PSAR-Align: improving multiple sequence alignment using probabilistic sampling
Jaebum Kim1,2,* and Jian Ma3,4,*

1Department of Animal Biotechnology, 2UBITA Center for Biotechnology Research (CBRU), Konkuk University, Seoul 143-701, Korea, 3Department of Bioengineering and 4Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

Associate Editor: Ziv Bar-Joseph

Received on July 29, 2013; revised on October 8, 2013; accepted on October 28, 2013

1 INTRODUCTION
Multiple sequence alignment (MSA) is one of the most important foundations for cross-species comparative genomic analysis (Kumar and Filipski, 2007; Notredame, 2007). Although many algorithms for MSA have been developed (Kemena and Notredame, 2009), MSA is still error-prone. For example, it was estimated that at least 10% of the human-mouse whole-genome alignment is misaligned at the UCSC Genome Browser and the number increases for other species (Prakash and Tompa, 2007).

We previously developed a novel measure, called PSAR (Kim and Ma, 2011), which can assess the reliability of an MSA based on its agreement with probabilistically sampled suboptimal alignments (SAs). SAs provide additional information that cannot be obtained by the optimal alignment alone, especially when the optimal alignment is not far superior to the SAs.

In this article, we introduce a new realignment method, PSAR-Align, which refines a given MSA based on a probabilistic framework that takes advantage of the SAs of the given MSA. Briefly, PSAR-Align (i) samples SAs from the given MSA, (ii) estimates posterior probabilities of aligning two residues from two different sequences and (iii) generates a revised MSA using an expected accuracy-based alignment algorithm (Bradley et al., 2009; Do et al., 2005; Paten et al., 2009; Roshan and Livesay, 2006).

2 METHODS
1.1 PSAR-Align algorithm
Given an input MSA, PSAR-Align first generates SAs by probabilistic sampling (Fig. 1A and B). Specifically, for each pair of one sequence and a remaining sub-alignment, PSAR-Align compares them based on a special pair hidden Markov model (pair-HMM) that emits columns of an MSA, which can be represented by dynamic programming matrix. To generate the SAs, PSAR-Align traces back through the dynamic programming matrix based on a probabilistic choice at each step that can take into account the relative score of a current path in comparison with neighboring paths (Kim and Ma, 2011). Then, for each pair of two residues \( x_i \) and \( y_j \) from two different sequences \( X \) and \( Y \) in the input MSA, their alignment in the sampled SAs are counted and converted to the posterior probability \( P(x_i \sim y_j \mid X, Y) \) (Fig. 1C and D), which is defined as follows:

\[
P(x_i \sim y_j \mid X, Y) \approx \sum_{S} \frac{1_{\{x_i \sim y_j \in S\}}}{|S|}
\]

where \( S \) is the set of the SAs, \(|S|\) is the total number of alignments in \( S \) and \( 1_{\{x_i \sim y_j \in S\}} \) is an indicator function that returns \( 1 \) only when \( x_i \) and \( y_j \) are aligned in an SA \( S \).

PSAR-Align uses these probabilities to generate the revised alignment by maximizing an expected accuracy of an MSA \( A \) \( \text{acc}(A) \) against the (unknown) true alignment. The expected accuracy is the sum of the posterior probabilities of aligned pairs of residues and unaligned (aligned with a gap) residues in a given MSA (Bradley et al., 2009), which is defined as follows:

\[
\text{acc}(A) = \sum_{X, Y} \left[ 2 \sum_{i,j, x_i \sim y_j \in A_{XY}} P(x_i \sim y_j \mid X, Y) + \sum_{i,j, x_i \sim y_j \in A_{XY}} P(\sim y_j \mid X, Y) \right]
\]

where \( A_{XY} \) is a pairwise alignment between two sequences \( X \) and \( Y \), \( P(x_i \sim y_j \mid X, Y) \) is the posterior probability of pairwise alignment mentioned earlier in the text and \( P(\sim y_j \mid X, Y) \) and \( P(\sim \sim \mid X, Y) \) are the posterior probabilities of aligning each residue with a gap that can be computed as follows:

\[
P(x_i \sim \sim \mid X, Y) = 1 - \sum_{j} P(x_i \sim y_j \mid X, Y)
\]

\[
P(\sim y_j \mid X, Y) = 1 - \sum_{i} P(x_i \sim y_j \mid X, Y)
\]

For the maximization of the expected accuracy, we used the sequence annealing algorithm in the FSA program (Bradley et al., 2009).

