GLProbs: Aligning multiple sequences adaptively

Yongtao Ye*
University of Hong Kong
ytye@cs.hku.hk

David W. Cheung*
University of Hong Kong
dcheung@cs.hku.hk

Yadong Wang
Harbin Institute of Technology
ydwang@hit.edu.cn

Siu-Ming Yiu*
University of Hong Kong
smyiu@cs.hku.hk

Qing Zhan
Harbin Institute of Technology
cyanzhan@gmail.com

Tak-Wah Lam*†
University of Hong Kong
twlam@cs.hku.hk

Hing-Fung Ting*†
University of Hong Kong
hfting@cs.hku.hk

ABSTRACT

This paper proposes a simple and effective approach to improve the accuracy of multiple sequence alignment. We use a natural measure to estimate the similarity of the input sequences, and based on this measure, we align the input sequences differently. For example, for inputs with high similarity, we consider the whole sequences and align them gobally, while for those with moderately low similarity, we may ignore the flank regions and align locally. To test the effectiveness of this approach, we have implemented a multiple sequence alignment tool call GLProbs, and compares its performance with a dozen leading alignment tools on three benchmark alignment databases. Our results shows that GLProbs has the best accuracy for almost all testings.

1. INTRODUCTION

The homogeneity of a set of biological sequences often implies functional similarity or divergence from a common ancestor, and the most common way to find out how homogenous the sequences are is to align them, i.e., to organize homologous positions across different sequences in columns. This process of multiple sequence alignment also helps biologists to isolate the most relevant regions in the sequences, and this is important to various analyses such as secondary structure prediction and phylogenetic trees construction. During the last two decades, there were a lot of software tools developed for multiple sequence alignment; however, all of them have their own weaknesses and perform poorly on some particular types of inputs. In particular, when sequence similarity falls below 25%, the accuracies of most multiple sequence alignment tools drop considerably.

*HKU-BGI Bioinfomatics Algorithms & Core Technology Research Laboratory, and Department of Computer Science.
†Joint corresponding authors.

ACM-BCB 2013 Washington, DC, USA

This paper proposes a simple, but surprisingly effective approach for improving the overall quality of multiple sequence alignment. We note that there are two major ways to align sequences, namely global alignment and local alignment. If the sequences in a family are similar, there would be few ambiguous positions and we should take the whole sequences into consideration and align them globally. If the sequences are not similar, they may still contain similar patterns (or motifs) over some local regions; in such case we should align them locally to find these similar local regions. Furthermore, sequences with different similarities have different features or properties that would help us align them with higher accuracies. This suggests an adaptive approach to improve the alignment quality: we first decide the similarity of the sequences, and if they are similar, we focus on global alignments, otherwise we focus on local alignments.

This approach has a fundamental difficulty: we need to know the correct multiple sequence alignment in order to determine how similar the sequences are. For example, the similarity scores given in benchmark alignment databases such as BAliBASE [27], OXBench [19], and SABmark [29] are computed based on some given multiple sequence alignments, whose correctness have been verified by biological methods. In this paper, we suggest a natural measure to estimate the similarity of a family of sequences without the need to know their correct multiple sequence alignment. Our measure is based on the percent identity (PID) of a pair of sequences, which is the percentage of the number of homologous positions in the alignment of these two sequences. To estimate the similarity of a family of sequences, we align every pair of its sequences, compute their PIDs, and use the average of these PIDs to estimate the similarity of the sequences. We believe that for a family of sequences with high similarity, most pairs of sequences in the family will be unambiguous and have high PIDs, and thus the average PID will be high; otherwise the average PID will be low.

To study the relevance of this "average PID" measure and the effectiveness of our approach, we have developed an adaptive alignment tool called GLProbs. We have tested it extenstively and compared its results with those of many leading multiple sequence alignment tools. We find that GLProbs's performance is significantly better. For example, as shown in Table 2, GLProbs has the best overall accuracies in all the testings.

Since the objective of our current study is not to develop

another alignment tool for the general public, we did not implement GLProbs from sketch. Instead, we developed GLProbs from the open source codes of MSAProbs-0.9.7. MSAProbs [15] is a tree-based progressive alignment tool based on the pair-Hidden Markov model (pair-HMM). Similar to other progressive alignment tools [11], MSAProbs aligns the sequences using several pairwise alignment steps, and the most related sequences are to be aligned first and the more distant ones are aligned later. To align two sequences x and y, the substitution scores used by MSAProbs are computed based on the posterior probabilities $\Pr(x_i \sim y_j \mid x, y)$, which is the probability that positions x_i and y_j of x and y will be matched under the condition that the pair-HMM has generated an alignment for x and y [6]. We will give more details on the implementation of GLProbs in Section 2. Below, we highlight some of its important features.

