299.
8. Simões-Pereira JMS: **A note on the tree realizability of a distance matrix.** *J Combin Theory* 1969, **6:**303–310.
9. Buneman P: **A note on the metric property of trees.** *J Combin Theory Ser B* 1974, **17:**48–50.
10. Chambers KE, McDaniell R, Raincrow JD, Deshmukh M, Stadler PF, Chiu C-h: **Hox cluster duplication in the basal teleost Hiodon alosoides (Osteoglossomorpha).** *Theory Biosci* 2009, **128:**109–120.
11. Eigen M, Winkler-Oswatitsch R, Dress AWM: **Statistical geometry in sequence space: a method of quantitative comparative sequence analysis.** *Proc Natl Acad Sci USA* 1988, **85:**5913–5917.
12. Nieselt-Struwe K: **Graphs in sequence spaces: a review of statistical geometry.** *Biophys Chem* 1997, **30:**111–131.
13. Holland BR, Huber KT, Dress AWM, Moulton V: **δ plots: A tool for analyzing phylogenetic distance data.** *Mol Biol Evol* 2002, **19:**2051–2059.
14. Huson D, Steel M: **Distances that perfectly mislead.** *Syst Biol* 2004, **53:**327–332.
15. Nieselt-Struwe K, von Haeseler A: **Quartet-mapping, a generalization of the Likelihood-Mapping procedure.** *Mol Biol Evol* 2001, **18:**1204–1219.
16. Strimmer K, von Haeseler A: **Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment.** *Proc Natl Acad Sci USA* 1997, **94:**6815–6819.
17. Stadler PF, Fried C, Prohaska SJ, Bailey WJ, Misof BY, Ruddle FH, Wagner GP: **Evidence for independent Hox gene duplications in the hagfish lineage: A PCR-based gene inventory of *Eptatretus stoutii*.** *Mol Phylog Evol* 2004, **32:**686–692.
18. Raincrow JD, Dewar K, Stocsits C, Prohaska SJ, Amemiya CT, Stadler PF, Chiu C-h: **Hox clusters of the bichir (Actinopterygii, *Polypterus senegalus*), highlight unique patterns of sequence evolution in gnathostome phylogeny.** *J Exp Zool* 2011, **316:**451–464.
19. Zhaxybayeva O, Hamel L, Raymond J, Gogarten JP: **Visualization of the phylogenetic content of five genomes using dekapentagonal maps.** *Genome Biol* 2004, **5:**20.
20. Hamel L, Zhaxybayeva O, Gogarten JP: **`PentaPlot`: A software tool for the illustration of genome mosaicism.** *BMC Bioinformatics* 2005, **6:**139.
21. Hendy M, Penny D: **A framework for the quantitative study of evolutionary trees.** *Syst Zool* 1989, **38:**297–309.
22. Bryant D: **Hadamard phylogenetic methods and the *n*-taxon process.** *Bull Math Biol* 2009, **71:**339–351.

23. Lento GM, Hickson RE, Chambers GK, Penny D: **Use of spectral analysis to test hypotheses on the origin of pinnipeds.** *J Mol Biol Evol* 1995, **12:**28–52.
24. Huber KT, Langton M, Penny V, Moulton D, Hendy M: **Spectronet: a package for computing spectra and median networks.** *Appl Bioinform* 2002, **1:**2041–2059.
25. White T, Hills SF, Gaddam R, Holland BR, Penny D: **Treeness triangles: Visualizing the loss of phylogenetic signal.** *Mol Biol Evol* 2007, **24:**2029–2039.
26. Ogden TH, Rosenberg M: **Multiple sequence alignment accuracy and phylogenetic inference.** *Syst Biol* 2006, **55:**314–328.
27. Landan G, Graur D: **Heads or tails: a simple reliability check for multiple sequence alignments.** *Mol Biol Evol* 2007, **24:**1380–1383.
28. Yang Z: **On the best evolutionary rate for phylogenetic analysis.** *Syst Biol* 1998, **47:**125–133.
29. Wägele J-W: *Foundations of Phylogenetic Systematics*. Munich, Germany: Verlag Dr Friedrich Pfeil; 2005.
30. Kück P, Mayer C, Wägele J-W, Misof B: **Long branch effects distort maximum likelihood phylogenies in simulations despite selection of the correct model.** *PLoS ONE* 2012, **7:**36593.
31. Björklund M: **Are third positions really that bad? a test using vertebrate cytochrome b.** *Cladistics* 1999, **15:**91–97.
32. Castresana J: **Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis.** *Mol Biol Evol* 2000, **17:**540–552.
33. Talavera G, Castresana J: **Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments.** *Syst Biol* 2007, **56:**564–577.
34. Hartmann S, Vision TJ: **Using ESTs for phylogenomics: Can one accurately infer a phylogenetic tree from a gappy alignment.** *BMC Evol Biol* 2008, **8:**95.
35. Roure B, Baurain D, Philippe H: **Impact of missing data on phylogenies inferred from empirical phylogenomic data sets.** *Mol Biol Evol* 2013, **30:**197–214.
36. Dress AWM, Flamm C, Fritzsch G, Grünewald S, Kruspe M, Prohaska SJ, Stadler PF: **Identification of homoplastic characters in multiple sequence alignments.** *Alg Mol Biol* 2008, **3:**7.
37. Bandelt HJ, Dress AWM: **A canonical decomposition theory for metrics on a finite set.** *Adv Math* 1992, **92:**47–105.
