Published online 16 May 2011 *Nucleic Acids Research, 2011, Vol. 39, No. 15* ***6359–6368***
doi:10.1093/nar/gkr334

PSAR: measuring multiple sequence alignment reliability by probabilistic sampling

Jaebum Kim[1] and Jian Ma[1,2,*]

[1]Institute for Genomic Biology and [2]Department of Bioengineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

Received January 14, 2011; Revised April 18, 2011; Accepted April 24, 2011

ABSTRACT

Multiple sequence alignment, which is of fundamental importance for comparative genomics, is a difficult problem and error-prone. Therefore, it is essential to measure the reliability of the alignments and incorporate it into downstream analyses. We propose a new probabilistic sampling-based alignment reliability (PSAR) score. Instead of relying on heuristic assumptions, such as the correlation between alignment quality and guide tree uncertainty in progressive alignment methods, we directly generate suboptimal alignments from an input multiple sequence alignment by a probabilistic sampling method, and compute the agreement of the input alignment with the suboptimal alignments as the alignment reliability score. We construct the suboptimal alignments by an approximate method that is based on pairwise comparisons between each single sequence and the sub-alignment of the input alignment where the chosen sequence is left out. By using simulation-based benchmarks, we find that our approach is superior to existing ones, supporting that the suboptimal alignments are highly informative source for assessing alignment reliability. We apply the PSAR method to the alignments in the UCSC Genome Browser to measure the reliability of alignments in different types of regions, such as coding exons and conserved non-coding regions, and use it to guide cross-species conservation study.

INTRODUCTION

Multiple sequence alignment (MSA) is the task of aligning three or more biological sequences, such as DNA, RNA, or proteins. Its purpose is to find homologous regions among the sequences and predict nucleotide or amino acid level relationships among them. MSA is of great importance in downstream analyses of biological sequences, such as phylogenetic analysis, the identification of patterns of sequence conservation, and protein structure prediction.

Despite recent progress (1–4), MSA is still a difficult task and error-prone. The alignment errors can directly affect the downstream analyses and may lead to incorrect biological conclusions. Many researchers have been using Pecan (5) alignments in the Ensembl Genome Browser (6) and Multiz (7) alignments in the UCSC Genome Browser (8) to conduct different types of comparative genomic analyses. However, they often do not ask how reliable the alignment is or do not know how to quantitatively measure the reliability, which is the main topic of this article.

Many studies have been conducted to find the extent, cause and effect of the alignment errors. For example, Wong *et al.* (9) analyzed the effect of alignment uncertainty on genomic analysis and showed how different alignment tools produce different alignments and how those different alignments lead to different results in the inference of phylogeny from an MSA. Lunter *et al.* (10) analyzed the pairwise alignments of DNA sequences at human–mouse divergence by simulation and reported that existing whole-genome alignments have >15% of misaligned bases. Landan and Graur (11) classified alignment errors and quantified their levels in pairwise and multiple alignments by simulation. They showed that only very close sequences (far less than 0.1 substitutions per site) can be accurately aligned, and sequences in moderate evolutionary distances produce an alignment with errors in almost half of its columns. Fletcher and Yang (12) tested the effect of alignment errors on the test of positive selection and found a high positive correlation between alignment errors and false positive predictions.

One line of efforts to address the alignment uncertainty problem is the development of methods that are alignment-free, or can take into account a set of additional alignments, instead of relying on a single fixed alignment. Fleissner *et al.* (13) proposed a simultaneous inference

*To whom correspondence should be addressed. Tel: +1 217 2446562; Fax: +1 217 2650246; Email: jianma@illinois.edu

method for a multiple alignment and a phylogeny. Bais *et al*. (14) introduced a tool that can predict conserved sites while simultaneously aligning sequences. Satija *et al*. (15) developed a method that performs phylogenetic footprinting on alignment samples obtained by a Markov chain Monte Carlo (MCMC) approach. These studies demonstrated how critical the use of additional alignments is to improve the prediction accuracy. However, this kind of method usually requires high computational power and may not be applicable to some biological problems.

An alternative direction is to estimate the quality of alignments and incorporate it into downstream analyses. Assessing MSAs is a difficult task because the 'true alignment' is unknown. To assess MSAs without using the true alignment, Prakash and Tompa (16) developed a tool, called StatSigMA, to check whether an MSA is contaminated with unrelated sequences, based on the statistics of local MSAs. They later extended the method to handle genome-wide alignments (17). Landan and Graur (18) showed that the accuracy of an alignment program on a data set can be computed without the true alignment by reasoning that good alignments should be unaffected by the orientation of the input sequences. They therefore defined the Heads or Tails (HoT) alignment quality score as the agreement between one alignment generated from original sequences and the other from their reversed sequences. Hall (19) reported that HoT alignment quality scores are highly correlated with the real alignment accuracy by simulation. However, Wise (20) found that the HoT score is not a reliable measure by examining pairwise alignments where an optimal alignment exists, suggesting that comparing just the original and the reversed alignments is not enough to take into account the variability in alignments. Recently, Landan and Graur (21) extended their previous method to incorporate co-optimal alternative alignments generated by progressive alignment tools. More recently, Penn *et al*. (22) developed a new method, called GUIDANCE, which originated from the observation that most alignment uncertainty results from the uncertainty of a guide tree used by progressive alignment methods. To this end, they constructed a set of perturbed trees by using the bootstrap method and then generated a set of MSAs conditioned on each perturbed tree. They then tested the agreement of an input MSA with the set of perturbed MSAs. In comparison with the HoT method, they found that GUIDANCE is a more accurate predictor of unreliable alignment regions. However, it is still unclear whether the heuristically chosen measures, such as the agreement between original and reversed alignments, and the consistency with the perturbed MSAs from different guide trees, are general enough to take into account all alignment errors.

