Efficient Mining Algorithm for Protein Sequence Alignment

Pyae Phyo Thu, Khin Thidar Lynn
University of Computer Studies, Mandalay, Myanmar
pyaephyothu.ucsm@gmail.com, lynnthidar@gmail.com

Abstract

Bioinformatics is the study and application of computational methods to life sciences data and the application of information technology to store, organize and analyze the vast amount of biological data. Protein sequence alignment is one of the crucial tasks of computational biology which forms the basis of many other tasks like protein structure prediction, protein function prediction and phylogenetic analysis. Alignment algorithms are needed to compare two or more sequences. Pairwise sequence alignment is concerned with comparing two Protein or DNA sequences - finding the global and local "optimum alignment" of the two sequences. Multiple sequence alignment (MSA) is a key step in elucidating evolutionary relationships, annotating newly sequenced segments, and understanding the relationship between biological sequences. So, pairwise alignment algorithm for protein sequences is proposed. Genetic algorithm with Bi-Directional Extension (BIDE)*, an efficient algorithm for mining frequent closed sequences, is used for multiple sequence alignment.*

1. Introduction

Bioinformatics is a newly emerging field and it is an integration of mathematical, statistical and computer methods to analyze biological data. Bioinformatics is defined as deriving knowledge from computational analysis of large volumes of biological and biomedical data. Bioinformatics methods are used in fundamental research on theories of evolution and in more practical considerations of protein design. The central challenge of computational structural biology is rationalize the mass of sequence information into biochemical and biophysical knowledge and to decipher the structural, functional and evolutionary clues encoded in the language of biological sequences [3].

Knowing the properties of biological sequence can be very valuable in analyzing data and making appropriate conclusions. In this context, appropriate characterization of the biological sequence structures and exploitation of biosequence properties consider important step to develop and create powerful algorithms in Bioinformatics. Biodata, or more precisely molecular biological data DNA, RNA and proteins, create organism body. Biodata are rich of information and have many properties. Some of the related properties are listed briefly:

- Small alphabet, biosequence alphabet (DNA, RNA and proteins) regards small when compared with transaction sequences (e.g. market-basket analysis). Biosequence typically requires an alphabet of size less than 21; DNA and RNA consist of four alphabets and proteins consist of 20 alphabets [11], [12].
- Long sequences, biological sequences carry full detail information about organism species in the genes. Biosequences are long, for example chromosome 1 of the human sized 243 megabytes and human genome sized more than 3 gigabytes. Therefore, long sequences considered an important property of biological sequence data set [13], [14], [15].
- Mutation, it is the most outstanding property that distinguishes between biosequences and transactional sequences. Occurrences of patterns are not always identical; some copies may be approximated. The biosequence pattern usually allows nontrivial numbers of insertions, deletions, and other mutations. The instances of the pattern usually differ from the model in a few positions. Mutation represents a real challenge of sequential pattern mining [13], [15], [16].

Bioinformatics is the application of information technology to store, organize and analyze the vast amount of biological data which is available in the form of sequences and structures of proteins and nucleic acids. It includes protein sequence analysis, drug discovery, gene prediction and genome analysis. One of the most central methods in bioinformatics is the alignment of two protein or DNA sequences. Biological databases can be broadly classified in to sequence and structure databases. Sequence databases are applicable to both nucleic acid sequences and protein sequences, whereas structure database is applicable to only Proteins. There are several bioinformatics tasks and applications on sequence and structural analysis of biological data. The most important one is computational sequence analysis.

Sequence alignment is a fundamental procedure conducted in any biological study that compares two or more biological sequences (whether DNA, RNA, or protein). Determination of protein/peptide sequences is a basic requirement for biomedical research, including cancer research. It is absolutely essential for characterizing and identifying proteins or peptides [4].

Sequence alignment plays an important role in drug design, its help in detecting any abnormal changes in protein sequences to which the drug is subjected. As a result proper drug can be designed with reduced side effects. Aligning plays an important role in drug design, forensics, DNA defects etc. The demand for faster and optimizing algorithm would also be at high peak for bioinformatics due to increasing need of better drugs and treatment.

2. Related Work

Biological sequences databases are growing exponentially resulting in extensive demands on the implementation of new fast and efficient sequence alignment algorithms. Most of the work in the sequence alignment field has been primarily intended on providing new fast and efficient alignment methods. Alexander Chan analyzed the complexities of pairwise sequence alignment algorithms [1].

