Available online at www.ijpab.com

ISSN: 2320 – 7051
Int. J. Pure App. Biosci. **2 (3):** 139-144 (2014)

Research Article

INTERNATIONAL JOURNAL OF PURE & APPLIED BIOSCIENCE

Keyword based Iterative Approach to Multiple Sequence Alignment

Himanshu S Mazumdar[1], Ankita C Baravaliya[2] and Maulika S Patel[3]*

[1]Head, Research and Development Center, Dharmsinh Desai University,
Nadiad, Gujarat, India, http://www.ddu.ac.in

[2]Student, Department of Information Technology, Dharmsinh Desai University,
Nadiad, Gujarat, India, http://www.ddu.ac.in

[3]Head, Department of Computer Engineering, G H Patel College of Engineering & Technology,
Vallabh Vidyanagar, Gujarat, India, http://www.gcet.ac.in

*Corresponding Author E-mail: maulika.sandip@gmail.com

ABSTRACT

In many research applications large number of similar looking peptide sequences needs to be analyzed for study of small differences using visual alignment technique called Multiple Sequence Alignment. For better understanding of proteins and their functions, it is necessary to align the strong bonds of each sequence and observe the changes in weak bonds. Multiple sequence alignment identifies and quantifies similarities and differences among several proteins visually or graphically. The dissimilarities in multiple sequences can be due to evolutionary processes such as mutation, insertion or deletion of amino acid residues. In multiple sequence alignment, most of the technique uses pair wise alignment method which is time consuming and computationally intensive. Performance of the algorithm presented here is found more efficient compared to recently reported techniques.

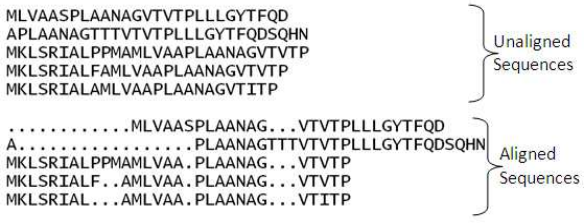
Keywords- *keyword, Occurring Frequency, backtracking, threshold value, iterative alignment.*

INTRODUCTION

Proteins are polymers of amino acids and can be viewed as a sequence of characters using 20 letter alphabet set excluding {B, O, U, J, X, Z} wherein each alphabet corresponds to an amino acid. A protein folds to form a stable tertiary structure. Each structure is capable of performing a unique function. Proteins are basic building blocks of cell and involved in various tasks. The string nature of protein helps in utilizing powerful algorithms of text processing to compare and align for similarity and minute changes. By aligning two or more protein sequences, it is possible to identify any relation between the sequences in terms of structure or function.

Multiple Sequence Alignment (MSA) is an essential tool for phylogeny inference, protein structure and thereby its function prediction. It is also used for other common tasks in sequence analysis like motif finding, sequence similarity search, structure prediction, identifying evolutionarily or structurally related positions or remote homologs. Aligning three or more sequences of different length can be difficult. From the outcome of MSA, sequence homology can be predicted and phylogenetic analysis can be done to evaluate the correlation between evolutionary origins. Figure 1 explains the basic meaning of MSA with an example. Similarity among sequences shows strong bonds and differences represents weak bonds.

Fig.1: An illustration of Multiple Sequence Alignment



The protein sequences chosen for alignment are assumed to have an evolutionary association by which they share an ancestry. For alignment of a pair of sequences, one must have an idea about which alignment is better than another i.e. a measure of the quality and accuracy of an alignment. Although there are many programs available for pair-wise sequence alignment, the most widely acknowledged tools use variations of the dynamic programming method[1]. Pair-wise Alignment uses align cost, gap cost and penalty and select the alignment which has minimum total cost. Extending these methods to multiple sequences causes a number of problems, among which are how to measure the cost of a multiple alignment and how to choose gap costs reliable with the measure chosen[2].

