Pro-malign: Multiple Sequence Alignment Algorithm using Approached Profile

Ahmed Mokaddem*, Amine Bel Haj, and Mourad Elloumi
Laboratory of Technologies of Information and Communication, and Electrical Engineering, University of Tunis, Tunis, Tunisia.

* Corresponding author: Email: moka.ahmed@yahoo.fr
Manuscript submitted November 30, 2017; accepted January 10, 2018.
10.17706/jsw.13.1.57-65

Abstract: In this paper, we present a new progressive multiple sequence alignment algorithm called Pro-malign. Our algorithm introduces the approached profile sequence a new definition of profile that aim to conserve similar residues in the alignment. The approached profile sequence is used to compute a new distance called approached profile distance. We present also a new score function between profiles and a new refinement algorithm. We assess our program Pro-malign on different datasets extracted from different benchmarks of protein sequences and we compare the scores obtained to other scores of the most efficient multiple sequence alignment programs.

Key words: Multiple sequence alignment, algorithms, refinement, profiles.

1. State of The Art

Multiple Sequence Alignement (MSA) is an important task in bioinformatics. Indeed, It allows the comparison of biological sequences (i.e., DNA, RNA, or protein). It can reveal structural and functionnal informations from a set of biologic sequences. It can help also used to extract conserved regions, construct a evolutionary tree, detect interactions between sequences and assemblly of genome assembly. Produce accurate alignments still an important way for most reserch study. Multiple sequence algrnment consists in optimisation the number of matches between the sequeces. Multiple Sequence Alignement is a NP-complet problem [1]. There are two principle approaches to resolve this problem:

1) *Iterative* approach: by using this approach, we iteratively apply a set of modifcations to a random initial multiple sequence alignment. These modifications are repeated until a *convergence*, i.e., no improvement made on the current alignment. . We can also fixed the iteration number. Several algorithms adopting iterative approach and using different methods to modify the initial multiple sequence alignment are defined Among iteratif algorithms, we mention [2]-[5].
2) *Progressive* approach: by using this approach, the multiple sequence alignment is gradully built following an order definied by a guide tree. Algorithms using this approach works in four steps *:*

a) In the first step, we compute distances between each pair of sequences and we store these distances in a matrix called distance matrix. Different distances were defined. Among these distances, we mention [6]-[12].
b) The second step conssists to construct a guide tree using the distance matrix of the first step. The guide tree defines the branching order for aligning sequences. UPGMA [13] and Neighbor-Joining [14] are the two algorithms used for guide tree construction.

c) In the third step, we align sequences following the branching order of the guide tree. In this step, we used profile for aligning alignment [15].

d) The last step or Refinement step:, we improve the multiple sequence alignment score by applying iteratively a set of operations, i.e., iteratively construction of the guide tree, randomly subdiving the multiple alignment on two profiles then realinging these profiles. Differents refinement techniques have been developed [16].

Among multiple sequence alignment algorithms using progressive approach, we mention CLUSTALW [17], T-COFFEE [18], MAFFT [7], MUSCLE [6] ProbCons [8], Gramalign [9], MSAProbs [10], Motalign [19], GLProbs [12] and Clustal Omega [20].

2. Preliminaries

Let $f = \{w_1, w_2, \ldots, w_N\}$ a set of N sequences, "-" is the symbol to represent a *gap* in a sequence. $|w_i|$ is the length of a sequence w_i, i.e., the characters number in this sequence. A profile is a set of sequences aligned in the progressive process.

The *profile sequence* is a *consensus* sequence constructed from a multiple sequence alignment by selecting, for every column of the alignment, the character that its appearance frequency is greater than the average of sequences number.

3. Approached Profile Distance

We define a new distance called *approached profile distance* using the same approach defined by Mokaddem and Elloumi [11]. By adopting this approach, we assign a distance to each pair of sequences, from an initial set of sequences, after comparing the pairwise alignment of these sequences to each sequences of the set. In order to compute our new distance, we introduce a new definition i.e., the *approached profile sequence*. Our new definition promote identity and similarity conservation in pairwise sequence alignment instead of identity in profile. Indeed, when we construct a profile of pairwise sequence alignment we select for each column the residue that appear in each column, otherwise we select a gap. However, similar characters can appear in the columns. These similarities are ignored and replaced by a gap character. In our case, we choose to conserve this similarity in the profile and we present a new profile sequence called the *approached profile sequence*, which consist to select for columns, that similar characters are aligned, the character that have the maximum occurrence number in the two sequences. We construct the *approached profile sequence* as follows: Let w_1 and w_2 be two sequences. First, we construct a pairwise sequence alignment using the Needleman and Wunsch algorithm [21]. Then, for each column of the pairwise alignment:

a) if two residues aligned in the same column are identical, we select this character.

b) Otherwise, the two residues are similar; we select the one, which appears the most in the two sequences. This residue represents the current column of the two sequences. Two residues are similar if they belong to the same compressed groups of residues [9].

c) Else, we select the gap character.

