Published in final edited form as:
Int J Data Min Bioinform. 2010 ; 4(1): 60–71.

Alignment of multiple proteins with an ensemble of Hidden Markov Models

Jia Song,
College of Electrical Engineering, Zhejiang University, China

Chunmei Liu,
Department of Systems and Computer Science, Howard University, Washington, DC 20059, USA

Yinglei Song,
Department of Mathematics and Computer Science, University of MD Eastern Shore, Princess Anne, MD, USA

Junfeng Qu, and
Department of Information Technology, Clayton State University, Morrow, GA, USA

Gurdeep S. Hura
Department of Mathematics and Computer Science, University of MD Eastern Shore, Princess Anne, MD, USA

Jia Song: sjia@zju.edu.cn; Chunmei Liu: chunmei@scs.howard.edu; Yinglei Song: ysong@umes.edu; Junfeng Qu: jqu@clayton.edu; Gurdeep S. Hura: gshura@umes.edu

Abstract

In this paper, we developed a new method that progressively construct and update a set of alignments by adding sequences in certain order to each of the existing alignments. Each of the existing alignments is modelled with a profile Hidden Markov Model (HMM) and an added sequence is aligned to each of these profile HMMs. We introduced an integer parameter for the number of profile HMMs. The profile HMMs are then updated based on the alignments with leading scores. Our experiments on BaliBASE showed that our approach could efficiently explore the alignment space and significantly improve the alignment accuracy.

Keywords

HMM; profile hidden Markov models; alignment; progressive

1 Introduction

The alignment of multiple protein sequences is an important problem in bioinformatics. In particular, given a set of protein sequences and a biological score function, the goal is to find an alignment of the sequences that optimises the value of the score function. Software tools for Multiple Sequence Alignment (MSA) have been extensively used to analyse protein sequences in biological research. For example, the phylogenetic relationships of a set of homologous sequences can often be inferred from an accurate alignment of these sequences

Correspondence to: Jia Song, sjia@zju.edu.cn; Chunmei Liu, chunmei@scs.howard.edu.

(Phillips et al., 2000). Moreover, MSA has been used to identify the motifs with certain biological functions in a set of homologous sequences (Thompson et al., 1999).

The optimal alignment of two protein sequences can be efficiently computed in quadratic time using a dynamic programming algorithm (Needleman and Wunsch, 1970). In general, MSA refers to the alignment of three or more protein sequences. The dynamic programming algorithm for aligning two sequences can be straightforwardly extended to align L sequences in time $O((2^L - 1)n^L)$ (Gupta et al., 1995; Lipman et al., 1989). However, this algorithm cannot be directly used to align multiple sequences in practice due to its high computational complexity. Therefore, a variety of heuristic methods and software tools (Thompson et al., 1994) have been developed to avoid the direct optimisation of the score function while efficiently generating accurate alignment results. Most of these methods belong to four categories: iterative, progressive, anchor-based and probabilistic.

Alignment tools using iterative methods include Muscle (Edgar, 2004) and DI-ALIGN (Morgenstern et al., 1998). Iterative methods start with an initial alignment and then iteratively refine the alignment until it cannot be further improved. In a single iteration, the alignment can be improved with a stochastic or deterministic approach. ClustalW (Thompson et al., 1994; Thomsen et al., 2003), T-coffee (Notredame et al., 2000), Treealign (Hein, 1989) and POA (Lee et al., 2002) use progressive methods for sequence alignment. In particular, these tools align the sequences in a set by repeatedly selecting two sequences from the set and replacing them with their alignments. The process terminates when the set contains only one ‘sequence’, which consists of a multiple alignment of all sequences in the set. MAFFT (Katoh et al., 2002), Align-m (Walle et al., 2004), L-align (Huang and Miller, 1991), PRRP (Gotoh, 1996), HSA (Zhang and Kahveci, 2006) are anchor-based alignment tools. These tools start the alignment by finding conserved local motifs in the sequences and use them as the anchors of the alignment. Regions between two anchors are then aligned to form an overall alignment of the sequences. Probcons (Chen, 2003), Hmmt (Eddy, 2003), SAGA (Notredame and Higgins, 1996) use probabilistic methods and models to describe the evolution of homologous sequences. Based on the substitution probabilities obtained from available multiple alignments, the alignment can be constructed by maximising the overall probability of substitutions.