In this article, we present a new alignment reliability score, called Probabilistic Sampling-based Alignment Reliability (PSAR) score. Instead of relying on heuristic assumptions, we directly generate suboptimal alignments from an input MSA, and compute the agreement of the input MSA with the suboptimal alignments. In order to prevent any bias from a predefined phylogenetic tree, we construct the suboptimal alignments without using a fixed phylogenetic tree. Instead, we approximate the alignments based on pairwise comparisons between each single sequence and the sub-alignment of the input MSA where the chosen sequence is left out. The main rational of this strategy is that in many applications the evolutionary tree of the sequences under consideration is unknown. The PSAR score is compared to the GUIDANCE score using simulation-based benchmarks. We find that the performance of the PSAR score is superior to the GUIDANCE score, indicating that the suboptimal alignments are more informative sources for assessing alignment reliability than the perturbed MSAs induced from different guide trees. We apply the PSAR method to the Multiz (7) alignments of human chromosome 22 with 10 other primate species and characterize different sequence regions in terms of alignment reliability. In addition to the computation of the alignment reliability score, the PSAR method can generate the set of suboptimal alignments. These suboptimal alignments can be used to reduce spurious results induced from alignment errors in many biological analyses that are highly dependent on an MSA. As an example, we measure the variability of the phastCons (23) conservation scores of unreliable alignment regions in the Multiz alignment of human chromosome 22.

METHODS

Alignment reliability and suboptimal alignments

MSA programs generate a single alignment as their output, which is called an 'optimal alignment' based on a certain scoring scheme. However, there could be many alternative alignments, called 'suboptimal alignments', with equal or very similar scores to the optimal alignment. For example, suppose the alignment in Figure 1A is an output alignment by an MSA program. Here, the placement of the gap in the third sequence is not trivial and the alignment in Figure 1E is the equally likely alignment. If we completely trust the MSA program and use the alignment in Figure 1A alone, we may lose the important information that can be obtained from the suboptimal alignment in Figure 1E. Therefore, it is necessary to quantitatively measure the reliability of such regions of alignments and take it into account in downstream analyses. We reason that the alignment reliability is naturally reflected in the suboptimal alignments as the variability of alignments.

Probabilistic sampling of suboptimal alignments

The PSAR score is computed based on suboptimal alignments that are sampled from the posterior probability distribution of alignments. Because the direct computation of the posterior probability distribution of MSAs is intractable (24), it is approximated by pairwise comparisons between each sequence and the rest of an input MSA. Specifically, given an input MSA, PSAR selects one sequence at a time and makes a sub-alignment by leaving the chosen sequence out of the MSA. To compare the left-out sequence with the sub-alignment,

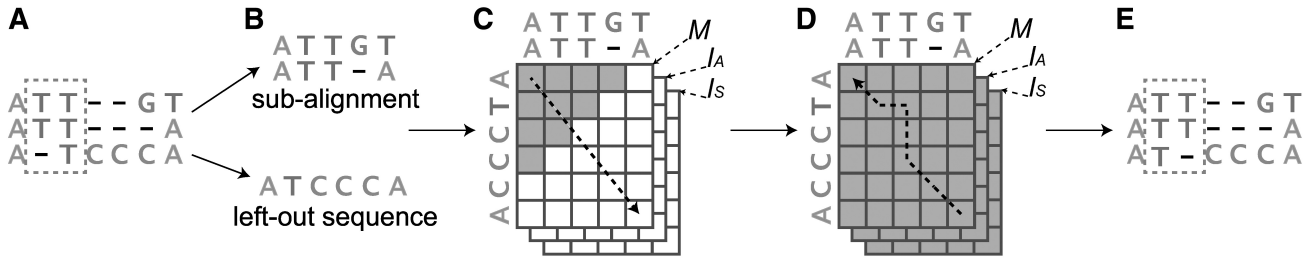

Figure 1. Procedure of probabilistic sampling. (**A**) Given an input MSA, (**B**) PSAR first chooses one sequence and makes a sub-alignment by leaving the chosen sequence out. Every gap in the left-out sequence and gaps spanning entire columns in the sub-alignment are removed. (**C**) Then, the pre-processed left-out sequence and sub-alignment are compared by dynamic programming (DP) based on a pair-HMM in Figure 2. Three DP tables, one for each pair-HMM state, are shown and they are filled from the top-left cell to the bottom-right cell. (**D**) Finally, PSAR probabilistically samples suboptimal alignments by tracing back from the bottom-right cell multiple times. An example of the sampled alignment is shown in (**E**).

all gaps in the left-out sequence and all columns in the sub-alignment that consist of only gaps are removed. For example, given the input MSA in Figure 1A, the third sequence is chosen as the left-out sequence. Figure 1B shows the pre-processed left-out sequence and sub-alignment. Alignment errors are largely due to the inaccurate creation and positioning of gaps (11). Therefore, we can use the size of gaps as the total number of sampling trials to obtain the amount of samples that is correlated with the alignment variability.

The pairwise comparison of the pre-processed left-out sequence and sub-alignment is based on a special type of a pair hidden Markov model (pair-HMM) that emits columns of an MSA given the left-out sequence and the sub-alignment. In other words, each column in the sub-alignment is considered as a distinct position of a sequence and it is emitted together with the character in the left-out sequence (in a matched case) or alone (in an inserted case). The pair-HMM (Figure 2) of PSAR has three states, a match state M and two insert states I_S and I_A for the left-out sequence and the sub-alignment, respectively. The match state M emits both a character in the left-out sequence and a column in the sub-alignment. This is represented by positioning the emitted character and column at the same column in the final alignment. The insert state I_A only emits a column in the sub-alignment. For the left-out sequence, a gap is inserted at the corresponding column in the final alignment. Similarly, the insert state I_S only emits a character in the left-out sequence. This is represented by placing gaps for every sequence in the sub-alignment at the corresponding column in the final alignment. In the pair-HMM, the transition probabilities among the states are estimated from the data by using the Baum–Welch algorithm (24). We note that fixed values can also be used for the transition probabilities instead of estimating them. This is particularly useful when the given alignment is believed to be not accurate enough for the parameter estimation, and the transition probabilities that are estimated from more informative alignments can be provided. The emission probabilities of the states M and I_A are computed based on Felsenstein's algorithm (25) by assuming a star topology. For example, suppose an MSA is composed of N sequences S_1 through S_N. Then, the emission

