Efficient Methods for Multiple Sequence Alignment with Guaranteed Error Bounds (1993; Gusfield)

Francis Y.L. Chin, The University of Hong Kong, www.cs.hku.hk/~chin
S.M. Yiu, The University of Hong Kong, www.cs.hku.hk/~smyiu

entry editor: M. Y. Kao

INDEX TERMS: Multiple sequence alignment, approximation algorithms, sum of pairs score, tree alignment score

SYNONYMS: Multiple string alignment, multiple global alignment

1 PROBLEM DEFINITION

Multiple sequence alignment is an important problem in computational biology. Applications include finding highly conserved subregions in a given set of biological sequences and inferring the evolutionary history of a set of taxa from their associated biological sequences (e.g., see [6]). There are a number of measures proposed for evaluating the goodness of a multiple alignment, but prior to this work, no efficient methods are known for computing the optimal alignment for any of these measures. The work of Gusfield [5] gives two computationally efficient multiple alignment approximation algorithms for two of the measures with approximation ratio of less than 2. For one of the measures, they also derived a randomized algorithm, which is much faster and with high probability, reports a multiple alignment with small error bounds. To the best knowledge of the entry authors, this work is the first to provide approximation algorithms (with guarantee error bounds) for this problem.

Notations and Definitions Let X and Y be two strings of alphabet Σ. The pairwise alignment of X and Y maps X and Y into strings X' and Y' that may contain spaces, denoted by '_', where (1) $|X'| = |Y'| = \ell$; and (2) removing spaces from X' and Y' returns X and Y, respectively. The score of the alignment is defined as $d(X', Y') = \sum_{i=1}^{\ell} s(X'(i), Y'(i))$ where $X'(i)$ (and $Y'(i)$) denotes the i-th character in X' (and Y') and $s(a, b)$ with $a, b \in \{\Sigma \cup \{\text{'\_'}\}\}$ is the distance-based scoring scheme that satisfies the following assumptions.

(a) $s(\text{'\_'}, \text{'\_'}) = 0$;

(b) triangular inequality: for any three characters, x, y, z, $s(x, z) \leq s(x, y) + s(y, z)$.

Let $\chi = \{X_1, X_2, ..., X_k\}$ be a set of $k > 2$ strings of alphabet Σ. A multiple alignment A of these k strings maps $X_1, X_2, ..., X_k$ to $X_1', X_2', ..., X_k'$ that may contain spaces such that (1) $|X_1'| = |X_2'| = ... = |X_k'| = \ell$; and (2) removing spaces from X_i' returns X_i for all $1 \le i \le k$. The multiple alignment A can be represented as a $k \times \ell$ matrix.

The sum of pairs (SP) measure The score of a multiple alignment A, denoted by $SP(A)$, is defined as the sum of the scores of pairwise alignments induced by A, that is, $\sum_{i<j} d(X_i', X_j') = \sum_{i<j} \sum_{p=1}^{\ell} s(X_i'[p], X_j'[p])$ where $1 \le i < j \le k$.

Problem 1 Multiple Sequence Alignment with Minimum SP score
Input: A set of k strings, a scoring scheme s.
Output: A multiple alignment A of these k strings with minimum $SP(A)$.

The tree alignment (TA) measure In this measure, the multiple alignment is derived from an evolutionary tree. For a given set χ of k strings, let $\chi' \supseteq \chi$. An evolutionary tree $T_{\chi'}$ for χ is a tree with at least k nodes, where there is a one-to-one correspondence between the nodes and the strings in χ'. Let $X_u' \in \chi'$ be the string for node u. The score of $T_{\chi'}$, denoted by $TA(T_{\chi'})$, is defined as $\sum_{e=(u,v)} D(X_u', X_v')$ where e is an edge in $T_{\chi'}$ and $D(X_u', X_v')$ denotes the score of the *optimal* pairwise alignment for X_u' and X_v'.