- Given a family of sequences as input, GLProbs will first align every pair of the sequences and compute their PIDs. The average $\overline{\text{PID}}$ will help GLProbs decide how to align the sequences, or more precisely, how to compute the posterior probabilities $\Pr(x_i \sim y_j \mid x, y)$ as follows:
  - $\overline{\text{PID}} > 40\%$. It uses the standard three-state global pair-HMM (see [8, 21]) to generate the posterior probabilities.
  - $25\% < \overline{\text{PID}} \leq 40\%$. It uses the local pair-HMM shown in Figure 1(a) to generate the posterior probabilities.
  - $\overline{\text{PID}} \leq 25\%$. In this case, the sequences in the family are so different that there may not even be any conserved local regions, and we are not sure what the right way is to align them. Thus, for this case, we resort to consensus; GLProbs computes more than one posterior probabilities using different models, and then uses their root-mean-squares to compute the subsitution scores.
- The local pair-HMM shown in Figure 1(a) has also been used in some earlier alignment tools such as ProDA [18] and CONTRAlign(local model) [7]. We note that these alignment tools may return poor alignments even for family with moderately low similarity. The main reason is that they also assign scores to leading and tailing flanking regions (i.e., the unaligned segments at the beginning, and at the end of the local alignment). To make the alignment process focus on the local conserved regions, we need to remove the "noises" of the flanking regions. To this end, GLProbs applies the standard technique of coupling the local pair-HMM with a random pair-HMM as shown in Figure 1(b), and using the log-odds ratios derived from the two models to determine the posterior probabilities (for details, see [1, 8]).
- We used some common substitution matrices and the standard unsupervised EM method to determine respectively the state emissions and state transitions probabilities for the pair-HMMs, except that we determine the state transition probability η (see Figure 1(b)) somewhat differently. We observe that (i) the similarity of the sequences has a notable effect on the length of the flanking regions (the less similar the sequences, the longer the flanking regions in their alignment), and (ii) the transition probability η has significant effect on the length of the flanking regions of the local alignments generated by GLProbs. Therefore, we use different values of η for families with different similarities. We have prepared a set of different η by trainings on families with different ranges of similarity. Then, when aligning a family of sequences, we use its $\overline{\text{PID}}$ to help us choose a suitable η so that GLProbs would handle the flanking regions more appropriately.

We have compared GLProbs with many leading multiple sequence alignment tools including ClustalW [28], T-Coffee [16], Mafft [13], Align_m [29], MUSCLE [10], ProbCons [6], Probalign [21], COBALT [17], MSAProbs [15], ClustalΩ [22] and CONTRAlign(local model) [7] using the benchmark alignments databases BAliBASE, OXBench and SABmark. GLProbs achieves the highest alignment accuracy and is statistically ranked as the best. In particular, GLProbs outperformed the other tools significantly for divergent sequences. For example, GLProbs gets a 10% improvement of TC score over ClustalW for families of sequences in OXbench with similarity between 0-20% (see Figure 2). We have also compared these tools on two biological applications, namely secondary structure prediction and phylogenetic analysis, and our results show that GLProbs had better performances as well. Details of these empirical comparisons will be given in Section 3. For verification of our results, GLProbs can be downloaded via the link `http://glprobs.sourceforge.net`, and the benchmark alignments data can be accessed from `http://www.drive5.com/bench`.

2. METHODS

2.1 The transition probability η

As mentioned earlier, we determine the transition probability η, which is the probability for State RX and for State RY to leave themselves, somewhat differently. GLProbs uses different values of η for families with different similarities so that the more divergent the sequences the longer the flanking regions. Note that GLProbs needs the value of η when the $\overline{\text{PID}}$ of the input sequences is in the range $[0, 40\%]$. We partitioned this range into 6 different subranges, namely [0-15%], (15%,20%], (20%,25%], (25%,30%], (30%,35%], (35%,40%], and for each of these subranges R_i, we prepare a data set D_i comprising families of sequences obtained from the benchmeark alignment database SAmarks whose similarities fall in that subrange. Then, we applied the unsupervised EM method [8] on each D_i to determine the value of η, which will be used in the alignment process when the input's $\overline{\text{PID}}$ falls in R_i.

2.2 GLProbs Algorithm

We now describe the alignment algorithm of GLProbs. Given the input sequences, GLProbs produces their multiple sequence alignment using the following six steps.

Step 1: Determine the Model

For every pair x, y of the input sequences, GLProbs finds their Viterbi pairwise alignment to compute their PID, which

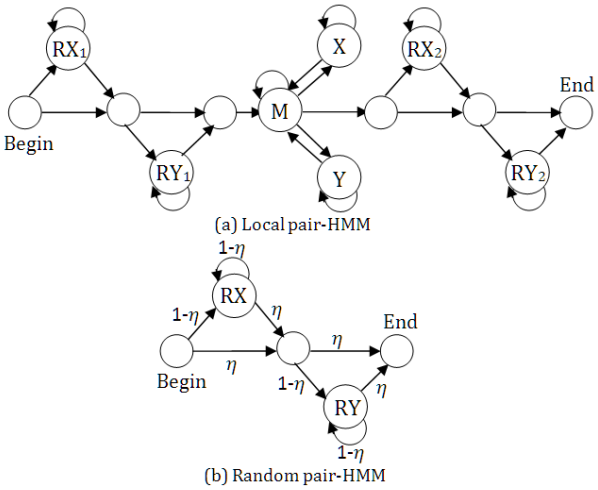


Figure 1: (a) In the model, State M emits two characters, one for sequence x and the other for sequence y, and they correspond to two characters being aligned together. State X emits a character in sequence x that is aligned to a gap, and similarly state Y emits a character in sequence y that is aligned a gap. States RX_1 and RY_1 emit two unaligned flanking subsequences on the left of the local alignment. Similarly, states RX_2 and RY_2 emit two unaligned flanking subsequences on the right of the local alignment. (b) State RX and RY emit two sequences in turn, independently to each other. Each of them has a probability η to leave itself.

is defined to be

$$\text{PID} = \frac{N\_Identity}{L\_Alignment}$$

where $N\_Identity$ is the number of identities in the pairwise alignment, and $L\_Alignment$ is the length of alignment. As mentioned in Section 1, if the average $\overline{\text{PID}}$ is greater than 40%, GLProbs uses the global pair-HMM, and if $\overline{\text{PID}}$ is in (25%-40%], it use the local pair-HMM to generate the posterior probabilites. We now explain how to handle the case when $\overline{\text{PID}}$ is smaller than or equal to 25%.