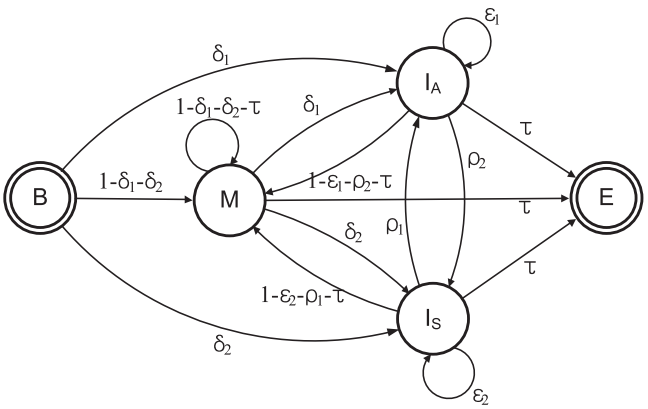

Figure 2. Pair-HMM of the PSAR method. This is a special type of pair-HMM, which is a generative model of an MSA that is constructed from a sequence S, and a sub-alignment A. The state M emits one character in the sequence S and one column in the sub-alignment A. The state I_S emits one character in S, whereas the state I_A emits one column in A. The state B is a begin state and E is an end state.

probability of a column that contains characters $S_1[i_1]$ through $S_N[i_N]$ is computed as:

$$P(S_1[i_1], \ldots, S_N[i_N]) = \sum_{\alpha \in \{A,C,G,T\}} \pi_\alpha \prod_{j=1}^{N} P(S_j[i_j]|\alpha)$$

where $S_j[i_j]$ is the i_j^{th} character of a sequence S_j, π_α is the background probability of a nucleotide α, and $P(S_j[i_j]|\alpha)$ is the probability of a nucleotide substitution from α to $S_j[i_j]$. Gaps are treated as missing data. PSAR uses the continuous-time Felsenstein model (25) as a nucleotide substitution model to describe $P(S_j[i_j]|\alpha)$. We note that instead of estimating the values of the parameters of the Felsenstein model, we used fixed values that were empirically verified because we cannot estimate them without knowing a phylogenetic tree (see 'Discussion' section).

To sample suboptimal alignments, PSAR first constructs dynamic programming (DP) tables by using the forward algorithm (24) based on the pair-HMM. Figure 1C shows an example for the left-out sequence and the sub-alignment in Figure 1B. Because the pair-HMM of PSAR consists of three states M, I_A,

and I_S, three DP tables (one for each state) are constructed. In each DP table, the cell at the i-th row and j-th column stores the combined probability of all alignments, called the forward probability, between the segment of the left-out sequence up to the i-th position and the segment of the sub-alignment up to the j-th column. Once the computation of the forward probabilities is done, PSAR traces back through the DP tables based on a probabilistic choice at each step (24). In other words, instead of choosing the highest scoring path, PSAR probabilistically takes a path based on its relative score in comparison with its neighboring paths. For example, the forward probability of a cell (i,j) in the DP tables for the states M, I_A and I_S are:

$$\begin{aligned} f^M(i,j) &= \Pr(i,j)[(1-\delta_1-\delta_2-\tau)f^M(i-1,j-1) \\ &\quad + (1-\epsilon_1-\rho_2-\tau)(f^{I_A}(i-1,j-1)) \\ &\quad + (1-\epsilon_2-\rho_1-\tau)(f^{I_S}(i-1,j-1))] \\ f^{I_A}(i,j) &= \Pr(j)[\delta_1 f^M(i,j-1)+\epsilon_1 f^{I_A}(i,j-1) \\ &\quad + \rho_1 f^{I_S}(i,j-1))] \\ f^{I_S}(i,j) &= \pi_{S_i}[\delta_2 f^M(i-1,j)+\epsilon_2 f^{I_S}(i-1,j) \\ &\quad + \rho_2 f^{I_A}(i-1,j)] \end{aligned}$$

where $\Pr(i, j)$ is the emission probability of both the character at the i-th position of the left-out sequence and the column at the j-th position of the sub-alignment, $\Pr(j)$ is the emission probability of the column at the j-th position of the sub-alignment, and π_{S_i} is the background probability of the nucleotide at the i-th position of the left-out sequence. Then, suppose we are at the cell $M(i,j)$ in the course of the traceback. PSAR chooses the next cell among $M(i-1, j-1)$, $I_A(i-1, j-1)$ and $I_S(i-1, j-1)$ based on the following probabilities that represent their relative strength:

$$\Pr(M(i-1,j-1)) = \frac{\Pr(i,j)(1-\delta_1-\delta_2-\tau)f^M(i-1,j-1)}{f^M(i,j)}$$

$$\Pr(I_A(i-1,j-1)) = \frac{\Pr(i,j)(1-\epsilon_1-\rho_2-\tau)f^{I_A}(i-1,j-1)}{f^M(i,j)}$$

$$\Pr(I_S(i-1,j-1)) = \frac{\Pr(i,j)(1-\epsilon_2-\rho_1-\tau)f^{I_S}(i-1,j-1)}{f^M(i,j)}$$

where $M(i,j)$, $I_A(i,j)$, and $I_S(i,j)$ represent the cells at the i-th row and j-th column in the DP tables for the states M, I_A, and I_S, respectively. The probabilistic choices for the states I_A and I_S can be similarly defined. The probabilistic sampling is repeated for different left-out sequences in the input MSA, and only distinct alignments among samples are collected. Figure 1D shows an example of the traceback through those three DP tables, and the corresponding alignment sample is shown in Figure 1E.

The probabilistic sampling can be controlled to generate suboptimal alignments that are very close to an optimal alignment, which is the highest scoring alignment that can be obtained from the comparison of the left-out sequence with the sub-alignment. As an option, the current implementation of PSAR provides a parameter that controls the degree of sampling. Specifically, at each step, the probabilistic sampling is done only when the relative strength of the best choice is lower than a given threshold. For example, the probabilistic choice based on the threshold t at the cell $M(i,j)$ can be defined as:

IF $\max[\Pr(M(i-1,j-1)), \Pr(I_A(i-1,j-1)),$
$\Pr(I_S(i-1,j-1))] \geq t$
THEN
Choose the cell with the maximum score
ELSE
Probabilistically choose the next cell based on the relative strength of each cell
ENDIF

A lower threshold produces an alignment sample that is closer to the optimal alignment, and the threshold of 0 means no probabilistic sampling. In that case, PSAR only produces the optimal alignment.