We present below the difference between profile sequence and approached profile sequence.

Pairwise sequence alignment

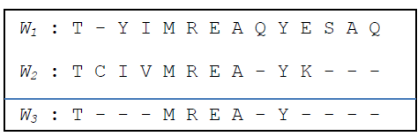

Fig. 1. Profile sequence.

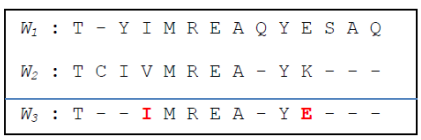

Fig. 2. Approached profile.

We used the approached profile to compute the new distance that aim to attribute a most import weight to the sequence that their pairwise alignment conserve the maximum number of similarity and identity in the other sequence of the set.

To calculate the approached profile distance between two sequences, we proceed in the following way:

d) During the first step, we align the two sequences wi and wj and then constructs the approached profile sequence using the definition before.

e) During the second step, we compare the approached profile sequence with the other N-2 sequences of the initial and we compute the approached profile using the formula (1) below.

$$D(w_1, w_2) = \frac{\sum_{k=1}^{N} [1-(\alpha+\beta)/(|w_p|+|w_k|)]}{N-2} \quad (1)$$

where w_p is the approached profile sequence of w_1 and w_2, N is the number of sequences, α is the number of occurrence of the different residues of w_p appearing in w_k, β is the occurrence of residues in the sequence w_k appearing in w_p.

4. MSC Score Function

In this section, we present our new score function, called MSC, between pairs of columns in order to align alignment in the progressive process. By adopting our new score function MSC; we assign a higher score to the two columns of the two alignments that are the most similar. In order to assign a score for a pair of columns c_1 and c_2, we operate as follows:

(i) First, we compute the SP [22] scores of the two columns.

(ii) Then, we find the major residue, i.e., the residue that has more occurrences in both columns. Then, we find similar residue to major residue in both columns. We compute the MSC score between two columns c_1 and c_2 using the following formula:

$$\text{MSC}(c_1, c_2) = (\alpha+\beta) * \text{SP}(c_1, c_2) \quad (2)$$

α is the occurrence number of the *major* residue in both columns, β is the occurrence of similar residue to the *major* residue.

Otherwise, i.e., no major residue, we find the number of similar residues that have the highest number of occurrences by using the compressed groups of residues [9]. In this case, MSC score is computed as follows:

$$\text{MSC}(c_1, c_2) = (\beta) * \text{SP}(c_1, c_2) \quad (3)$$

β is the highest occurrence number of similar residues.

5. Refinement Algorithm

In this section, we present our refinement algorithm that aim to ameliorate the multiple sequence alignment score. Our algorithm operates as follows: First, we create from the multiple sequence alignment two new families of sequences. Then, we construct the multiple sequence alignment for each family. Finally, we align the multiple sequence alignment for each family by adapting the Needleman and Wunsch algorithm to aligning multiple sequence alignment using the MSC score function between columns instead of the substitution matrix.

The process used to create the two families in the first step:

1) First, we construct a distance matrix from the initial multiple sequence alignment using our new approached profile distance.

2) Then, we select the two most divergent sequences, i.e., having the maximum approached profile distance in the distance matrix. We assign each sequence in a different family. Thus, we obtain two new families sequences formed each of them by on sequence. These two sequences are be called *original sequence*.

3) Then, we compute the SP scores between each sequence of the initial set and the original sequence of each family using the pairwise alignment of these two sequences extracted from the current multiple sequence alignment.

4) Finally, we add each sequence to the family, that the corresponding SP score with the original sequence is greater. Thus, we obtain two new families of sequences.

6. Pro-malign Algorithm

The algorithm *Pro-malign* operates as follows:

1) During the first step, we compute the approached profile distances between each pair of sequences and we store these distances in the distances matrix.
2) During the second step, we adopt the UPGMA algorithm to construct a guide tree using the distance matrix of the first step.
3) During the third step, we follow the branching order obtained by the guide tree to construct the initial multiple sequence alignment by using our MSC score function and the adaptation of the Needleman and Wunsch algorithm to align profile.
4) During the fourth step, we apply our refinement algorithm.