So far, most of the existing alignment tools use methods designed to optimise the score function associated with an alignment and can only generate a single alignment. However, it has been pointed out recently that the alignment that optimises the score function may not be the one that is biologically most desirable. Therefore, a method that can generate a list of alignments with leading scores is more useful in practice. The major goal of this paper is to develop such a method. In particular, our method starts with the two sequences with the maximum similarity. A set of alignments with leading scores between the two can be computed with a dynamic programming algorithm. For each alignment in the set, a profile Hidden Markov Model (HMM) (Eddy, 2003) can be constructed to describe the statistical distribution of amino acids for each column in the alignment. An *ensemble* of profile HMMs can thus be constructed from the set of alignments with leading scores. The remaining sequences in the set are then progressively aligned to the profile HMMs in the ensemble one by one. After a given sequence is aligned to all profile HMMs in the ensemble, the statistical parameters in these profile HMMs are updated based on the alignments with leading scores. The process terminates when the set does not contain unaligned sequences. The alignments with leading scores are then reported as the possible alignments of the sequences.

We have developed a software tool, PALIGN to implement this method and tested its performance on BaliBASE benchmarks. We also compared its accuracy and efficiency with those of other alignment tools, including ClustalW, Probcons, Muscle, and T-coffee. Our

experiments showed that PALIGN can achieve accuracy comparable with that of other tools on sequences with high or medium similarity and significantly improved accuracy on sequences in the twilight zone (with similarity lower than 25%). In addition, the number of profile HMMs in the ensemble can be used as a parameter and its value can be changed in different runs of the program, the efficiency and alignment accuracy of PALIGN can thus be adjusted based on the needs of the user. Our experiments also showed that, including additional secondary structure information into this alignment approach can improve the alignment accuracy.

2 Models and methods

The method progressively constructs the alignments in the following steps. Firstly, the order to align sequences to the profile HMMs in the ensemble is determined by computing the optimal pairwise alignment between each pair of sequences in the set. In particular, the two sequences in the pair with the maximum alignment score are selected to construct the initial profile HMMs in the ensemble. These two sequences are also called *base sequences*. We initialise a sequence set S' to contain the two base sequences and iteratively enlarge S' by computing the average alignment score between each sequence that is not in S' and the sequences in S', the one with the maximum average alignment score is then included in S'. Note that this procedure also determines the order by which the sequences will be aligned to the profile HMMs in the ensemble. Secondly, assume the number of profile HMMs in the ensemble is k, we use a dynamic programming algorithm to compute the alignments with k maximum alignment scores in time $O(k \log_2 kn^2)$, where n is the length of the sequence. Each of the alignment can be described with a profile HMM that contains the distribution of amino acids in each column of the alignment.

Thirdly, following the order determined in the first step, the remaining sequences are then aligned to the profile HMMs in the ensemble one by one. For a given sequence, we align it to each of the profile HMMs in the ensemble and for each pair of sequence and profile HMM, we use a similar dynamic programming algorithm to compute the alignments with k maximum alignment scores. We thus can get k^2 alignments in total for the sequence, we then select the alignments with k maximum scores from the available k^2 alignments and the profile HMMs in the ensemble are then updated based on these alignments. This process is repeatedly applied to each sequence while it is aligned to the profile HMMs in the ensemble until no unaligned sequences exist in the set. As the last step, a list of k alignments with the maximum scores are reported as the result of alignment. Figure 1(a)–(c) sketches the three stages for this alignment method.

2.1 Ensemble initialisation

We modify the quadratic time dynamic programming algorithm for computing the optimal pairwise alignment to obtain the alignments with k maximum alignment scores. To simplify the notation, we assume that the alignment scores for two amino acids are stored in a matrix M and the penalty for a contiguous gap region of length d is Pd, where P is a constant. Our algorithm can be slightly changed to cope with the affine gap penalty function, where an open penalty and an extension penalty are needed to compute the overall penalty arising from a contiguous gap region.