LITERATURE REVIEW

MSA is one of the most elementary problems in computational molecular biology. The best known scheme for finding an optimal alignment uses dynamic programming in which execution time increases exponentially with the number of sequences to be aligned[1]. Heuristic techniques have gained popularity to reduce the time complexity of alignment. Some of the methods in this group include DiAlign[3], ClustalX (i.e. Windows version of ClustalW)[4], T-Coffee (Tree-based Consistency Objective Function For alignment Evaluation)[5], MAFFT (Multiple Sequence Alignment by Fast Fourier Transform)[6] and MUSCLE (Multiple Sequence Alignment by Log Expectation)[7]. These methods make use of pairwise alignments as a basis for multiple alignments. Popular techniques for MSA can be categorized as:

- Progressive Alignments
- Iterative Alignments

Progressive Alignments are the most widely used technique in which sequences are added one after another for alignment. First it generates the guide tree which shows relationship between sequences and according to that it adds further sequences. This approach has benefit of speed and simplicity both with rational sensitivity[2].

ClustalW[8] is a tool that computes pair-wise alignments for all against all sequences and stores the scores in a matrix. A guided tree is built which will suggest the order in which pairs of sequences are to be aligned and combined with previous alignments to obtain an MSA.

In T-Coffee[5], sequences are aligned in a progressive manner using a consistency-based objective function. It uses results from ClustalW[4] and other similar programs as input sequences and produces more accurate alignment of distantly related proteins.

COBALT generates constraints and uses them to create a multiple alignment. It is an example of progressive alignment method[9].

MAFFT implements a combination of both, progressive and iterative heuristics. It assumes that the input sequences are all homologous that are descended from a common ancestor. It also provides the facility of adding unaligned sequences in already align sequences, parallel processing and alignments with different constraints[6].

In *Iterative Alignments* method, optimal alignment is achieved by improving the alignments in each iteration. Every such improvement is considered as iteration[2]. The method for improvement can be either stochastic or deterministic. When no more enhancements are observed, the iterative procedure can be terminated.

DiAlign is an iterative program that first focuses on local alignments between sub-segments or substrings or sequence motifs without having a gap penalty. The alignment of individual sub-segment is then achieved with a matrix representation like dot-matrix plot in a pair-wise alignment[3].

MSA is an important bioinformatics tool. No perfect method exists for assembling MSA and all the available methods do approximations (heuristics). BAliBASE is a standard database consisting of reference alignments used to compare the quality of alignments produced by the various alignment programs. It is suggested in[10] that different alignment methods react in different ways for aligning sequences in BAliBASE.

PROPOSED METHOD

We have developed a methodology for multiple sequence alignment, in which there is no need to build any matrix or a directed tree. A keyword, occurring frequency, threshold value and keyword group are

used to generate an order of group of keywords to be used for alignment. This process includes backtracking and keyword-pair frequency. The terms referred above are explained below.

Keyword: A substring of length 5 obtained from given sequences, which occurs not more than once in any sequence and its frequency is above threshold value.

Occurring frequency: It is number of times a particular keyword occurs in the input sequences.

Threshold value: A predefined minimum frequency value to select a keyword for alignment.

Keyword group: It is collection of most occurring keywords that appear in same order in different sequences.

Ordered Pair: It is non-conflicting pair of keywords occurs in sequences.

Step by step method of proposed MSA algorithm:

1. Load the homologous protein sequences.
2. Find list of keywords based on its frequency and threshold value.
3. Find list of ordered pairs of keywords with their occurrence frequency. Sort the list in descending order of frequency.
4. Make keyword groups using backtracking.
5. Align all sequences by largest group using padding dots.
6. Find the keywords' position of largest group and all other groups.
7. Generate an order of keywords for alignment.
8. Align original sequence using the order of keywords and padding dots.
9. Remove Gap-only columns and display the output.

Above steps are explained with an example in Figure 2. The number of input sequences is five and the threshold value is two.