Analogously, the multiple alignment of χ under the TA measure can also be represented by a $|\chi'| \times \ell$ matrix, where $|\chi'| \ge k$, with a score defined as $\sum_{e=(u,v)} d(X_u', X_v')$ (e is an edge in $T_{\chi'}$), similar to the multiple alignment under the SP measure in which the score is the summation of the alignment scores of all pairs of strings. Under the TA measure, since it is always possible to construct the $|\chi'| \times \ell$ matrix such that $d(X_u', X_v') = D(X_u', X_v')$ for all $e = (u, v)$ in $T_{\chi'}$ and we are usually interested in finding the multiple alignment with the minimum TA value, so $D(X_u', X_v')$ is used instead of $d(X_u', X_v')$ in the definition of $TA(T_{\chi'})$.

Problem 2 Multiple Sequence Alignment with Minimum TA score
Input: A set of k strings, a scoring scheme s.
Output: An evolutionary tree T for these k strings with minimum $TA(T)$.

2 KEY RESULTS

Theorem 1. Let A^* be the optimal multiple alignment of the given k strings with minimum SP score. They provide an approximation algorithm (the center star method) that gives a multiple alignment A such that $\frac{SP(A)}{SP(A^*)} \le \frac{2(k-1)}{k} = 2 - \frac{2}{k}$.

The center star method is to derive a multiple alignment which is consistent with the optimal pairwise alignments of a center string with all the other strings. The bound is derived based on the triangular inequality of the score function. The time complexity of this method is $O(k^2 \ell^2)$, where ℓ^2 is the time to solve the pairwise alignment by

dynamic programming and k^2 is needed to find the center string, X_c, which gives the minimum value of $\sum_{i \neq c} D(X_c, X_i)$.

Theorem 2. Let A^* be the optimal multiple alignment of the given k strings with minimum SP score. They provide a randomized algorithm that gives a multiple alignment A such that $\frac{SP(A)}{SP(A^*)} \leq 2 + \frac{1}{r-1}$ with probability at least $1 - \left(\frac{r-1}{r}\right)^p$ for any $r > 1$ and $p \geq 1$.

Instead of computing $\binom{k}{2}$ optimal pairwise alignments to find the best center string, the randomized algorithm only considers p randomly selected strings to be candidates for the best center string, thus this method needs to compute only $(k - 1)p$ optimal pairwise alignments in $O(k\, p\, \ell^2)$ time where $1 \leq p \leq k$.

Theorem 3. Let T^* be the optimal evolutionary tree of the given k strings with minimum TA score. They provide an approximation algorithm that gives an evolutionary tree T such that $\frac{TA(T)}{TA(T^*)} \leq \frac{2(k-1)}{k} = 2 - \frac{2}{k}$.

In the algorithm, they first compute all the $\binom{k}{2}$ optimal pairwise alignments to construct a graph with every node representing a distinct string X_i and the weight of each edge (X_i, X_j) as $D(X_i, X_j)$. This step determines the overall time complexity $O(k^2\, \ell^2)$. Then, they find a minimum spanning tree from the graph. The multiple alignment has to be consistent with the optimal pairwise alignments represented by the edges of this minimum spanning tree.

3 APPLICATIONS

Multiple sequence alignment is a fundamental problem in computational biology. In particular, multiple sequence alignment is useful in identifying those common structures, which may only be weakly reflected in the sequence and not easily revealed by pairwise alignment. These common structures may carry important information for their evolutionary history, critical conserved motifs, common 3D molecular structure, as well as biological functions.

More recently, multiple sequence alignment is also used in revealing non-coding RNAs (ncRNAs) [3]. In this type of multiple alignment, we are not only align the underlying sequences, but also the secondary structures (refer to Chapter 16 of [10] for a brief introduction of secondary structure of a RNA) of the RNAs. Researchers believe that ncRNAs that belong to the same family should have common components giving a similar secondary structure. The multiple alignment can help to locate and identify these common components.

4 OPEN PROBLEMS

A number of open problems related to the work of Gusfield remain open. For the SP measure, the center star method can be extended to the q-star method ($q > 2$) with approximation ratio of $2 - q/k$ ([1, 7], Section 7.5 of [8]). Whether there exists an approximation algorithm with better approximation ratio or with better time complexity is still unknown. For the TA measure, to be the best knowledge of the entry authors, the approximation ratio in Theorem 3 is currently the best result.