We assume the two sequences are s_1 and s_2 and their lengths are m and n respectively. The ith character in a sequence s is denoted with $s[i]$ and the subsequence in s between i and j is denoted with $s[i…j]$. To find the alignments with the k maximum alignment scores, we maintain a $k \times m \times n$ dynamic programming table A to store the intermediate results of the dynamic programming. In particular, $A[l][i][j]$ stores the lth largest alignment score between

subsequences $s_1[1\ldots i]$ and $s_2[1\ldots j]$. For a number set X, we use $X(t)$ to denote the tth largest number in M and define

$$P_{tij}=A[t][i-1][j-1]+M[s_1[i-1]][s_2[j-1]] \tag{1}$$

$$L_{tij}=A[t][i-1][j]+P \tag{2}$$

$$R_{tij}=A[t][i][j-1]+P \tag{3}$$

$$S_{ij}=\bigcup_{t=1}^{k}\{P_{tij},L_{tij},R_{tij}\}. \tag{4}$$

It is then straightforward to see that the recursive relation for the dynamic programming is as follows.

$$A_{lij}=S_{ij}(l) \tag{5}$$

where $1 \le i \le m$, $1 \le j \le n$ and $1 \le l \le k$. Such a dynamic programming procedure can be performed in time $O(k \log_2 knm)$ since we can sort all elements in S_{ij} to find the k largest ones. The values of $A[1][m][n]$, $A[2][m][n]$,…, $A[k][m][n]$ are the k largest alignment scores between s_1 and s_2. An additional table that stores the selection made to compute the value of each cell in A can be used to determine the k alignments.

Based on the k alignments, we are able to construct an ensemble of k profile HMMs. Each of the profile HMM in the set corresponds to one of the k alignments. A profile HMM contains two states D_i and M_i for column i in the corresponding alignment. A deletion state D_i does not emit any amino acids and is used to describe the gaps in column i, a matching state M_i emits an amino acids and is used to describe the amino acids in column i. The probabilities of emission and transition for each state can be computed from the alignment as well.

2.2 Ensemble update

The remaining unaligned sequences in the set are then aligned to the profile HMMs in the ensemble one by one in the order determined in the first step. The alignment between a sequence and a profile HMM can be performed with a dynamic programming algorithm similar to the one for pairwise alignments. In particular, assume the sequence to be aligned to a profile HMM is s, we maintain a $k \times m$ dynamic programming table A_B for each state B in the profile HMM, where m is the length of the sequence. $A_B[l][i]$ is the lth largest alignment score for cases where $s[i]$ is aligned to state B. To simplify the notation, we use $T(B)$ to denote the set of states with nonzero probabilities to transit to B, $e_B(a)$ to denote probability for state B to emit amino acids a and $T(C, B)$ to define the probability for a transition from state C to B. In addition, we assume the set of amino acids is Σ and consider $Y \in T(M_j)$, $Z \in T(D_j)$ for matching state M_j and deleting state D_j respectively, we define

$$P_{M_j}(Y,l,i)=\sum_{a\in\sum} T(Y,M_j)e_{M_j}(a)M[s[i]][a]+A_Y[l][i-1]+(1-T(Y,M_j))P \tag{6}$$

$$Q_{M_j}(l,i)=\bigcup_{Y\in T(M_j)}\{P_{M_j}(Y,l,i)\} \tag{7}$$

$$S_{M_j}(i)=\bigcup_{1\le l\le k} Q_{M_j}(l,i) \tag{8}$$

$$P_{D_j}(Z,l,i)=(1-T(Z,M_j))P+A_Z[l][i] \tag{9}$$

$$Q_{D_j}(l,i)=\bigcup_{Z\in T(D_j)}\{P_{D_j}(Z,l,i)\} \tag{10}$$

$$S_{D_j}(i)=\bigcup_{1\le l\le k} Q_{D_j}(l,i). \tag{11}$$

It is then straightforward to see that the recursive relationship for the dynamic programming is as follows.

$$A_{M_j}(l,i)=S_{M_j}(i)(l) \tag{12}$$

$$A_{D_j}(l,i)=S_{D_j}(i)(l). \tag{13}$$

Since the number of states in $T(B)$ is a constant for state B in a profile HMM, the computation time of this dynamic programming procedure is $O(k \log_2 knL)$, where L is the number of states in the profile HMM.

For each pair of a sequence and a profile HMM, the above dynamic programming algorithm can provide k alignments with leading scores, we thus can obtain k^2 such alignments in total for the ensemble. We then pick k alignments with leading alignment scores from them. The parameters in each profile HMM can then be updated based on these k alignments. The overall alignment process needs time $O(k \log_2 kn^3)$ since we need to progressively align the remaining $n-2$ sequences to the profile HMMs in the ensemble.