RESULTS AND DISCUSSION

Figure 3 shows the snapshot which contains the sequences for alignment. Figure 4 shows the image format of these sequences. Each color represents one keyword. This format of representation is useful for viewing and analysis of the alignment results. Alignment of sequences in text format is shown in Figure 5. Figure 6 shows the aligned output in the image format. The gray color displays the dissimilarity and empty columns. Colors are not unique to keyword that is one-to-many mapping is done among color and keywords. The aligned sequences are sorted based on the number of keywords present in the sequences. These keywords can be highlighted for further analysis.

Fig. 2: Methodology is explained with Example

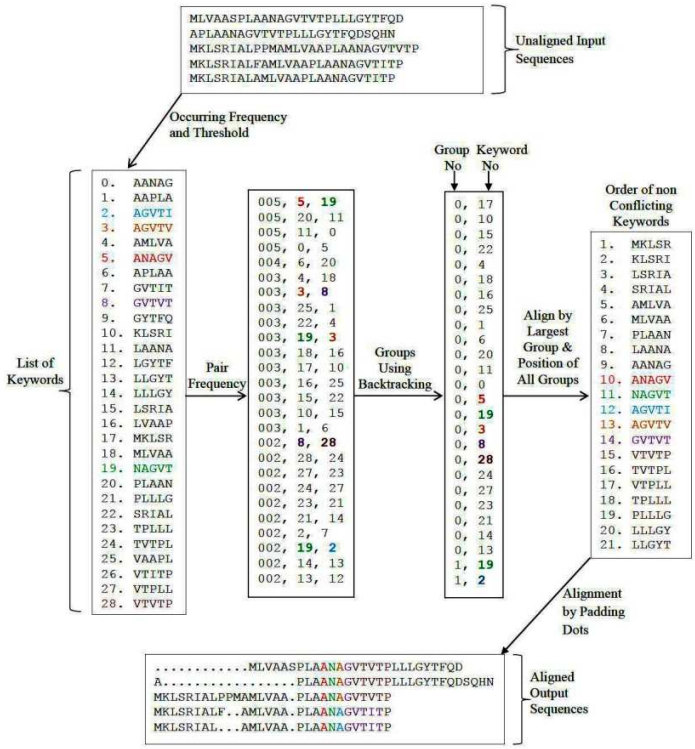


Fig. 3: Protein Sequences without Alignment

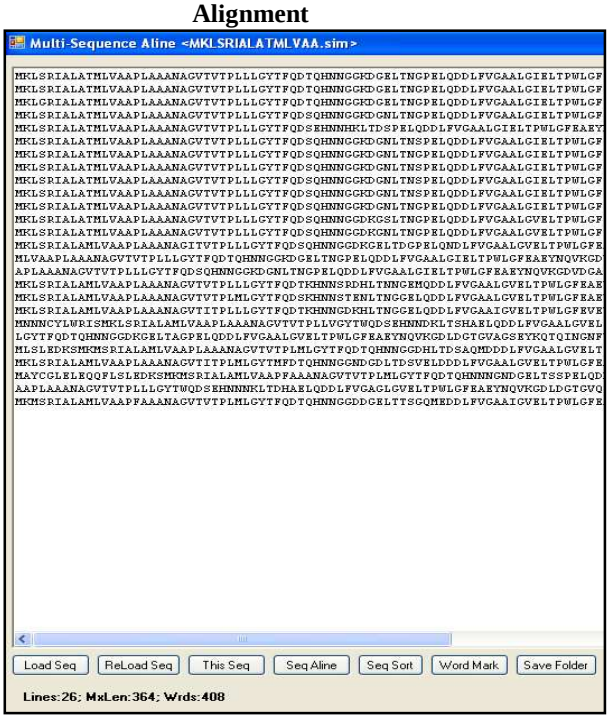


Fig. 4: Protein Sequences in Image Format without Alignment

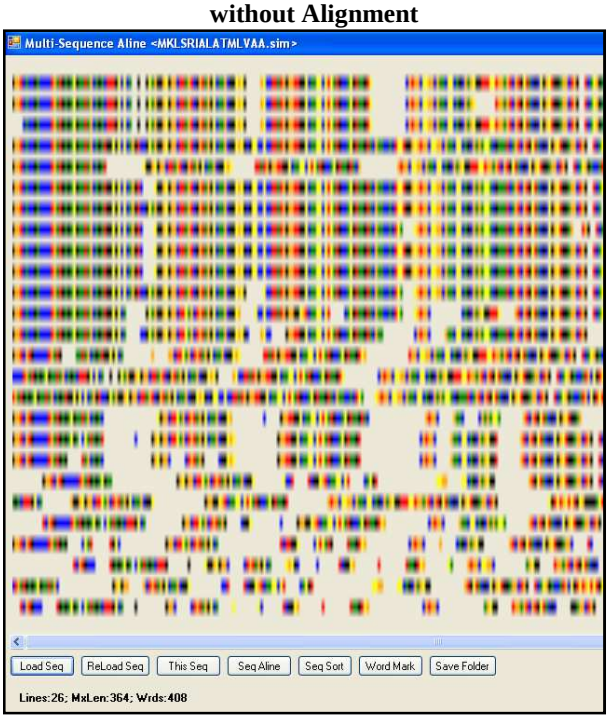


Fig. 5: Aligned Output in Text Format

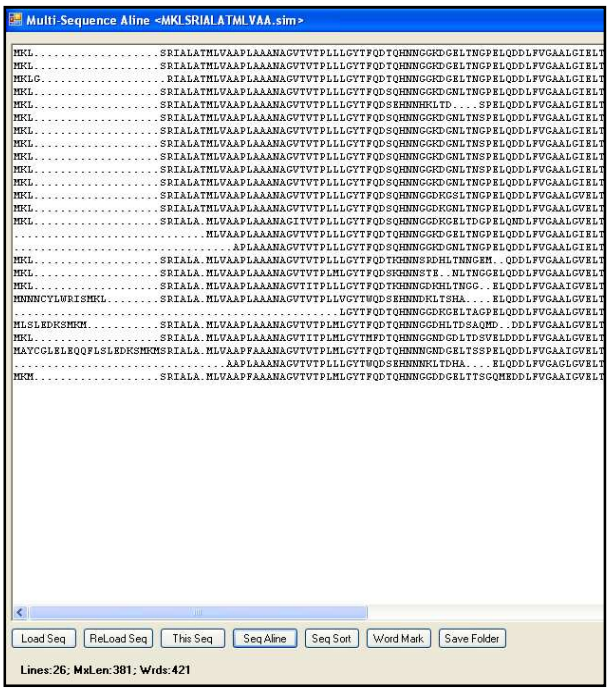
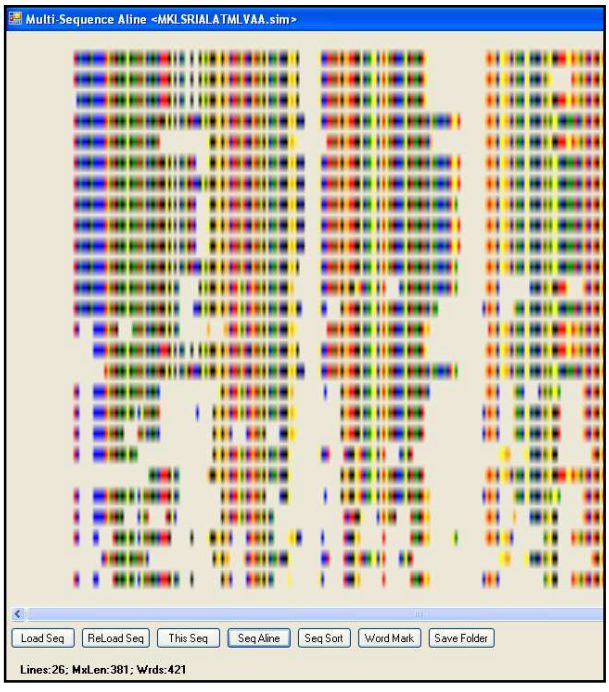


Fig. 6: Aligned Output in Image Format



We have tested our tool with ClustalX and the time taken for alignment of sequences taken from BAliBASE data base is shown in Table 1.