7. An Illustrative Example

We used a set of test sequences family, which is applied by several algorithms in the literature. Let be a set of 4 sequences.

w_1: TYIMREAQYESAQ ; w_2:TCIVMREAYE; w_3: YIMQEVQQER; w_4 : WRYIAMREQYES

During the first step, we compute the *approached profile distance* between each pair of sequences

w_1			
w_2	0.36		
w_3	0.43	0.75	
w_4	0.35	0.45	0.56

Then, we construct a guide tree using the distance matrix as showing in Fig. 3.

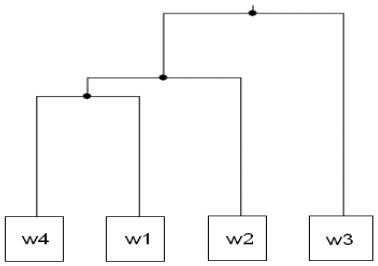

Fig. 3. Guide tree.

Then, we construct the multiple sequence alignment following the branching order of the guide tree and we obtain the following initial multiple sequence alignment as shown in Fig. 4.

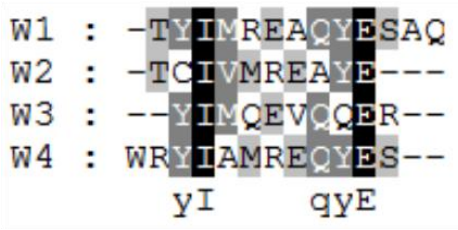

Fig. 4. Multiple sequence alignment before refinement.

Finally, we apply the refinement step algorithm and we obtain the final multiple sequence alignment below in Fig. 5.

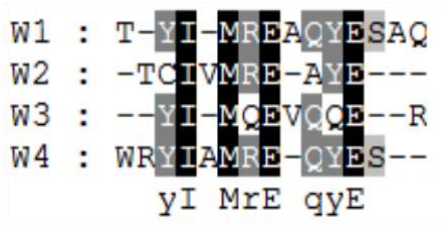

Fig. 5. Multiple sequence aligNMent after refinement.

8. Experimental Study

In order to assess *Pro-malign* performances, we used different datasets extracted from different benchmarks for protein sequences. We used the following benchmarks: BALIBASE [23], OXBENCH [24], PREFAB [6] and HOMSTRAD [25]. We compared the results of *Pro-malign* with the most efficient and the most used multiple sequence alignment programs namely, CLUSTALW [17], MUSCLE [6], MAFFT [7], Clustal Omega [20] and T-COFFEE [18] using the same datasets. In order to compare alignments, we used the *Sum of Pairs Score* (SPS) [26] and the *Column score* (CS) [26] for BALIBASE. The results of SPS and CS obtained with a datasets from BALIBASE are respectively represented in Table 2 and Table 3. However, we used Q and TC scores for OXBENCH, PREFAB and HOMSTRAD. TC score and Q score correspond respectively to the CS and SPS of the BALIBASE benchmark. We used the *bali_score* [26] program that generates SPS and CS scores based on predetermined reference alignments.

For OXBENCH, PREFAB and HOMSTRAD benchmarks, we computed the Q and TC scores, using the *Q-score* [6] program. Table 1 represents the average of TC and Q of several datasets extracted from OXBENCH. Table 4 and table 5 represents respectively Q and TC scores obtained with a datasets extracted from HOMSTRAD.

For PREFAB benchmark, the comparison is realized between two pairwise sequences alignments instead of two multiple sequences alignments. Thus, Q and TC scores presented in Table 6 are the same.

We used *GeneDoc* [27] software to convert FASTA format to files with MSF format in order to compute the SPS and CS of each alignment. We used the following parameters:

- Gap opening Penalty=10, Gap extension Penalty= 3,
- Substitution Matrix: Blosum62 [28], Blosum80 [28] and VTML200 [29].