3 Testing results

3.1 Without secondary structure information

We implemented this method in a computer program, PALIGN and tested its performance on some of the BaliBASE benchmarks. (available at http://bips.u-strasbg.fr/fr/Products/Databases/BAliBASE/) In particular, sequences are

aligned to maximise the SP score of the alignment and we used PAM250 as the score matrix and a gap open penalty of value −19.814 and a gap extension penalty −1.396. We performed alignments on these benchmarks with ensembles containing a variety of different numbers of profile HMMs and compared the maximum alignment scores that can be achieved. Table 1 provides the maximum alignment scores PALIGN has achieved on three benchmarks of low similarity (with identity less than 25%), three benchmarks of medium similarity (with identity in between 20% and 40%), and three benchmarks of high similarity (with identity larger than 35%). From the table, we are able to see that, from the perspective of maximisation, PALIGN can achieve the maximum alignment score on most of the tested benchmarks using ensembles containing no more than four profile HMMs. Moreover, ensembles containing more profile HMMs may lead to alignments with lower alignment scores. We think the greedy strategy employed to update the ensembles during the progressive alignment can partly explain this.

In addition to evaluating the effect of maximisation with ensembles of different sizes, we also measured the computation time PALIGN needs to perform the alignments using these ensembles. Table 2 shows the computation time needed by PALIGN on ensembles of different sizes. From the table, we can see that PALIGN can perform multiple alignments in a few seconds.

We then use the program provided by BaliBASE to evaluate the accuracy of PALIGN and compared it with that of other alignment tools, including MULTALIGN (Corpet, 2006) DIALIGN Morgenstern et al. (1998), MULTAL (Taylor, 1988) and HMMT Eddy (2003). Comparison is performed on benchmarks of low, medium and high similarity values. To achieve the maximum accuracy, we ran PALIGN using ensembles of size 1–5 and selected the alignment with the highest BaliBASE score in the generated lists of alignments. Table 3 shows the BaliBASE accuracy scores of PALIGN and other alignment tools on some benchmarks. The BaliBASE score is computed by comparing an alignment obtained by running a software tool with the reference alignment of the benchmark. A higher score indicates an alignment closer to the reference alignment. From the table, we can see that PALIGN achieves significantly better accuracy than MULTALIGN, MULTAL and HMMT in benchmarks in twilight zone (sequences with identity lower than 25%). PALIGN also achieves comparable or slightly better accuracy than other alignment tools on benchmarks with medium and high similarity.

3.2 With secondary structure information

Previous work has shown that, secondary structure information of the sequences can significantly improve the alignment accuracy (Zhang and Kahveci, 2006). Our model can be modified slightly to include the secondary structure information. In particular, a secondary structure unit is associated with each amino acid residue in a protein sequence. In general, there are three types of secondary structure units for protein sequences, they are C (coil), E (β sheet), and H (α helix).

To integrate secondary structure information into the model, we include the secondary structure matching score in the overall alignment score. In other words, the overall alignment score is the weighted sum of the matching score of sequence content and the that of secondary structure units. We can use a 3×3 score matrix to compute the secondary structure matching score.

We used a simple score matrix to compute the secondary structure matching score, where the score of a match is 1 and a mismatch is 0. A weight of 0.8 is allocated to the matching score of sequence content and secondary structure matching score has a weight of 0.2. We

use the same penalty scores for gaping opening and extension for both types of matching scores.

We applied this modified approach to the same benchmark sets. Table 4 shows the alignment accuracy of this approach on these sets. We can see from the table that a significant amount of improvement is achieved on most of the benchmark sets. The improvement is more significant on sequence sets in twilight zone. This may suggest that structure information is more important for aligning sequences with low sequence identity.

4 Conclusions

In this paper, we develop a new method that can efficiently and accurately align multiple protein sequences by exploring the alignment space. This method constructs and maintains an ensemble of profile HMMs and progressively align the remaining unaligned sequences to the profile HMMs in the ensemble and updates the profile HMMs in the ensemble based on the alignments with leading scores. Our experiments have demonstrated that this method, which has been implemented in an alignment tool PALIGN, can achieve significantly better accuracy than other alignment tools on sequences of low similarity. The accuracy of PALIGN is comparable or slightly better than these tools on sequences with medium and high similarity. In addition, we have integrated the secondary structure information into this approach, our experiments show that the alignment accuracy can be further improved while taking into account structure information.