Needleman and Wunsch proposed an algorithm based on dynamic programming for global alignment of two sequences. This algorithm first calculates a scoring matrix for the two given amino acid sequences A and B, by placing one sequence along row side and other column side. The size of matrix is (k+1)*(m+1) (K and M are lengths of the two sequences). The optimal score at each matrix position is calculated by adding the current match score to previously scored position and subtracting gap penalties, which may evaluate to either a positive, negative or 0 value. An alignment is computed using this scoring matrix by trace back procedure, inserting gaps in the sequences so as to get optimal alignment of the sequences [8].

Smith and Waterman proposed an algorithm to find a pair of segments one from each of two long sequences such that there is no other pair of segments with greater similarity (homology). In this local alignment algorithm, similarity measure allowed arbitrary length deletions and insertions [7].

Anitha and Poorna suggested an algorithm for global alignment between two DNA sequences using Boolean algebra and compare the performance of the algorithm with Needleman-Wunsch algorithm [9].

Bandyopadhyay et. al proposed direct comparison methods to obtain global and local alignment between the two sequences; the method proposed an alternate scoring scheme based on fuzzy concept [10].

Chang et al. established fuzzy PAM matrix using fuzzy logic and then estimated score for fitness function of genetic algorithm using fuzzy arithmetic. Their experimental results evidenced fuzzy logic useful in dealing with the uncertainties problem, and applied to protein sequence alignment successfully [11]. The sole aim of the researchers has been to develop efficient alignment algorithms based on different and latest techniques.

There are a number of MSA methods in the literature. A direct method is dynamic programming, which is used to find the global optimal solution. It makes use of a substitution matrix and a gap penalty. However, this method is not practical in that the time complexity increases exponentially as the number of sequences increases. An alternative is the progressive approach, which uses a heuristic search. The most popular progressive alignment method is ClustalW [17]. However, progressive methods heavily rely on initial alignment. So sometimes it compromises accuracy at the cost of improved efficiency. Other MSA methods include iterative methods and Hidden Markov models.

The methods mentioned above lack the ability to effectively search the huge solution space. Thus in recent decades genetic algorithms (GAs) have been proposed to solve MSA problems [18, 19].

3. Protein Sequence Alignment

Sequence alignment has become an essential part of biological science. There are now many different techniques and implementations of methods to perform alignment of sequences. There are two types of sequence alignment: local alignment and global alignment. The local alignment seeks for segments of the two sequences that match well. The global alignment attempts to match both the sequences to each other from end to end. Sequence alignment is the procedure of comparing two (pair-wise alignment) or more (multiple-alignment) sequences by searching for a series of individual characters or character patterns that are in the same order in both sequences. Two sequences are aligned by writing them across a page in two rows. Identical or similar characters are placed in the same column, whereas non-identical characters are either placed in the same column as a mismatch or are opposite a gap in the other sequence. The basis for comparison of proteins using the similarity of their sequences is that the proteins are related by evolution; they have a common ancestor. Analysis of evolutionary relationships between protein sequences depends critically on sequence alignments. An optimal

alignment is operationally defined as the pairwise alignment with the highest alignment score for a given scoring scheme [2].

Proteins are built from an alphabet of twenty smaller molecules, known as amino acids. The primary structure is the sequence of amino acids in the polypeptide chain. Ala, Arg, Asn, Asp, Cys, Glu, Gln, Gly, His, Ile, Leu, Lys, Met, Phe, Pro, Ser, Thr, Trp, Tyr and Val represent the three letter abbreviation of amino acids. A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y and V represent the one letter abbreviation of amino acids. In sequence alignment, one letter abbreviation of amino acids is used. Each protein is characterized by its sequence. To predict the biological functions of one protein and the roles of its residues, we usually compare the sequence of this protein with similar protein sequences whose functions have been examined experimentally. Protein sequence alignment is the task of identifying evolutionarily or structurally related positions in a collection of amino acid sequences. As databases of protein sequences and properties increase in size, it becomes more and more reliable to depend on previously classified proteins to determine the structure and function of a novel protein.

The protein sequence analysis in bioinformatics is done by comparing the sequences residue wise. There are two types of protein sequence alignment.

