Article: Methods

A Method of Alignment Masking for Refining the Phylogenetic Signal of Multiple Sequence Alignments

Vaibhav Rajan*
Research Scientist, Xerox Research Centre India,
Bangalore, India

*This research was conducted when the author was at the
Laboratory for Computational Biology and Bioinformatics,
Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland.

vaibhav.rajan@xerox.com

A Method of Alignment Masking for Refining the Phylogenetic Signal of Multiple Sequence Alignments

Vaibhav Rajan[*]
Research Scientist, Xerox Research Centre India, Bangalore, India

Inaccurate inference of positional homologies in multiple sequence alignments and systematic errors introduced by alignment heuristics obfuscate phylogenetic inference. Alignment masking, the elimination of phylogenetically uninformative or misleading sites from an alignment before phylogenetic analysis is a common practice in phylogenetic analysis. While masking is often done manually, automated methods are necessary in order to handle the much larger datasets being prepared today. In this study we introduce the concept of subsplits and demonstrate their use in extracting phylogenetic signal from alignments. We design a clustering approach for alignment masking where each cluster contains similar columns—similarity being defined on the basis of compatible subsplits; our approach then identifies noisy clusters and eliminates them. Trees inferred from the columns in the retained clusters are found to be topologically closer to the reference trees. We test our method on numerous standard benchmarks (both synthetic and biological datasets) and compare its performance with other methods of alignment masking. We find that our method can eliminate sites more accurately than other methods, particularly on divergent data, and can improve the topologies of the inferred trees in likelihood based analyses. Software availability: upon request from the author.

Introduction

A multiple sequence alignment is usually the first step in a phylogenetic analysis. An alignment is a statement of homology—each position (also called site, character or column) in an alignment represents homologous character states in the sequences. These homologous *columns*, as we will call them, are then used to infer a phylogeny through various inference procedures. While methods of simultaneous inference of multiple sequence alignments and phylogenies are being actively developed (Fleissner et al. 2005; Lunter et al. 2005; Redelings and Suchard 2005; Liu et al. 2009), current practice is to infer the two in successive steps—the so-called two-phase approach.

The quality of a multiple sequence alignment has great impact on the final inferred tree (Kjer 1995; Morrison and Ellis 1997; Xia et al. 2003; Ogden and Rosenberg 2006; Smythe et al. 2006; Wang et al. 2011). Finding homologous characters, and hence finding accurate multiple sequence alignments, is hard because of the heterogeneity of evolutionary signal in the sequences due to differing relative branch lengths (Rokas et al. 2003), or processes such as hybridization, lineage sorting, horizontal transfer, or recombination (Takahashi et al. 2001; Lerat et al. 2003; Doyle et al. 2004), or heterotachy and nonstationarity of substitution processes (Felsenstein 2004; Susko et al. 2005; Inagaki and Roger 2006). When modelled as an optimization problem, multiple sequence alignment is NP-hard (Wang and Jiang 1994) and cannot be solved optimally for more than a few sequences. As a result, a large number of heuristic methods have been developed, such as Clustal (Thompson et al. 1994), T-Coffee (Notredame et al. 2000), MAFFT (Katoh et al. 2002), Opal (Wheeler and Kececioglu 2007), and many more, each with its own tradeoffs. These heuristics may introduce their own systematic errors into the homologies inferred. The assessment of the quality of alignments produced by these heuristics is itself an active area of research (Morrison and Ellis 1997; Lassmann and Sonnhammer 2005; Prakash and Tompa 2005; Ahola et al. 2006; Kjer et al. 2007); most assessment approaches have focused on optimization scores or accuracy of inferred homologies rather than accuracy of inferred trees. Errors in inferring homologies are detrimental to the quality of reconstructed trees (Felsenstein 2004; Susko et al. 2005). Identifying conflicting signals in the data and removing the corresponding columns thus form an important preprocessing step for any phylogenetic analysis. While this identification should take place at each step of the analysis, we focus here on identifying conflicting signals after the alignment has been performed.

Different regions of the genome, often evolving at different rates, may lead to incompatible hypotheses of phylogenies. The most divergent parts are often the most misleading since multiple substitutions obfuscate their phylogenetic signal. Columns considered to be phylogenetically uninformative or noisy could be removed before inferring the tree. While many authors consider such removal, called *alignment masking*, to be beneficial (Rodrigo et al. 1994; Swofford et al. 1996; Grundy and Naylor 1999; Castresana 2000; Lytynoja and Milinkovitch 2001), others think that there is loss of information upon removing any part of the sequence (Lee 2001; Aagesen 2004). Whether or not masking is beneficial may depend on the data being analysed. Previous studies have shown that in many cases identifying and removing noisy columns can improve phylogenetic inference, resolve deep divergence and correct systematic biases (Castresana 2000; Talavera and Castresana 2007; Dress et al. 2008; Misof and Misof 2009; Cummins and McInerney 2011). It has also been shown to improve other inference methods that rely on accurate alignments such as positive selection inference using codon models (Privman et al. 2012).

In this paper we present a new approach to refining the phylogenetic signal of alignments based on a hitherto unused notion of subsplits, that can extract more information from an alignment than before. The first step of the method creates clusters of columns; in each cluster columns contain similar phylogenetic signal, similarity being defined on the basis of compatible subsplits. These clusters can

Key words: alignment masking, site removal, subsplit, split, compatibility, SR, phylogeny inference.

[*] This research was conducted when the author was at the Laboratory for Computational Biology and Bioinformatics, Ecole Polytechnique Fédérale de Lausanne, Switzerland.

E-mail: vaibhav.rajan@xerox.com

Mol. Biol. Evol. 24(1):1–13. 2008
doi:10.1093/molbev/msl161
Advance Access publication October 25, 2008

be directly used for further investigation into conflicting phylogenetic signals in the data. This step does not assume any model of evolution nor an existing phylogeny. We then use the rate of evolution as a criterion for eliminating noisy clusters instead of eliminating columns individually. The combined procedure, called SR (for Signal Refinement), is compared with three previous methods of alignment masking on synthetic and real datasets. Our experiments show that SR can retain columns with higher likelihood and lower homoplasy more accurately than the other methods, particularly on divergent data. Moreover refinement using SR significantly improves the phylogenies inferred from the alignments. We find that trees from the refined alignments are topologically closer to the reference trees in a variety of simulated and real datasets.

Alignment masking is often performed manually by researchers studying the sequences. With larger and larger sequences being studied today, automated methods have become necessary. Previous methods of alignment masking have adopted many different strategies. In studies such as (Ruiz-Trillo et al. 1999; Rodriguez-Ezpeleta et al. 2007) columns with the highest rate of evolution, with the rate inferred by a likelihood model that assumes variable rates over columns, are eliminated before tree reconstruction. The rate of evolution can also be approximated without assuming a model of evolution by using splits as described in (Cummins and McInerney 2011) and can be used to identify and remove noisy columns. The software NOISY (Dress et al. 2008) automates the task of removing strongly randomized columns, assumed to be homoplastic, based on assessing the distribution of character states along a cyclic ordering of the taxa. These methods are effective if the rate threshold in the first two methods and the reliability score threshold in NOISY can be determined accurately. Castresana designed a method called GBLOCKS (Castresana 2000; Talavera and Castresana 2007) to identify conserved blocks in an alignment and exclude sections that are variable beyond a threshold. Many parameters have to be given by the user to set this threshold and their method has been tested mainly for protein sequences. Fernandes, Nelson and Beverley (Fernandes et al. 1993) used a method based on pairwise sequence comparisons in sliding windows to detect conserved regions. Their method did not deal with gaps and was later extended to do so by Misof and Misof (Misof and Misof 2009). The latter method, called ALISCORE, identifies random similarity (as opposed to phylogenetically biased similarity) in alignments based on Monte Carlo resampling within a sliding window.

Other methods compare several multiple sequence alignments to eliminate unstable regions in the alignment (Gatesy et al. 1993). This has been extended to assess the reliability of positional homology by studying the consistency of different alignments of similar sequences (Notredame et al. 2000; Lytynoja and Milinkovitch 2001; Lassmann and Sonnhammer 2005). Consistency criteria have also been used in bootstrap-like approaches in (Landan and Graur 2007; Kim and Ma 2011). These measures are conservative and can easily be misled by varying the parameters in the aligners used—see (Lutzoni et al. 2000; Kjer et al. 2007) for discussion and criticisms. Other methods for distinguishing conserved regions from non-conserved ones such as (Pesole et al. 1992) have not been developed for phylogenetic analysis.

New Approaches

We introduce the concept of *subsplits* and observe that there is valuable phylogenetic signal not just in the splits but in all the subsplits as well. Often in alignments, particularly those of divergent species, there are not enough compatible splits. In such data, compatible subsplits can give us clues about the phylogenetic signal. However, using this information in practical methods is challenging as the complete machinery of maximum compatibility (or clique) analysis would be hopelessly time-consuming.

So, instead of using clique analysis directly, we design a new approach that can use subsplits to extract phylogenetic signal. We view the problem as a clustering problem. Our goal is to cluster the columns such that columns with roughly equivalent phylogenetic signal are in the same cluster. In this way we hope to classify the columns such that all 'noisy' columns form identifiable clusters that can be eliminated. There are three key elements in our method:

1. A measure of distance between columns
2. A clustering method
3. A criterion to eliminate one or more noisy clusters

We call this method **SR**: Signal Refinement. The new distance measure is based on an estimate of the number of compatible subsplits between columns. We present the performance of the method with two different clustering algorithm. We denote by **SRk** (**S**ignal **R**efinement with **k**-means clustering) the method that uses *k*-means clustering and by **SRap** (**S**ignal **R**efinement with **A**ffinity **P**ropagation clustering) the method that uses affinity propagation.

Results

To test the hypothesis that subsplits contain valuable phylogenetic signal, we first conduct clique analyses on small alignments and compute the support values of maximal cliques of both splits and subsplits. This analysis is done only to show that subsplits can be useful phylogenetic indicators and is not related to the performance or efficiency of the Signal Refinement (SR) methods.

In the next set of experiments we compare the properties of the refined alignments, in terms of their Normalized Consistency Indices and Normalized Log-Likelihoods, when various methods of alignment masking are used: NOISY, GBLOCKS, ALISCORE and SR.

In the final set of experiments we compare the Maximum-Likelihood phylogenies inferred from alignments before and after refinement for a number of simulated and real datasets.

Clique analysis of subsplits

Table 1 shows the result of five representative samples from the EAG datasets. These are MAFFT alignments of

datasets containing 10 taxa, 1000 columns and five different values of average branch lengths. For each dataset we report the number of cliques of splits found and the number of cliques of subsplits found. It is not surprising to find far more cliques of subsplits since the number of subsplits is exponentially larger than the number of splits.

We measure the signal for a tree in an alignment by counting the support that each edge (which can be viewed as a bipartition of taxa or a split) receives in the alignment. The support of a split or a subsplit in an alignment is simply the number of columns in which the split or the subsplit appears. The inferred ML tree has greater support from subsplits than from the direct splits, again a direct consequence of the large number of subsplits. The same observation with respect to the true tree appears more dramatic in these datasets since there is no support from the splits at all.

Results from other datasets are similar and lead us to postulate that subsplits may enable us to capture phylogenetic signals in divergent species that may be lost if only splits are used. Using subsplits as a basis for measuring compatibility of columns, we design our clustering-based approach for signal refinement.

Comparison of Alignment Masking Algorithms

We show aggregate results for five representative datasets from the EAG datasets: Clustal alignments, each containing 20 taxa. The five datasets differ in their average branch lengths. All other results in the EAG dataset, with Clustal and other aligners, are similar.

Figure 1 shows the Normalized Consistency Indices for these five datasets for each of the seven methods and Random1 and Random2. There are only five discrete points for each method, the connecting lines are drawn only for clarity. The top subfigure shows the curves for our Signal Refinement methods SRk, SRap, with the corresponding control methods Random1, Random2 (that eliminate a random number of columns; the number of columns eliminated by Random1 and Random2 is the same as the number eliminated by SRap and SRk respectively) and the original unrefined alignment (MSA). The curves for Random1 and Random2 are almost identical to that of the original alignment. This shows that a random deletion of columns does not help in eliminating homoplastic columns. The elimination has to be guided by a principled search and both SRk and SRap can do that, with SRk performing better.

In the bottom we repeat the curves for SRk and SRap along with those from the other five methods. The best CI values are achieved by SRk and SRap. Thus, SRap and SRk extract columns that have a better fit to their inferred trees. The columns returned by all three variants of GBLOCKS have smaller CI than the original alignment. ALISCORE does not remove any columns in these datasets and so its curve is the same as that of the original alignment. Even in other datasets that we tested, ALISCORE typically removes very few columns and its performance is similar to or worse than that of GBLOCKS. So, in the following experiments we do not report the performance of ALISCORE. NOISY does reasonably well in our experiments, especially on more divergent datasets.

Figure 2 shows the Normalized Log-Likelihood Scores. The top subfigure shows the curves for SRk, SRap, Random1, Random2 and the original unrefined alignment (MSA). Once again, curves for Random1 and Random2 are almost identical to that of the original alignment. The log-likelihood of refined alignments from SRk is significantly higher than that of its randomized counterpart Random2. The performance of SRap is similar except on datasets with less divergence.

In the bottom we repeat the curves for SRk and SRap along with those from the other five methods. The log-likelihood scores of SRk, followed by that of NOISY are higher than that of the original alignment for all datasets. The log-likelihood for SRap is lower only for the least divergent dataset but otherwise trails that of NOISY. Compared to SRap and SRk, the log-likelihoods of all the other methods reduce more steadily with increase in divergence.

In most of the ISG datasets GBLOCKS removes all the columns during refinement yielding null alignments. So, we compare the performance of our Signal Refinement methods only with NOISY on the ISG datasets. Figure 3 shows the improvements in CI (top) and log-likelihood (bottom) as percentage increase from the values for the original alignments, averaged over 300 datasets. The improvements when SRk is used can be upto 27% increase in CI and from 7 to 35% increase in likelihood, depending on the dataset. Even the lower end of the range is better than the average performance of NOISY, Random1 and Random2. SRap also performs, on average and in most cases, better than NOISY. We also verify that our signal refinement methods give much better results than random elimination of columns.

Table 2 and figure 4 show the results for the Arthropods dataset for MAFFT and Clustal alignments.

Table 2 shows the lengths of the alignments used: number of columns for the original alignments and as percentages of these lengths for the refined alignments. We observe that NOISY, GBLOCKS-N and SRap eliminate roughly 40% of the columns. GBLOCKS-A and GBLOCKS-H eliminate less than 20% of the columns and SRk retains only about 30% of the columns.

The normalized consistency indices (top) and log-likelihood scores (bottom) are shown in Figure 4. The log-likelihood axis is inverted and so a shorter bar represents a higher likelihood. The general trend is similar to what is seen in the artificial datasets: The columns selected by SRk have the highest CI. SRap has the second highest CI and is followed by all the other methods that are comparable. With respect to likelihood also, SRk and SRap are better than all the other methods.

Thus we observe that the two methods of alignment masking based on our clustering framework can identify columns with high likelihood and compatibility better than the existing methods.

Phylogenetic Analysis

The main purpose of our signal refinement tool is to enable better inference of phylogenies and we now show that by identifying columns with high likelihood and inter-column compatibility using SR, we can indeed produce

better trees. In general we observe that SRk is not as stable as SRap, its variance in all the metrics we measure is high, perhaps due to the large number of columns eliminated. The improvements with SRap are more consistent and we use and recommend SRap for tree inference. In the following when we use the term SR, we refer to SRap.

Various summary statistics for the results on 300 ISG datasets are shown in figures 5-14. In each of the figures, the RF distance is used as the distance measure in the top subfigure and matching distance is used in the bottom subfigure. We find that in a majority of the datasets trees inferred from SR-refined alignments are closer to the respective reference trees than trees from both NOISY-refined alignments and the original alignments. These count statistics are detailed in figures 5-8.

Figures 5 and 6 show three pairwise-comparisons of trees with respect to the respective reference trees for Clustal and MAFFT alignments respectively. In each group of (three) bars, we compare two sets of trees. In the first group, we compare trees obtained from SR-refined alignments to the trees from the original alignments (SR vs MSA). The second and third groups of bars show similar comparisons for trees from NOISY-refined alignments and original alignments (NOISY vs MSA) and trees from SR-refined alignments and NOISY-refined alignments (SR vs NOISY) respectively.

For a single pairwise comparison, denoted by 'X vs Y', the three bars show:

- the proportion of instances in which trees from X-refined alignments are closer to the reference tree (denoted by "X better")
- the proportion of instances in which both trees from X-refined alignments and the trees from Y-refined alignments are equidistant from the reference trees (denoted by "X equals Y"), and
- the proportion of instances in which trees from the Y-refined alignments are closer to the reference trees (denoted by "Y better").

In the top subfigure of figure 5, where RF distances are used, we observe that in most ($\sim$ 80%) of the instances both SR and NOISY produce trees that are as close to the reference tree as the tree from the original alignment. In 16% of the instances SR-refined trees are better than the trees from the original alignments, and in 18% of the instances SR-refined trees are better than the trees from NOISY-refined alignments. The second group of bars show that the effect of NOISY is less beneficial: only in 2% of the instances trees inferred from NOISY-refined alignments are better than the trees from the original alignment. Matching distance, due to its higher resolution, shows the differences more clearly (in the bottom sub-figure). In more than 50% of the instances SR-refined alignments yield better trees than both trees from the original alignments and NOISY-refined alignments and in 88% of the instances the trees are not worse than the trees from the original alignments. NOISY can improve the tree, compared to the tree from the original alignment, only in 12% of the cases.

Figure 6 shows the same statistics when MAFFT alignments are used. The results are similar for RF distances. The bottom subfigure (using matching distance) shows that a higher percentage of instances (68%) show improvement in the inferred tree due to SR-based refinement.

Figure 7 shows a three-way comparison of inferred trees with respect to the reference trees. Each bar shows the proportion of instances (as percentages) where the trees obtained were closer to the reference tree than trees obtained from other methods. More formally, let d_{SR}, d_N and d_{MSA} denote the distances of the inferred trees from the reference tree, obtained from the SR-refined alignment, NOISY-refined alignment and the original alignment respectively. Each group of bars shows the number of instances where $d_{SR} < d_N$ *and* $d_{SR} < d_{MSA}$ (first bar), $d_N < d_{SR}$ *and* $d_N < d_{MSA}$ (second bar), and $d_{MSA} < d_{SR}$ *and* $d_{MSA} < d_N$ (third bar).

In the first group of bars (for MAFFT alignments) in the top subfigure (results where RF distance is used) we see that in 16% of the instances trees from SR-refined alignments were better than *both* trees from NOISY-refined and the original alignments. The corresponding counts when matching distance is used, is shown in the bottom subfigure: in 64% (45%) of the instances trees from SR-refined alignments were better than *both* trees from NOISY-refined and the original MAFFT (Clustal) alignments. Note that the counts do not add up to 100 since there are instances when all three alignments yield trees equidistant to the reference tree.

Figure 8 shows the previous three-way comparison but here we also include, in the count, instances where the inferred trees are equidistant to the trees being compared against. That is, each group of bars show the number of instances where $d_{SR} \leqslant d_N$ *and* $d_{SR} \leqslant d_{MSA}$ (first bar), $d_N \leqslant d_{SR}$ *and* $d_N \leqslant d_{MSA}$ (second bar), and $d_{MSA} \leqslant d_{SR}$ *and* $d_{MSA} \leqslant d_N$ (third bar). Thus we see that SR-refined trees obtain the best tree (i.e., closest to the reference tree among the three inferred trees) from both Clustal and MAFFT alignments in more than 95% of the instances when RF distance is used as a metric and in more than 78% of the instances when matching distance is used as a metric.

The distribution and extent of improvement are shown in figures 9-14. We measure the relative improvement, $(d_{MSA} - d_{SR})/d_{MSA}$, for those instances where $d_{SR} < d_{MSA}$. The amount of improvement depends on the dataset and varies from 1% to 10% with an overall average of 3% for MAFFT alignments and 3.7% for Clustal alignments where the improvement is measured by the percentage decrease in the matching distance to the reference tree.

Figure 9 shows a pointwise comparison of d_{SR} versus d_{MSA} for all 300 isg datasets. Each point, corresponding to a single dataset, is a pair (x,y) where x is the distance of the tree inferred from an SR-refined alignment and y is the distance of the tree inferred from the MAFFT alignment (both distances are from the true tree). The diagonal, drawn as a line, indicates those points where both the inferred trees are equidistant from the true tree. The points **above** the diagonal are those where $(x < y)$: the tree inferred from the SR-refined alignment is closer to the true tree than the tree inferred from the original alignment. The points **below** the

diagonal are those where $(y < x)$: the tree inferred from the SR-refined alignment is farther from the true tree than the tree inferred from the original alignment. As discussed earlier in the count statistics, majority of the points lie above the diagonal. The scatterplot clarifies both the distribution and extent of improvement due to refinement.