Acknowledgments

We thank the anonymous reviewers for their constructive comments on the paper. CL's work was supported in part by grant 2 G12 RR003048 from the RCMI Program, Division of Research Infrastructure, National Center for Research Resources, NIH, and New Faculty Startup Award from Howard University.

References

Chen, L. Multiple protein structure alignment by deterministic annealing. IEEE Computer Society Bioinformatics Conference; Stanford, CA. 2003. p. 609

Corpet F. Multiple sequence alignment with hierarchical clustering. Nucleic Acids Research. 2006; 16(22):10881–10890. [PubMed: 2849754]

Eddy, S. Multipel alignment using hidden markov models. The Third International Conference on Intelligent Systems for Molecular Biology; Brisbane, Australia. 2003. p. 114-120.

Edgar RC. Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 2004; 32(5):1792–1797. [PubMed: 15034147]

Gotoh O. Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. Journal of Molecular Biology. 1996; 264(4):823–838. [PubMed: 8980688]

Gupta S, Kececiogo J, Schaffer A. Improving the practical space and time efficiency of the shortest paths approach to sum-of-pairs multiple sequence alignment. Journal of Computational Biology. 1995; 2(3):459. [PubMed: 8521275]

Hein J. A new method that simultaneously aligns and reconstructs ancestral sequences for any number of homologous sequences, when the phylogeny is given. Molecular Biology and Evolution. 1989; 6(6):649–668. [PubMed: 2488477]

Huang X, Miller W. A time efficient, linear space local similarity algorithm. Advances in Applied Mathematics. 1991; 12:337–357.

Katoh K, Misawa K, Kuma K, Miyata T. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. Nucleic Acids Research. 2002; 30(14):3059–3066. [PubMed: 12136088]

Lee C, Grasso C, Sharlow M. Multiple sequence alignment using partial order graphs. Bioinformatics. 2002; 18(3):452–464. [PubMed: 11934745]

Lipman D, Altschul S, Kececioglu J. A tool for mulitple sequence alignment. Proceedings of the National Academy of Sciences. 1989; 86(12):4412–4415.

Morgenstern B, Frech K, Dress A, Werner T. Dialign: finding local similarities by multiple sequence alignment. Bioinformatics. 1998; 14(3):290–294. [PubMed: 9614273]

Needleman S, Wunsch C. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology. 1970:443–453. [PubMed: 5420325]

Notredame C, Higgins D. Saga: sequence alignment by genetic algorithm. Nucleic Acids Research. 1996; 24(8):1515–1524. [PubMed: 8628686]

Notredame C, Higgins D, Heringa J. T-coffe:a novel method for fasta and accurate multiple sequence alignment. Journal of Molecular Biology. 2000; 302(1):205–217. [PubMed: 10964570]

Phillips A, Janies D, Wheeler W. Multiple sequence alignment in phylogenetic analysis. Molecular Phylogenetics and Evolution. 2000; 16(3):317–330. [PubMed: 10991785]

Taylor W. A flexible method to align large number of biological sequences. Journal of Molecular Evolution. 1988; 28:161–169. [PubMed: 3148736]

Thompson J, Higgins D, Gibson T. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research. 1994; 22(22):4673–4680. [PubMed: 7984417]

Thompson J, Plewniak H, Poch O. A comprehensive comparison of multiple sequence alignment programs. Nucleic Acids Research. 1999; 27(13):2682–2690. [PubMed: 10373585]

Thomsen R, Fogel G, Krink T. Improvement of cluastal-derived sequence alignments with evolutionary algorithms. Congress on Evolutionary Computation. 2003; 1:312–319.

Walle I, Lasters I, Wyns L. Align-m: a new algorithm for multiple alignment of highly divergent sequences. Bioinformatics. 2004; 20(9):1428–1435. [PubMed: 14962914]

Zhang X, Kahveci T. A new approach for alignment of multiple proteins. Pacific Symposium on Biocomputing. 2006; 11:339–350. [PubMed: 17094251]

Biographies

Jia Song is currently a PhD student in the Department of Control Theory and Control Engineering at Zhejiang University. Since 2007 she has been a Lecturer in the Department of Electronic Engineering at Suzhou Vocational College. She received her Master Degree from College of Electrical Engineering, Zhejiang University in 2005.