For this case, GLProbs uses the global pair-HMM and the local pair-HMM to generate respectively the posterior probabilities $u = \Pr_{\text{global}}(x_i \sim y_j \mid x, y)$ and $v = \Pr_{\text{local}}(x_i \sim y_j \mid x, y)$. In addition, it also uses the *double affine pair-HMM* proposed in [5] to generate the probabilities $w = \Pr_{\text{affine}}(x_i \sim y_j \mid x, y)$.[1] Then, the posterior probabilities used for this case is given by

$$\Pr(x_i \sim y_j \mid x, y) = \sqrt{(u^2 + v^2 + w^2)/3}.$$

Table 1 summarizes our scheme.

Step 2: Compute the Pairwise Distances

Given the posterior probabilities obtained in Step 1 as substitution scores, GLProbs applies the Needleman-Wunsch algorithm (without gap penalty) to compute, for every pair x, y of the input sequences, the maximum expected accuracy

$$E(x, y) = \max_a\{\textstyle\sum_{x_i \sim y_j \in a} \Pr(x_i \sim y_j \mid x, y)\},$$

where the maximum is over all possible alignments a of x and y (for details, see [8]). Then, it determines their pairwise distance d_{xy}, which is given by

$$d_{xy} = 1 - \frac{E(x, y)}{\min\{|x|, |y|\}}.$$

[1]The double affine pair-HMM is similar to the three-state global model, except that there is an extra pair of gap states for long insertions and deletions.

Table 1: The scheme for determining the model

Category	$\overline{\text{PID}}$	Model	Posterior Probability
Divergent	$\leq 25\%$	Combination of global, local and double affine pair-HMMs	$\sqrt{\frac{u^2+v^2+w^2}{3}}$
Medium	25%-40%	Local pair-HMM	v
Similar	>40%	Global pair-HMM	u

Step 3: Construct a Guide Tree

Based on the pairwise distances computed in Step 2, GLProbs applies the greedy linear heuristic UPGMA [23] to construct a guide tree. During the construction, we use the following definition of *distance between two clusters of sequences*: For any two clusters of sequences C_k and C_l, if $C_k = C_i \cup C_j$ is the union of the two disjoint clusters C_i and C_j, then the distances d_{kl} between the C_k and C_l is defined recursively to be

$$d_{kl} = \frac{d_{il}|C_i| + d_{jl}|C_j|}{|C_i| + |C_j|}$$

where $|C_i|$ and $|C_j|$ are the number of sequences in cluster C_i and C_j, respectively.

Step 4: Transform the Probabilities for Consistency

GLProbs adjusts the posterior probabilities $\Pr(x_i \sim y_j \mid x, y)$ for every pair of input sequences x and y by considering the similarity of x and y to other sequences in the input. To be precise, let P_{xz} and P_{zy} be the posterior probabilities matrices for sequences x, z and sequences z, y, respectively. Then the adjusted posterior probabilities matrix P'_{xy} for x, y is given as follows:

$$P'_{xy} \leftarrow \tfrac{1}{|S|} \textstyle\sum_{z \in S} P_{xz} P_{zy}$$

where S is the set of input sequences. These adjusted posterior probabilities will be used to determine the substitution scores in the following step.

Step 5: Weighted Progressive Alignment

This step obtain a multiple sequence alignment of the input sequences by performing weighted profile-profile alignments iteratively following the order suggested by the guide tree. The weighting scheme used is similar to that used in ClustalW [28], which avoids biased sampling of sequences.

Step 6: Final Refinement

The purpose of this step is to try randomly to improve the accuracy of the alignment by correcting the mis-aligned positions. It is executed only for input with $\overline{\text{PID}}$ less than 70%

because we find from our empirical testings that the refinement does not help for sequences with high similarity.

During this step, we iteratively divides the multiple sequence alignment into two random groups (each sequence will be assigned to the two groups with equal probability), and re-align them using the standard unweighted profile-profile procedure to see if we can make any improvement. Given an input family of N sequences, we stop the iterations when one of the following conditions is ture:

- There are $2N$ iterations in which we cannot find any improvement.
- We have iterated $10N$ times.

3. RESULTS

To evaluate how good GLProbs is, we have compared it with other leading multiple sequence alignment tools by using them to align families of sequences obtained from some popular benchmark alignment databases and then comparing the sum-of-pairs score (SP) and total column score (TC) of their alignments. We have also used their alignemnts to perform two biological analyses, namely secondary structure prediction and phyogenetic analysis; by comparing the accuracies of these biological analyses, we get some more evidences on which tools are better.

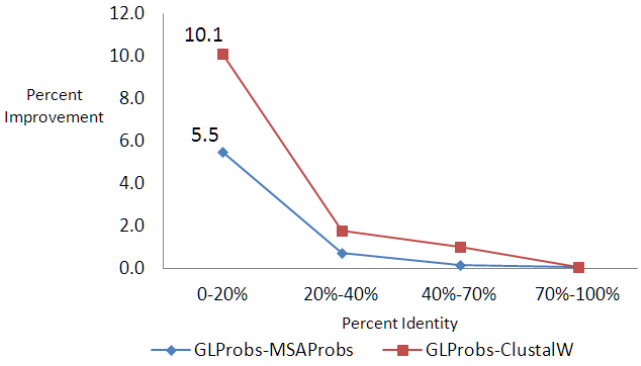
3.1 Alignment Accuracy

We have compared GLProbs with the following multiple sequence alignment tools: T-Coffee 9.03, ClustalW 2.1, Clustal Omega 1.1.0., COBALT, MAFFT 7.031, Align_m 2.3, MUSCLE 3.8.31, Contralign(local model) 2.01, ProbCons 1.12, Probalign 1.4 and MSAProbs 0.9.7. We have used these tools (with default parameters) to align families of sequences obtained from the following three benchmark alignment databases: OXBench1.3, SABmark1.65 and BAliBASE3.0 (downloaded from [9]).