Figure 10 shows a similar comparison when NOISY is used for refinement instead of SR. We observe that the extent of improvement (compared to that with SR) does not change much. However, far more number of datasets are below the diagonal: those where refinement does not improve tree inference. Figure 11 shows a comparison of SR with respect to NOISY. Points above the diagonal (again, far more in number) are those datasets where trees inferred from SR-refined alignments are closer to the true tree, compared to trees inferred from NOISY-refined alignments. Similar plots, shown in figures 12-14, for Clustal alignments also display the same trends.

Figure 15 shows a comparison of the RF (top subfigure) and matching distances (bottom subfigure) of the inferred trees from the reference tree for the Eukaryotes dataset. The trees are inferred from fourteen different alignments which include the original MAFFT and Clustal alignments and refined alignments from six different methods as described in the previous section. Using GBLOCKS or ALISCORE does not improve the final phylogeny in any case. NOISY and SR, both yield small improvements in the inferred tree when MAFFT is used for aligning the sequences and the improvement is higher with SR. Clustal alignments appeared to be less affected by NOISY and SR-refined trees were closer to the reference tree by a small margin.

Figure 16 shows a similar comparison of distances of the inferred trees from the reference tree for the Arthropods dataset. Since ALISCORE removes very few columns during refinement (see table 2), the tree inferred from ALISCORE-refined alignment is very similar to the tree from the original alignment. Masking using GBLOCKS does not improve the inferred tree. For the MAFFT alignment, masking does not improve the quality of the inferred tree with any of the masking methods. Only NOISY and SR show improvements in the final tree for the Clustal alignments. The trees obtained from the original Clustal alignment and after masking using SR are shown in figures 18 and 19 respectively. Figure 17 shows the reference tree we used.

These results demonstrate the beneficial effects of masking, using SR, for tree inference on a variety of datasets. In our experiments, trees inferred from SR-refined alignments are better than or equivalent to the trees inferred from the original alignment in most cases. The amount of improvement depends on the dataset and the aligner used.

Discussion

We have designed a new approach to alignment masking based on the idea of extracting phylogenetic information from subsplits. The two Signal Refinement methods, SRk and SRap, are compared with three previous methods of alignment masking and alignments refined using SRk and SRap have better consistency indices and likelihood scores on many synthetic datasets and real datasets. We also show, empirically, that masking using SRap results in better tree inference in many instances.

We have intentionally not reported bootstrap scores to assess the robustness of inferred trees from the refined alignments. The bootstrap support values of trees from refined alignments are usually lower than those from the original alignments, even when the tree topologies are the same. The high support values for trees from the original alignments is attributed to the systematic biases caused by the aligners that are stronger in the divergent regions of the sequences that are removed by the refinement algorithms. See (Lake 1991; Higgins et al. 2005; Talavera and Castresana 2007) for a detailed discussion.

Castresana had reported that branch lengths and maximum-likelihood pairwise distances decrease after eliminating potentially noisy columns (Castresana 2000) and claimed that this is because the problem of saturation is completely alleviated. We also observe a reduction in branch lengths and pairwise distances when SR methods are applied. But in both these studies it is not clear what the optimal divergence levels are and hence not possible to determine the exact point where saturation starts obscuring branch-length information. Current methods can refine the signal-to-noise ratio in alignments with respect to the topology of the phylogeny, but further work is required to extend these methods to refine branch-length estimates.

There are limits to what alignment masking can achieve. The point at which removal of columns would no longer improve the inferred phylogenies or may actually lead to important loss of phylogenetic signal is data-dependent and can only be determined experimentally. In an ideal case when the model describes the data perfectly, we may expect the best possible inference from the data. However the most commonly used models do not describe the data well enough (see discussion in (Keane et al. 2006)). Alignment masking techniques can provide experimental tools to extract data that can fit the models better. As an experimental tool, SR provides a framework for automatically clustering columns based on phylogenetic similarity for further analysis of the alignments.

While the performance of any alignment masking method depends considerably on the data being analyzed, SRk and SRap have some clear advantages over other methods. The scoring function used in ALISCORE assumes that sequence variation is independent and identically distributed and the number of expected random matches has a Poisson distribution. These assumptions are not true in many cases. Further, if random similarity occurs in less than $\sim 20\%$ of the sequences then ALISCORE fails to identify such random regions. GBLOCKS was designed for moderately divergent sequences and its effective use depends on the careful setting of several parameters. NOISY relies on the assumption that pairwise distances give rise to robust circular split systems and performs reasonably well on all but the most divergent datasets. Our methods are particularly useful with divergent species where subsplits are perhaps the only remaining phylogenetic signal. They provide an easily usable, model-free view of clusters of columns that can be valuable for a phylogenetic analysis, especially in a post-processing phase. If the rate-based

criterion is used for cluster selection, then we recommend the application of SRap with a threshold value that is experimentally determined for the given dataset.

The more phylogenetically accurate an alignment is, the lesser is the improvement by any masking method. Tree reconstruction methods can also tolerate a large number of errors as we see in our ML analyses and also reported earlier by (Ogden and Rosenberg 2006). However, as sequence lengths and divergence increases, misaligned sequences may obfuscate the phylogenetic signal in a number of ways. They may confuse masking algorithms too, especially those that rely on sequence similarity—indeed, GBLOCKS recommends methods to detect misaligned fragments (such as those described in (Thompson et al. 1997)) before its application. Compatibility-based methods, including ours, have been found to be more robust in such conditions, while being more conservative since they return a smaller number of compatible columns in the presence of a large number of errors.

The clustering approach of SR is promising as it is able to select clusters of columns that yield better trees in many cases. But we still do not know how to characterize these clusters to be able to choose the best columns for tree reconstruction. Further study is needed in this direction to give more insight into the kind of data where the method works, or does not work. This, we presume, could also shed light on varying bootstrap values as columns are removed. The use of SRap in a framework such as SATE (Liu et al. 2009), where alignment and phylogenies are simultaneously inferred is also worth exploring.

The two steps in the SR methods – the clustering step that clusters the columns using a subsplit-based similarity measure and the selection step that uses likelihood-based column rates to provide a criterion for selection, are independent and indeed any other criterion could be used to choose the desired clusters. The rate-based criterion to select clusters works well in practice and we show that we can accurately select those columns that have high likelihood and high fit to an inferred phylogenetic tree. In turn, this results in improved phylogeny inference in a wide variety of datasets.

Material and Methods

Background: Splits and compatibility

Splits and compatibility based on splits have been studied extensively (Meacham and Estabrook 1985; Pisani and Wilkinson 2002; Semple and Steel 2003; Felsenstein 2004). The PICA manual (Wilkinson 2001) gives an excellent review. Here we recall some concepts that we will use in subsequent sections.

A column in an alignment matrix consists of elements from a fixed set of states. In DNA data, this set is $\{A,C,G,T,-\}$, where $-$ stands for a gap. Any column can be viewed as a partition of the set of taxa where each partition contains taxa with the same state. For example, a column $i = ATTTT--TAA$ is the partition $\{\{1,9,10\},\{2,3,4,5,8\},\{6,7\}\}$ since taxa 1,9 and 10 contain A, taxa 6 and 7 contain gaps and the remaining taxa contain T. Mathematically, a partition of a set A is a set of nonempty subsets of A such that every element of A is in exactly one of the subsets. We call each set in a partition, a *partition-set*.

Let $N = \{1,2,...,n\}$ and index the taxa and the rows of the alignment matrix from 1 to n. Let C be the set of character states. A column j defines a function $f_j : N \mapsto C$ such that $f_j(i) = A_{ij}$, where A_{ij} is the character state in the i^{th} row of the j^{th} column. Let P_j be a partition of N based on the column j, where each partition-set p_c is $f_j^{-1}(c)$, $c \in C$, that is, each partition-set contains those taxa that have the same character state in the column. A *split* from a column j is a bipartition of N, $A|\overline{A}$ where $A \in P_j$ and $\overline{A} = N - A$, the complement of A.

A split abstracts the phylogenetic information of a column and is independent of the data used in the matrix. For example, columns $j = GCCCCAACGG$, containing different DNA states, and $k = -DDDDRRD--$, containing amino acids, both have exactly the same splits as the example above, $i = ATTTT--TAA$. A split provides the fundamental signal for a clade in an inferred tree in any reconstruction method. Two splits $X|\overline{X}$ and $Y|\overline{Y}$ are *compatible* if one of the four intersections $X \cap Y$, $X \cap \overline{Y}$, $\overline{X} \cap Y$ and $\overline{X} \cap \overline{Y}$ is empty. Two columns are compatible if all their splits are pairwise compatible. Compatible splits are non-conflicting hypotheses of clades that can co-exist on the same tree. In a parsimony analysis, compatible columns are those that can fit the same tree without homoplasy.

The maximum compatibility approach to phylogenetic reconstruction (Estabrook et al. 1977) consists of finding the largest compatible subset of a collection of splits $\mathbf{C}$ from the given data and then reconstructing a phylogeny, either partially or fully, from this subset. Finding this subset is equivalent to finding the largest clique in the compatibility graph of $\mathbf{C}$, where each node represents a split and an edge joins two nodes if the corresponding splits are compatible. This problem is NP-hard (Day and Sankoff 1986) although a few tractable variants of the problem have been designed (Semple and Steel 2003). Note that every clique, not just the largest clique, in the compatibility graph corresponds to a hypothetical phylogeny, though not always fully resolved. We will use cliques of splits to measure the phylogenetic signal in an alignment. We define the *support* of a split $A|\overline{A}$ to be the number of times the partition-set A or $\overline{A}$ appears in the alignment matrix. The support of a clique of compatible splits is the sum of the supports of each split in the clique.

Compatibility analysis has been used to to identify conflicting phylogenetic signals by identifying incompatible columns (Meacham 1994; Pisani and Wilkinson 2002)) based on the notion that fast-evolving or homoplastic columns are unlikely to have compatible splits with slowly evolving columns. Recently, Pisani (Pisani 2004) proposed a compatibility-based randomization test to identify and remove fast-evolving columns in order to diagnose and counter the effects of long-branch attraction. Cummins and McInerney (Cummins and McInerney 2011) have also developed a compatibility-based method of inferring the rate of evolution of columns. We design a new approach based on the novel idea of subsplits that extends this suite of methods.

A novel clustering approach

Our key observation is that the partition-sets (and their corresponding splits) alone do not extract all the information in an alignment. For example, column $X = ATTTTTAA$ with partition-sets $\{1,7,8\}$ and $\{2,3,4,5,6\}$ and column $Y = AAATTTAT$ with partition-sets $\{1,2,3,7\}$ and $\{4,5,6,8\}$ have no compatible splits. But together these two columns can support, without conflict, the hypotheses for clades $\{4,5,6\}$ and $\{1,7\}$ (and the corresponding splits: $\{4,5,6\}|\{1,2,3,7,8,9\}$ and $\{1,7\}|\{2,3,4,5,6,8,9\}$) which are the common subsets of the partition-sets in the two columns.

We define a *partition-subset* to be a non-empty subset of a partition-set of a column. A *subsplit* from a partition P_j is the bipartition $A|\overline{A}$, where $A \subset P \in P_j$, that is, A is a subset of a partition-set P belonging to the partition P_j. Note that every partition-subset P corresponds to a subsplit $P|\overline{P}$. A *trivial* partition-subset has cardinality one: it represents a single taxon. The definitions of support and compatibility are analogous to those for splits since mathematically both splits and subsplits are just bipartitions of sets. The difference only lies in the way they are derived from a column in the alignment.

We observe that there is valuable phylogenetic signal not just in the splits but in all the subsplits as well. We design a new approach that can use subsplits to extract phylogenetic signal efficiently. We view the problem as a clustering problem. Our goal is to cluster the columns such that columns with roughly equivalent phylogenetic signal are in the same cluster. In this way we hope to classify the columns such that all 'noisy' columns form identifiable clusters that can be eliminated. There are three key elements in our method:

1. A measure of distance between columns
2. A clustering method
3. A criterion to eliminate one or more noisy clusters

We now describe each of these. We define a similarity coefficient between any two columns i and j to be,

$$S_{ij} = \Sigma_{s_i \in i} \Sigma_{s_j \in j} \frac{C(s_i,s_j) - I(s_i,s_j)}{C(s_i,s_j) + I(s_i,s_j)}$$

where s_i is a partition-set in column i and s_j is a partition-set in column j and the summations run over all partition-sets in columns i and j respectively; $C(s_i,s_j)$ and $I(s_i,s_j)$ are defined as follows.

- $C(s_i,s_j) = (2^{|s_i \cap s_j|} - 1 - |s_i \cap s_j|) + (2^{|s_i - s_j|} - 1 - |s_i - s_j|) + (2^{|s_j - s_i|} - 1 - |s_j - s_i|)$
- $I(s_i,s_j) = (2^{|s_i|} - 1 - |s_i|) + (2^{|s_j|} - 1 - |s_j|) - C(s_i,s_j)$.

$C(s_i,s_j)$ is the number of non-trivial partition-subsets in the sets: $s_i \cap s_j$, $s_i - s_j$ and $s_j - s_i$. Any subsplit from a partition-subset in $s_i - s_j$ is compatible with any subsplit from a partition-subset in s_j, since they are disjoint. $(2^{|s_i - s_j|} - 1 - |s_i - s_j|)$ counts the number of such non-trivial partition-subsets. Similarly, $(2^{|s_j - s_i|} - 1 - |s_j - s_i|)$ counts the number of such non-trivial partition-subsets in s_j. To this we add the number of non-trivial partition-subsets that are present in both the columns, $(2^{|s_i \cap s_j|} - 1 - |s_i \cap s_j|)$. $I(s_i,s_j)$ counts all the remaining non-trivial partition-subsets. The difference $C(s_i,s_j) - I(s_i,s_j)$ is a conservative estimate of the difference between the number of compatible subsplits and the number of incompatible subsplits in the two columns. Since we only compute the cardinalities of the exponentially growing sets, this is an easily computable measure of similarity and works well in practice. The distance measure, D_{ij}, between columns i and j is then $D_{ij} = 1 - S_{ij}$.

To eliminate noisy clusters, we choose a criterion based on the rate of evolution of columns. We use a parametric method of inferring the rate of evolution assuming the GTR model (Lanave et al. 1984). These rates can be determined approximately (without performing the entire likelihood analysis) using Yang's method (Yang 1994). A tree topology is used as input to this procedure but it has been shown that the method is robust with respect to the correctness of this topology (Sullivan et al. 1996). A distance-based estimate, that can be computed very fast, does sufficiently well in practice. In our experiments we use the Neighbor-Joining tree (Saitou and Nei 1987). The rates are discrete approximations of the rates from a Gamma distribution that are binned into a finite number (a parameter selected by the user) of categories. In our experiments, we set the number of categories to be 10. Thus each column has a rate category in the range $[1,10]$. We define the rate of a cluster of columns to be the average rate category of columns in the cluster. Then, as explained below, we eliminate one or more clusters that have higher rate than the rest of the clusters.

Many general-purpose clustering algorithms have been designed that can be used in this framework. In some algorithms, such as *k*-means, the number of clusters (k) have to be specified in advance. In such algorithms it would be ideal to obtain two clusters so that one of them can be discarded and the other retained for phylogenetic inference. If more than two clusters are specified then a strategy is needed to select this number and to select the number of clusters that should be eliminated. Some other clustering algorithms do not need the number of clusters as input and can return an arbitrary number of clusters depending on the data (e.g. affinity propagation (Frey and Dueck 2007)). We favor these kind of algorithms since the number of clusters should indeed be dependent on the data. Among the algorithms we tested we found that affinity propagation and *k*-means clustering give us the best results. We denote by **SRk** (**S**ignal **R**efinement with **k**-means clustering) the method that uses *k*-means clustering and by **SRap** (**S**ignal **R**efinement with **A**ffinity **P**ropagation clustering) the method that uses affinity propagation. Note that SRk returns exactly two clusters (we set $k = 2$) and this can be used to perform alignment masking by eliminating the cluster with the higher rate. With SRap, a threshold on the cluster rate is selected and all clusters with rate higher than this threshold are eliminated.

Data and Experiments

Data Simulations We tested the alignment masking methods and their effects on phylogenetic inference on two artificial nucleotide datasets: one that simulates highly divergent DNA and another that simulates DNA coding sequences with the mutational spectrum of *E. coli*. All the datasets can be downloaded from our website (http://lcbb.epfl.ch/softwares/SR.html).

The first set was taken from the iSGv2.1.0 benchmark datasets (iSG Datasets 2010) that represent highly diverged sequence collections. These datasets are considered to be very difficult instances for phylogenetic inference. The datasets used in the corresponding simulations are given in the appendix. The level of sequence divergence aimed at in these datasets is the average normalized Hamming distance (ANHD) between sequences approaching saturation (ANHD = 0.75) for nucleotide sequences, with varying levels of gappiness in the true alignments. The parameters chosen for the simulations are the same as those in (Liu et al. 2009): the same nucleotide frequencies, indel length distributions, and GTR+Gamma parameters, for non-coding DNA sequences. The datasets are simulated using indel-Seq-Gen v.2.1.0 (Strope et al. 2009), with the guide trees generated by r8s (Sanderson 2003). We chose 10 replicates for each of the 30 model conditions for our experiments analyzing 300 datasets in total; each model condition is defined by a distribution of gap lengths (short, medium, or long), a probability of indel occurrence, average root-to-tip tree length, and the number of taxa. The original datasets have 5,000 taxa each. For our experiments we sample 100 taxa, uniformly at random, from the dataset (While running the experiments on 5000 taxa is possible it would be very time-consuming to test on a large number of datasets). The phylogenies corresponding to these 100 taxa were extracted from the true tree (on which the initial simulations were conducted). We will refer to these datasets as the *ISG datasets*.

The second dataset is created using the software EvolveAGene (Hall 2008) that simulates evolution by separating mutation from selection. The proportion of base substitutions and indels, i.e. the mutational spectrum, is that of *E. coli*. It allows the user to specify the way selection operates on the mutations. There are many options provided to control the effects of selection on portions of the sequence or branches of the input tree. See (Hall 2008) for more details. We created datasets containing trees that have branch lengths varying from almost equal short branch lengths to a mixture of short and long branches that are difficult to infer. One of the input parameters, the average branch length, represents the average number of changes per column. The actual length of each branch is a random number between 0 and twice the average branch length. We used five different values for the average branch length: 0.1, 0.2, 0,3, 0.4 and 0.5. The other parameters were set to their default values. For each value of average branch length, we created datasets with roughly 1,000, 5,000 and 10,000 columns (the exact number of columns differs by a few nucleotides and depends on the simulation), for 10 and 20-taxon trees. These relatively small datasets were used for clique analysis (which is infeasible for larger datasets) and to compare the masking algorithms. We will refer to these datasets as the *EAG datasets*.

Real datasets To study the performance of the alignment masking methods, we use the mitochondrial protein-coding genes from 46 species of arthropods (*Crustacea, Hexapoda, Chelicerata, Myriapoda*). The accession numbers of the sequences are given in the appendix. The arthropod phylogeny has been studied for decades and has been difficult to resolve. For our reference phylogeny we use the taxonomic phylogeny obtained from NCBI Nucleotide database (Sayers et al. 2010) for each species used. The taxonomic classification does not give us a fully resolved binary tree. We resolve polytomies and also corroborate the phylogeny with the phylogenies reported in (Regier et al. 2010; Giribet et al. 2001; Cameron et al. 2007; Black and Piesman 1994; Kristensen 1991; Tree of Life web project 1995). The reference tree is shown in figure 11. Beside the genus and species names of each taxon we also show the subphylum and class names.

We also use a curated alignment from Robin Gutell's comparative RNA database (Cannone et al. 2002). The rRNA alignments in this database are highly reliable since they are obtained from the RNA secondary structure. The alignment we use contains 117 ribosomal RNA sequences (each with 9,079 sites) of the 23S gene sampled from eukaryotes. The average gap length is 12.6 and the percentage of indels is 59.7. We use the cleaned alignment and reference tree provided by (Warnow 2009).

Evaluation Metrics We use two metrics to measure the quality of an alignment: the normalized consistency index and the normalized likelihood score. The consistency index, as defined in (Farris 1969), is defined with respect to a given tree. It measures the degree of correlation between a column and a phylogeny and is an indicator of the amount of homoplasy in the column. Let $r(i)$ be the number of states in a column i; for DNA data with gaps that we use, the maximum possible value of r is 5. $r(i)-1$ is the minimum number of mutations needed to fit this column on to a tree. Let $l(i)$ be the inferred minimum number of mutations on the given tree for this column. This is also the parsimony score of this column and can be computed using Fitch's algorithm (Fitch 1971). The consistency index of the column i is the ratio $(r(i)-1)/l(i)$. It is 1 when the column has the most parsimonious evolution and reduces with increase in homoplasy. The consistency index of an alignment is the sum of the consistency indices of each of its columns. The *Normalized Consistency Index* (CI) of an alignment is its consistency index normalized by the number of columns. A high consistency index indicates low homoplasy in the dataset: given two alignments, the one with a higher consistency index typically has a lower parsimony score. In our experiments we do find that datasets with better consistency index have better parsimony scores and so we report only the consistency index in our comparisons. It has been shown that the consistency index is negatively correlated with the number of taxa (Archie 1989). Since we always compare alignments with the same number of taxa, we do not need to normalize the index by the number of taxa. The log-likelihood of an alignment, given a tree

and assuming a model of evolution, is also dependent on the number of columns. So, we normalize it by the number of columns to get the *Normalized Log-Likelihood score*. We assume the GTR model of evolution in our likelihood computations.