Chunmei Liu received her PhD in Computer Science at the University of Georgia. She is an Assistant Professor at the Department of Systems and Computer Science, Howard University. Her research interests include Bioinformatics, Computational Biology, Algorithms, and Graph Theory. She has published multiple research papers in national and international journals and conference proceedings.

Yinglei Song received his PhD in Computer Science from The University of Georgia. He is currently Assistant Professor in the Department of Mathematics and Computer Science at The University of Maryland, Eastern Shore. His research interests are theory of computing, bioinformatics, and parameterised complexity.

Junfeng Qu is with Department of Information Technology at Clayton State University. He received his PhD in Computer Science from the University of Georgia. Prior to joining Clayton State, he worked at the University of Georgia, Georgia Institute of Technology and Sharp USA. His research interest is focused on data mining, bioinformatics, database design, data warehouse, algorithm, and software engineering. He is working collaboratively with Georgia Institute of Technology, CDC etc. on Biological Information Data Analysis and Mining.

Gurdeep S. Hura received his BE (BS) in Electronics and Tele-communication Engineering from Government Engineering College, Jabalpur University (India) in 1972, ME (MS) in Controls and Guidance Systems from University of Roorkee (India) in 1975 and PhD from University of Roorkee (India) in 1985, respectively. He is currently Professor and Chair in MCS Department at The University of Maryland Eastern Shore. His research interests include Petri net modelling and their applications, computer networks and network security, software engineering, distributed systems.

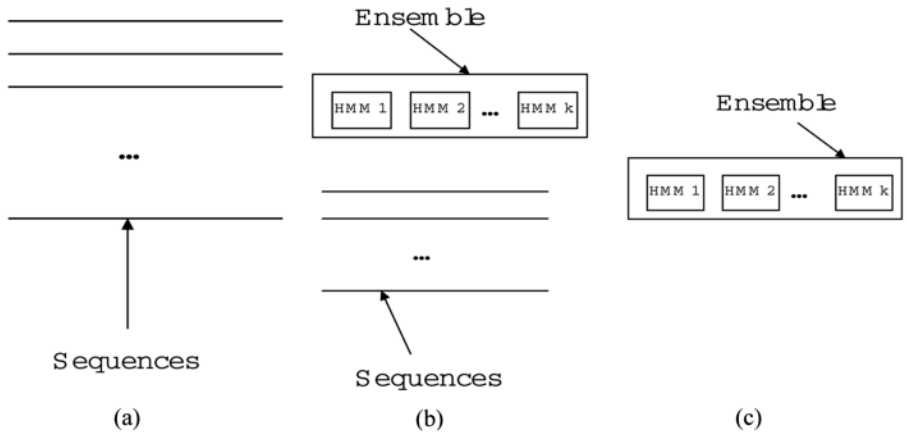

Figure 1.

An illustration of the alignment method: (a) firstly, given a set of sequences, the order to include the sequences in the alignment is determined; (b) secondly, an ensemble with *k* profile HMMs is constructed and progressively updated while the remaining unaligned sequences are aligned to the profile HMMs in the ensemble and (c) thirdly, *k* alignments with leading scores can be generated from the profile HMMs in the ensemble when no sequences in the set are unaligned

Table 1

The SP alignment scores achieved with ensembles containing different number of profile HMMs. Parameter is the number of the profile HMMs in the ensemble; the first three benchmarks are of low similarity, the next three are of medium similarity and the last three are of high similarity

Parameter	1	2	3	4	5	6	7	8	9	10
1aboA	−49.67	−47.37	−59.67	−58.92	−54.66	−52.12	−54.86	−53.11	−59.13	−51.47
1idy	−35.14	−35.54	−32.40	−31.05	−34.54	−35.30	−30.80	−28.55	−28.25	−29.04
1r69	−66.53	−60.32	−44.40	−45.09	−46.04	−49.54	−46.25	−55.77	−50.12	−56.77
1aab	50.82	39.08	44.55	51.68	46.82	51.96	51.62	51.62	29.14	42.48
1fjlA	12.21	27.90	15.13	19.04	9.64	4.41	9.04	15.44	11.24	10.69
1hfh	86.10	86.95	86.76	81.41	83.16	80.16	75.51	69.51	61.56	43.32
1aho	74.68	77.79	96.97	90.22	75.48	74.48	74.04	77.24	71.64	71.49
1csp	112.44	108.94	113.79	111.89	105.64	105.89	111.14	107.39	102.14	101.39
1dox	50.42	61.01	63.15	62.01	80.75	73.95	70.08	68.42	51.62	55.62