Table 1. Scores Obtained with Oxbench

Scores	MUSCLE	CLUSTALW	MAFFT	Pro-malign
TC	75,61	75,63	74,87	**76,43**

Q	61,68	63,82	61,45	**64,22**

Table 4. Q Scores Obtained with Homstrad

Datasets	Clustal mega	T-COFFEE	MAFFT	MUSCLE	Pro-malign
rep	782	764	782	761	**799**
ANK	231	236	775	808	**830**
UBQ	1	986	1	1	**1**
protg	493	539	536	493	**571**
xia	**981**	**981**	977	**981**	979
MIF	**982**	965	965	965	965
ACPS	303	**484**	342	382	470
Invasin	670	**758**	747	747	718
S_100	905	873	875	**986**	**986**
RRF	897	896	**900**	**900**	899

Table 5. TC Scores Obtained with Homstrad

Datasets	Clustal Omega	T-COFFEE	MAFFT	MUSCLE	Pro-malign
rep	533	520	533	507	**587**
ANK	7,94	0	690	714	**754**
UBQ	1	973	1	1	**1**
protg	15,6	109	109	15,6	**156**
xia	**947**	942	939	**947**	942
MIF	**947**	**947**	**947**	**947**	**947**
ACPS	18,2	282	191	164	**327**
Invasin	557	**670**	639	639	660
S_100	745	673	694	979	**979**
RRF	796	801	801	801	**801**

Table 6. Q/TC Scores Obtained with Prefab

Datasets	Clustal Omega	MAFFT	MUSCLE	Pro-malign
1prtF_2bosA	0	0	106	**288**
1cmbA_1mjoB	919	919	919	**919**
1fmb_2hpeA	929	929	929	**949**
1evsA_1lki	579	566	179	**738**
1aqzA_9rnt	**660**	**660**	553	585
1debA_1fe6A	0	0	**231**	**231**
1eaiC_1ate	929	911	911	**964**
1hmcB_1jli	167	156	117	**729**
1exzA_1hmcB	539	**626**	**626**	530
1f53A_1g6eA	711	**855**	737	803
1d2iA_1es8A	906	**922**	**922**	**922**

In several datasets extracted from different benchmarks cases, we obtained better scores than those of typical multiple alignment programs. Thus, our program can gives the best multiple sequences alignment for several datasets from different benchmarks for protein sequences.

9. Conclusion and Perspectives

In this paper, we presented the approached profile sequence a new consensus sequence that conserve not only identity but also similarity between pairwise alignment. We used the approached profile to compute a new distance approached profile distance. We present also the MSC score function between columns for aligning alignment and finally a new refinement algorithm. We integrate our new methods in a new multiple sequence alignment algorithm, called *Pro-malign*. We benchmarked *Pro-malign* on different datasets extracted from different benchmarks of protein sequences. We obtained good results for different datasets.

As a future work, we plan to apply our algorithm on different other benchmarks of protein, DNA and RNA sequences; we plan to improve experimental results by improving refinement step to detect error on the alignment or regions.

References

[1] Wang, L., & Jiang, T. (1994). On the complexity of multiple sequence alignment. *J. Comput. Biol*.

[2] Notredame, C., & Higgins, D. (1996). SAGA: Sequence alignment by genetic algorithm. *Nucl. Acids. Res.*

[3] Riaz, T., Wang, Y., & Li., K. B. (2004). Multiple sequence alignment using tabu search. *Proceedings of the Second Conference on Asia-Pacific Bioinformatics*.

[4] Lee Z. J., Su, S. F., Chuang, C. C., & Liu, K. H. (2008). Genetic algorithm with ant colony optimization (ga-aco) for multiple sequence alignment. *Applied Soft Computing*.

[5] Gupta, R., Agarwal, P., & Soni, A. k. (2014). TSGA-MSA: Trace sequence algorithm for alignment of MSA. *Proceedings of the International Conference on Advances in Computer Engineering and Applications ICACEA*.

[6] Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy high throughput. *Nucleic Acids Research*.

[7] Katoh, K., & Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics*.

[8] Do, C. B., Mahabhashyam, M. S., Brudno, M., & Batzoglou. S. (2005). PROBCONS: Probabilistic consistency-based multiple sequence alignment, *Genome Res*.

[9] Russell, D. J. (2014). GramAlign: Fast alignment driven by grammar-based phylogeny. *Multiple Sequence Alignment Methods*.

[10] Liu, Y. Chaochmidt, Bertil, M., & Douglas L. (2010). MSAProbs: Multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities. *Bioinformatics*. *26(16)*, 1958-1964.

[11] Mokaddem, A., & Elloumi, M. (2012). New distances for improving progressive alignment algorithm, *Proceedings of the 2nd International Conference on Advances in Computing and Information Technolog*.

[12] Yongtao, Y. E., Cheung, D. W., Wang, Y., *et al*. (2013). GLProbs: Aligning multiple sequences adaptively. *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*.