The OXBench benchmark testing is shown in Table 2, in which the results are divided into five categories according to the similarities of the input families. For example, the two columns under the category "ALL(0-100%)" show the average SP and TC scores over all the input families used in the test, while the two under "(0-20%)" are those for families with similarities between 0 and 20%. Note that GLProbs achieves the highest (average) SP and TC scores for all five categories. In particular, as shown in Figure 2, for the category (0-20%), which corresponds to families of divergent sequences, GLProbs achieves an improvement of 5.5% in TC score over MSAProbs (with rank second) and 10.1% over ClustalW (the most widely used). On the other hand, when sequences are very similar (the (70-100%) category), there is not much difference among the tools, and a simple model (e.g. Probalign uses the simple three-state global pair-HMM) has already rather good accuracies.

Table 3 shows the average SP and TC scores for the SABmark1.65 and BAliBASE3.0 benchmarks, in which the results are divided into two categories according to the similarity of the input sequences. Again, for these two benchmark databases, GLProbs achieves the highest SP and TC scores on almost all test data, except that in the (0-30%) category for BAliBASE, Probalign has a better TC score. It is reported by Subramanian et al. [24] that "BAliBASE is heavily biased toward globally related protein families", and

Figure 2: The improvements of GLProbs over MSAProbs and ClustalW of TC score on OXBench



we wonder whether this is the reason why Probalign, which is based on global pair-HMM, has a better TC score.

Table 4 compares the running times of the tools on a single-core processor. We note that we did not pay much efforts in optimizing GLProbs for efficiency because our focus is on accuracies.

In Table 5, we have calculated the P-value using Friedman rank test [12] for revealing the statistical significance for GLProbs to other alignment tools on OXBench, SABmark and BAliBASE. We note that GLProbs achieves statistically significant accuracy improvement over all the other tools on both SP and TC scores.

3.2 Application to Biological Analyses

We have also tested how useful the tools are in real applications by applying them to two biological analyses, namely secondary structure prediction and phylogentic analysis.

Secondary Structure Prediction

Secondary structure is the general three-dimensional form of local segments of biopolymers such as proteins and nucleic acids (DNA/RNA). For proteins, a prediction is to identify regions alpha helices, beta strands or 'coil'.

In our test, we made use of the common tool JPRED3 [4] to predict secondary structures from multiple sequence alignments given by the alignment tools on the same sets of sequences. The first sequence in the alignment is the one that JPRED uses to project its secondary structuret. We tested two query sequences, Lipid Binding Protein (PDB ID: 1U27) [26] and Nuclear Receptor (PDB ID: 1LBD) [2] whose structure have been completely determined. We compared the results to the known structure with the number of residues whose prediction does not agree with reference (mismatch) quantitatively; the fewer the number of mismatches, the better the alignment.

Firgure 3 shows the secondary structure predictions of 1U27 and 1LBD from different alignments of the same sets of sequences which can be downloaded from Aidan Budd et al. [3]. Sample details are described in PanelA: 100 PH-domain sequences with less than 25% identity and 148 SR-domain sequences with 25%-40% identity. PanelB and PanelC are the secondary structure predictions of 1U27 and 1LBD respectively. Table 6 reports the results that the

Table 2: Mean SP and TC scores on OXBench

	ALL(0-100%)		0%-20%		20%-40%		40%-70%		70%-100%	
	SP	TC	SP	TC	SP	TC	SP	TC	SP	TC
GLProbs	**90.38**	**82.14**	**47.29**	**22.95**	**78.41**	**60.82**	**94.72**	**88.29**	**99.32**	**98.14**
ClustalΩ	88.91	79.99	39.09	16.38	75.31	56.14	93.79	86.73	99.26	97.74
MSAProbs	90.07	81.75	44.83	22.08	77.84	59.82	94.56	88.03	99.26	98.09
COBALT	88.96	79.73	39.41	15.33	75.91	56.79	93.65	86.10	99.14	97.50
Probalign	89.97	81.68	43.58	20.51	77.26	59.38	94.69	88.20	**99.32**	**98.14**
CONTRAlign	89.34	79.87	44.76	17.83	76.54	56.71	93.62	86.01	99.28	97.84
ProbCons	89.68	80.86	44.15	20.30	77.05	58.34	94.22	87.26	99.14	97.62
MUSCLE	89.50	80.67	45.64	21.90	76.97	58.56	93.72	86.51	99.20	97.81
Align_m	86.95	76.06	28.36	12.74	70.62	49.76	91.61	82.02	99.06	97.18
MAFFT	88.00	77.96	37.82	13.27	73.19	51.89	93.07	84.83	99.09	97.52
T-Coffee	89.52	80.50	43.99	19.11	76.66	57.89	94.14	86.94	99.07	97.40
ClustalW	89.43	80.16	42.94	18.23	77.05	57.26	93.75	86.31	99.24	97.89

Columns show the average sum of pairs scores (SP) and total column scores (TC) multiplied by 100. The best results in each column are shown in bold.