We use the Robinson-Foulds (RF) distance (Robinson and Foulds 1981) and the matching distance (Lin et al. 2011) to measure the topological distance between two trees. The former is widely used but has many undesirable properties: it is poorly distributed and thus affords little discrimination while also lacking robustness in the face of very small changes—reattaching a single leaf elsewhere in a tree of any size can instantly maximize the distance. These problems have been discussed in (Lin et al. 2011) and the matching distance has been designed to overcome these problems.

To understand the matching distance recall that each branch in a phylogeny is a bipartition of taxa or a split. The two sets of the split are the taxa of the two subtrees formed by removing that branch. Given two trees, the matching distance 'matches' each branch of one tree with its closest branch in the other tree—the degree of closeness being determined by the number of leaves that have to be moved across the branch to make the bipartitions equal; this number is called the weight of the matched pair of branches. Thus, if two branches have the same bipartition of leaves, they will be matched and the associated weight will be zero. The matching distance is the sum of these weights for every matched pair of branches. It is worth noting that the matching distance penalizes changes in deeper branches more than changes in branches closer to the leaves. This measure has finer resolution than the RF distance and so, it is particularly helpful in comparing trees when the RF distance of the two (different) trees from a third (reference) tree is the same. We use both measures to compute the distance between an inferred tree and the reference tree.

Experimental Setup We restrict our study to maximum-likelihood (ML) trees that are inferred using RAxML (Stamatakis 2006) (assuming the GTRGAMMA model of evolution).

Clique Analysis: To test our hypothesis that subsplits may contain valuable phylogenetic signal we perform clique analyses on the EAG datasets that contain 10 taxa. Using MAFFT alignments, we compute all the splits and all the subsplits. We then compute all maximal cliques of splits and all maximal cliques of subsplits, using the Bron-Kerbosch algorithm (Bron and Kerbosch 1973) (implemented in gPy (gPy 2008)) and their support values. Such a computation is prohibitively time-consuming for larger datasets and so we restricted it to datasets with 10 taxa.

Performance of Masking Algorithms: To compare the performance of various alignment masking algorithms, experiments are conducted as follows. For each dataset we first run an alignment algorithm. We use two different aligners: MAFFT and Clustal and then infer ML trees from each of them. These aligners were chosen because they employ different heuristics (although they have many similarities) and thus yield alignments of fairly different qualities.

We use different methods of alignment masking to obtain refined alignments: NOISY, GBLOCKS and ALISCORE along with our Signal Refinement methods SRk and SRap. For NOISY and ALISCORE, the default parameters are used. For GBLOCKS, we set the minimum block length to 5, as recommended by the authors for DNA alignments. We test all three different values for the allowed gap positions—none, half and all—we will call these three different settings GBLOCKS-N (GN), GBLOCKS-H (GH) and GBLOCKS-A (GA) respectively. All other parameters are set to their default values.

SRk uses k-means clustering with $k = 2$, thus yielding exactly two clusters. The cluster with higher average rate of columns is eliminated. SRap uses affinity propagation where the number of clusters need not be specified in advance. We set a default rate threshold of 6: clusters with rate higher than 6 are eliminated. Both the clustering algorithms are implemented using the scikit library (Scikits 2007) in Python. As control experiments, we also create two other refined alignments where the columns are chosen uniformly at random from the original alignment. Random1 and Random2 contain as many columns as those in the alignments refined by SRap and SRk respectively.

We reconstruct ML trees from each of the refined alignments. The normalized consistency index and normalized likelihoods are recorded for the original alignments and the trees inferred from them along with those for the refined alignments and the trees inferred from them. Metrics for the random alignments provide useful comparison points for SRk and SRap.

Phylogenetic Analysis: We compare ML trees inferred from alignments before and after refinement for the ISG datasets, the Eukaryotes dataset and the Arthropods dataset. In the case of the Eukaryotes dataset, the sequences are extracted from the reference alignment.

MAFFT and Clustal are used to align the sequences. The alignments are refined using SRap, GBLOCKS, NOISY and ALISCORE as described in the previous section. From each alignment (before and after refinement) we infer trees using RAxML and compare the trees to the respective reference trees.

Acknowledgments

I thank Bernard Moret, Xiuwei Zhang and Karl Kjer for many useful discussions and comments on the manuscript.

Literature Cited

Aagesen L. 2004. The information content of an ambiguously alignable region, a case study of the trnL intron from the Rhamnaceae. Organisms Diversity and Evolution 4(1-2):35 – 49.

Ahola V, Aittokallio T, Vihinen M, Uusipaikka E. 2006. A statistical score for assessing the quality of multiple sequence alignments. BMC Bioinformatics 7(1):484.

Archie JW. 1989. Homoplasy excess ratios: New indices for measuring levels of homoplasy in phylogenetic sys-

tematics and a critique of the consistency index. Systematic Zoology 38(3):pp. 253–269.

Black WC, Piesman J. 1994. Phylogeny of hard- and soft-tick taxa (Acari: Ixodida) based on mitochondrial 16s rDNA sequences. Proceedings of the National Academy of Sciences 91(21):10034–10038.

Bron C, Kerbosch J. 1973. Algorithm 457: finding all cliques of an undirected graph. Communications of the ACM 16:575–577.

Cameron SL, Lambkin CL, Barker SC, Whiting MF. 2007. A mitochondrial genome phylogeny of diptera: whole genome sequence data accurately resolve relationships over broad timescales with high precision. Systematic Entomology 32(1):40–59.

Cannone JJ, Subramanian S, Schnare MN, et al.(14 co-authors). 2002. The Comparative RNA Web (CRW) site: An online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics 3(15).

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular Biology and Evolution 17(4):540–552.

Cummins CA, McInerney JO. 2011. A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. Systematic Biology 60(6):833–844.

Day WHE, Sankoff D. 1986. Computational complexity of inferring phylogenies by compatibility. Systematic Zoology 35(2):pp. 224–229.

Doyle JJ, Doyle JL, Rauscher JT, Brown AHD. 2004. Diploid and polyploid reticulate evolution throughout the history of the perennial soybeans (Glycine subgenus Glycine). New Phytologist 161(1):121–132.

Dress A, Flamm C, Fritzsch G, Grunewald S, Kruspe M, Prohaska S, Stadler P. 2008. Noisy: Identification of problematic columns in multiple sequence alignments. Algorithms for Molecular Biology 3(1):7.

Estabrook GF, Strauch Joseph GJ, Fiala KL. 1977. An application of compatibility analysis to the Blackiths' data on orthopteroid insects. Systematic Zoology 26(3):pp. 269–276.

Farris JS. 1969. A successive approximations approach to character weighting. Systematic Biology 18(4):374–385.

Felsenstein J. 2004. Inferring Phylogenies. Sinauer Assoc., Sunderland, MA.

Fernandes AP, Nelson K, Beverley SM. 1993. Evolution of nuclear ribosomal RNAs in kinetoplastid protozoa: perspectives on the age and origins of parasitism. Proceedings of the National Academy of Sciences 90(24):11608–11612.

Fitch WM. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Systematic Biology 20(4):406–416.

Fleissner R, Metzler D, von Haeseler A. 2005. Simultaneous statistical multiple alignment and phylogeny reconstruction. Systematic Biology 54(4):548–561.

Frey BJ, Dueck D. 2007. Clustering by passing messages between data points. Science 315(5814):972–976.

Gatesy J, DeSalle R, Wheeler W. 1993. Alignment-ambiguous nucleotide sites and the exclusion of systematic data. Molecular Phylogenetics and Evolution 2(2):152 – 157.

Giribet G, Edgecombe GD, Wheeler WC. 2001. Arthropod phylogeny based on eight molecular loci and morphology. Nature 413.

gPy. 2008. gPy: A collection of python modules for graphical models. `http://www-users.cs.york.ac.uk/jc/teaching/agm/gPy/`.

Grundy WN, Naylor GJ. 1999. Phylogenetic inference from conserved sites alignments. Journal of Experimental Zoology 285(2):128–139.

Hall BG. 2008. Simulating DNA coding sequence evolution with EvolveAGene 3. Molecular Biology and Evolution 25(4):688–695.

Higgins DG, Blackshields G, Wallace IM. 2005. Mind the gaps: Progress in progressive alignment. Proceedings of the National Academy of Sciences of the United States of America 102(30):10411–10412.

Inagaki Y, Roger AJ. 2006. Phylogenetic estimation under codon models can be biased by codon usage heterogeneity. Molecular Phylogenetics and Evolution 40(2):428 – 434.

iSG Datasets. 2010. iSGv2.1.0 benchmark datasets. `http://bioinfolab.unl.edu/~cstrope/iSG/benchmark/`.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research 30(14):3059–3066.

Keane T, Creevey C, Pentony M, Naughton T, Mclnerney J. 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. BMC Evolutionary Biology 6(1):29.

Kim J, Ma J. 2011. PSAR: measuring multiple sequence alignment reliability by probabilistic sampling. Nucleic Acids Research 39(15):6359–6368.

Kjer KM, Gillespie JJ, Ober KA. 2007. Opinions on multiple sequence alignment, and an empirical comparison of repeatability and accuracy between POY and structural alignment. Systematic Biology 56(1):133–146.

Kjer KM. 1995. Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: An example of alignment and data presentation from the frogs. Molecular Phylogenetics and Evolution 4(3):314 – 330.

Kristensen NP. 1991. Phylogeny of extant hexapods. In Naumann ID, Carne PB, Lawrence JF, Nielsen ES, Spradberry JP, Taylor RW, Whitten MJ, Littlejohn MJ, editors, Insects of Australia: A Textbook for Students and Research Workers. Volume I and II, 125–140.

Lake JA. 1991. The order of sequence alignment can bias the selection of tree topology. Molecular Biology and Evolution 8(3):378–385.

Lanave C, Preparata G, Sacone C, Serio G. 1984. A new method for calculating evolutionary substitution rates. Journal of Molecular Evolution 20:86–93. 10.1007/BF02101990.

Landan G, Graur D. 2007. Heads or tails: A simple reliability check for multiple sequence alignments. Molecular Biology and Evolution 24(6):1380–1383.

Lassmann T, Sonnhammer ELL. 2005. Automatic assessment of alignment quality. Nucleic Acids Research 33(22):7120–7128.

Lee MS. 2001. Unalignable sequences and molecular evolution. Trends in Ecology and Evolution 16(12):681 – 685.

Lerat E, Daubin V, Moran NA. 2003. From gene trees to organismal phylogeny in prokaryotes:the case of the -proteobacteria. PLoS Biol 1(1):e19.

Lin Y, Rajan V, Moret B. 2011. A metric for phylogenetic trees based on matching. In Chen J, Wang J, Zelikovsky A, editors, Bioinformatics Research and Applications, volume 6674 of *Lecture Notes in Computer Science*, 197–208. Springer Berlin / Heidelberg.

Liu K, Raghavan S, Nelesen S, Linder CR, Warnow T. 2009. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. Science 324(5934):1561–1564.

Lunter G, Miklos I, Drummond A, Jensen J, Hein J. 2005. Bayesian coestimation of phylogeny and sequence alignment. BMC Bioinformatics 6(1):83.

Lutzoni F, Wagner P, Reeb V, Zoller S. 2000. Integrating ambiguously aligned regions of DNA sequences in phylogenetic analyses without violating positional homology. Systematic Biology 49(4):628–651.

Lytynoja A, Milinkovitch MC. 2001. SOAP, cleaning multiple alignments from unstable blocks. Bioinformatics 17(6):573–574.

Meacham CA. 1994. Phylogenetic relationships at the basal radiation of angiosperms: Further study by probability of character compatibility. Systematic Botany 19(4):pp. 506–522.

Meacham CA, Estabrook GF. 1985. Compatibility methods in systematics. Annual Review of Ecology and Systematics 16:pp. 431–446.

Misof B, Misof K. 2009. A Monte Carlo approach successfully identifies randomness in multiple sequence alignments : A more objective means of data exclusion. Systematic Biology 58(1):21–34.

Morrison DA, Ellis JT. 1997. Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of apicomplexa. Molecular Biology and Evolution 14(4):428–441.

Notredame C, Higgins DG, Heringa J. 2000. T-coffee: a novel method for fast and accurate multiple sequence alignment. Journal of Molecular Biology 302(1):205 – 217.

Ogden TH, Rosenberg MS. 2006. Multiple sequence alignment accuracy and phylogenetic inference. Systematic Biology 55(2):314–328.

Pesole G, Attimonelli M, Preparata G, Saccone C. 1992. A statistical method for detecting regions with different evolutionary dynamics in multialigned sequences. Molecular Phylogenetics and Evolution 1(2):91 – 96.

Pisani D. 2004. Identifying and removing fast-evolving sites using compatibility analysis: An example from the Arthropoda. Systematic Biology 53(6):978–989.

Pisani D, Wilkinson M. 2002. Matrix representation with parsimony, taxonomic congruence, and total evidence. Systematic Biology 51(1):151–155.

Prakash A, Tompa M. 2005. Statistics of local multiple alignments. Bioinformatics 21(suppl 1):i344–i350.

Privman E, Penn O, Pupko T. 2012. Improving the performance of positive selection inference by filtering unreliable alignment regions. Molecular Biology and Evolution 29(1):1–5.

Redelings BD, Suchard MA. 2005. Joint bayesian estimation of alignment and phylogeny. Systematic Biology 54(3):401–418.

Regier JC, Shultz JW, Zwick A, Hussey A, Ball B, Wetzer R, Martin JW, Cunningham CW. 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature 463.

Robinson D, Foulds L. 1981. Comparison of phylogenetic trees. Mathematical Biosciences 53(1-2):131 – 147.

Rodrigo AG, Bergquist PR, Bergquist PL. 1994. Inadequate support for an evolutionary link between the metazoa and the fungi. Systematic Biology 43(4):578–584.

Rodriguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Systematic Biology 56(3):389–399.

Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425(6960):798–804.

Ruiz-Trillo I, Riutort M, Littlewood DTJ, Herniou EA, Bagu J. 1999. Acoel flatworms: Earliest extant bilaterian metazoans, not members of platyhelminthes. Science 283(5409):1919–1923.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4(4):406–425.

Sanderson MJ. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics 19(2):301–302.

Sayers EW, Barrett T, Benson DA, et al. (40 co-authors). 2010. Database resources of the national center for biotechnology information. Nucleic Acids Research 38(suppl 1):D5–D16.

Scikits. 2007. Scikits.learn: machine learning in python. `http://scikit-learn.sourceforge.net/index.html`.

Semple C, Steel M. 2003. Phylogenetics. Oxford University Press, UK.

Smythe AB, Sanderson MJ, Nadler SA. 2006. Nematode small subunit phylogeny correlates with alignment parameters. Systematic Biology 55(6):972–992.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22(21):2688–2690.

Strope CL, Abel K, Scott SD, Moriyama EN. 2009. Biological sequence simulation for testing complex evolutionary hypotheses: indel-Seq-Gen version 2.0. Molecular Biology and Evolution 26(11):2581–2593.

Sullivan J, Holsinger K, Simon C. 1996. The effect of topology on estimates of among-site rate variation. Journal of Molecular Evolution 42:308–312.

Susko E, Spencer M, Roger A. 2005. Biases in phylogenetic estimation can be caused by random sequence segments. Journal of Molecular Evolution 61:351–359. 10.1007/s00239-004-0352-9.

Swofford D, Olsen G, Waddell P, Hillis D. 1996. Phylogenetic inference. In Hillis D, Mable B, Moritz C, editors, Molecular Systematics, 407–514. Sinauer Assoc., Sunderland, MA.

Takahashi K, Terai Y, Nishida M, Okada N. 2001. Phylogenetic relationships and ancient incomplete lineage sorting among cichlid fishes in Lake Tanganyika as revealed by analysis of the insertion of retroposons. Molecular Biology and Evolution 18(11):2057–2066.

Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Systematic Biology 56(4):564–577.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. 1997. The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research 25(24):4876–4882.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22(22):4673–4680.

Tree of Life web project. 1995. Arthropoda. `http://tolweb.org/Arthropoda/2469/1995.01.01`.

Wang L, Jiang T. 1994. On the complexity of multiple sequence alignment. Journal of Computational Biology 1(4):337–348.

Wang LS, Leebens-Mack J, Wall P, Beckmann K, de Pamphilis C, Warnow T. 2011. The impact of multiple protein sequence alignment on phylogenetic estimation. IEEE/ACM Transactions on Computational Biology and Bioinformatics 8(4):1108 – 1119.

Warnow T. 2009. Empirical datasets with reference topologies. `http://www.cs.utexas.edu/~phylo/datasets/phylogeny-topology.html`.

Wheeler TJ, Kececioglu JD. 2007. Multiple alignment by aligning alignments. Bioinformatics 23(13):i559–i568.

Wilkinson. 2001. PICA user manual. Department of Zoology, The Natural History Museum, London .

Xia X, Xie Z, Kjer KM. 2003. 18S ribosomal RNA and tetrapod phylogeny. Systematic Biology 52(3):283–295.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. Journal of Molecular Evolution 39:306–314.

Appendix

Accession numbers of taxa used in the Arthropods dataset:

1. NC_000844,
2. NC_000857,
3. NC_000875,
4. NC_001322,
5. NC_001566,
6. NC_001620,

7. NC_001709,
8. NC_001712,
9. NC_002010,
10. NC_002074,
11. NC_002084,
12. NC_002184,
13. NC_002609,
14. NC_002629,
15. NC_002651,
16. NC_002660,
17. NC_002697,
18. NC_002735,
19. NC_003057,
20. NC_003058,
21. NC_003081,
22. NC_003343,
23. NC_003344,
24. NC_003367,
25. NC_003368,
26. NC_003372,
27. NC_003395,
28. NC_003970,
29. NC_003979,
30. NC_004251,
31. NC_004357,
32. NC_004370,
33. NC_004371,
34. NC_004454,
35. NC_004465,
36. NC_004529,
37. NC_004622,
38. NC_004816,
39. NC_005037,
40. NC_005291,
41. NC_005292,
42. NC_005293,
43. NC_005306,
44. NC_005333,
45. AY191995,
46. AY191994

The names of the datasets used in the ISG datasets are below. The details of the simulations can be found in (iSG Datasets 2010).

1. 5000_1.5_5_0.0008,
2. 5000_1.5_5_0.001,
3. 5000_1.5_5_1e-07,
4. 5000_1.5_5_4e-05,
5. 5000_1.5_5_8e-05,
6. 5000_2.5_5_0.0001,
7. 5000_2.5_5_0.0002,
8. 5000_2.5_5_0.0005,
9. 5000_2.5_5_1e-07,
10. 5000_2.5_5_5e-05,
11. 5000_2_30_2.5e-05,
12. 5000_3.5_30_2e-05,
13. 5000_3.5_30_2e-06,
14. 5000_3.5_30_7e-05,
15. 5000_3_30_0.0001,
16. 5000_4_31_0.0001,
17. 5000_4_31_1e-08,
18. 5000_4_31_3e-05,
19. 5000_4_31_5e-05,
20. 5000_4_31_5e-06,
21. 5000_7_30_1.5e-06,
22. 5000_7_30_3e-06,
23. 5000_8_30_1.5e-06,
24. 5000_8_30_1e-09,
25. 5000_8_30_3e-06,
26. 5000_8_31_5e-07,
27. 5000_9_31_1.5e-06,
28. 5000_9_31_1e-09,
29. 5000_9_31_3e-07,
30. 5000_9_31_5e-07

Tables

	#cliques	#cliques	Support: ML tree edges		Support: True tree edges	
Rate	splits	subsplits	splits	subsplits	splits	subsplits
0.1	1	1575	18	45	0	24
0.2	1	15	24	42	0	18
0.3	5	1836	63	219	0	120
0.4	4	1350	171	361	0	149
0.5	3	1800	23	77	6	30

Table 1: Results of clique analysis on EAG datasets containing 10 taxa and 1000 columns. The EAG datasets simulate DNA coding sequences with the mutational spectrum of *E. Coli*. The alignments were created using MAFFT and the trees were obtained using RAxML. The first two columns show the number of cliques found when splits and subsplits are used respectively, the latter being much larger since there are exponentially higher number of subsplits than splits. The support of a split or a subsplit is the number of columns in which the split or subsplit appears. Both the inferred ML tree and the true tree used for the simulation have higher support from subsplits than splits. Such comparisons in this and other datasets lead us to design methods of refining the phylogenetic signal of an alignment using subsplits. Note that this analysis is only to illustrate the potential use of subsplits in alignment masking and improved phylogeny inference and is not related to the performance of our Signal Refinement (SR) methods.