Fig. 7: Comparison with BAliBASE

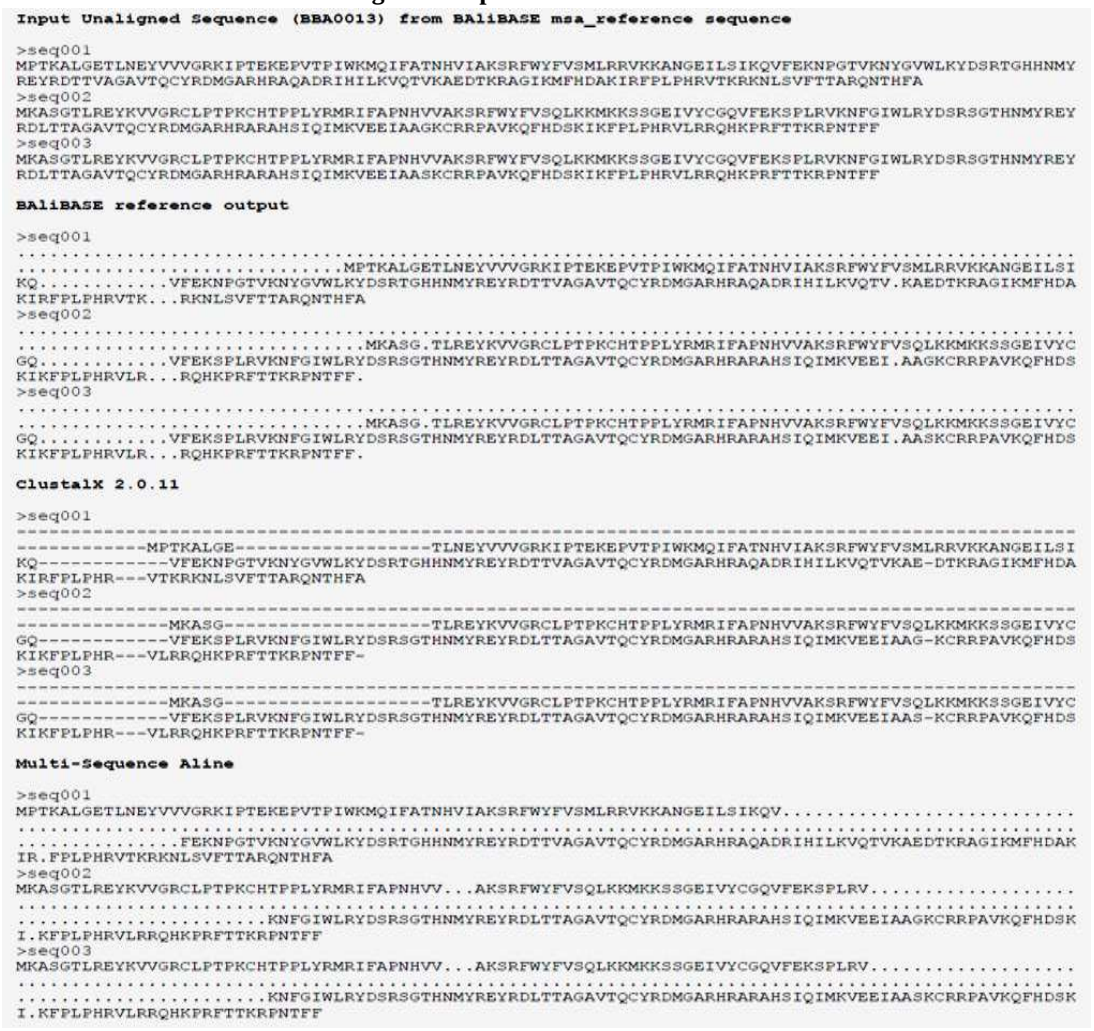


Table 1: Comparison of Time Taken by ClustalX and Multi Sequence Aline Tool

TFA File No.	ClustalX 2.0.11 (Character Based) (min : sec . ms)	Multi Sequence Aline Tool (Segment Based) (min : sec . ms)
1.	00:08.00	00:02.00
2.	00:09.91	00:01.50
3.	00:17.21	00:04.00
4.	00:40.80	00:21.70
5.	00:03.95	00:01.35
6.	00:14.90	00:06.03
7.	00:02.43	00:00.90
8.	00:18.20	00:05.85
9.	00.05.35	00.01.80
10.	00:18.88	00:03.45