[13] Sneath, P., & Sokal, R. (1973). *Numerical Taxonomy*. Freeman (Publish.).

[14] Saitou, N., & Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol*.

[15] Wheeler, T. J., & Kececioglu, J. D. (2007). Multiple alignment by aligning alignments,. *Bioinformatics*, *23(13)*.

[16] Wallace, I. M., O'Sullivan, O., & Higgins, D. G (2005). Evaluation of iterative alignment algorithms for multiple alignments, Bioinformatics.

[17] Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTALW: Improving the sensitivity of

progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleid Acids Research*.

[18] Notredame, C., Heringa, J., & Higgins, D. (2000). T-coffee: A novel method for fast and accurate multiple sequence alignments. *J. Molecular Biology*.

[19] Mokaddem, A. & Elloumi, M. (2013). Motalign: Multiple sequence alignment algorithm using new distances and score function. *Proceedings of the 4th International Workshop on Biological Knowledge Discovery and Data Mining*.

[20] Sievers, F., & Higgins, D. G. (2014). Clustal omega, accurate alignment of very large numbers of sequences.

[21] Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino-acid sequence of two proteins. *Journal of Molecular Biology*.

[22] Bonizzoni, P., & Gianluca. D. V. (2001). The complexity of multiple sequence alignment with SP-score that is a metric. *Theoretical Computer Science*.

[23] Thompson, J. D., Koehl, P., Ripp, R., & Poch, O. (2005). BALIBASE 3.0: Latest developments of the multiple sequence alignment benchmarks. *Proteins*.

[24] Raghava, G. P., Searle, S. M., Audley, P. C., Barber, J. D., & Barton G. J. (2003) *OXBENCH:* A benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics*.

[25] Stebbings, L. A., & Mizuguchi, K. (2004). HOMSTRAD: Recent developments of the homologous protein structure alignment database. *Nucleic Acids Research*.

[26] Thompson, J. D., Plewniak, F., & Poch O. (1999). A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.*.

[27] Nicholas, K. B., Nicholas, J. R., & Hugh, B. (1997). GeneDoc: A tool for editing and annotating multiple sequence alignments. *Distributed by the Autho.*

[28] Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the Natl. Acad. Sci.*

[29] Muller, T., Spang, R., & Vingron, M. (2002). Estimating amino acid substitution models: A comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method. *Mol Biol Evol*, *19(1)*, 8-13.

Ahmed Mokaddem was born in Ariana, in 1982. He obtained a master degree in 2008 from the University of Tunis, Tunisia. He received the PhD degree from the same University of Tunis in 2014.

He is a computer science teacher in Tunis. His main research areas include computer science, string matching and Bioinformatics. He published articles on conference: *PAAA: A Progressive Iterative Alignment Algorithm based on Anchors,* in Proc. PRIB'11. *New Distances for Improving Progressive Alignment Algorithm*, in Proc. ACITY'12 and *Motalign: Multiple Sequence Alignment Algorithm Using New Distances and Score Function*, in Proc. BIOKDD'13.

Dr. Mokaddem is member of the BioInformatics Group (BIG) of the Laboratory of Technologies of Information and Communication, and Electrical Engineering (LaTICE), Tunisia.

Amine Bel Hadj was born in Marsa, in 1990. He obtained a master degree in 2015 from the University of Tunis, Tunisia.

He is a web developer in private company. His main task is research and software development.

Mr Bel Hadj is member of the Laboratory of Technologies of Information and Communication, and Electrical Engineering (LaTICE), Tunisia

Mourad Elloumi obtained his master's degree in 1989 from the University of Aix-Marseilles III, France. He obtained the Habilitation for conducting research in computer science, from the University of Manouba, Tunisia.

He is professor at the Faculty of Economic Sciences and Management of Tunis (FSEGT), University of Tunis-El Manar, Tunisia. His main research areas include computer science; string matching; bioinformatics, knowledge discovery and data mining, pattern. He published different papers and books: Recognition in Computational Molecular Biology: Techniques and Approaches (New Jersey, USA, Wiley, 2016), Biological Knowledge Discovery Handbook: Preprocessing, Mining and Postprocessing of Biological Data(New Jersey, USA, Wiley, 2014) and Algorithms in Computational Molecular Biology : Techniques, Approaches and Applications (New Jersey, USA, Wiley, 2011)

Prof. Elloumi is the head of the BioInformatics Group (BIG) of the Laboratory of Technologies of Information and Communication, and Electrical Engineering (LaTICE) Tunis, Tunisia.