Aligner	Length	Percentage of retained columns						
		SRap	SRk	NOISY	GA	GN	GH	ALISCORE
MAFFT	11855	61.87	33.03	65.09	83.81	62.16	81.41	98.99
Clustal	11202	54.88	31.52	69.10	93.92	63.46	90.72	99.72

Table 2: Lengths (number of columns) of the original alignments from two different aligners and percentages of retained columns in the refined alignments for the Arthropods dataset containing 46 taxa. Abbreviations – SRap: Signal Refinement using affinity propagation clustering, SRk: Signal Refinement using k-means clustering, GA: GBLOCKS with allowed gap positions set to All, GN: GBLOCKS with allowed gap positions set to None, GH: GBLOCKS with allowed gap positions set to Half.

Figure Captions

Figure 1

Normalized consistency indices of refined and original alignments each containing 20 taxa, from the EAG datasets. The EAG datasets simulate DNA coding sequences with the mutational spectrum of *E. Coli*. The alignments were created using Clustal and the trees were obtained using RAxML. MSA refers to the original alignment, the others are alignments after eliminating columns using various masking techniques. The abbreviations are – SRap: Signal Refinement using affinity propagation clustering, SRk: Signal Refinement using k-means clustering, GA: GBlocks with allowed gap positions set to All, GN: GBlocks with allowed gap positions set to None, GH: GBlocks with allowed gap positions set to Half. The top subfigure shows the curves for alignments refined by SRk, SRap, Random1 and Random2 (that eliminate a random set of columns; the number of columns eliminated by Random1 and Random2 is the same as the number eliminated by SRap and SRk respectively) and the original unrefined alignment (MSA). The curves for Random1 and Random2 are almost identical to that of the original alignment showing that a random deletion of columns does not help in eliminating homoplastic columns. The elimination has to be guided by a principled search and both SRk and SRap can do that, with SRk performing better. The bottom subfigure repeats the curves for SRk, SRap and the original alignment, and shows curves for refined alignments from NOISY and GBlocks. The best CI values are achieved by SRk and SRap. Thus, SRap and SRk extract columns that have a better fit to their inferred trees.

Figure 2

Normalized log-likelihoods of refined and original alignments each containing 20 taxa, from the EAG datasets. The EAG datasets simulate DNA coding sequences with the mutational spectrum of *E. Coli*. The alignments were created using Clustal and the trees were obtained using RAxML. MSA refers to the original alignment, the others are alignments after eliminating columns using various masking techniques. The abbreviations are – SRap: Signal Refinement using affinity propagation clustering, SRk: Signal Refinement using k-means clustering, GA: GBlocks with allowed gap positions set to All, GN: GBlocks with allowed gap positions set to None, GH: GBlocks with allowed gap positions set to Half. The top subfigure shows the curves for alignments refined by SRk, SRap, Random1 and Random2 (that eliminate a random set of columns; the number of columns eliminated by Random1 and Random2 is the same as the number eliminated by SRap and SRk respectively) and the original alignment. The log-likelihood of refined alignments from SRk is significantly higher than

that of its randomized counterpart Random2. The performance of SRap is similar except on datasets with less divergence. The bottom subfigure repeats the curves for SRk, SRap and the original alignment, and shows curves for refined alignments from NOISY and GBlocks. The log-likelihood scores of SRk, followed by that of NOISY are higher than that of the original alignment for all datasets. The log-likelihood for SRap is lower only for the least divergent dataset but otherwise trails that of NOISY. Compared to SRap and SRk, the log-likelihoods of all the other methods reduce more steadily with increase in divergence.

Figure 3

Percentage increase in normalized consistency indices of refined alignments with respect to the original alignments (top) and percentage increase in normalized log-likelihoods of refined alignments with respect to the original alignments (bottom). The points and bars represent the mean and standard deviation respectively for 300 of the ISG datasets that simulates highly divergent DNA. The alignments were created using MAFFT and the trees were obtained using RAxML. The abbreviations are – SRap: Signal Refinement using affinity propagation clustering, SRk: Signal Refinement using k-means clustering. Random1 and Random2 are control experiments that eliminate a random set of columns; the number of columns eliminated by Random1 and Random2 is the same as the number eliminated by SRap and SRk respectively. The improvements when SRk is used can be upto 27% increase in CI and from 7 to 35% increase in likelihood, depending on the dataset. Even the lower end of the range is better than the average performance of NOISY, Random1 and Random2. SRap also performs, on average and in most cases, better than NOISY. We also verify that our signal refinement methods give much better results than random elimination of columns.

Figure 4

Normalized consistency indices (top) and normalized log-likelihoods (bottom) of the original alignment and refined alignments from different refinement methods and aligners for the Arthropods dataset containing 46 species. The MAFFT alignment contains 11855 columns and the Clustal alignment contains 11202 columns. The details of the columns removed by each of the aligners is given in table 3. The abbreviations are – SRap: Signal Refinement using affinity propagation clustering, SRk: Signal Refinement using k-means clustering, GA: GBlocks with allowed gap positions set to All, GN: GBlocks with allowed gap positions set to None, GH: GBlocks with allowed gap positions set to Half. Random1 and Random2 are control experiments that eliminate a random set of columns; the number of columns eliminated by Random1 and Random2 is the same as the number eliminated by SRap and SRk respectively. Note that the log-likelihood scale is inverted: the shortest bar has the highest likelihood. The columns selected by SRk have the highest CI. SRap has the second highest CI and is followed by all the other methods that are comparable. With respect to likelihood also, SRk and SRap are better than all the other methods. Also our

signal refinement methods give much better results than random elimination of the same number of columns.

Figure 5

Three pairwise-comparisons of trees with respect to the respective reference trees for Clustal alignments on the ISG datasets that simulates highly divergent DNA. RF distance is used as the distance measure in the top subfigure and matching distance is used in the bottom subfigure. In each group of (three) bars, we compare two sets of trees. In the first group, we compare trees obtained from SR-refined alignments to the trees from the original alignments (SR vs MSA). The second and third groups of bars show similar comparisons for trees from NOISY-refined alignments and original alignments (NOISY vs MSA) and trees from SR-refined alignments and NOISY-refined alignments (SR vs NOISY). For a group, denoted by 'X vs Y', the three bars show:

- the proportion of instances in which trees from X-refined alignments are closer to the reference tree (denoted by "X better")
- the proportion of instances in which both trees from X-refined alignments and the trees from Y-refined alignments are equidistant from the reference trees (denoted by "X equals Y"), and
- the proportion of instances in which trees from the Y-refined alignments are closer to the reference trees (denoted by "Y better").

Abbreviations: SR: Signal Refinement using affinity propagation clustering, MSA: Original alignment with no refinement.

The top subfigure uses RF distances and we see that in 16% of the instances SR-refined trees are better (closer to the reference trees) than the trees from the original alignments, and in 18% of the instances SR-refined trees are better than the trees from NOISY-refined alignments. We also see that only in 2% of the instances trees inferred from NOISY-refined alignments are better than the trees from the original alignment. In the bottom subfigure where matching distance is used, we see that in more than 50% of the instances SR-refined alignments yield better trees than both trees from the original alignments and NOISY-refined alignments and in 88% of the instances the trees are not worse than the trees from the original alignments. NOISY can improve the tree, compared to the tree from the original alignment, only in 12% of the cases.

Figure 6

Three pairwise-comparisons of trees with respect to the respective reference trees for MAFFT alignments on the ISG datasets that simulates highly divergent DNA. RF distance is used as the distance measure in the top subfigure and matching distance is used in the bottom subfigure. In each group of (three) bars, we compare two sets of trees. In the first group, we compare trees obtained from SR-refined alignments to the trees from the original alignments (SR vs MSA). The second and third groups of bars show similar comparisons for trees from NOISY-refined alignments and original alignments (NOISY vs MSA) and trees

from SR-refined alignments and NOISY-refined alignments (SR vs NOISY). For a group, denoted by 'X vs Y', the three bars show:

- the proportion of instances in which trees from X-refined alignments are closer to the reference tree (denoted by "X better")
- the proportion of instances in which both trees from X-refined alignments and the trees from Y-refined alignments are equidistant from the reference trees (denoted by "X equals Y"), and
- the proportion of instances in which trees from the Y-refined alignments are closer to the reference trees (denoted by "Y better").

Abbreviations: SR: Signal Refinement using affinity propagation clustering, MSA: Original alignment with no refinement.

With respect to RF distance (in the top subfigure) we see that in 16% of the instances SR-refined trees are better than the trees from the original alignments and in 76% of the instances SR-refined trees are not worse than trees from the original alignments. The bottom subfigure (using matching distance) shows that 68% of the instances show improvement in the inferred tree due to SR-based refinement. Compared to NOISY-refined trees, SR-refined trees are better in 75% of the instances. In both cases, we see that NOISY does not match the performance of SR.

Figure 7

Three-way comparison of inferred trees with respect to the reference trees for ISG datasets (that simulates highly divergent DNA) on MAFFT and Clustal alignments. RF distance is used as the distance measure in the top subfigure and matching distance is used in the bottom subfigure. Each bar shows the proportion of instances (as percentage) where the trees obtained were closer to the reference tree than trees obtained from other methods. Let d_{SR}, d_N and d_{MSA} denote the distances of the inferred trees, from an SR-refined alignment, NOISY-refined alignment and the original alignment respectively, with respect to the reference tree. Each group of bars shows the number of instances where $d_{SR} < d_N$ *and* $d_{SR} < d_{MSA}$ (first bar), $d_N < d_{SR}$ *and* $d_N < d_{MSA}$ (second bar), and $d_{MSA} < d_{SR}$ *and* $d_{MSA} < d_N$ (third bar). Abbreviations: SR: Signal Refinement using affinity propagation clustering, MSA: Original alignment with no refinement

In the first group of bars (for MAFFT alignments) in the top subfigure (results where RF distance is used) we see that in 16% of the instances trees from SR-refined alignments are better than *both* trees from NOISY-refined and the original alignments. The corresponding counts when matching distance is used, is shown in the bottom subfigure: in 64% (45%) of the instances trees from SR-refined alignments were better than *both* trees from NOISY-refined and the original MAFFT (Clustal) alignments.

Figure 8

Three-way comparison of inferred trees with respect to the reference trees for ISG datasets (that simulates highly divergent DNA) on MAFFT and Clustal

alignments. RF distance is used as the distance measure in the top subfigure and matching distance is used in the bottom subfigure. Each bar shows the proportion of instances (as percentage) where the trees obtained were closer to or equidistant to the reference tree than trees obtained from other methods. Let d_{SR}, d_N and d_{MSA} denote the distances of the inferred trees, from an SR-refined alignment, NOISY-refined alignment and the original alignment respectively, with respect to the reference tree. Each group of bars shows the number of instances where $d_{SR} \leq d_N$ and $d_{SR} \leq d_{MSA}$ (first bar), $d_N \leq d_{SR}$ and $d_N \leq d_{MSA}$ (second bar), and $d_{MSA} \leq d_{SR}$ and $d_{MSA} \leq d_N$ (third bar). Abbreviations: SR: Signal Refinement using affinity propagation clustering, MSA: Original alignment with no refinement

We see that SR-refined trees obtain the best tree (i.e., closest to the reference tree among the three inferred trees) from both Clustal and MAFFT alignments in more than 95% of the instances when RF distance is used as a metric and in more than 78% of the instances when matching distance is used as a metric.

Figure 9

Comparison of RF (top subfigure) and matching distance (bottom subfigure) of the inferred tree before and after refining the alignment with SR. Each point, corresponding to a single dataset, is a pair (x, y) where x is the distance of the tree inferred from an SR-refined MAFFT alignment and y is the distance of the tree inferred from the (unrefined) MAFFT alignment (both distances are from the true tree). There are 300 datasets in total.

The diagonal shows the points where x and y are equal: both the inferred trees are equidistant from the true tree. The points above the diagonal are those where $(x < y)$: the tree inferred from SR-refined alignment is closer to the true tree than the tree inferred from the original alignment. The points below the diagonal are those where $(y < x)$: the tree inferred from SR-refined tree is farther from the true tree than the tree inferred from the original alignment.

Majority of the points lie above the diagonal. The exact count statistics are shown in the previous plots. The scatterplot clarifies both the distribution and extent of improvement due to refinement.

Figure 10

Comparison of RF (top subfigure) and matching distance (bottom subfigure) of the inferred tree before and after refining the alignment with NOISY. Each point, corresponding to a single dataset, is a pair (x, y) where x is the distance of the tree inferred from a NOISY-refined MAFFT alignment and y is the distance of the tree inferred from the MAFFT alignment (both distances are from the true tree). There are 300 datasets in total.

The diagonal shows the points where x and y are equal: both the inferred trees are equidistant from the true tree. The points above the diagonal are those where $(x < y)$: the tree inferred from NOISY-refined alignment is closer to the true tree than the tree inferred from the original alignment. The points below the diagonal are those where $(y < x)$: the tree inferred from NOISY-refined alignment is farther from the true tree than the tree inferred from the original

alignment.

The exact count statistics are shown in the previous plots. The scatterplot clarifies both the distribution and extent of improvement due to refinement.

Figure 11

Comparison of RF (top subfigure) and matching distance (bottom subfigure) of the inferred tree after refining with SR and NOISY. Each point, corresponding to a single dataset, is a pair (x, y) where x is the distance of the tree inferred from an SR-refined MAFFT alignment and y is the distance of the tree inferred from the NOISY-refined MAFFT alignment (both distances are from the true tree). There are 300 datasets in total.

The diagonal shows the points where x and y are equal: both the inferred trees are equidistant from the true tree. The points above the diagonal are those where $(x < y)$: the tree inferred from SR-refined alignment is closer to the true tree than the tree inferred from NOISY-refined alignment. The points below the diagonal are those where $(y < x)$: the tree inferred from SR-refined alignment is farther from the true tree than the tree inferred from NOISY-refined alignment.

Majority of the points lie above the diagonal. The exact count statistics are shown in the previous plots. The scatterplot clarifies both the distribution and extent of improvement due to refinement.

Figure 12

Comparison of RF (top subfigure) and matching distance (bottom subfigure) of the inferred tree before and after refining the alignment with SR. Each point, corresponding to a single dataset, is a pair (x, y) where x is the distance of the tree inferred from an SR-refined Clustal alignment and y is the distance of the tree inferred from the (unrefined) Clustal alignment (both distances are from the true tree). There are 300 datasets in total.

The diagonal shows the points where x and y are equal: both the inferred trees are equidistant from the true tree. The points above the diagonal are those where $(x < y)$: the tree inferred from SR-refined alignment is closer to the true tree than the tree inferred from the original alignment. The points below the diagonal are those where $(y < x)$: the tree inferred from SR-refined tree is farther from the true tree than the tree inferred from the original alignment.

Majority of the points lie above the diagonal. The exact count statistics are shown in the previous plots. The scatterplot clarifies both the distribution and extent of improvement due to refinement.

Figure 13

Comparison of RF (top subfigure) and matching distance (bottom subfigure) of the inferred tree before and after refining the alignment with NOISY. Each point, corresponding to a single dataset, is a pair (x, y) where x is the distance of the tree inferred from a NOISY-refined Clustal alignment and y is the distance of the tree inferred from the Clustal alignment (both distances are from the true tree). There are 300 datasets in total.

The diagonal shows the points where x and y are equal: both the inferred trees are equidistant from the true tree. The points above the diagonal are those where $(x < y)$: the tree inferred from NOISY-refined alignment is closer to the true tree than the tree inferred from the original alignment. The points below the diagonal are those where $(y < x)$: the tree inferred from NOISY-refined alignment is farther from the true tree than the tree inferred from the original alignment.

The exact count statistics are shown in the previous plots. The scatterplot clarifies both the distribution and extent of improvement due to refinement.

Figure 14

Comparison of RF (top subfigure) and matching distance (bottom subfigure) of the inferred tree after refining with SR and NOISY. Each point, corresponding to a single dataset, is a pair (x, y) where x is the distance of the tree inferred from an SR-refined Clustal alignment and y is the distance of the tree inferred from the NOISY-refined Clustal alignment (both distances are from the true tree). There are 300 datasets in total.

The diagonal shows the points where x and y are equal: both the inferred trees are equidistant from the true tree. The points above the diagonal are those where $(x < y)$: the tree inferred from SR-refined alignment is closer to the true tree than the tree inferred from NOISY-refined alignment. The points below the diagonal are those where $(y < x)$: the tree inferred from SR-refined alignment is farther from the true tree than the tree inferred from NOISY-refined alignment.

Majority of the points lie above the diagonal. The exact count statistics are shown in the previous plots. The scatterplot clarifies both the distribution and extent of improvement due to refinement.

Figure 15

Comparison of the RF (top subfigure) and matching distances (bottom subfigure) of the inferred trees from fourteen different alignments with respect to the reference tree for the Eukaryotes dataset containing 117 ribosomal RNA sequences. These alignments include the original MAFFT and Clustal alignments and refined alignments from six different methods.

Using GBLOCKS or ALISCORE does not improve the final phylogeny in any case. NOISY and SR, both yield small improvements in the inferred tree when MAFFT is used for aligning the sequences and the improvement is more with SR. Clustal alignments appeared to be less affected by NOISY and SR-refined trees are closer to the reference tree by a small margin.

Figure 16

Comparison of the RF (top subfigure) and matching distances (bottom subfigure) of the inferred trees from fourteen different alignments with respect to the reference tree for the Arthropods dataset containing mitochondrial protein-coding genes from 46 species of Arthropods. These alignments include the orig-

inal MAFFT and Clustal alignments and refined alignments from six different methods.

Since ALISCORE removes very few columns during refinement (see table 3), the tree inferred from ALISCORE-refined alignment is very similar to the tree from the original alignment. Masking using GBLOCKS does not improve the inferred tree. For the MAFFT alignment, masking does not improve the quality of the inferred tree with any of the masking methods. Only NOISY and SR show improvements in the final tree for the Clustal alignments, and the tree from the SR-refined alignment is the closest to the reference tree. The trees obtained from the original Clustal alignment and after masking using SR are shown in figures 12 and 13 respectively. Figure 11 shows the reference tree we used.

Figure 17

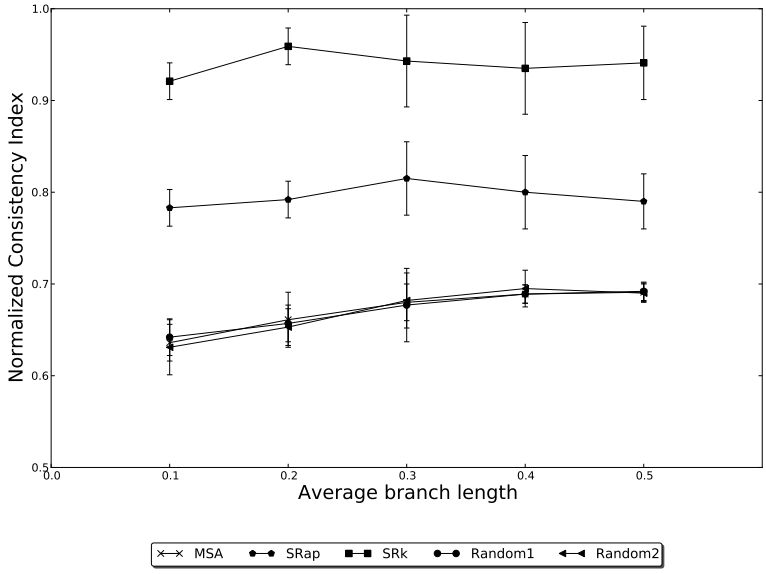
Reference phylogeny of 46 species of Arthropods. For our reference phylogeny we use the taxonomic phylogeny obtained from NCBI Nucleotide database for each species used. The taxonomic classification does not give us a fully resolved binary tree. We resolve polytomies and also corroborate the phylogeny with the phylogenies reported in several references (see text for details). Beside the genus and species names of each taxon we also show the subphylum and class names.

Figure 18

Phylogeny of 46 species of Arthropods inferred using RAxML from a Clustal alignment of mitochondrial protein-coding genes. Beside the genus and species names of each taxon we also show the subphylum and class names.

Figure 19

Phylogeny of 46 species of Arthropods inferred using RAxML from an SR-refined Clustal alignment of mitochondrial protein-coding genes. Beside the genus and species names of each taxon we also show the subphylum and class names. This tree, obtained from the refined alignment, is closer to the reference tree than the tree from the original alignment shown in figure 12.