Computation of alignment reliability scores

To obtain the reliability score of an input MSA, we compute the following measure, inspired by the GUIDANCE method (26), by comparing the input MSA with each sampled MSA:

- Pair score: each pair of aligned characters in the input MSA is assigned a score 1 if the pair is also aligned in the sample MSA, and 0 otherwise.

Then, similar to the GUIDANCE method (26), we define the PSAR score for each pair of aligned characters (PSAR pair score) and for each column (PSAR column score) in the input MSA as the average of the pair scores over all sampled MSAs and the average of the PSAR pair scores over all pairs in that column, respectively.

Computational complexity

The time complexity of the probabilistic sampling is $O(L^2NS)$, where L is the alignment length, N is the number of sequences, and S is the number of sampling trials. This is because for each of S sampling trials and for each of N different left-out sequences, it needs to compare one sequence and one sub-alignment, whose length is $O(L)$. The reliability score computation requires $O(LN^2S)$ time because for each of S samples and for each pair of sequences $O(N^2)$, it has to examine $O(L)$ alignment columns. The memory complexity of our method is $O(LNS)$ because it needs to store $O(S)$ alignments of length $O(L)$ for N sequences.

Simulation-based benchmarking

We evaluated the performance of PSAR based on simulated sequences. The simulation-based benchmarks have the advantage of providing the true alignment that can be directly compared with predicted results. The simulation-based approach, however, highly depends on simulation parameters that determine the underlying

processes and rates of sequence evolution. To overcome the limitation of the simulation approach, we used a simulation method that is based on the entire spectrum of parameter values estimated from real data (27). We simulated 1000 sequences of length 500 bp in the phylogenetic tree of five *Drosophila* species, *D. melanogaster*, *D. simulans*, *D. yakuba*, *D. ananassae* and *D. pseudoobscura*, by using the entire distributions of the values of parameters, such as substitution rate, the ratio of substitutions to insertions and deletions and the ratio of insertions to deletions, inferred from *Drosophila* non-coding sequence alignments (27). We call this data set 'insect benchmark'.

To perform the assessment on species in different taxa, we also generated 1000 sequences of length 500 bp in the phylogenetic tree of five mammalian species, human, chimpanzee, gorilla, orangutan and rhesus by using the Dawg simulation program (28). In this case, we used a single value for each parameter that is estimated from real mammalian sequences (7, 29). The original branch lengths of the phylogenetic tree (7, 29) were scaled up by five times to generate enough mutations in the 500 bp-long sequences as compared to the real data. We call this data set 'mammal benchmark'. We compared our program PSAR with the GUIDANCE program (26) by computing the true and false positive rates of classifying pairs of aligned characters in an input MSA on multiple cutoff scores. We reported the Receiver Operating Characteristic (ROC) curves with the Area Under the Curve (AUC) scores. We used three different MSA programs, Pecan v0.7 (5), MAFFT v6.818b (30) and ClustalW v2.0.10 (31), to generate input MSAs. The current version of the GUIDANCE program provides three MSA programs, MAFFT, ClustalW and Prank (32), as an alignment option to generate perturbed MSAs that are used to compute the alignment certainty scores. However, we only tested the GUIDANCE program with MAFFT and ClustalW options because the running time of Prank was too long.

Application to genome-wide alignments

We downloaded the Multiz (7) alignments of 45 vertebrate genomes with human chromosome 22 (GRCh37/hg19 assembly) from the UCSC Genome Browser (8). The downloaded alignments consist of multiple fragments with varying size and with different number of species being aligned. We then created the fragments of the Multiz alignments of 10 primate genomes for human, chimp, gorilla, orangutan, rhesus, baboon, marmoset, tarsier, mouse lemur and bushbaby, by filtering out the sequences of non-primate species. We note that the choice of 10 primate species was arbitrary, and it is straightforward to use other species, such as more diverged species. The final alignment fragments amount to 67% of the human chromosome 22. For each fragment, we ran the PSAR program with a pre-processing step based on the length of the fragment. Specifically, if the fragment length is longer than 1 Kbp, we first identified highly conserved regions without gaps, called 'anchors', by using the Gblocks program (33) with default options, and next extracted inter-anchor regions. To avoid making too short inter-anchor regions, we appended a certain amount of flanking anchor regions to both ends of the inter-anchor regions and made the minimum length of the inter-anchor regions 500 bp. We ran the PSAR program for only inter-anchor regions by assuming that the alignments in the anchor regions are highly reliable. This computation took roughly 3 days on Intel Xeon 2.80 GHz machine with parallel execution of 10 processors (see 'Discussion' section).

Together with the source code of the PSAR program, the PSAR scores for human chromosome 22 are available on our supplementary website (http://bioen-compbio.bioen.illinois.edu/psar) in the wiggle (WIG) format that can be directly visualized by the UCSC Genome Browser. The PSAR scores of other chromosomes will be available on the website in the near future.

RESULTS

Performance evaluation

We evaluated the performance of our method PSAR in comparison with the GUIDANCE method (26), using the simulated data (see 'Methods' section). The GUIDANCE method computes the alignment certainty score by using perturbed MSAs (alignment samples) that are generated based on perturbed phylogenetic trees. This evaluation focused on measuring how accurately each method classifies pairs of aligned characters in an input MSA into reliable or unreliable classes. We first identified all aligned characters in the input MSA. Among them, by varying cutoff scores, we then counted the number of true positive pairs that are labeled as a reliable one as well as aligned in the true alignment, and the number of false positive pairs that are also labeled as a reliable one but not aligned in the true alignment. We note that the results from different MSA programs are not comparable because the different MSA programs usually generate different MSAs that have different number of aligned characters. For a fair comparison, we constructed the same number of alignment samples for PSAR and GUIDANCE.

We first performed the evaluation on the insect benchmark (see 'Methods' section). When we used the Pecan alignments as input MSAs (Figure 3A), PSAR outperformed GUIDANCE with both alignment options (see 'Methods' section). PSAR classified pairs of characters with the AUC score of 0.905, whereas GUIDANCE achieved the AUC scores of 0.871 and 0.851 with MAFFT and ClustalW options, respectively. In addition, the true positive rate of PSAR is degraded more slowly than GUIDANCE as the false positive rate decreases. For example, when the false positive rate is 0.1, the true positive rate of PSAR is still around 0.8, while GUIDANCE's score drops down to below 0.7. When the false positive rate is lower than 0.2, two different options of the GUIDANCE program show very similar performance.