Table 3: Mean SP and TC scores on SABmark and BAliBASE

	SABmark						BAliBASE					
	ALL(0-60%)		0%-30%		30%-60%		ALL(0-60%)		0%-30%		30%-60%	
	SP	TC	SP	TC	SP	TC	SP	TC	SP	TC	SP	TC
GLProbs	**61.42**	**41.36**	**53.40**	**30.57**	**91.09**	**81.26**	**83.20**	**67.59**	**70.08**	46.24	**94.00**	**85.15**
ClustalΩ	55.02	35.47	46.08	24.68	88.12	75.40	75.96	59.38	59.83	38.00	89.22	76.95
MSAProbs	60.27	40.02	52.03	29.02	90.76	80.74	82.35	66.83	68.54	45.78	93.71	84.13
COBALT	56.71	36.00	48.68	26.08	86.44	72.74	76.08	57.49	60.42	36.88	88.97	74.43
Probalign	59.53	38.63	51.16	27.70	90.51	79.04	82.53	67.27	69.83	**46.81**	93.82	84.09
CONTRAlign	57.45	35.59	49.91	25.97	85.36	71.19	77.59	58.10	62.29	36.50	90.16	75.85
Probcons	59.69	39.17	51.46	28.20	90.14	79.75	81.55	65.22	67.32	43.05	93.25	83.44
MUSCLE	54.51	33.47	46.37	23.60	84.59	69.99	75.60	58.27	57.23	33.07	90.71	78.98
Align_m	46.19	31.07	35.57	19.58	85.50	73.56	71.45	56.04	52.08	34.79	87.38	73.51
MAFFT	52.63	32.57	43.51	21.58	86.37	73.22	72.46	52.58	53.33	22.76	88.19	73.00
T-Coffee	59.14	39.53	50.77	28.54	90.12	80.18	80.82	64.93	65.79	42.43	93.18	83.43
ClustalW	51.92	31.37	43.86	21.86	81.74	66.56	69.63	49.21	51.57	26.01	84.47	68.28

Columns show the average sum of pairs scores (SP) and total column scores (TC) multiplied by 100. The best results in each column are shown in bold. We also classified SABmark into "Twilight Zone" and "Superfamily" subsets, and BAliBASE into BB11 and BB12 subsets, so that they are compatible with the testing datum from other papers [21], [15] and [22] (see Table 8 at the end of the paper).

Table 4: Running times (mm:ss) on OXBench, SABmark and BAliBASE

	GLProbs	MSAProbs	COBALT	Probalign	CONTRAlign	ProbCons	MUSCLE	Align_m	MAFFT	T-Coffee	ClustalΩ	ClustalW
OXBench	3:39	4:04	4:08	2:10	10:19	1:48	0:19	21:14	0:19	15:05	0:12	0:22
SABmark	3:28	1:58	4:34	1:01	4:56	1:12	0:46	5:32	0:22	4:36	0:18	0:14
BAliBASE	4:14	3:02	2:03	1:47	6:37	1:41	0:37	7:09	0:14	5:18	0:21	0:21

Tests were carried out on an Intel i7 single-core 3.20 GHz processor with 64GB RAM.

Table 5: P-values on OXbench, SABmark and BAliBASE

	ClustalΩ	MSAProbs	COBALT	Probalign	CONTRAlign	ProbCons	MUSCLE	Align_m	MAFFT	T-Coffee	ClustalW
SP	< 0.0001	0.0040	< 0.0001	0.0008	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
TC	< 0.0001	0.0028	< 0.0001	0.0014	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001

Entries show the P-values between GLProbs and other tools using Friedman rank test. The difference is considered significant if P-values < 0.05.

multiple alignments of GLProbs predicts the most accurate secondary structures to 1U27 and 1LBD.

Table 6: Mismatches of secondary structure predictions

	GLProbs	MSAProbs	COBALT	MUSCLE	MAFFT	T-Coffee	ClustalW
1U27	**19**	22	22	23	29	23	30
1LBD	**39**	41	47	44	44	42	45

The minimum mismatches in each row are shown in bold.

Phylogenetic Analysis

Finding a correct evolutionary tree is essential for the study of evolutionary relationships among groups of organisms. We tested the alignment tools on six protein sequences families from TreeFam database [14] with phylogeny reconstruction using the Maximum Likelihood approach of MEGA5 [25]. In Figure 4, we show the phylogenetic trees of one of the six families, namely TF105801 with 27 sequences. We also measured the distances between the constructed trees and the reference by Robinson-Foulds metric [20]; the smaller the distance, the closer to the true phylogenetic for a tested tree, which also suggests better performance of a multiple sequence alignment. Table 7 compares the distances and indicates that the phylogenetic trees derived from GLProbs are the most accurate ones in almost all tests..

Table 7: Distances of different computed phylogenetic trees

TreeFamID	GLProbs	ClustalΩ	MSAProbs	MUSCLE	T-Coffee
TF105801(27)	**0.58**	0.63	0.79	0.67	0.79
TF105629(88)	**0.62**	0.68	0.67	0.66	0.65
TF105311(70)	**0.64**	0.66	0.67	0.67	0.70
TF101222(48)	0.69	0.78	**0.67**	**0.67**	0.76
TF105820(86)	**0.72**	0.75	0.82	**0.72**	0.82
TF105063(133)	**0.82**	0.84	0.85	0.83	0.84

The minimum distances in each row are shown in bold. The numbers of sequences in each queried family are shown in parentheses.

4. CONCLUSIONS

The major insight of GLProbs is to use different methods to align families of sequences with different similarities. As a hindsight after our study, it seems obviously not very reasonable to model all kinds of sequences in the same way. This suggests that taking different kinds of biological feature into consideration will be important to tools for other problems (e.g. motif finding).

We would also like to point out that even though we have not optimized GLProbs for efficiency, GLProbs' running time is still comparable with many tools because when the input sequences are quite similar (more than 25% identity), GLProbs only applies one simple model, local or global, both of which only need three-states computation.

5. ACKNOWLEDGMENTS

Our program is developed based on the open source code of MSAProbs (available at `http://msaprobs.sourceforge.net`), and we thank the developers of MSAProbs. Also, we thank Ruibang Luo for many helpful discussions.

6. REFERENCES

[1] M. Borodovsky and S. Ekisheva. *Problems and Solutions in Biological sequence analysis*. Cambridge University Press, Cambridge, UK, 2006.

[2] W. Bourguet and M. Ruff *et al*. Crystal structure of the ligandbinding domain of the human nuclear receptor rxr-alpha. *Nature*, 375:377–382, 1995.