Normalized Consistency Index
Average branch length
MSA
SRap
SRk
Random1
Random2

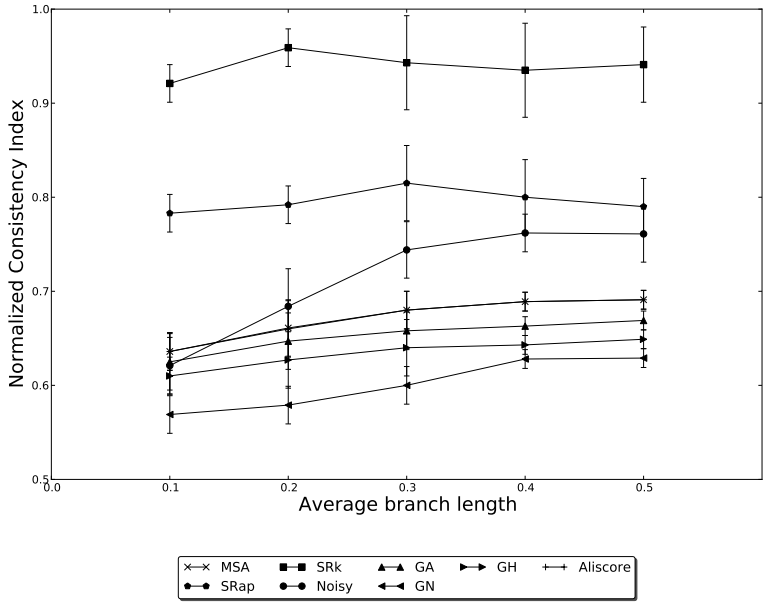
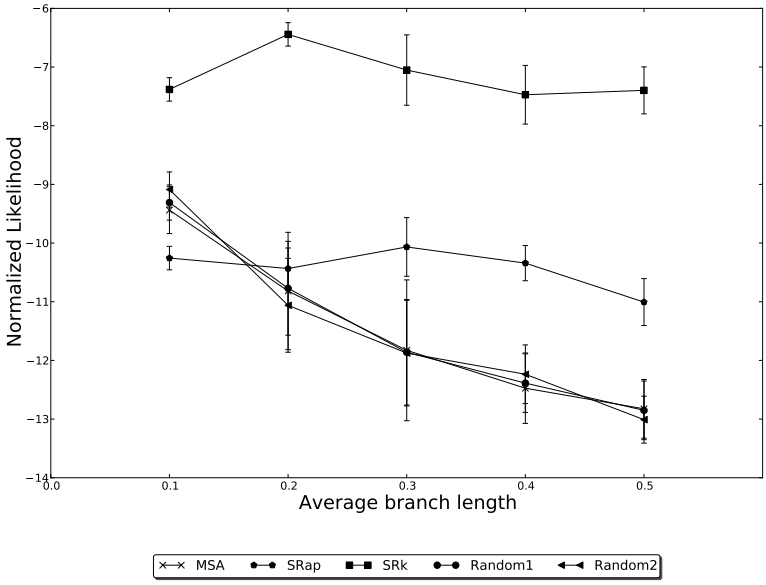

Normalized Consistency Index
Average branch length
MSA
SRap
SRk
Noisy
GA
GN
GH
Aliscore

Figure 1


Normalized Likelihood
Average branch length
MSA
SRap
SRk
Random1
Random2

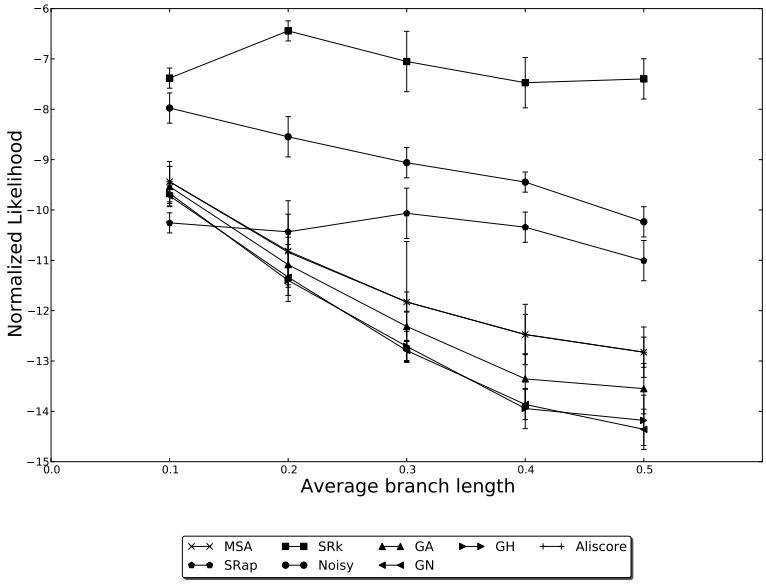
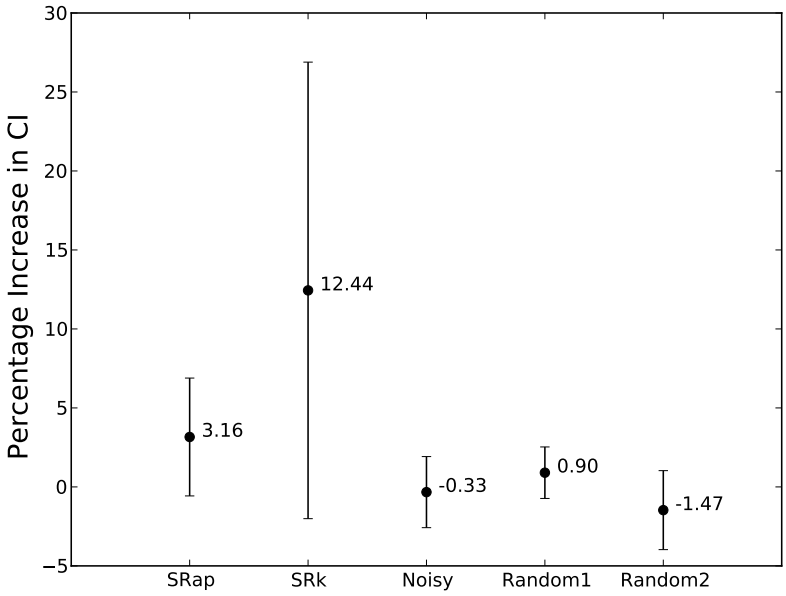

Normalized Likelihood
Average branch length
MSA
SRap
SRk
Noisy
GA
GN
GH
Aliscore

Figure 2


Percentage Increase in CI
30
25
20
15
10
5
0
−5
12.44
3.16
-0.33
0.90
-1.47
SRap
SRk
Noisy
Random1
Random2

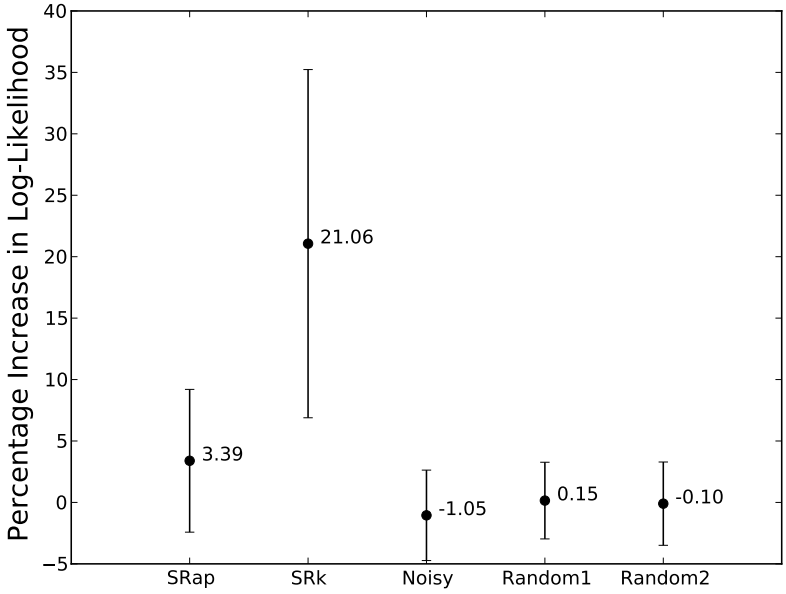
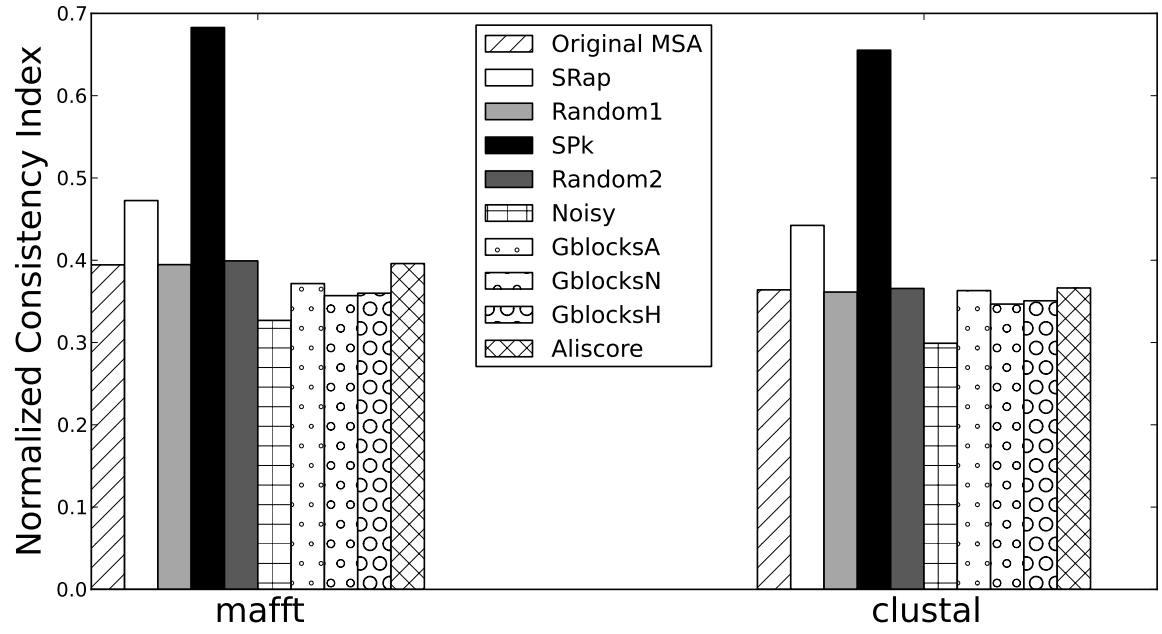

Percentage Increase in Log-Likelihood
40
35
30
25
20
15
10
5
0
−5
21.06
3.39
-1.05
0.15
-0.10
SRap
SRk
Noisy
Random1
Random2

Figure 3


Normalized Consistency Index
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Original MSA
SRap
Random1
SPk
Random2
Noisy
GblocksA
GblocksN
GblocksH
Aliscore
mafft
clustal

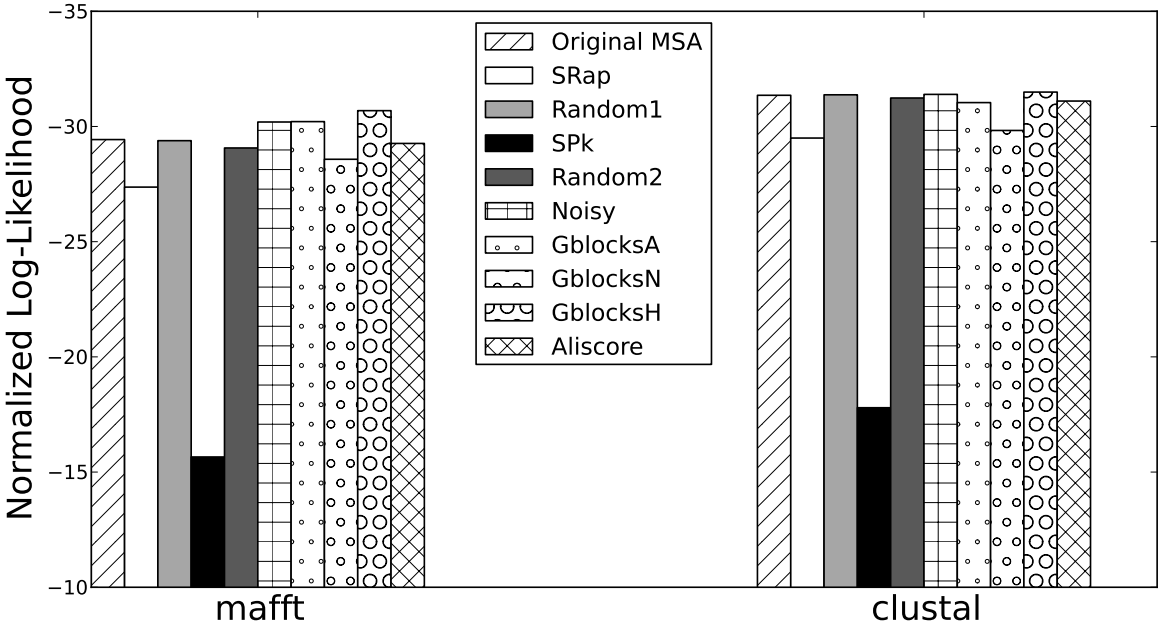
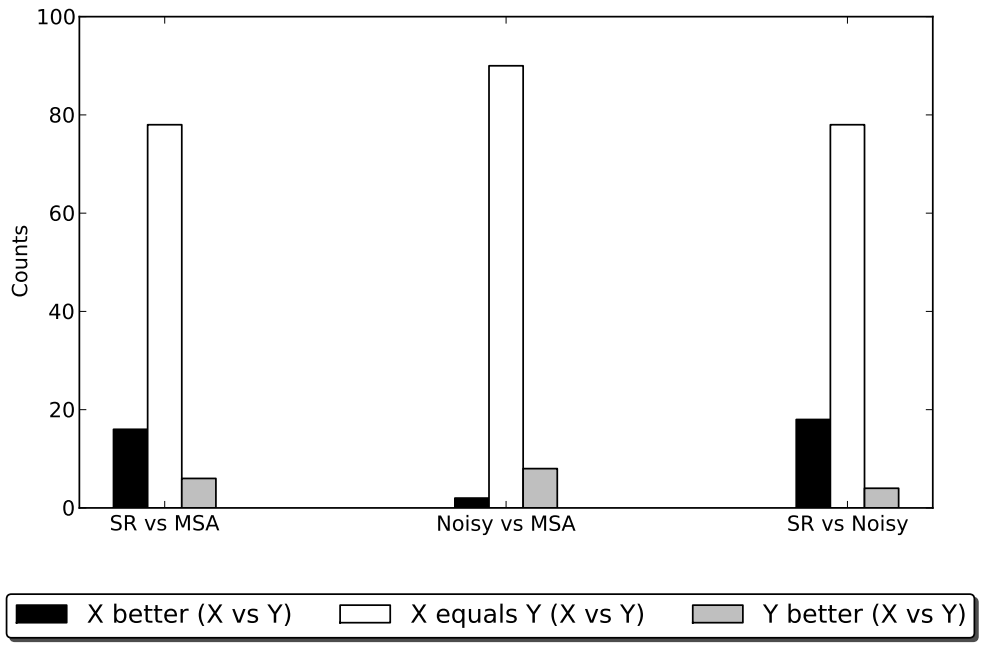

Normalized Log-Likelihood
−10
−15
−20
−25
−30
−35
Original MSA
SRap
Random1
SPk
Random2
Noisy
GblocksA
GblocksN
GblocksH
Aliscore
mafft
clustal

Figure 4


100
80
60
40
20
0
Counts
SR vs MSA
Noisy vs MSA
SR vs Noisy
X better (X vs Y)
X equals Y (X vs Y)
Y better (X vs Y)

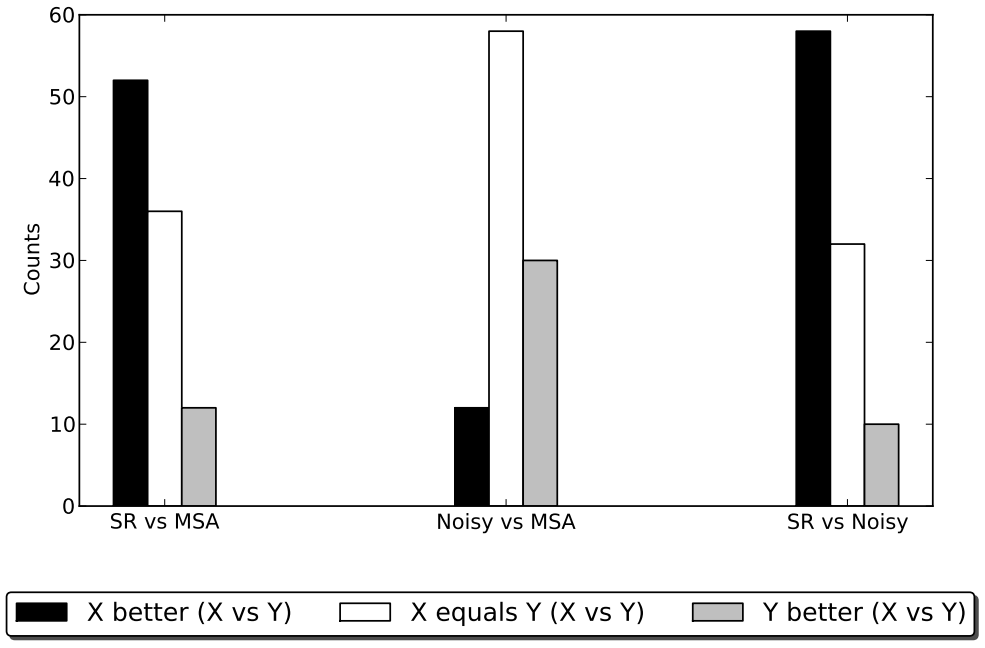
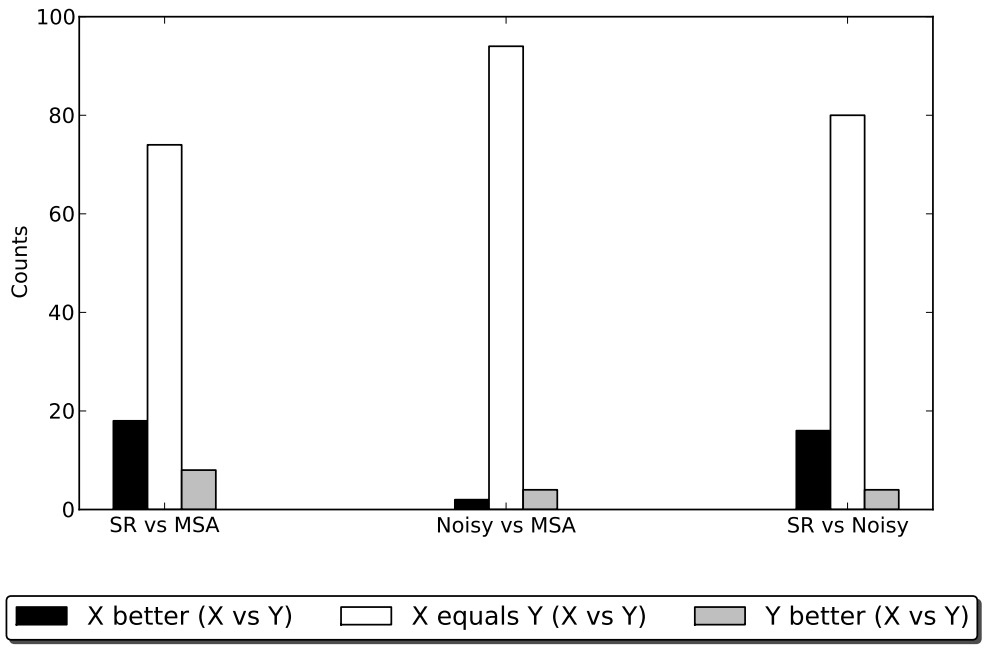

60
50
40
30
20
10
0
Counts
SR vs MSA
Noisy vs MSA
SR vs Noisy
X better (X vs Y)
X equals Y (X vs Y)
Y better (X vs Y)

Figure 5


100
80
60
40
20
0
Counts
SR vs MSA
Noisy vs MSA
SR vs Noisy
X better (X vs Y)
X equals Y (X vs Y)
Y better (X vs Y)