(i) Pairwise Sequence Alignment
(ii) Multiple Sequence Alignment

For pairwise sequence analysis or for multiple sequence analysis, the protein sequences are compared with amino acid by amino acid. Proteins are essential parts of organisms and participate in virtually every process within a cell. One method of determining homology between two proteins is through a pair-wise sequence alignment of their primary structures. It has been found that two proteins that are homologous, such that they were evolutionarily derived from a common protein, tend to align well with a large number of identical or highly similar residues in similar positions along the sequences.

The system architecture of the research work is as follows:

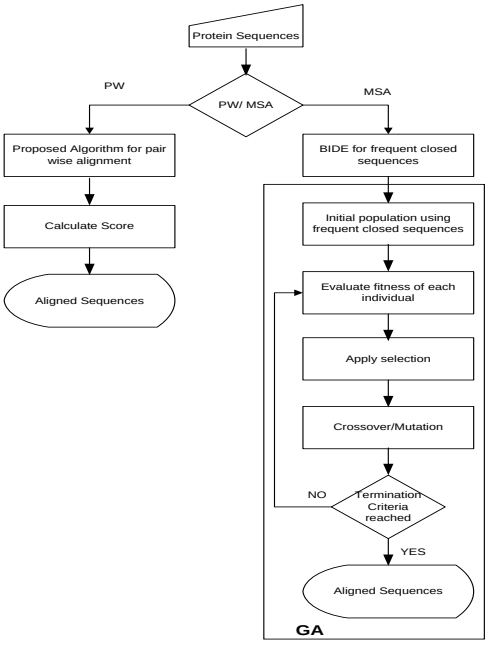

Figure 1: System architecture

3.1. Proposed Algorithm for Pairwise Sequence Alignment

Alignments are one of the most basic and important ways to measure similarity between two or more sequences. Why are two sequences compared? Database searches are useful for finding homologues. Database searches don't provide precise comparisons. More precise tools are needed to analyze the sequences in detail including– Dot plots for graphic analysis and – Local or global alignments for residue/residue analysis. The alignment of two sequences is called a pairwise alignment. In general, a pairwise sequence alignment is an optimization problem which determines the best transcript of how one sequence was derived from the other. In order to give an optimal solution to this problem, all possible alignments between two sequences are computed using a proposed algorithm and a dynamic programming approach (Sequence Manipulation Suite). Scoring schemes allow the comparison of the alignments such that the one with the best score can be picked. Figure 2 describes the process of proposed pairwise algorithm and dynamic programming algorithm.

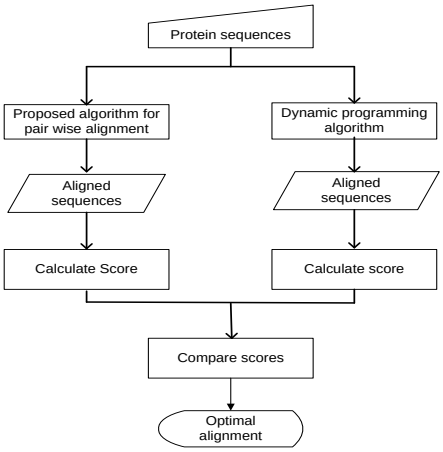

Figure 2: System flow diagram for pairwise sequence alignment

Proposed Algorithm

Input: Two sequences s and t
Output: Aligned sequences
Method:

(1) Set n to be the length of the first sequence.
Set m to be the length of the second sequence.

End for
End for

(2) Retain aligned sequences.

(3) Initialize the first sequence s to 0…n.
(4) Initialize the second sequence t to 0…m.
(5) for (i= 1; i<=n; i++)
(6) for(j=1; j<=m; j++)

If the character of s[i-1] equals to the character of t[j-1],
T[i][j]=T[i-1][j-1].
Else T[i][j]=Math.min (Math.min (T[i-1][j],T[i][j-1]), T[i-1][j-1])+1;

Alignment algorithms try to find the best alignment between two sequences given the scoring system. An alignment program is used to compare the sequence homology between two protein or DNA sequences. These programs find the best match between the two sequences. Occasionally gaps need to be introduced to make the two sequences alignment. A simple scoring scheme is used with a constant gap penalty (G) of -2, a mismatch score (MM) of -1 and a match score (M) of 2.

The proposed algorithm has higher score than the dynamic programming algorithm (SMS). And it can reduce the space and time complexities. The proposed algorithm results the optimal alignment. Table 1 shows the results of sequence alignments by using proposed pairwise algorithm and dynamic programming algorithm.