[3] A. Budd and D. Judge. Applications of msa- practical. `http://www.embl.de/~seqanal/MSAcambridgeGenetics2007/MSAapplications/MSAappnMSA2007.htm`.

[4] C. Cole, J. Barber, and G. Barton. The JPRED 3 secondary structure prediction server. *Nucleic Acids Res*, 36:W197–W201, 2008.

[5] C. Do, M. Brudno, and S. Batzoglou. ProbCons: Probabilistic consistency-based multiple alignment of amino acid sequences. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pages 703–708, San Jose, CA, 2004. AAAI Press.

[6] C. Do *et al*. ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res*, 15:330–340, 2005.

[7] C. Do *et al*. CONTRAlign: discriminative training for protein sequence alignment. In *In Proceedings of the Tenth Annual International Conference on Computational Molecular Biology (RECOMB)*, pages 2–5. Venice, Italy, April 2006.

[8] R. Durbin *et al*. *Biological sequence analysis*. Cambridge University Press, Cambridge, UK, 1998.

[9] R. Edga. Bench. `http://www.drive5.com/bench`.

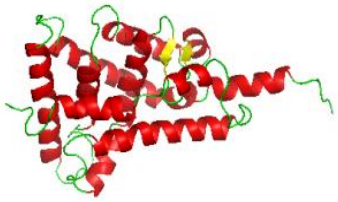
[10] R. Edgar. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32:1792–1797, 2004.

[11] D. Feng and R. Doolittle. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.*, 15:351–360, 1987.

[12] M. Friedman *et al*. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.*, 32:675–701, 1937.

Figure 3: Secondary structure predictions of 1U27 and 1LBD

A

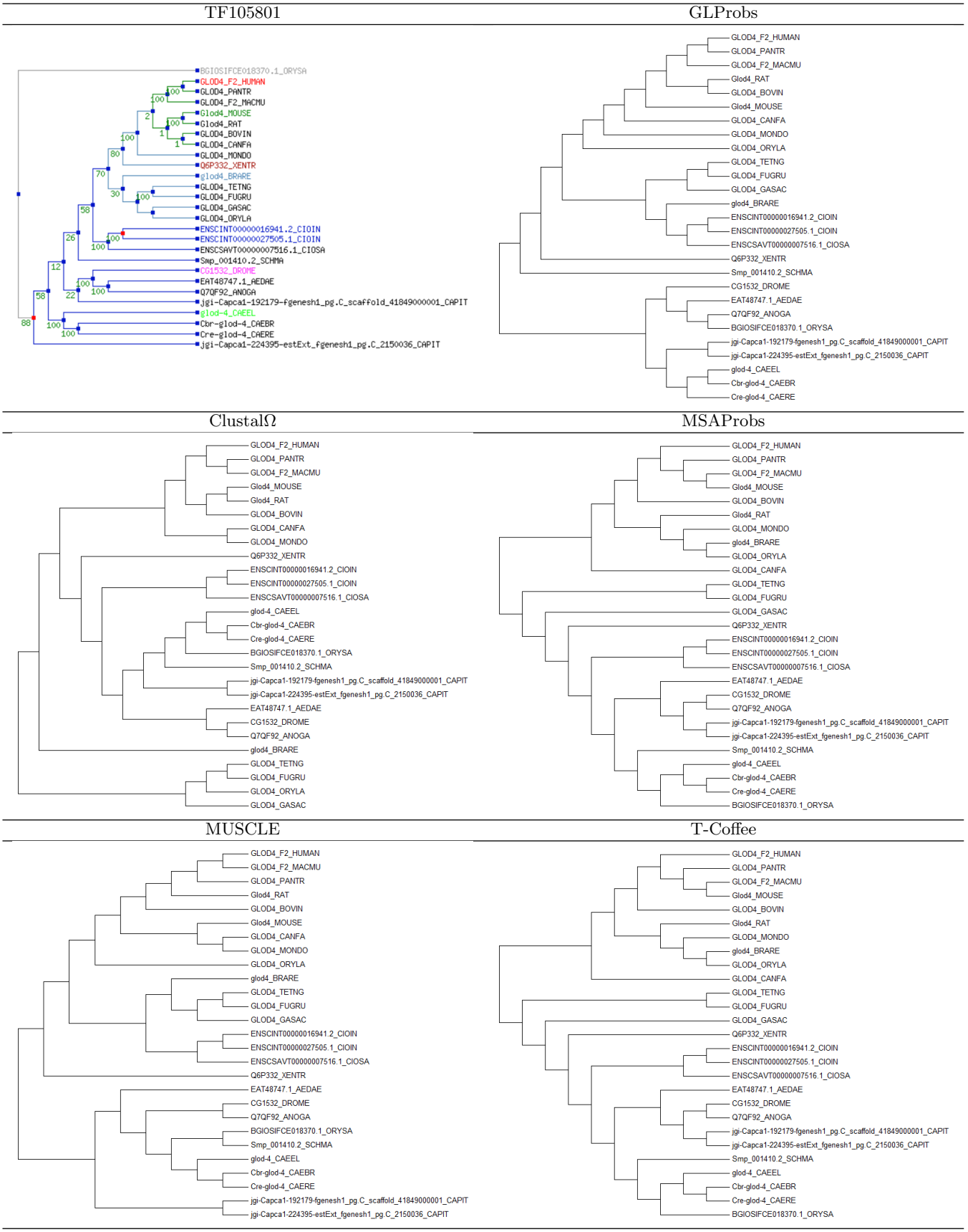
Query Sequence	Sequence length	Number of sequences in MSA	Average percent identity of MSA	Domain	3D Structure
1U27	129	100	<25%	PH	
1LBD	238	148	25%-40%	SR	