The GUIDANCE program requires an MSA program to construct perturbed MSAs, and Pecan is not supported

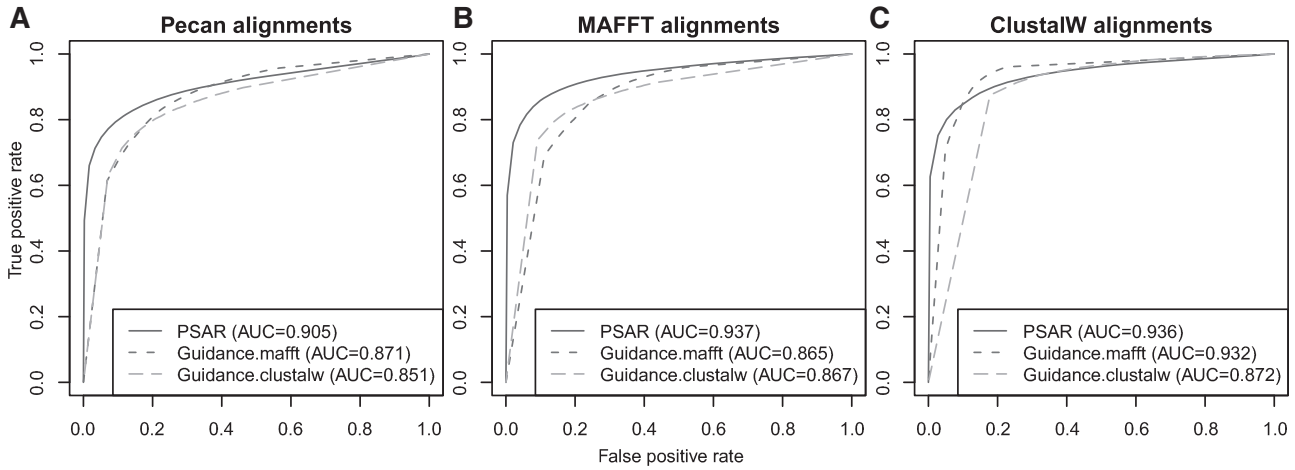

Figure 3. Performance of PSAR in comparison with GUIDANCE on the insect benchmark (see 'Methods' section). Three MSA programs, Pecan (A), MAFFT (B), and ClustalW (C), were used to generate input MSAs. ROC curves are reported and AUC scores are shown in parentheses in legend. The GUIDANCE program was run with two MSA programs, MAFFT and ClustalW ('Guidance.mafft' and 'Guidance.clustalw' in legend, respectively), to generate perturbed MSAs.

in its current version. Therefore, the use of different MSA programs for generating the input MSA and perturbed MSAs in GUIDANCE may introduce a bias. To address this problem, we repeated the same experiment with two additional MSA programs, MAFFT and ClustalW, which are used in the GUIDANCE program to generate perturbed MSAs. With the MAFFT alignments as input MSAs (Figure 3B), we found very similar patterns as from the Pecan alignments, except that the AUC score of PSAR improved to 0.937. On the other hand, when the ClustalW alignments were provided as input (Figure 3C), the AUC score of GUIDANCE with MAFFT option was dramatically increased to 0.932, whereas the score increase of GUIDANCE with CluatalW option was marginal.

Overall, the PSAR program was found to be superior to the GUIDANCE program across three different input alignments (PSAR's AUC scores of greater than 0.9) and the performance of PSAR was more robust than GUIDANCE to different input alignments. In addition, the minimum false positive rate of GUIDANCE was much higher than that of PSAR. This made the shape of the GUIDANCE's curves straight in small false positive rate regions.

We next conducted the same experiment with the mammal benchmark (see 'Methods' section). As Figure 4 shows, the classification accuracy of both programs generally decreased, except for the AUC of PSAR with ClustalW alignments. However, the AUC score of PSAR was still better than that of GUIDANCE across all settings, and the performance of PSAR was degraded more slowly than GUIDANCE. In this data set, the minimum false positive rate that the GUIDANCE program can achieve was worse than the rate from the insect benchmark (0.07 in the insect benchmark and 0.25 in the mammal benchmark by the MAFFT option with Pecan alignments as input), resulting in more obvious straight lines in the ROC curves.

Fraction of unreliable alignments in human chromosome 22

To measure the quality of the precomputed Multiz alignments that are widely used in many applications, we obtained the Multiz alignments of human chromosome 22 from the UCSC Genome Browser and computed the PSAR scores for each alignment column (see 'Methods' section). The human chromosome 22 is 51 Mbp long and its sequence positions can be partitioned into four disjoint types, coding exons, conserved non-coding regions, non-coding regions with repetitive elements and the rest of non-coding regions, using the annotations available in the UCSC Genome Browser. The annotation for the conserved non-coding regions in the UCSC Genome Browser was created by running the phastCons program (23). As shown in Figure 5A, coding exons and conserved non-coding regions each comprise only 2% of the human chromosome 22. Among the non-coding regions, almost half of them are repetitive elements.

For each different type of regions, we identified unreliable positions based on different PSAR score cutoffs (from 0.5 to 1.0) and computed its fraction over the total length of the region (Figure 5B). As expected, the alignments of conserved regions, such as coding exons and conserved non-coding regions, were highly reliable. For example, when the PSAR score cutoff was 0.9, only 1.3% and 1.8% were aligned unreliably for coding exons and conserved non-coding regions, respectively. On the other hand, other non-coding regions were found to have relatively high fraction of unreliably-aligned positions. Almost 5% of their positions were classified as unreliable regions with the PSAR score cutoff 0.9, and this fraction is more than 2-fold higher than the conserved regions. In addition, the unreliably-aligned positions in non-conserved non-coding regions were accumulated more quickly than conserved regions as the PSAR score cutoff was increased.