The current version of PSAR-Align was implemented mainly in C++ with additional Perl scripts, and the expected accuracy
maximization step was implemented on top of the source code of FSA (Bradley et al., 2009).

2.2 Evaluation
We assessed the performance of PSAR-Align by using a simulated benchmark generated by Dawg (Cartwright, 2005). The benchmark mimics non-coding DNA sequences of five mammalian species (human, mouse, rat, dog and cow), whose phylogenetic tree was obtained from the UCSC Genome Browser (Meyer et al., 2013). The benchmark consists of 1000 replicates of ~1 kb-long sequences, and ClustalW (Thompson et al., 2002), MAFFT (Katoh et al., 2005), MAVID (Bray and Pachter, 2004), MUSCLE (Edgar, 2004) and Pecan (Paten et al., 2008) were used to generate the input MSA. Two evaluation measures were used: (i) alignment sensitivity, which is the fraction of aligned and unaligned (aligned with a gap) residues in the true alignment that agree with the predicted alignment and (ii) alignment specificity, which is the fraction of aligned and unaligned (aligned with a gap) residues in the predicted alignment that agree with the true alignment.

2.3 Computational complexity
The time and memory complexities of alignment sampling are $O(L^2 S)$ and $O(LN)$, respectively, where $L$ is the alignment length, $N$ is the number of sequences and $S$ is the number of sampling trials (Kim and Ma, 2011). The pairwise posterior probability computation requires $O(N^2 L)$ time and memory complexity, and the maximization of the expected accuracy was done efficiently by FSA (Bradley et al., 2009). In the evaluation, a single run of PSAR-Align for each input MSA took ~3 min in an Intel (R) Xeon 2.67 GHz machine with 64 GB memory.

3 RESULTS
We evaluated PSAR-Align by using simulated sequences of five mammalian species (see Section 2). By using ClustalW, MAFFT, MAVID, MUSCLE and Pecan, input MSAs were generated and fed into PSAR-Align, which resulted in a refined MSA. The original (by the aforementioned five programs) and revised (by PSAR-Align) MSAs were compared with true MSAs, which were known from the simulation. We used two evaluation measures: alignment sensitivity and specificity (see Section 2). As shown in Table 1, the alignment specificity of the original MSA by all five programs increased in the PSAR-Align MSA. The amount of increases ranges from 0.662 (Pecan) to 2.591 (ClustalW). Similar differences were also observed for alignment sensitivity, which showed an increase ranging from 0.554 (ClustalW) to 2.602 (ClustalW). In the case of Pecan, alignment specificity of the revised alignment by PSAR-Align was slightly higher than the original, but the opposite pattern was observed from alignment sensitivity. Our evaluation results indicate that (i) PSAR-Align can be used to improve MSAs from different types of MSA programs and (ii) Pecan is a high-quality MSA program based on our evaluation datasets.

4 CONCLUSION
We have developed a new alignment refinement tool, PSAR-Align, which is a realignment algorithm based on probabilistically sampled SAs. The performance of PSAR-Align was evaluated by simulation-based benchmarks. This tool will be useful for comparative genomics studies using MSA.

Funding: National Research Foundation of Korea Grant (2012R1A1A1015186 to J.K.) and National Institutes of Health grant (HG006464 to J.M.).

Conflict of Interest: none declared.

<table>
<thead>
<tr>
<th>Table 1. Benchmark results of PSAR-Align</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input MSA</td>
</tr>
<tr>
<td>------------</td>
</tr>
<tr>
<td>Original$^b$</td>
</tr>
<tr>
<td>ClustalW</td>
</tr>
<tr>
<td>MAFFT</td>
</tr>
<tr>
<td>MAVID</td>
</tr>
<tr>
<td>MUSCLE</td>
</tr>
<tr>
<td>Pecan</td>
</tr>
</tbody>
</table>

Note: Better scores between original and PSAR-Align are shown in bold.
$^a$Input MSA to PSAR-Align.
$^b$Revised MSA of the original by PSAR-Align.
$^c$Average across 1000 replicates with 95% confidence interval in parentheses.

REFERENCES