B

```
1U27_seq   MGHHHHHHGSPDREGWLLKLGGGRVKTWKRRWFILTDNCLYYFEYTTDKEPRGIIPLENLSIREVDDPRKPNCFELYIPNNKG
1U27       ------------EEEEEEEE-------EEEEEEEE---EEEEE---------EEEE----EEEE---------EEEEE------
GLProbs    ----------EEEEEEEEEEE--------EEEEEEE--EEEEEE--------EEE-----EEE---------EEEEEE-----
MSAProbs   ----------EEEEEEEEEEE-------EEEEEEEE--EEEEEE---------EE----EEEE---------EEEEEE-----
COBALT     ----------EEEEEEEEEEEE------EEEEEEE---EEEEE-----------------EEE--------EEEEEE------
MUSCLE     -----------EEEEEEEE--------EEEEEEEEE--EEEEEE-------------EEEEEE--------EEEEEE------
MAFFT      -----------EEEEEEEEE-------EEEEEEEEE--EEEEEE---------EEEEEE-------------------------
T-Coffee   ----------EEEEEEEEEE---------EEEEEEE--EEEEEE---------EEE---EEEEE--------EEEEEE-----
ClustalW   ----HHH----EEEEEEEEE---------EEEEEEE----EE------------EE----EEE---------------------

1U27_seq   QLIKACKTEADGRVVEGNHMVYRISAPTQEEKDEWIKSIQAAVSVD
1U27       -----EEE-----EEE----EEEEE--HHHHHHHHHHHHHHHHH--
GLProbs    --EEEE--------------EEEEE---HHHHHHHHHHHHHHHH--
MSAProbs   -EEE---------------EEEEEE---HHHHHHHHHHHHHHHH--
COBALT     -EEE---------------EEEEEE---HHHHHHHHHHHHHHHH--
MUSCLE     EEEEEEEEEE----------EEEEE---HHHHHHHHHHHHHHHH--
MAFFT      -------EEEEEEE------EEEEEE--HHHHHHHHHHHHHHHH--
T-Coffee   -EEHHH--HH---------EEEEEE---HHHHHHHHHHHHHHHH--
ClustalW   --------EEEEE-------EEEEE---HHHHHHHHHHHHHHHH--

C
1LBD_seq   SANEDMPVERILEAELAVEPKTETYVEANMGLNPSSPNDPVTNICQAADKQLFTLVEWAKRIPHFSELPLDDQVILLRAGWNE
1LBD       -------HHHHHHHHHHH---HHHHHHH-----------HHHHHHHHHHHHHHHHHHHHH--------HHHHHHHHHH-HHH
GLProbs    --------HH--HHHHH-------------------HHHHHHHHHHHHHHHHHHHHHHHH----------HHHHHHH---HH
MSAProbs   ----------HHHH----------------------HHHHHHHHHHHHHHHHHHHHHHHH----------HHHHHHH---HH
COBALT     ---------EEEE----------------------------HHHHHHHHHHHHHHHHHHH----------HHHHHHH---HH
MAFFT      -------HHHHHHHH----------HHH-----------HHHHHHHHHHHHHHHHHHHHH----------HHHHHHH--HHH
MUSCLE     -----EEEEEE-----------HHHHHHHH----------HHHHHHHHHHHHHHHHHHHH----------HHHHHHH--HHH
T-Coffee   -----EEEEEEEEH----------------------HHHHHHHHHHHHHHHHHHHHHHHH----------HHHHHHH--HHH
ClustalW   ------HHHHHH----------------------HHHHHHHHHHHHHHHHHHHHHHHH----------HHHHHHH---HH

1LBD_seq   LLIASFSHRSIAVKDGILLATGLHVHRNSAHSAGVGAIFDRVLTELVSKMRDMQMDKTELGCLRAIVLFNPDSKGLSNPAEVE
1LBD       HHHHHHHHHHH---EEE-----EEEHHHHHH---HHHHHHHHHHHHH-------HHHHHHHHHHH----------HHHHH
GLProbs    HHHHHHHHH-------EEE----EE---HHHH---HHHHHHHHHHHHHH------HHHHHHHHHHHH---------HHHHH
MSAProbs   HHHHHHHHH-----EEEE----EE-----------HHHHHHHHHHHHH------HHHHHHHHHHH---------HHHHH
COBALT     HHHHHHHHH-------EEE----E-----------HHHHHHHHHHHHH------H-HHHHHHHHH----------HHHHH
MAFFT      HHHHHHHHH-----EEEE--------HHHHHHHH----HHHHHHHHHH------HHHHHHHHHHH---------HHHHH
MUSCLE     HHHHHHHHH-------EEE---------H------HHHHHHHHHHHHHH------HHHHHHHHHHH---------HHHHH
T-Coffee   HHHHHHHHH-------EEE----EE-HHHHHH----HHHHHHHHHHHHH------HHHHHHHHHHH---------HHHHH
ClustalW   HHHHHHHHHH----EEEE----EE-----------HHHHHHHHHHHHH------HHHHHHHHHHH---------HHHHH

1LBD_seq   ALREKVYASLEAYC KHKYPEQPGRFAKLLLRLPALRSIGLKCLEHLFFFKLIGDTPIDTFLMEMLEAPHQMT
1LBD       HHHHHHHHHHHHHHH--------HHHHHHHHHHHHHHHHHHHH-HHHHHH-HHHHHHHHHHH------
GLProbs    HHHHHHHHHHHHHH-------HHHHHHHHHHHHHHHHHHHHHHHHEE-------HHHHHHHH------
MSAProbs   HHHHHHHHHHHHHHH-------HHHHHHHHHHHHHHHHHHHHHHHHHH------HHHHHHHH------
COBALT     HHHHHHHHHHHHHHH-------HHHHHHHHHHHHHHHHHHHHHHHHH-------HHHHHHHH------
MAFFT      HHHHHHHHHHHHHHH-------HHHHHHHHHHHHHHHHHHHHHHHEEEE------HHHHHHHH------
MUSCLE     HHHHHHHHHHHHHHH-------HHHHHHHHHHHHHHHHHHHHHHHEE-------HHHHHHHH------
T-Coffee   HHHHHHHHHHHHHHH-------HHHHHHHHHHHHHHHHHHHHHHHH-------HHHHHHHH------
ClustalW   HHHHHHHHHHHHHHH--------HHHHHHHHHHHHHHHHHHHHHHH----------HHHHHHH-----
```

1U27_seq and 1LBD_seq are the query protein sequences. 1U27 and 1LBD denote the solved secondary structures. The names of aligners stand for the secondary structure predicted through the MSAs constructed by themselves.'E','H' and '-' represent extended, helical and other types of secondary structure respectively.