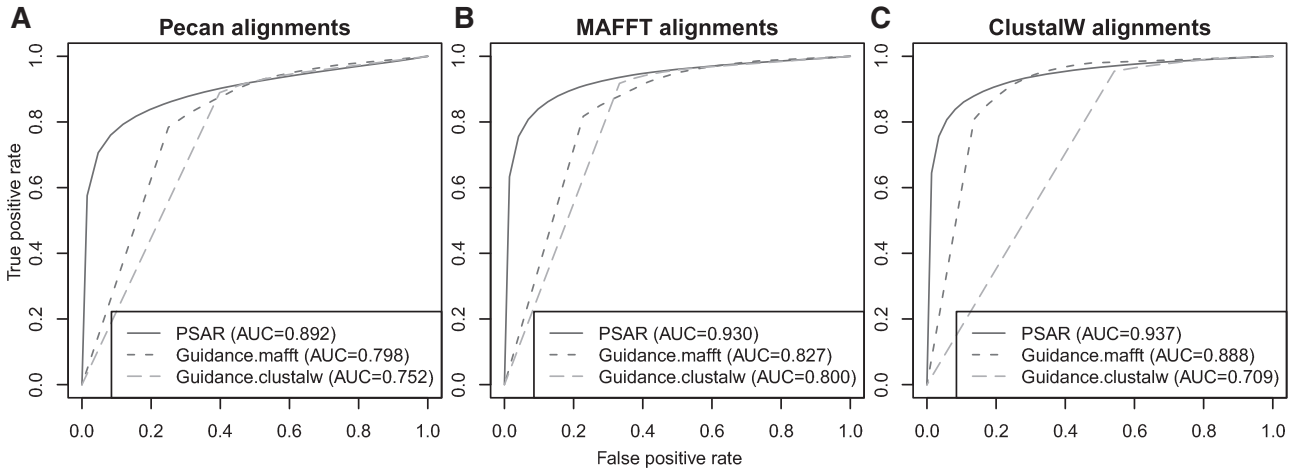

Figure 4. Performance of PSAR in comparison with GUIDANCE on the mammal benchmark (see 'Methods' section). Three MSA programs, Pecan (**A**), MAFFT (**B**), and ClustalW (**C**), were used to generate input MSAs. ROC curves are reported and AUC scores are shown in parentheses in legend. The GUIDANCE program was run with two MSA programs, MAFFT and ClustalW ('Guidance.mafft' and 'Guidance.clustalw' in legend, respectively), to generate perturbed MSAs.

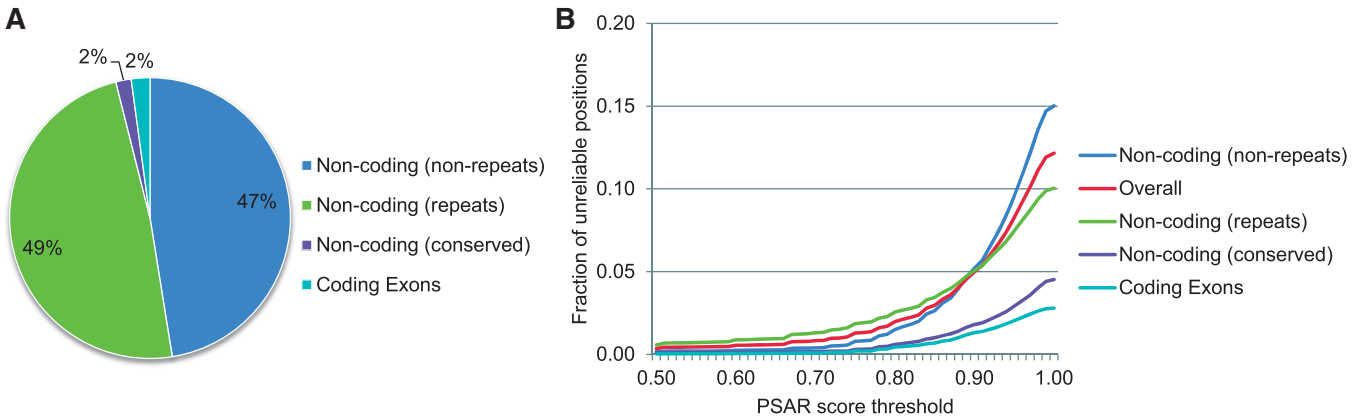

Figure 5. Fraction of unreliable alignments in human chromosome 22. (**A**) Fraction of different types of regions, such as coding exons ('Coding Exons'), conserved non-coding regions ['Non-coding (conserved)'], non-coding regions with repetitive elements ['Non-coding (repeats)'], and the rest of non-coding regions ['Non-coding (non-repeats)']. (**B**) Fraction of unreliably-aligned positions for each different type of regions as a function of the PSAR score cutoff. 'Overall' represents the union of all types of regions (the whole chromosome 22).

Evolutionary conservation in unreliable alignment regions

As a by-product, the PSAR method can produce suboptimal alignments, and they can be used to improve the reliability of any downstream analyses in unreliable alignment regions. There could be many ways to take advantage of the suboptimal alignments. Here, we applied the suboptimal alignments that were generated by the quality controlling parameter $t = 0.6$ (see 'Methods' section) to show the variability of evolutionary conservation scores of the phastCons program (23) in unreliable alignment regions.

As shown in Figure 5B, the alignment reliability of most evolutionarily conserved non-coding regions, which were predicted by the phastCons program, is reasonably good. However, there are suspicious alignment regions where the phastCons conservation scores are high. In this case, it is necessary to investigate whether the erroneous alignments led to incorrect conservation measurement. To this end, we first estimated the parameters of the phastCons program based on its instructions (http://compgen.bscb.cornell.edu/phast/phastCons-HOWTO.html) by using the Multiz alignments of the whole human chromosome 22. For each alignment fragment with unreliable positions (PSAR score < 0.7), we then computed the phastCons conservation scores for each suboptimal alignment. We summarized the conservation scores from multiple suboptimal alignments by showing their mean and standard deviation with the total number of suboptimal alignments. Figure 6 shows an example region with unreliable positions. We found in Figure 6 that positions with low PSAR scores (the first row) were likely to have variable phastCons conservation scores (the third row). As the alignment at the bottom in Figure 6 shows, the equally likely placement of gaps probably increased the variability of alignment in that region, and this resulted in the low PSAR and variable conservation scores.