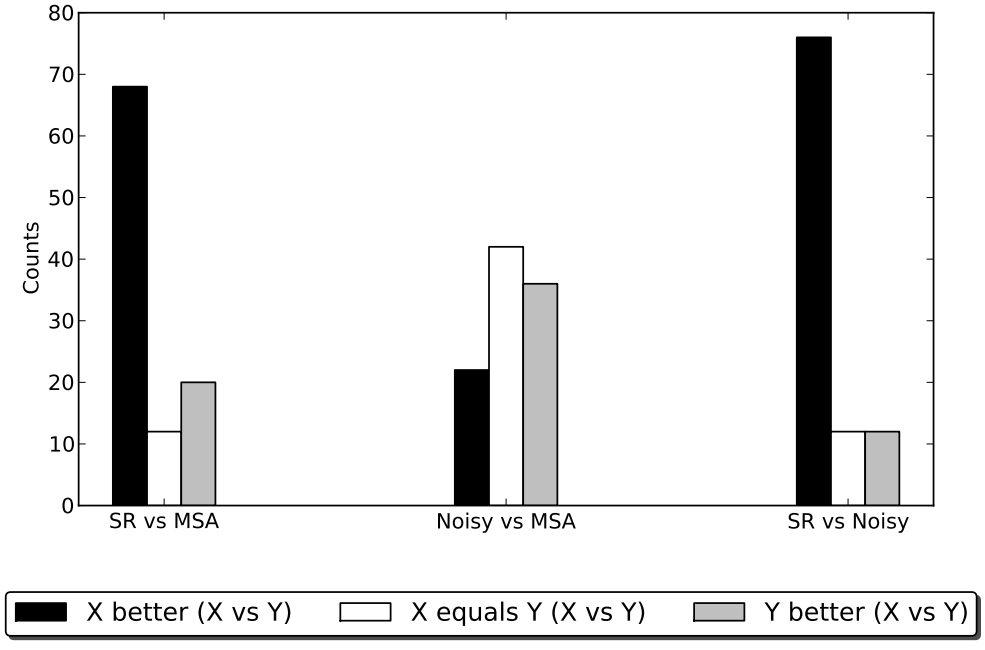
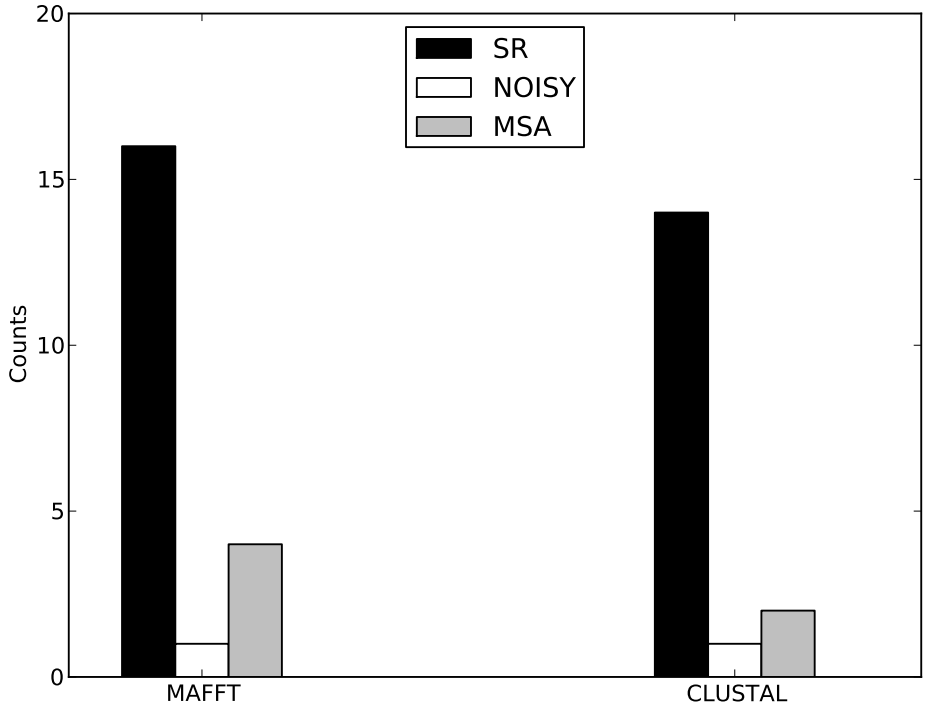

80
70
60
50
40
30
20
10
0
Counts
SR vs MSA
Noisy vs MSA
SR vs Noisy
X better (X vs Y)
X equals Y (X vs Y)
Y better (X vs Y)

Figure 6


SR
NOISY
MSA
Counts
0
5
10
15
20
MAFFT
CLUSTAL

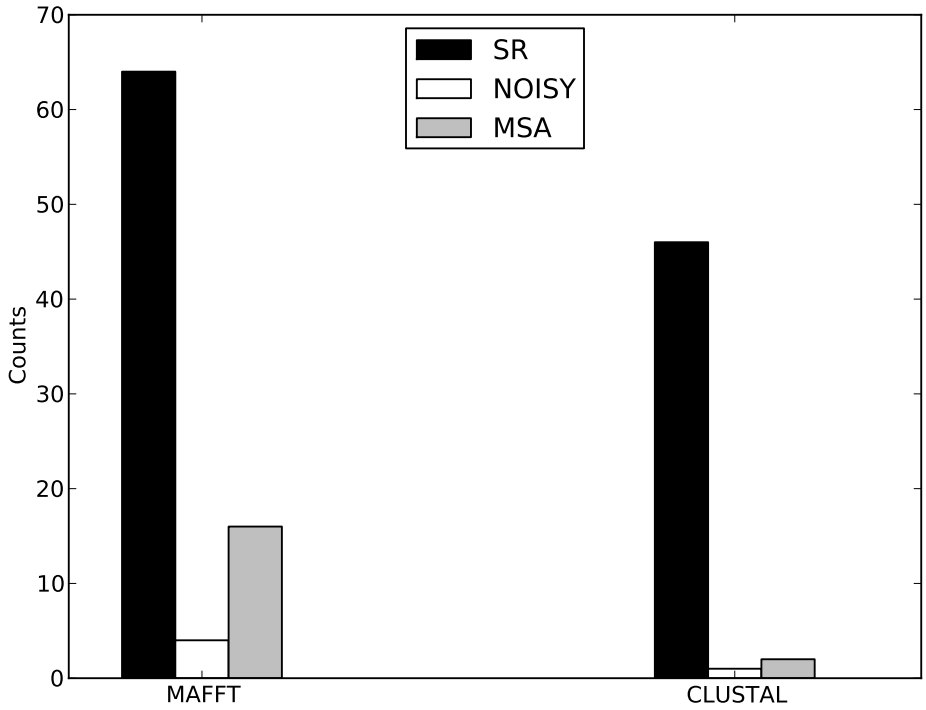
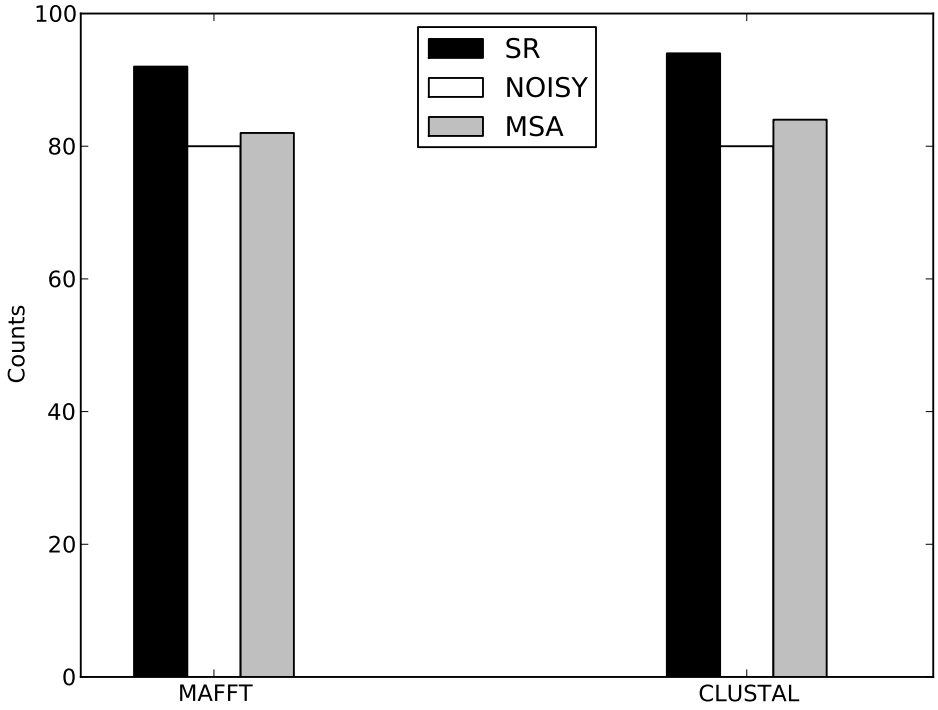

SR
NOISY
MSA
Counts
0
10
20
30
40
50
60
70
MAFFT
CLUSTAL

Figure 7


100
80
60
40
20
0
Counts
SR
NOISY
MSA
MAFFT
CLUSTAL

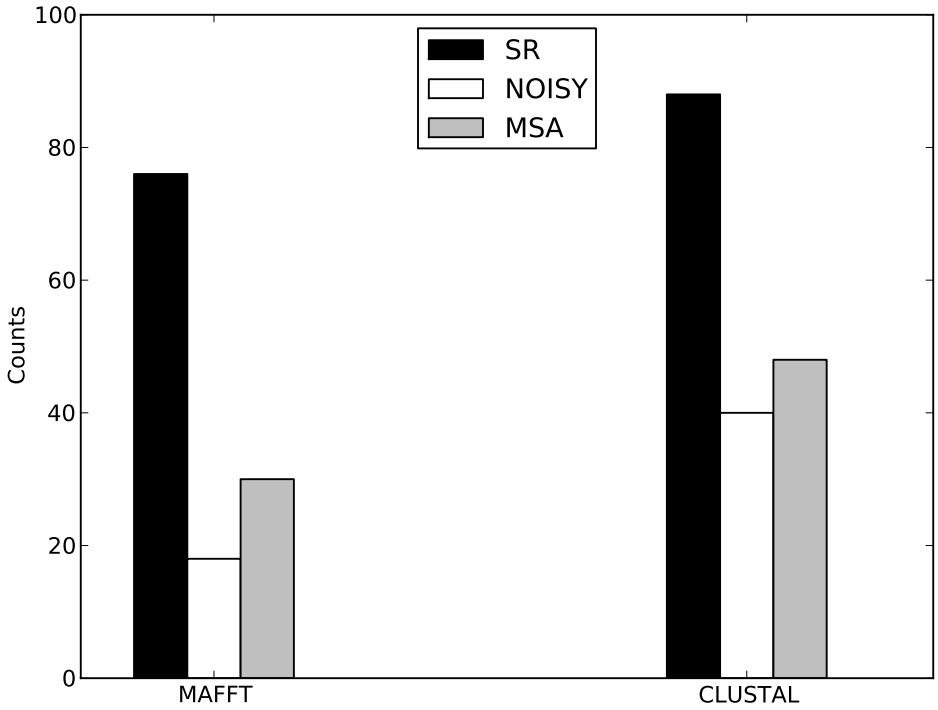
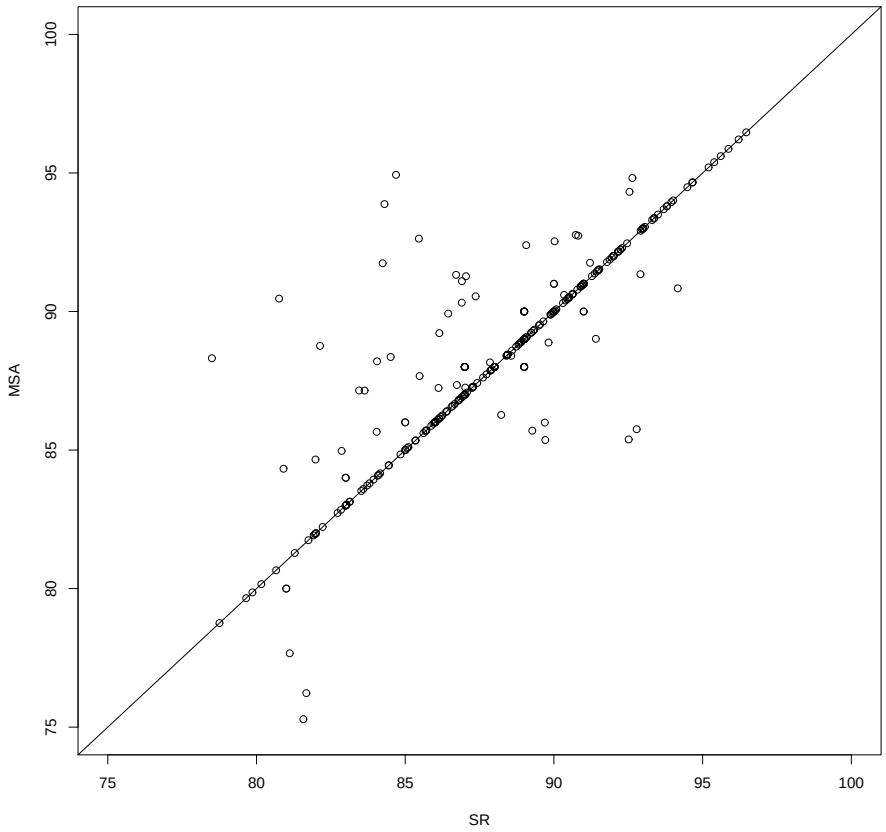

100
80
60
40
20
0
Counts
SR
NOISY
MSA
MAFFT
CLUSTAL

Figure 8



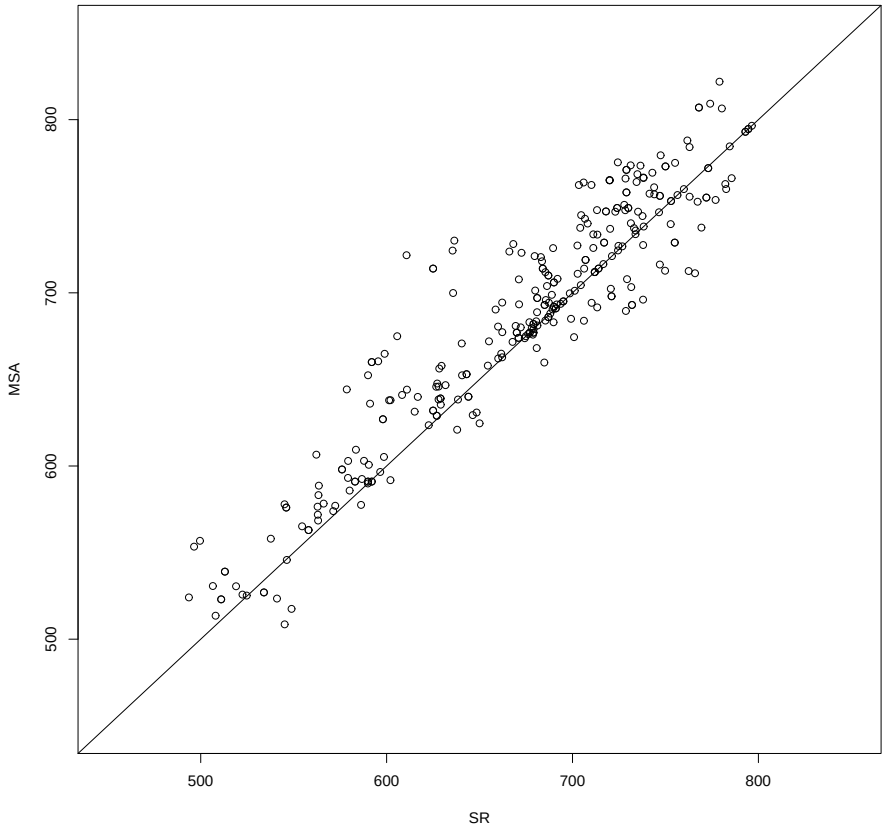
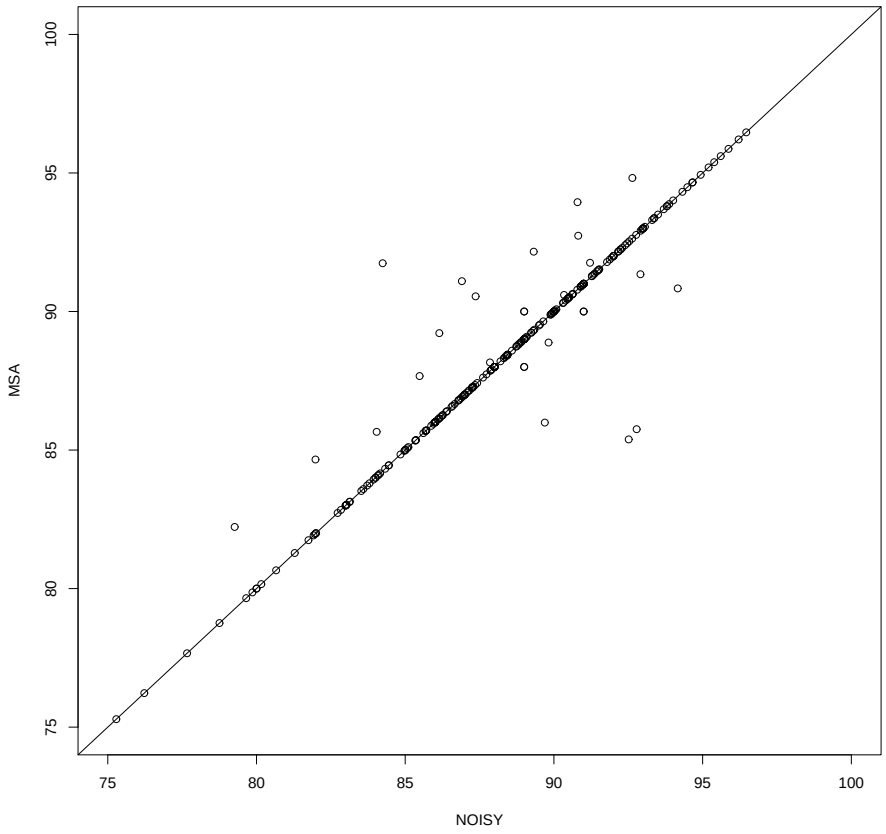


Figure 9



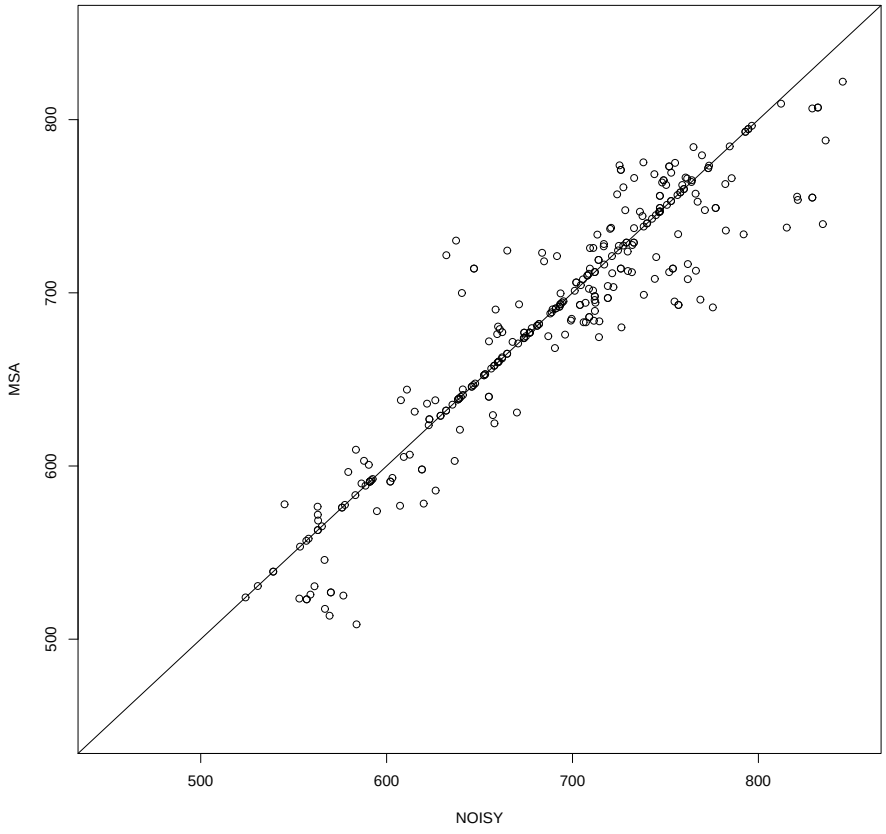
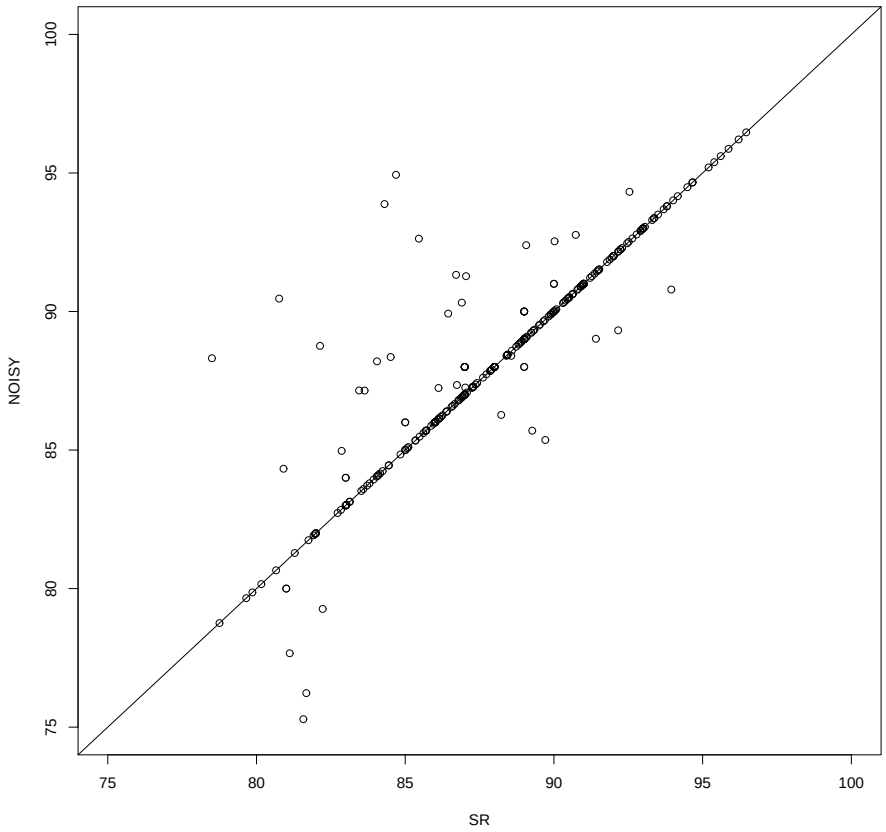