Figure 4: Phylogenetic trees of TF105801



TF105801 is the reference phylogenetic tree. The name of the tools stand for the phylogenetic trees computed through the MSAs constructed by themselves.

Table 8: Mean SP and TC scores on SABmark and BAliBASE

	SABmark						BAliBASE					
	ALL		Twilight Zone		Superfamily		ALL		BB11		BB12	
	SP	TC	SP	TC	SP	TC	SP	TC	SP	TC	SP	TC
GLProbs	**61.42**	**41.36**	**44.35**	**24.30**	**67.27**	**47.21**	**83.20**	**67.59**	**69.72**	44.68	**94.84**	**87.38**
ClustalΩ	55.02	35.47	35.55	18.10	61.69	41.42	75.96	59.38	59.01	36.21	90.60	79.38
MSAProbs	60.27	40.02	42.97	22.88	66.20	45.90	82.35	66.83	68.13	44.02	94.63	86.52
COBALT	56.71	36.00	39.25	19.58	62.69	41.64	76.08	57.49	59.29	34.58	90.58	77.27
Probalign	59.53	38.63	42.42	22.64	65.39	44.11	82.53	67.27	69.50	**45.34**	94.63	86.20
CONTRAlign	57.45	35.59	39.01	17.69	63.77	41.73	77.59	58.10	61.78	35.60	91.23	77.52
ProbCons	59.69	39.17	42.81	22.78	65.47	44.79	81.55	65.22	66.99	41.68	94.12	85.54
MUSCLE	54.51	33.47	34.69	16.96	61.29	39.13	75.60	58.27	57.15	32.06	91.53	80.89
Align_m	46.19	31.07	25.72	16.28	53.21	36.14	71.45	56.04	51.88	33.06	88.36	75.88
MAFFT	52.63	32.57	31.72	15.17	59.79	38.53	72.46	52.58	52.96	26.19	89.30	75.38
T-Coffee	59.14	39.53	41.66	23.29	65.13	45.10	80.82	64.93	65.63	41.36	93.94	85.29
ClustalW	51.92	31.37	31.45	15.09	58.93	36.95	69.63	49.21	50.06	22.99	86.52	71.84

Columns show the average sum of pairs scores (SP) and total column scores (TC) multiplied by 100. The best results in each column are shown in bold. SABmark: The "Twilight Zone" represents different SCOP folds subsets, where each subset contains sequences within no more than 25% identity; The "Superfamily" contains different SCOP superfamilies, which are no more than 50% identity. BAliBASE: BB11 consists of very distant sequences with <20% identity and BB12 consists of medium to divergent sequences with identities from 20% to 40%.

[13] K. Katoh and K. Misasa *et al*. MAFFT: A novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res*, 30:3059–3066, 2002.

[14] H. Li and A. Coghlan *et al*. TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.*, 34:572–580, 2006.

[15] Y. Liu, B. Schmidt, and D. Maskell. MSAProbs: multiple sequence alignment based on pair hidden markov models and partition function posterior probabilities. *Bioinformatics*, 26:1958–1964, 2010.

[16] C. Notredame, D. Higgins, and J. Heringa. T-Coffee: A novel method for multiple sequence alignments. *J. Mol. Biol.*, 302:205–217, 2000.

[17] J. Papadopoulos and R. Agarwala. COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics*, 23:1073–9, 2007.

[18] T. Phuong *et al*. Multiple alignment of protein sequences with repeats and rearrangements. *Nucleic Acids Res*, 34:5932–42, 2006.

[19] G. Raghava *et al*. OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics*, 4:47, 2003.

[20] D. R. Robinson and L. R. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53:131–147, 1981.

[21] U. Roshan and D. Livesay. Probalign: multiple sequence alignment using partition function posterior probabilities. *Bioinformatics*, 22:2715–2721, 2006.

[22] F. Sievers and A. Wilm *et al*. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Mol. Syst. Biol.*, 22:539, July 2011.

[23] P. Sneath and R. Sokal. *Numerical taxonomy*. Freeman, San Francisco, USA, 1973.

[24] A. Subramanian *et al*. DIALIGN-T: An improved algorithm for segment-based multiple sequence alignment. *BMC Bioinformatics*, 6:66, 2005.

[25] K. Tamura and D. Peterson *et al*. MEGA 5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, 28:2731–2739, 2011.

[26] T.C.Cronin, J.P.DiNitto, M.P.Czech, and D.G.Lambrightr. Structural determinants of phosphoinositide selectivity in splice variants of grp1 family ph domains. *EMBO J*, 23:3711–3720, 2004.

[27] J. Thompson, F. Plewniak, and O. Poch. BAliBASE: A benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics*, 15:87–88, 1999a.

[28] J. Thompson *et al*. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22:4673–4680, 1994.

[29] V. Walle *et al*. Align-m: a new algorithm for multiple alignment of highly divergent sequences. *Bioinformatics*, 20:1428–1435, 2004.