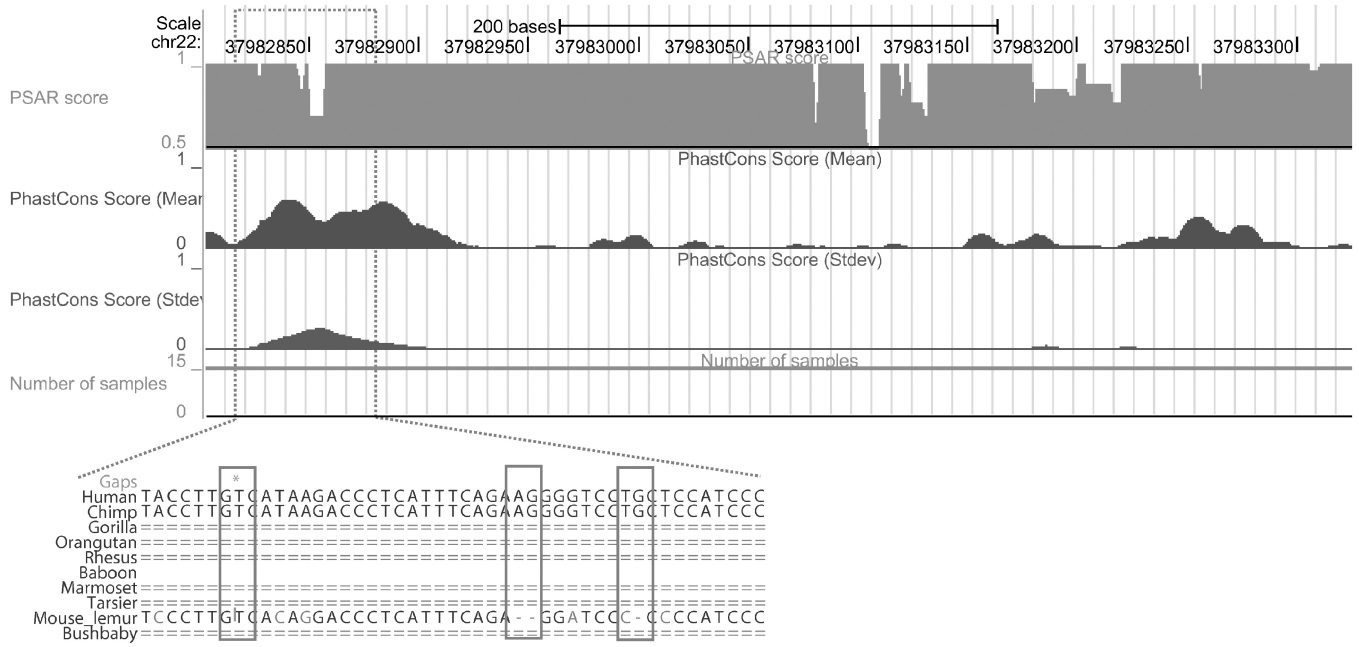

Figure 6. phastCons conservation scores and their variability in an example unreliable region (human chromosome 22:37,982,804-37,983,325). The first row shows the PSAR scores. The phastCons conservation scores were computed for each suboptimal alignment in this region, and the mean and standard deviation are shown in the second and the third rows, respectively. The alignment at the bottom is the Multiz alignment in the red-dotted box. The low PSAR and the variable phastCons conservation scores are probably attributed to the equally likely placement of gaps highlighted by red rectangles in the Multiz alignment.

DISCUSSION

MSA programs are generally based on heuristic methods. This makes it difficult to produce accurate MSAs, and alignment errors naturally cause many problems in downstream analyses. In this study, we propose a new method, called PSAR, to measure the reliability of an MSA. The PSAR method computes the reliability scores of pairs of characters or columns in an input MSA by investigating the consistency with suboptimal alignments generated through probabilistic sampling. We use simulation to show that the performance of PSAR is better and more robust to different input MSAs than the GUIDANCE program.

Our method ignores the phylogenetic relationships among sequences. This may lose important evolutionary information and the inclusion of it may produce more fine-tuned alignment samples. However, we note that the more alignment samples are tuned to a predefined phylogenetic tree, the more highly they are biased to the tree. This affects adversely on studies whose aim is to find the evolutionary relationship among species, such as a coalescence analysis to identify genomic segments with different genealogies (34).

In the evaluation of our method, we only used simulation-based benchmarks. This is because we focused on DNA sequences, and there are no benchmark databases made of DNA sequences and their manually-curated alignments. Another issue is how biologically realistic the simulated sequences are. In the case of the insect benchmark (see 'Methods' section), we used the value distributions of simulation parameters estimated from real data together with a simulation method that was validated to produce realistic benchmarks of MSAs (27). In the case of the mammal benchmark (see 'Methods' section), we will improve the parameter estimation in the near future.

The additional benefit of our method is that it not only measures the quality of an MSA, but also generates suboptimal alignments that are very close to an optimal alignment (see 'Methods' section). These suboptimal alignments can be directly used as additional data to complement the limitation of single MSA-based analyses. For example, when predicting evolutionarily conserved elements (23) or estimating a tree (35), we can use the set of suboptimal alignments and combine the separate results obtained from each of them. We showed in Figure 6 how this approach can be applied to the search for evolutionarily conserved regions. By using this approach, we can reduce the false positive predictions that may result from alignment errors.

There may be a circularity issue in our method because our method estimates the parameters of the pair-HMM on a given MSA, and therefore the quality of the MSA has a direct effect on the quality of the alignment samples. To alleviate this problem, we can use the parameter values that are estimated from reliable MSAs if the given MSA is believed not accurate enough for the parameter estimation.

One potential drawback of our method is that it may not be efficient enough to handle MSAs with a large number of species. As a future direction, we will incorporate heuristic approaches, such as a corner cutting

method (36) that ignores low probability regions in a DP table (typically, top-right and bottom-left corners), that are useful to dramatically reduce the total computation time. An additional limitation of our method is that the computation of the pair and column scores may be too simple. First, for each pair of aligned characters in an input MSA, we use the counting of its occurrences in alignment samples to compute the pair score. However, each alignment sample has a different probability and therefore the contribution of a different sample to the final pair score should be different. As a future direction, we will develop a weighted version of the pair score that can consider the probabilities of alignment samples. Second, we use a simple average for the PSAR column score even though it is a natural choice when we do not know the phylogenetic relationship among input sequences. However, this will be improved by using context-dependent phylogenetic information (34) in an attempt to avoid any bias from a fixed phylogenetic tree. The context-dependent phylogenetic information is also useful for estimating the parameter values of a nucleotide substitution model, which is impossible without a phylogenetic tree including branch lengths.