Table 2

The computation time (in seconds) needed on ensembles of different sizes on the benchmarks as in Table 1

Parameter	1	2	3	4	5	6	7	8	9	10
1aboA	0.39	0.79	1.58	2.75	4.44	6.69	9.44	13.03	18.07	22.84
1idy	0.29	0.59	1.11	1.96	3.16	4.72	6.69	9.26	12.13	15.57
1r69	0.31	0.62	1.23	2.16	3.49	5.58	7.37	10.07	13.44	17.33
1aab	0.34	0.68	1.31	2.32	3.66	5.54	7.98	10.96	14.37	18.37
1fjlA	0.62	1.18	2.21	3.83	6.11	9.19	13.13	17.85	23.69	31.02
1hfh	1.45	2.91	5.62	9.91	15.84	23.62	34.05	46.93	62.52	80.10
1aho	0.41	0.81	1.54	2.73	4.37	6.63	9.58	13.04	16.76	21.66
1csp	0.45	0.89	1.70	2.94	4.70	7.11	10.09	13.75	18.33	23.95
1dox	0.56	1.15	2.24	3.91	6.24	9.40	13.37	18.22	24.27	31.40

Int J Data Min Bioinform. Author manuscript; available in PMC 2011 April 12.

Table 3

The BaliBASE score achieved by PALIGN and a few other alignment tools on some benchmarks, including six benchmarks of low similarity (with identity less than 25%), three benchmarks of medium similarity (with identity in between 20% and 40%) and three benchmarks of high similarity (with identity larger than 35%). Ensemble Size (ES) is the size of the ensemble where the maximum BaliBASE score is achieved for a benchmark

Benchmarks	PALIGN	ES	DIALIGN	MULTALIGN	MULTAL	HMMT
1aboA	0.844	2	0.359	0.703	0.526	0.181
1r69	0.409	4	0.406	0.325	0.225	0.100
1tvxA	0.444	3	0.306	0.228	0.244	0.108
1ubi	0.434	3	0.000	0.488	0.428	0.140
1wit	0.886	3	0.851	0.842	0.763	0.549
2trx	0.505	5	0.728	0.500	0.235	0.292
1aab	0.967	4	1.000	1.000	1.000	0.214
1fjlA	0.908	2	1.000	1.000	1.000	0.436
1hfh	0.828	2	0.410	0.936	0.883	0.261
1aho	0.860	4	1.000	0.913	0.938	0.789
1csp	0.924	2	0.993	0.987	0.625	0.776
1dox	0.850	4	0.859	0.799	0.480	0.806

Int J Data Min Bioinform. Author manuscript; available in PMC 2011 April 12.

Table 4

The BaliBASE score achieved by the modified approach and a few other alignment tools on some benchmarks, including six benchmarks of low similarity (with identity less than 25%), three benchmarks of medium similarity (with identity in between 20% and 40%) and three benchmarks of high similarity (with identity larger than 35%). Ensemble Size (ES) is the size of the ensemble where the maximum BaliBASE score is achieved for a benchmark

Benchmarks	PALIGN	ES	DIALIGN	MULTALIGN	MULTAL	HMMT
1aboA	0.952	3	0.359	0.703	0.526	0.181
1r69	0.629	3	0.406	0.325	0.225	0.100
1tvxA	0.579	4	0.306	0.228	0.244	0.108
1ubi	0.624	2	0.000	0.488	0.428	0.140
1wit	0.736	4	0.851	0.842	0.763	0.549
2trx	0.622	3	0.728	0.500	0.235	0.292
1aab	0.921	4	1.000	1.000	1.000	0.214
1fjlA	0.933	3	1.000	1.000	1.000	0.436
1hfh	0.876	2	0.410	0.936	0.883	0.261
1aho	0.893	3	1.000	0.913	0.938	0.789
1csp	0.984	3	0.993	0.987	0.625	0.776
1dox	0.973	4	0.859	0.799	0.480	0.806