Figure 10



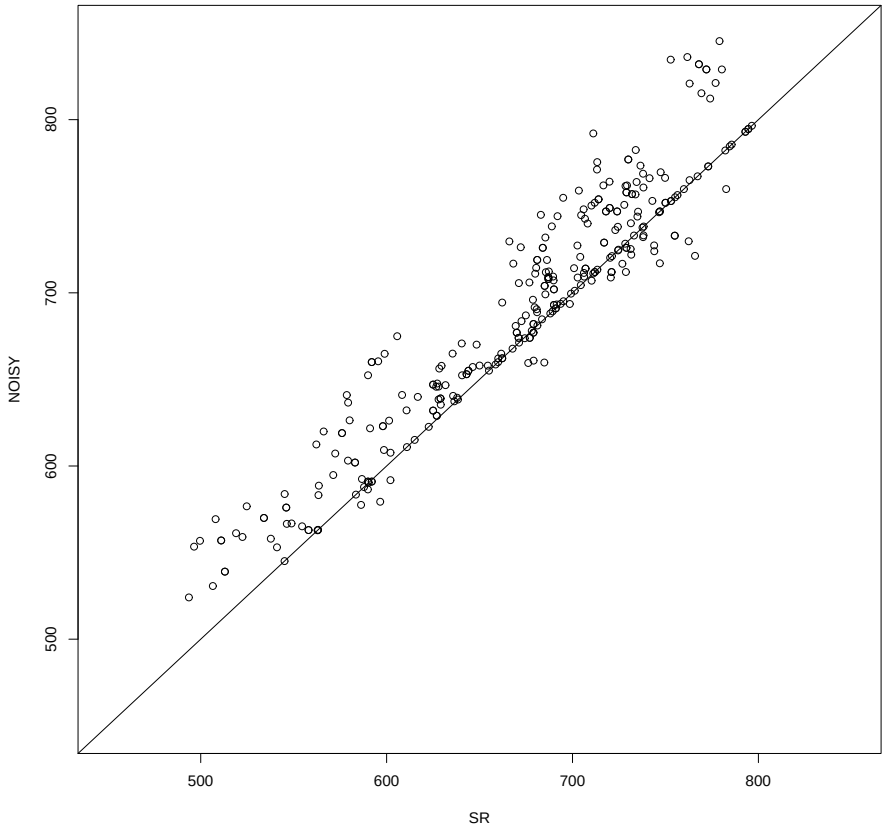
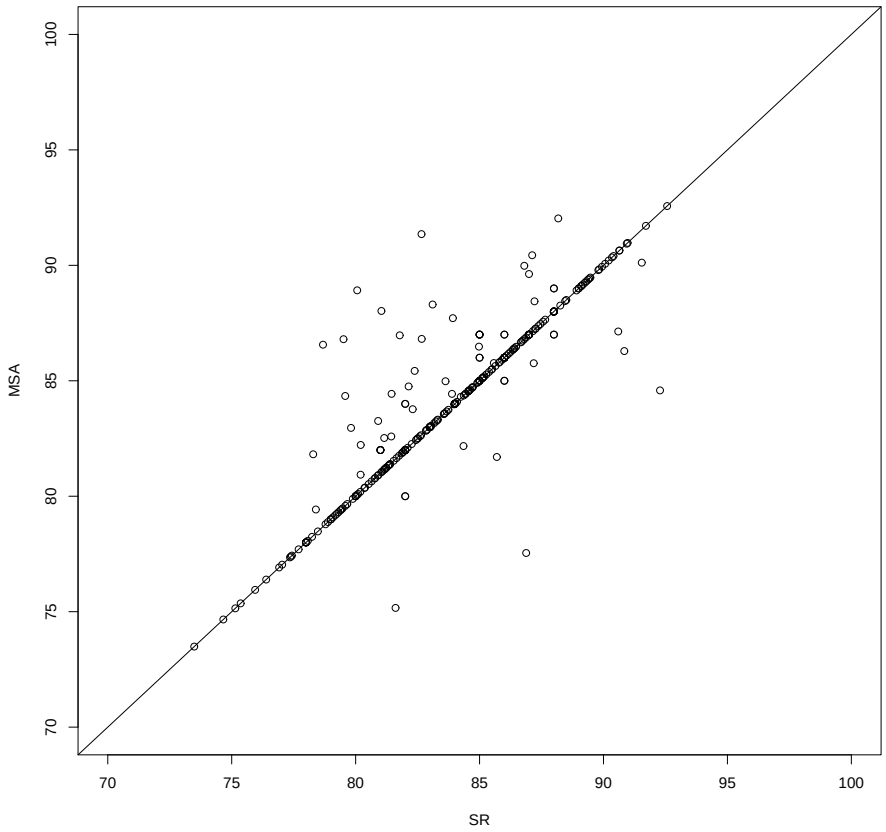


Figure 11



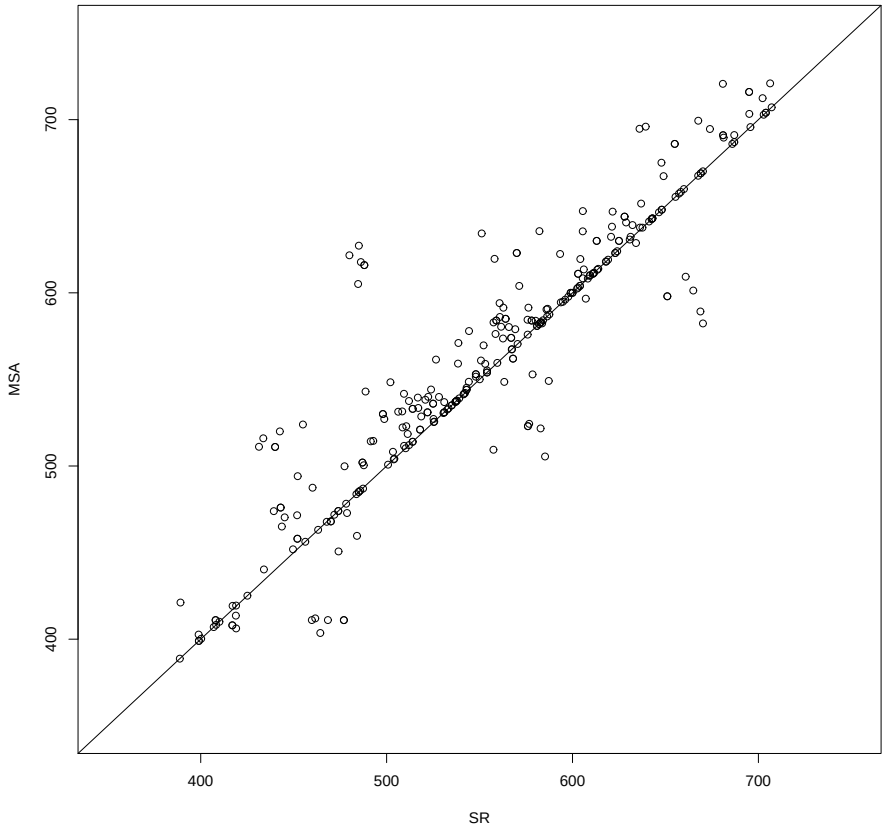
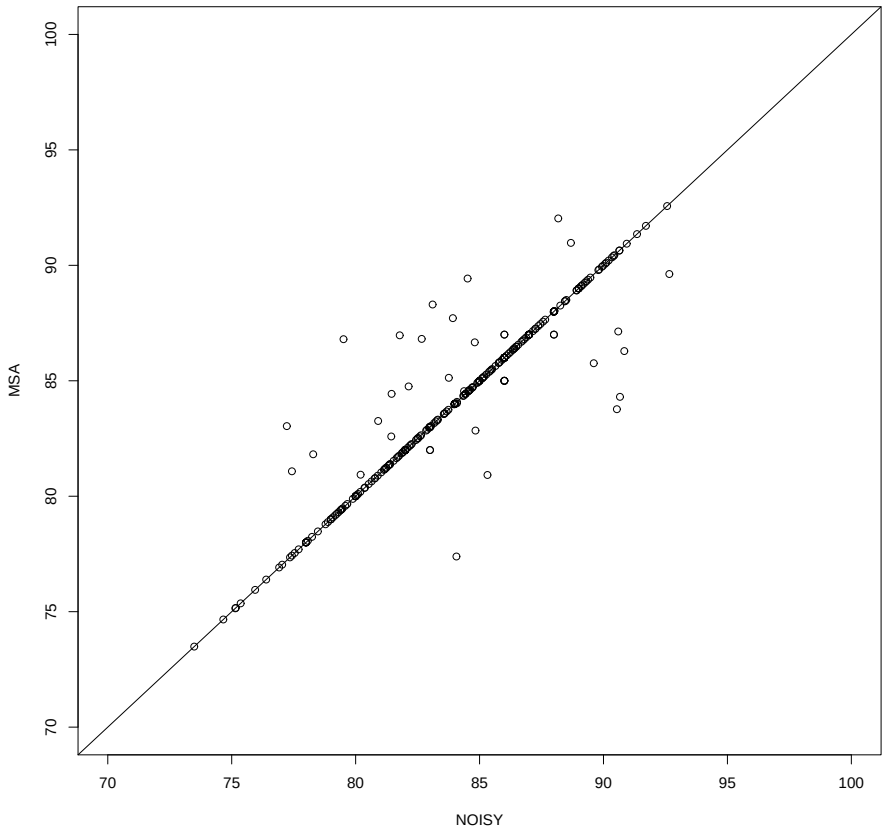


Figure 12



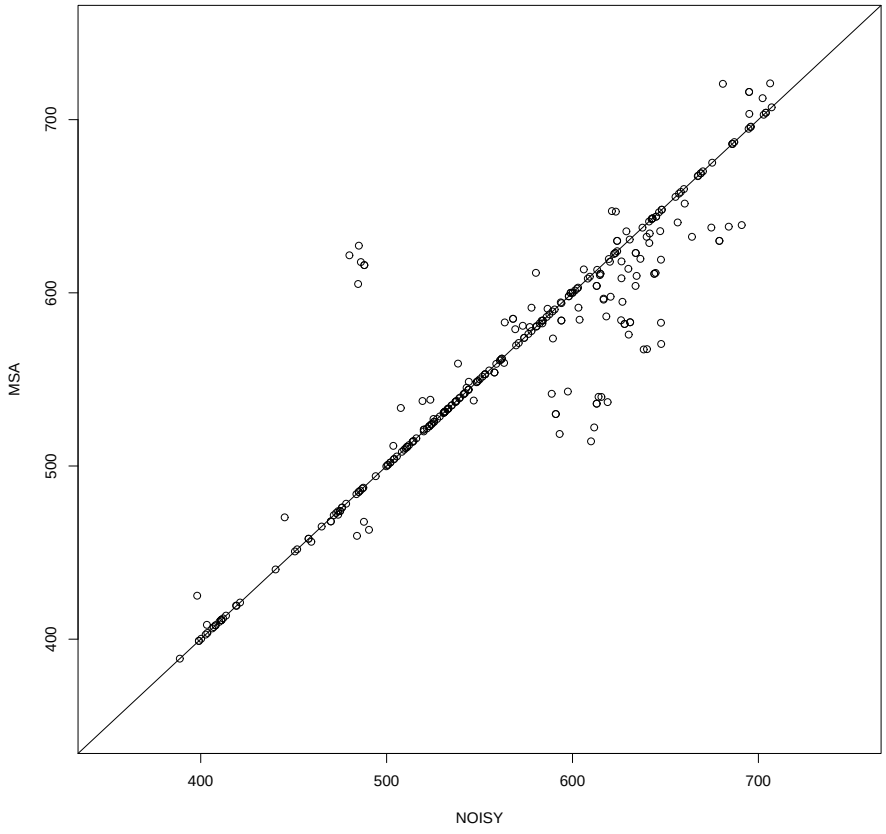
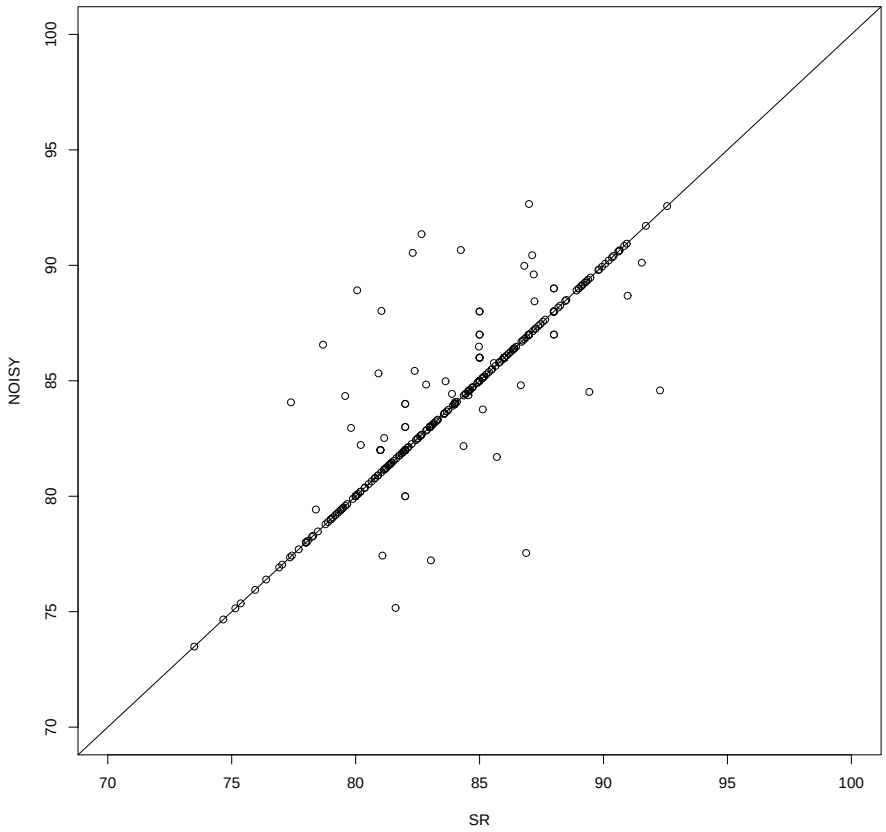


Figure 13



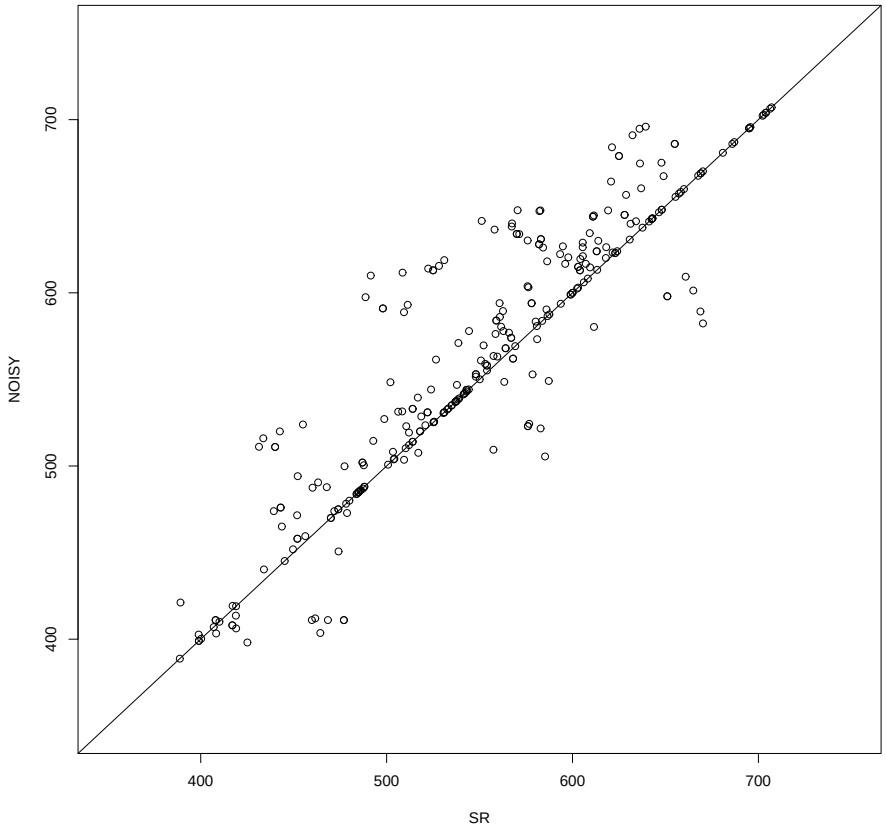


Figure 14

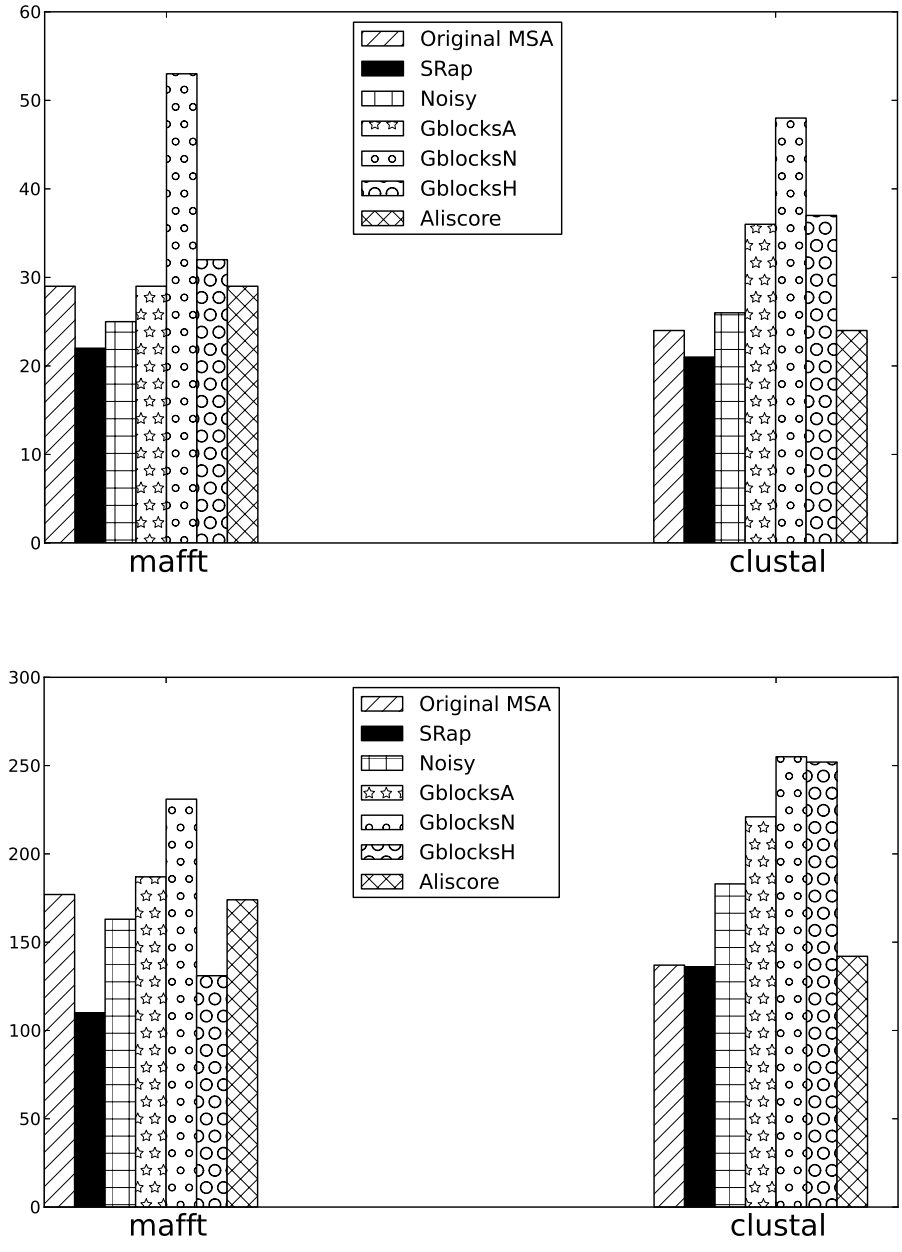

Original MSA
SRap
Noisy
GblocksA
GblocksN
GblocksH
Aliscore
0
10
20
30
40
50
60
mafft
clustal
0
50
100
150
200
250
300
mafft
clustal