38. Huson DH: **`SplitsTree`: analyzing and visualizing evolutionary data.** *Bioinformatics* 1998, **14:**68–73.
39. Semple C, Steel M: **Cyclic permutations and evolutionary trees.** *Adv Appl Math* 2004, **32:**669–680.
40. Bryant D, Moulton V: **Neighbor-net: An agglomerative method for the construction of phylogenetic networks.** *Mol Biol Evol* 2004, **21:**255–265.
41. Grünewald S, Forslund K, Dress AWM, Moulton V: **QNet: an agglomerative method for the construction of phylogenetic networks from weighted quartets.** *Mol Biol Evol* 2007, **24:**532–538.
42. Dress AWM, Flamm C, Fritzsch G, Grünewald S, Kruspe M, Prohaska SJ, Stadler PF: **`noisy` Software.** 2011. [http://www.bioinf.uni-leipzig.de/Software/noisy/]
43. Misof B, Misof K: **A Monte Carlo approach successfully identifies randomness of multiple sequence alignments: A more objective approach of data exclusion.** *Syst Biol* 2009, **58:**21–34.
44. Kück P, Meusemann K, Raupach M, von Reumont B, Wägele W, Misof B: **Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees.** *Frontiers Zool* 2010, **7:**10.
45. von Reumont BM, Meusemann K, Szucsich NU, Dell'Ampio E, Bartel D, Simon S, Letsch HO, Stocsits RR, Luan Y, Wägele JW, Pass G, Hadrys H, Misof B: **Can comprehensive background knowledge be incorporated into substitution models to improve phylogenetic analyses? a case study on major arthropod relationships.** *BMC Evol Biol* 2009, **9:**119.
46. Wägele J-W, Letsch H, Klussmann-Kolb A, Mayer C, Misof B, Wägele H: **Phylogenetic support values are not necessarily informative: the case of the Serialia hypothesis (a mollusk phylogeny).** *Frontiers Zool* 2009, **6:**12.
47. Schwarzer J, Misof B, Tautz D, Schliewen UK: **The root of the East African cichlid radiations.** *BMC Evol Biol* 2009, **9:**186.
48. Letsch HO, Kück P, Schmidt C, Fleck G, Stocsits RR, Misof B: **The impact of rRNA secondary structure consideration in alignment and tree reconstruction: simulated data and a case study on the phylogeny of hexapods.** *Mol Biol Evol* 2010, **27:**2507–2521.
49. Murienne J, Edgecombe GD, Giribet G: **Including secondary structure, fossils and molecular dating in the centipede tree of life.** *Mol Phylog Evol* 2010, **57:**301–313.
50. Meusemann K, von Reumont BM, Simon S, Roeding F, Kueck P, Ebersberger I, Strauss S, Walzl M, Pass G, Breuers S, Achter V, Wägele J-W, Hadrys H, Burmester T, von Haeseler A, Misof B: **A phylogenomic approach to resolve the arthropod tree of life.** *Mol Biol Evol* 2010, **27:**2451–2464.
51. Sanderson MJ, Driskell AC: **The challenge of constructing large phylogenetic trees.** *Trends Plant Sci* 2003, **8:**374–379.
52. Driskell AC, Ané C, Burleigh JG, McMahon MM, O'Meara BC, Sanderson MJ: **Prospects for building the tree of life from large sequence databases.** *Science* 2004, **306:**1172–1174.
53. Wiens JJ: **Missing data, incomplete taxa, and phylogenetic accuracy.** *Syst Biol* 2003, **52:**528–538.
54. Wiens JJ: **Missing data and the design of phylogenetic analyses.** *J Biomed Inform* 2006, **39:**34–42.
55. Alexe G, Alexe S, Crama Y, Foldes S, Hammer PL, Simeone B: **Consensus algorithms for the generation of all maximal bicliques.** DIMACS Technical Reports 2002-52, Rutgers University, Piscataway, NJ, USA, 2002. [http://dimacs.rutgers.edu/TechnicalReports/2002.html]
56. Sanderson MJ, Driskell AC, Ree RH, Eulenstein O, Langley S: **Obtaining maximal concatenated phylogenetic data sets from large sequence databases.** *Mol Biol Evol* 2003, **20:**1036–1042.
57. Yan C, Burleigh JG, Eulenstein O: **Identifying optimal incomplete phylogenetic data sets from sequence databases.** *Mol Phylogenet Evol* 2005, **30:**528–535.
58. Liu X, Li J, Wang L: **Modeling protein interacting groups by quasi-bicliques: complexity, algorithm, and application.** *IEEE/ACM Trans Comput Biol Bioinform* 2010, **7:**354–364.
59. Yannakakis M: **Node deletion problems on bipartite graphs.** *SIAM J Comput* 1981, **10:**310–327.
60. Peeters R: **The maximum edge biclique problem is NP-complete.** *Discrete Appl Math* 2003, **131:**651–654.
61. Chang W-C, Vakati1 S, Krause R, Eulenstein O: **Exploring biological interaction networks with tailored weighted quasi-bicliques.** *BMC Bioinformatics 2012* 2012, **13**(S10):16.
62. Misof B, Meyer B, von Reumont BM, Kück P, Misof K, Meusemann K: **Selecting informative subsets of sparse supermatrices increases the chance to find correct trees.** *BMC Bioinformatics* 2013, **14:**348.
63. Gribaldo S, Philippe H: **Ancient phylogenetic relationships.** *Theor Popul Biol* 2002, **61:**391–408.
64. Wake DB, Wake MH, Specht CD: **Homoplasy: from detecting pattern to determining process and mechanism of evolution.** *Science* 2011, **331:**1032–1035.

doi:10.1186/s13015-014-0022-4
Cite this article as: Misof *et al.*: ***A priori*** assessment of data quality in molecular phylogenetics. *Algorithms for Molecular Biology* 2014 **9**:22.