Another interesting direction related to this problem is the constrained multiple sequence alignment problem [9] which requires the multiple alignment to contain certain aligned characters with respect to a given constrained sequence. The best known result [2] is an approximation algorithm (also follows the idea of center star method) which gives an alignment with approximation ratio of $2 - 2/k$ for k strings.

For the complexity of the problem, Wang and Jiang [11] were the first to prove the NP-hardness of the problem with SP score under a *non-metric* distance measure over a 4 symbol alphabet. More recently, in [4], the multiple alignment problem with SP score, star alignment, and TA score have been proved to be NP-hard for all binary or larger alphabets under *any metric*. Developing efficient approximation algorithms with good bounds for any of these measures is desirable.

5 EXPERIMENTAL RESULTS

Two experiments have been reported in the paper showing that the worst case error bounds in Theorems 1 and 2 (for the SP measure) are pessimistic compared to the typical situation arising in practice.

The scoring scheme used in the experiments is: $s(a, b) = 0$ if $a = b$; $s(a, b) = 1$ if either a or b is a space; otherwise $s(a, b) = 2$. Since computing the optimal multiple alignment with minimum SP score has been shown to be NP-hard, they evaluate the performance of their algorithms using the lower bound of $\sum_{i<j} D(X_i, X_j)$ (recall that $D(X_i, X_j)$ is the score of the optimal pairwise alignment of X_i and X_j). They have aligned 19 similar amino acid sequences with average length of 60 of homeoboxs from different species. The ratio of the scores of reported alignment by the center star method to the lower bound is only 1.018 which is far from the worst case error bound given in Theorem 1. They also aligned 10 not-so-similar sequences near the homeoboxs, the ratio of the reported alignment to the lower bound is 1.162. Results also show that the alignment obtained by the randomized algorithm is usually not far away from the lower bound.

6 DATA SETS

The exact sequences used in the experiments are not provided.

7 CROSS REFERENCES

Statistical Multiple Alignment (2003; Hein, Jensen, Pedersen)

8 RECOMMENDED READING

[1] V. Bafna, E.L. Lawler, and P.A. Pevzner. Approximation algorithms for multiple sequence alignment. *Theoretical Computer Science*, Vol. 182, pp. 233-244, 1997.

[2] Francis Y.L. Chin, N. L. Ho, Tak Wah Lam, Prudence W.H. Wong. Efficient constrained multiple sequence alignment with performance guarantee. J. Bioinformatics and Computational Biology, Vol. 3(1), pp.1-18, 2005.

[3] Deniz Dalli, Andreas Wilm, Indra Mainz and Gerhard Stegar. STRAL: progressive alignment of non-coding RNA using base pairing probability vectors in quadratic time. Bioinformatics, Vol. 22(13), pp. 1593-1599, 2006.

[4] Isaac Elias. Setting the intractability of multiple alignment. *Proceedings of the 14th Annual International Symposium on Algorithms and Computation* (*ISAAC 2003*), pp. 352-363, 2003.

[5] Dan Gusfield. Efficient methods for multiple sequence alignment with guaranteed error bounds, *Bulletin of Mathematical Biology*, Vol. 55(1), pp. 141-154, 1993.

[6] J. Pevsner. Bioinformatics and functional genomics, *John Wiley*, 2003.

[7] P.A. Pevzner. Multiple alignment, communication cost, and graph matching. *SIAM J. on Applied Mathematics*, Vol. 52, pp. 1763-1779, 1992.

[8] P.A. Pevzner. Computational molecular biology: an algorithmic approach, *The MIT Press*, 2000.

[9] C.Y. Tang, C.L. Lu, M.D.-T. Chang, Y.-T. Tsai, Y.-J. Sun, K.-M. Chao, J.-M. Chang, Y.-H. Chiou, C.-M. Wu, H.-T. Chang, and W.-I. Chou. Constrained multiple sequence alignment tool development and its application to RNase family alignment. *Proceedings of the First IEEE Computer Society Bioinformatics Conference* (*CSB 2002*), pp. 127-137, 2002.

[10] Martin Tompa, Lecture notes, Department of Computer Science & Engineering, University of Washington, http://www.cs.washington.edu/education/courses/527/00wi/.

[11] L. Wang and T. Jiang. On the complexity of multiple sequence alignment. *J. Comp. Biol.*, Vol. 1, pp.337-48, 1994.