Figure 15

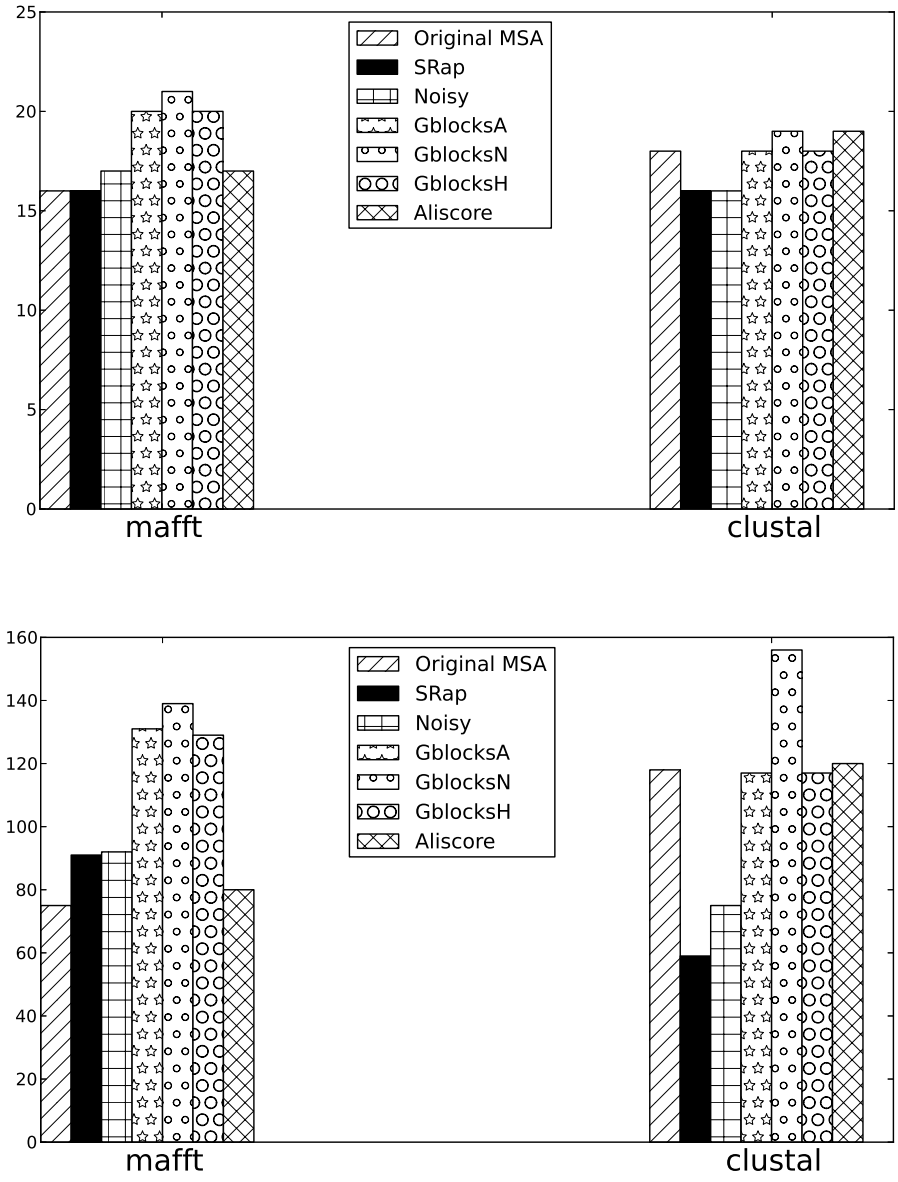

25
20
15
10
5
0
Original MSA
SRap
Noisy
GblocksA
GblocksN
GblocksH
Aliscore
mafft
clustal
160
140
120
100
80
60
40
20
0
Original MSA
SRap
Noisy
GblocksA
GblocksN
GblocksH
Aliscore
mafft
clustal

Figure 16

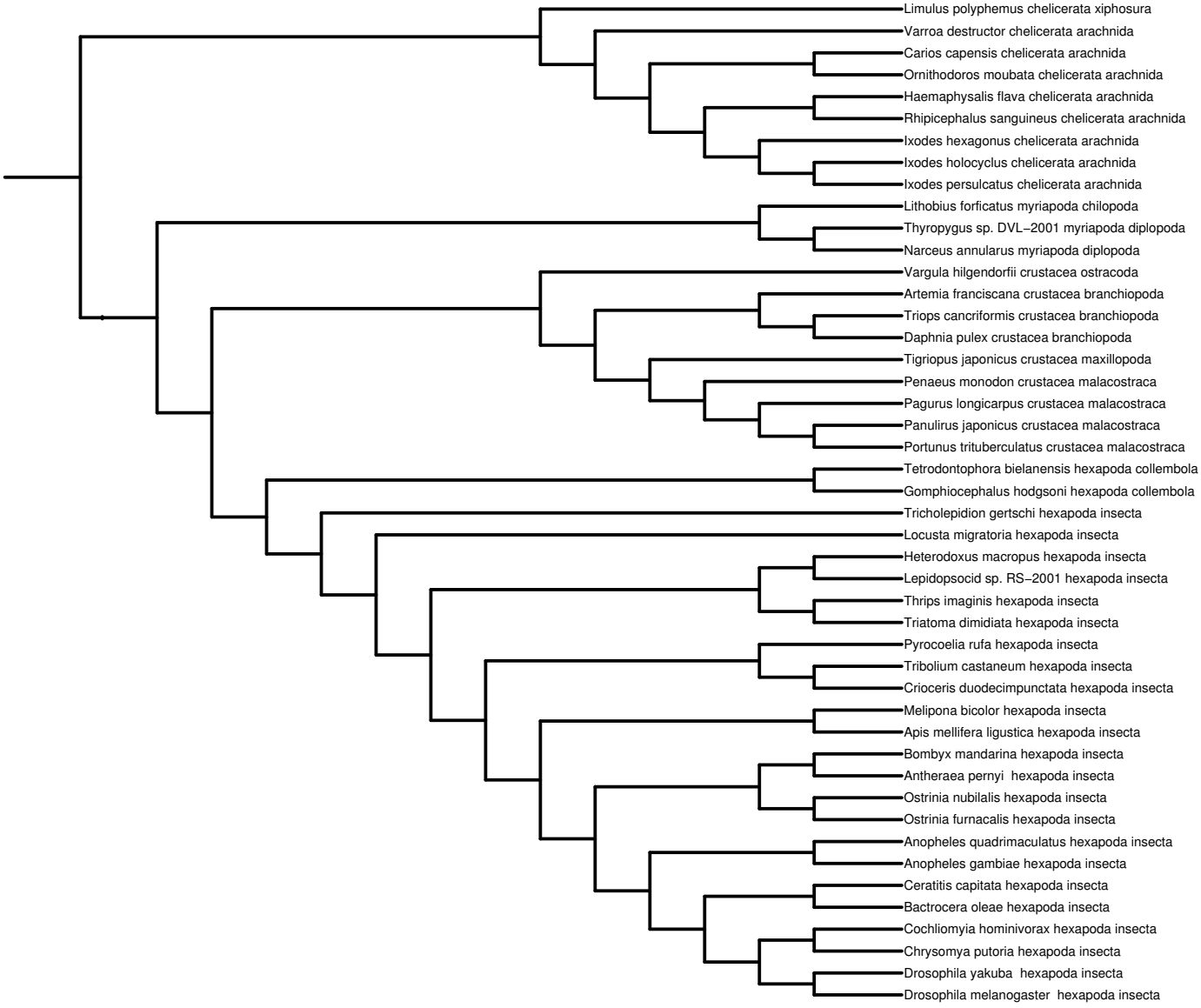

Limulus polyphemus chelicerata xiphosura
Varroa destructor chelicerata arachnida
Carios capensis chelicerata arachnida
Ornithodoros moubata chelicerata arachnida
Haemaphysalis flava chelicerata arachnida
Rhipicephalus sanguineus chelicerata arachnida
Ixodes hexagonus chelicerata arachnida
Ixodes holocyclus chelicerata arachnida
Ixodes persulcatus chelicerata arachnida
Lithobius forficatus myriapoda chilopoda
Thyropygus sp. DVL−2001 myriapoda diplopoda
Narceus annularus myriapoda diplopoda
Vargula hilgendorfii crustacea ostracoda
Artemia franciscana crustacea branchiopoda
Triops cancriformis crustacea branchiopoda
Daphnia pulex crustacea branchiopoda
Tigriopus japonicus crustacea maxillopoda
Penaeus monodon crustacea malacostraca
Pagurus longicarpus crustacea malacostraca
Panulirus japonicus crustacea malacostraca
Portunus trituberculatus crustacea malacostraca
Tetrodontophora bielanensis hexapoda collembola
Gomphiocephalus hodgsoni hexapoda collembola
Tricholepidion gertschi hexapoda insecta
Locusta migratoria hexapoda insecta
Heterodoxus macropus hexapoda insecta
Lepidopsocid sp. RS−2001 hexapoda insecta
Thrips imaginis hexapoda insecta
Triatoma dimidiata hexapoda insecta
Pyrocoelia rufa hexapoda insecta
Tribolium castaneum hexapoda insecta
Crioceris duodecimpunctata hexapoda insecta
Melipona bicolor hexapoda insecta
Apis mellifera ligustica hexapoda insecta
Bombyx mandarina hexapoda insecta
Antheraea pernyi hexapoda insecta
Ostrinia nubilalis hexapoda insecta
Ostrinia furnacalis hexapoda insecta
Anopheles quadrimaculatus hexapoda insecta
Anopheles gambiae hexapoda insecta
Ceratitis capitata hexapoda insecta
Bactrocera oleae hexapoda insecta
Cochliomyia hominivorax hexapoda insecta
Chrysomya putoria hexapoda insecta
Drosophila yakuba hexapoda insecta
Drosophila melanogaster hexapoda insecta

Figure 17

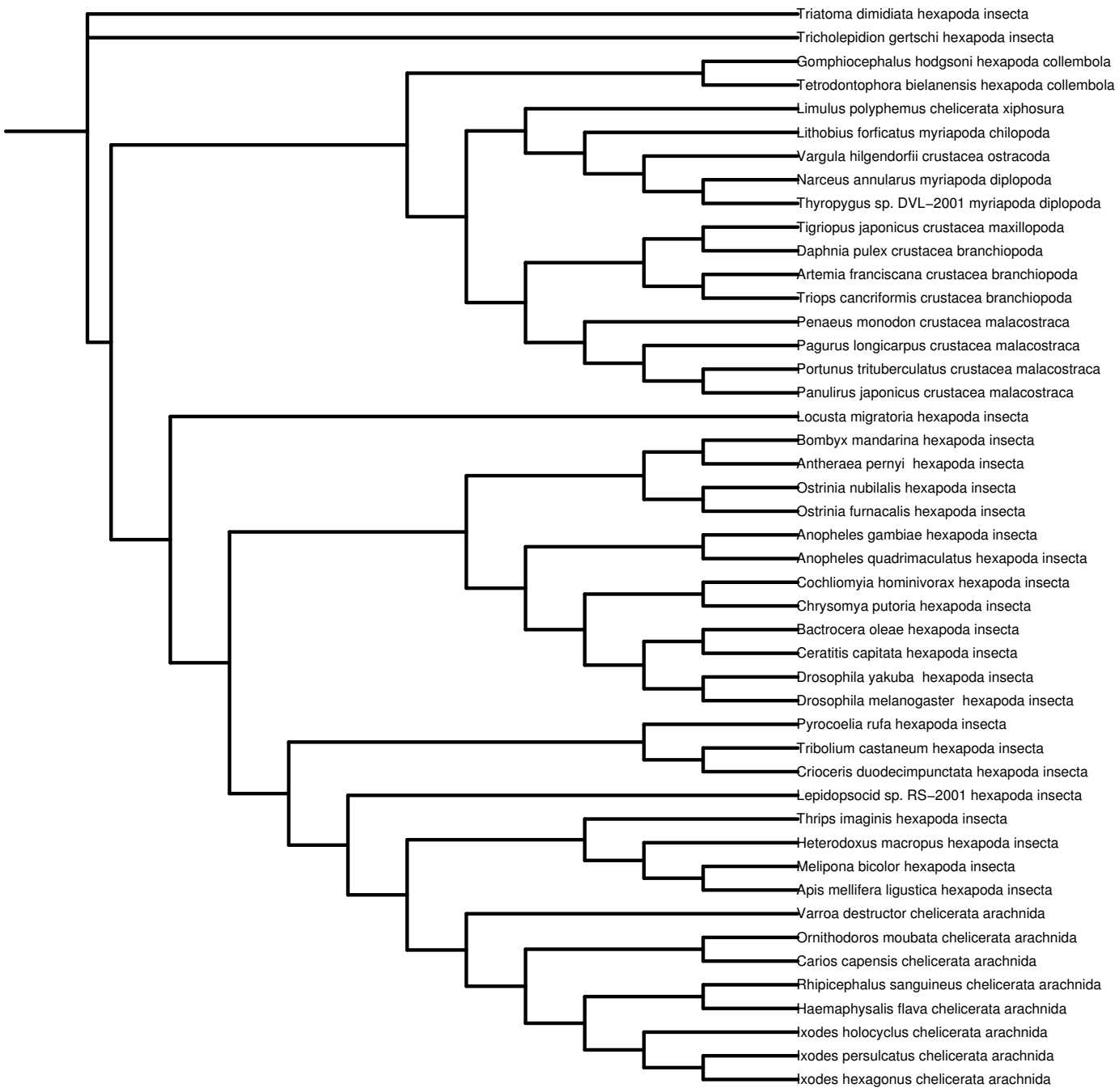

Triatoma dimidiata hexapoda insecta
Tricholepidion gertschi hexapoda insecta
Gomphiocephalus hodgsoni hexapoda collembola
Tetrodontophora bielanensis hexapoda collembola
Limulus polyphemus chelicerata xiphosura
Lithobius forficatus myriapoda chilopoda
Vargula hilgendorfii crustacea ostracoda
Narceus annularus myriapoda diplopoda
Thyropygus sp. DVL−2001 myriapoda diplopoda
Tigriopus japonicus crustacea maxillopoda
Daphnia pulex crustacea branchiopoda
Artemia franciscana crustacea branchiopoda
Triops cancriformis crustacea branchiopoda
Penaeus monodon crustacea malacostraca
Pagurus longicarpus crustacea malacostraca
Portunus trituberculatus crustacea malacostraca
Panulirus japonicus crustacea malacostraca
Locusta migratoria hexapoda insecta
Bombyx mandarina hexapoda insecta
Antheraea pernyi hexapoda insecta
Ostrinia nubilalis hexapoda insecta
Ostrinia furnacalis hexapoda insecta
Anopheles gambiae hexapoda insecta
Anopheles quadrimaculatus hexapoda insecta
Cochliomyia hominivorax hexapoda insecta
Chrysomya putoria hexapoda insecta
Bactrocera oleae hexapoda insecta
Ceratitis capitata hexapoda insecta
Drosophila yakuba hexapoda insecta
Drosophila melanogaster hexapoda insecta
Pyrocoelia rufa hexapoda insecta
Tribolium castaneum hexapoda insecta
Crioceris duodecimpunctata hexapoda insecta
Lepidopsocid sp. RS−2001 hexapoda insecta
Thrips imaginis hexapoda insecta
Heterodoxus macropus hexapoda insecta
Melipona bicolor hexapoda insecta
Apis mellifera ligustica hexapoda insecta
Varroa destructor chelicerata arachnida
Ornithodoros moubata chelicerata arachnida
Carios capensis chelicerata arachnida
Rhipicephalus sanguineus chelicerata arachnida
Haemaphysalis flava chelicerata arachnida
Ixodes holocyclus chelicerata arachnida
Ixodes persulcatus chelicerata arachnida
Ixodes hexagonus chelicerata arachnida

Figure 18

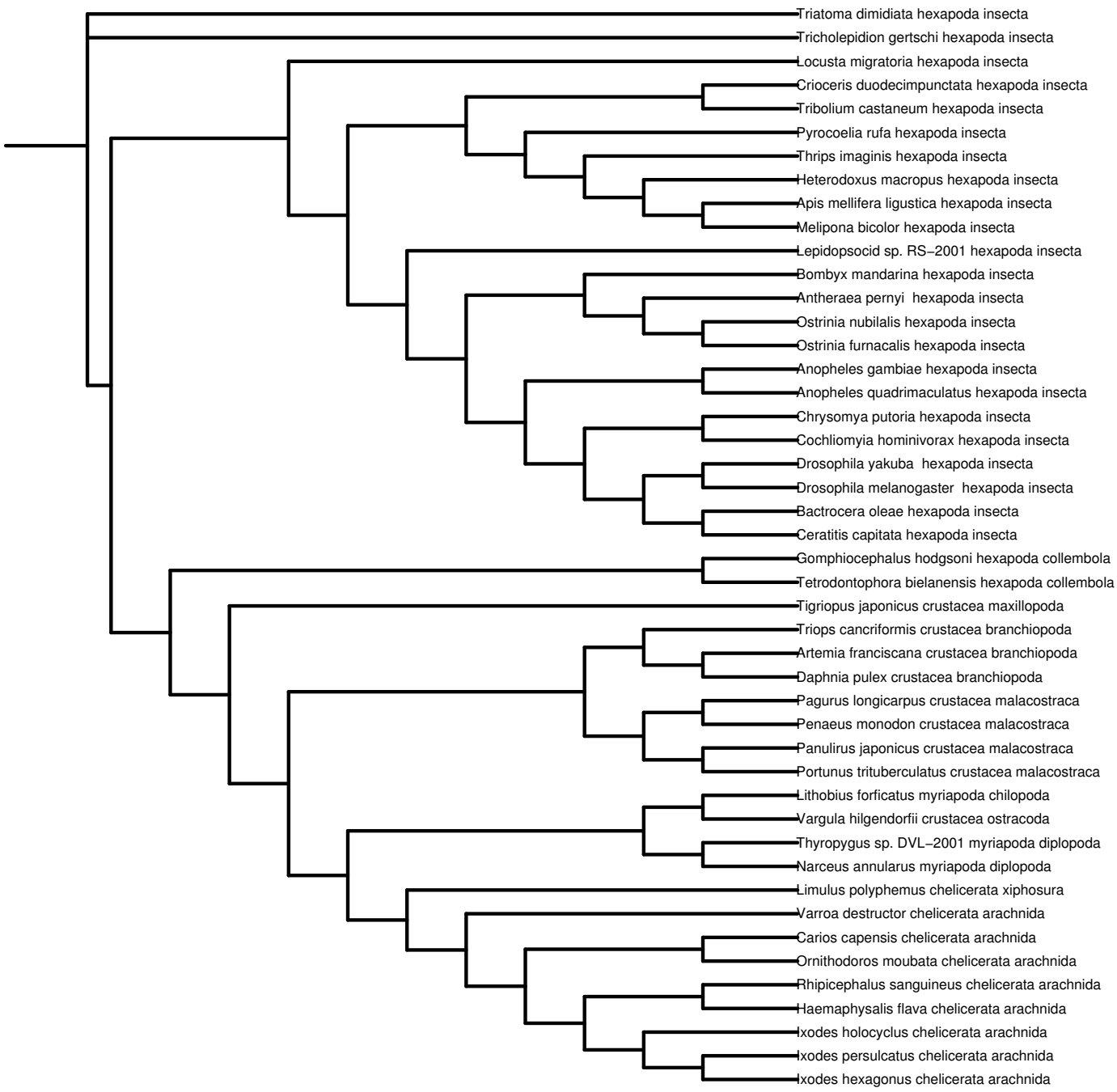

Triatoma dimidiata hexapoda insecta
Tricholepidion gertschi hexapoda insecta
Locusta migratoria hexapoda insecta
Crioceris duodecimpunctata hexapoda insecta
Tribolium castaneum hexapoda insecta
Pyrocoelia rufa hexapoda insecta
Thrips imaginis hexapoda insecta
Heterodoxus macropus hexapoda insecta
Apis mellifera ligustica hexapoda insecta
Melipona bicolor hexapoda insecta
Lepidopsocid sp. RS−2001 hexapoda insecta
Bombyx mandarina hexapoda insecta
Antheraea pernyi hexapoda insecta
Ostrinia nubilalis hexapoda insecta
Ostrinia furnacalis hexapoda insecta
Anopheles gambiae hexapoda insecta
Anopheles quadrimaculatus hexapoda insecta
Chrysomya putoria hexapoda insecta
Cochliomyia hominivorax hexapoda insecta
Drosophila yakuba hexapoda insecta
Drosophila melanogaster hexapoda insecta
Bactrocera oleae hexapoda insecta
Ceratitis capitata hexapoda insecta
Gomphiocephalus hodgsoni hexapoda collembola
Tetrodontophora bielanensis hexapoda collembola
Tigriopus japonicus crustacea maxillopoda
Triops cancriformis crustacea branchiopoda
Artemia franciscana crustacea branchiopoda
Daphnia pulex crustacea branchiopoda
Pagurus longicarpus crustacea malacostraca
Penaeus monodon crustacea malacostraca
Panulirus japonicus crustacea malacostraca
Portunus trituberculatus crustacea malacostraca
Lithobius forficatus myriapoda chilopoda
Vargula hilgendorfii crustacea ostracoda
Thyropygus sp. DVL−2001 myriapoda diplopoda
Narceus annularus myriapoda diplopoda
Limulus polyphemus chelicerata xiphosura
Varroa destructor chelicerata arachnida
Carios capensis chelicerata arachnida
Ornithodoros moubata chelicerata arachnida
Rhipicephalus sanguineus chelicerata arachnida
Haemaphysalis flava chelicerata arachnida
Ixodes holocyclus chelicerata arachnida
Ixodes persulcatus chelicerata arachnida
Ixodes hexagonus chelicerata arachnida

Figure 19