ACKNOWLEDGEMENTS

We would like to thank Webb Miller for the inspiration and helpful discussions. We also would like to thank anonymous reviewers for their insightful comments and suggestions. J.K. is an IGB fellow at the University of Illinois.

FUNDING

Faculty fellowship from the National Center for Supercomputing Applications at the University of Illinois (to J.M.) and National Science Foundation CAREER award IIS-1054309 (to J.M.). Funding for open access charge: National Science Foundation CAREER award IIS-1054309.

Conflict of interest statement. None declared.

REFERENCES

1. Blanchette,M. (2007) Computation and analysis of genomic multi-sequence alignments. *Annu. Rev. Genomics Hum. Genet.*, **8**, 193–213.
2. Notredame,C. (2007) Recent evolutions of multiple sequence alignment algorithms. *PLoS Comput. Biol.*, **3**, e123.
3. Pirovano,W. and Heringa,J. (2008) Multiple sequence alignment. *Methods Mol. Biol.*, **452**, 143–161.
4. Simossis,V., Kleinjung,J. and Heringa,J. (2003) An overview of multiple sequence alignment. *Curr. Protoc. Bioinformatics*, **Chapter 3**, Unit 3.7.
5. Paten,B., Herrero,J., Beal,K. and Birney,E. (2009) Sequence progressive alignment, a framework for practical large-scale probabilistic consistency alignment. *Bioinformatics*, **25**, 295–301.
6. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
7. Blanchette,M., Kent,W.J., Riemer,C., Elnitski,L., Smit,A.F., Roskin,K.M., Baertsch,R., Rosenbloom,K., Clawson,H., Green,E.D. *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**, 708–715.
8. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
9. Wong,K.M., Suchard,M.A. and Huelsenbeck,J.P. (2008) Alignment uncertainty and genomic analysis. *Science*, **319**, 473–476.
10. Lunter,G., Rocco,A., Mimouni,N., Heger,A., Caldeira,A. and Hein,J. (2008) Uncertainty in homology inferences: assessing and improving genomic sequence alignment. *Genome Res.*, **18**, 298–309.
11. Landan,G. and Graur,D. (2009) Characterization of pairwise and multiple sequence alignment errors. *Gene*, **441**, 141–147.
12. Fletcher,W. and Yang,Z. (2010) The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol. Biol. Evol.*, **27**, 2257–2267.
13. Fleissner,R., Metzler,D. and vonHaeseler,A. (2005) Simultaneous statistical multiple alignment and phylogeny reconstruction. *Syst. Biol.*, **54**, 548–561.
14. Bais,A.S., Grossmann,S. and Vingron,M. (2007) Simultaneous alignment and annotation of cis-regulatory regions. *Bioinformatics*, **23**, e44–e49.
15. Satija,R., Novak,A., Miklos,I., Lyngso,R. and Hein,J. (2009) BigFoot: Bayesian alignment and phylogenetic footprinting with MCMC. *BMC Evol. Biol.*, **9**, 217.
16. Prakash,A. and Tompa,M. (2005) Statistics of local multiple alignments. *Bioinformatics*, **21(Suppl. 1)**, i344–i350.
17. Prakash,A. and Tompa,M. (2007) Measuring the accuracy of genome-size multiple alignments. *Genome Biol.*, **8**, R124.
18. Landan,G. and Graur,D. (2007) Heads or tails: a simple reliability check for multiple sequence alignments. *Mol. Biol. Evol.*, **24**, 1380–1383.
19. Hall,B.G. (2008) How well does the HoT score reflect sequence alignment accuracy? *Mol. Biol. Evol*, **25**, 1576–1580.
20. Wise,M.J. (2010) No so HoT - heads or tails is not able to reliably compare multiple sequence alignments. *Cladistics*, **26**, 438–443.
21. Landan,G. and Graur,D. (2008) Local reliability measures from sets of co-optimal multiple sequence alignments. *Pac. Symp. Biocomput.*, 15–24.
22. Penn,O., Privman,E., Ashkenazy,H., Landan,G., Graur,D. and Pupko,T. (2010) GUIDANCE: a web server for assessing alignment confidence scores. *Nucleic Acids Res.*, **38(Suppl.)**, W23–W28.
23. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
24. Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge.
25. Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.
26. Penn,O., Privman,E., Landan,G., Graur,D. and Pupko,T. (2010) An alignment confidence score capturing robustness to guide tree uncertainty. *Mol. Biol. Evol.*, **27**, 1759–1767.
27. Kim,J. and Sinha,S. (2010) Towards realistic benchmarks for multiple alignments of non-coding sequences. *BMC Bioinformatics*, **11**, 54.
28. Cartwright,R.A. (2005) DNA assembly with gaps (Dawg): simulating sequence evolution. *Bioinformatics*, **21(Suppl. 3)**, iii31–iii38.
29. Pollard,K.S., Hubisz,M.J., Rosenbloom,K.R. and Siepel,A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**, 110–121.
30. Katoh,K., Asimenos,G. and Toh,H. (2009) Multiple alignment of DNA sequences with MAFFT. *Methods Mol. Biol.*, **537**, 39–64.
31. Larkin,M.A., Blackshields,G., Brown,N.P., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

32. Loytynoja,A. and Goldman,N. (2005) An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl Acad. Sci. USA*, **102**, 10557–10562.
33. Castresana,J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.*, **17**, 540–552.
34. Hobolth,A., Christensen,O.F., Mailund,T. and Schierup,M.H. (2007) Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet.*, **3**, e7.
35. Dutheil,J.Y., Ganapathy,G., Hobolth,A., Mailund,T., Uyenoyama,M.K. and Schierup,M.H. (2009) Ancestral population genomics: the coalescent hidden Markov model approach. *Genetics*, **183**, 259–274.
36. Hein,J., Wiuf,C., Knudsen,B., Møller,M.B. and Wibling,G. (2000) Statistical alignment: computational properties, homology testing and goodness-of-fit. *J. Mol. Biol.*, **302**, 265–279.