We have assessed the quality of alignment using the BAliBASE database. BAliBASE sequence BBA0013.tfa contains 19 sequences. Alignment is performed using our tool, ClustalX and the output is comparable with reference output given in BAliBASE (Figure 7). It is observed that the time taken by our tool is considerably less than ClustalX [Table 1]. Proposed algorithm takes time in generation of groups, alignment of keywords and keyword position finding if number of keywords is very large.

CONCLUSION

A novel technique for Multiple Sequence Alignment is proposed that computes the alignment of protein sequences based on a keyword set and its order. The results are compared with the ClustalX tool and also using the benchmark BAliBASE database. It is possible to adjust the threshold value depending on the number of sequences. The web-based version of the tool will be made available for research use on website www.rndddu.net/ProteinLab.aspx in near future.

REFERENCES

1. Yonatan Bilu, Agarwal Pankaj K. & Rachel Kolodny, "Faster Algorithms for Optimal Multiple Sequence Alignment Based on Pairwise Comparisons", IEEE/ACM Transactions On Computational Biology And Bioinformatics, **3(4):** October-December (2006)
2. C. Notredame,"Recent Progress in Multiple Sequence Alignment: A Survey", Pharmacogenomics, **3(1):** 131-144 (2002)
3. Burkhard Morgenstern, "Multiple Sequence Alignment with DIALIGN", Methods in Molecular Biology, **1079,**191-202, (2014)
4. Julie D. Thompson , Toby J. Gibson, Frederic Plewniak and Desmond G. Higgins, "The CLUSTAL_X windows interface of ClustalW flexible strategies for multiple sequence alignment aided by quality analysis tools", Nucleic Acids Research, **25(24):** Oxford University, October 28, (1997), UK
5. C. Notredame, D.G. Higgins & J. Heringa, "T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment, *J. Molecular Biology*, **302(1):** 205-217 (2000)
6. Daron M. Standley & Kazutaka Kutoh, "MAFFT Multiple Sequence Alignment SoftwareVersion 7: Improvements in Performance and Usability", Molecular Biology **3(0147):** 772-780, January 16, (2013), Japan.
7. Robert C. Edgar, "MUSCLE: multiple sequence alignment with high Accuracy and high throughput", Nucleic Acids Research, **32(5):** March 19, (2004), USA.
8. Julie D. Thompson, Desmond G. Higgins and Toby J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice", Nucleic Acids Research, **22(22):** (1994)
9. Jason S. Papadopoulos & Richa Agarwala, "COBALT: constraint-based alignment tool for multiple protein sequences", Bioinformatics, **23(9):** 1073-1079, (2007)
10. Julie D. Thompson, Frederic Plewniak and Olivier Poch, "BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs", Bioinformatics, **15(1):** 87-88 (1999)
11. BAliBASE Database: ftp://ftp-igbmc.u-strasbg.fr/pub/msa_reference/.
12. David J. Lipman, Stephen F. Altschul & John D. Kececioglut, "A tool for multiple sequence alignment (proteins/structure/evolution/dynamic programming)", Proc. Natl. Acad. Sci. **86,** 4412-4415, June (1989), USA.
13. Saul B. Needleman & Christian D. Wunsch, "A General Method Applicable to the Search for Similarities in the Amino Acid Sequences of Two Proteins," *J. Molecular Biology*, **48,** 443-453 (1970)
14. Altschul, S.F. et. al. "Basic Local Alignment Search Tool", *J. Molecular Biology*, **215** 403-410, (1990)
15. William R. Pearson & David J. Lipman, "improved tools for Biological Sequence Comparison", Biochemistry, **85,** 2444-2448, April (1998)