Table1: Comparison of the alignment results between proposed algorithm and dynamic programming approach (SMS)

Dataset	Sequence Length	Align Length		Match		Mismatch		Gap		Score	
		Proposed Algorithm	SMS	Proposed Algorithm	SMS	Proposed Algorithm	SMS	Proposed Algorithm	SMS	Proposed Algorithm	SMS
gi|6678177|ref|NP_033320.1| syntaxin-4 [Musmusculus]	298	311	314	137	134	139	139	35	41	65	47
gi|151554658|gb|AAI47965.1| STX3 protein [Bostaurus]	289										
gi|6678177|ref|NP_033320.1| syntaxin-4 [Musmusculus]	298	308	314	145	145	132	126	31	43	96	78
gi|37577287|ref|NP_001971.2| syntaxin-2 isoform 1 [Homo sapiens]	287										
gi|151554658|gb|AAI47965.1| STX3 protein [Bostaurus]	289	298	303	186	186	92	87	20	30	240	225
gi|37577287|ref|NP_001971.2| syntaxin-2 isoform 1 [Homo sapiens]	287										

3.2. Multiple Sequence Alignment

Multiple sequence alignment is an important problem in molecular biology, where it is used for constructing evolutionary trees from DNA sequences and for analyzing the protein structures to help design new proteins. To date, most multiple alignment methods are based on a dynamic programming approach. This approach however results in exponential time complexity, since it requires time proportional to the product of the sequence lengths. In general, multiple sequence alignment belongs to a class of hard optimization problems called combinatorial problems. One of the methods that have been developed recently to solve this type of problems is genetic algorithm [21]. Genetic algorithms create a "population" of random solutions and then use the concepts of natural selection, crossover and mutation to improve these solutions. Genetic algorithms have been used successfully in a wide variety of application areas to find solutions for hard optimization problems. They offer the advantage of operating on several solutions simultaneously, combining exploratory search through the solution space with exploitation of current results.

In this study, frequent closed sequence patterns are resulted by using BIDE algorithm. These sequence patterns are used by inserting gaps among suitable patterns for initial population in the genetic algorithm. In conventional genetic algorithm, all the individuals are randomly generated in the initial population. Since the candidate alignment is scaled to be 1.2 times as long as the longest sequence, gaps are inserted into each sequence to fill up the matrix. Crossover is used

to combine genes from the existing chromosomes and create new ones. Then, the best chromosomes are selected to form the next generation. This selection is based on a fitness function which assigns a fitness value to every chromosome. To compare different alignments, a fitness function is defined based on the number of matching symbols and the number and size of gaps. In biology, this fitness function is referred to as cost function and is given biological meaning by using different weights for different types of matching symbols and assigning gap costs when gaps are used. The ones with the best fitness value survive to give offsprings for the new generation, and the process is repeated until satisfactory solutions evolve. BIDE prunes the search space by using the BackScan pruning method and the ScanSkip optimization techniques. By using BIDE algorithm, GA will reduce the next obtained generation of populations, time complexity and space complexity and will have better fitness value as compared to their ancestors.

Table 2. An example sequence database SDB

Sequence identifier	Sequence
1	CAATC
2	ATCT
3	CATC
4	ATTCA

Table 2 shows the input sequence database in running example. The database has totally 3unique items, four input sequences (i.e., $|SDB|=4$). Suppose min_sup = 2.The complete set of frequent closed sequences, S_{fcs} = {AA:2, ATT:2, ATC:4, CA:3, CATC:2,CT:3}, consists of only six sequences, while the whole set of frequent sequences consists of 17 sequences, that is, S_{fs} ={A:4, AA:2, AT:4, ATT:2, ATC:4, AC:4, B:4, TT:2, TC:4, C:4, CA:3, CAT:2, CATC:2, CAC:2, CT:3, CTC:2,CC:2}. Obviously, S_{fcs} is more compact than S_{fs} . Also, if a frequent sequence,S_α, has the same support as that of one of its proper supersequence, S_β , S_α is absorbed by S_β. For example, frequent sequence CTC:2 is absorbed by sequence CATC:2, because ($CTC \sqsubset CATC$)and ($sup^{SDB}(CTC) = sup^{SDB}(CATC) = 2$).

Notice that in the above definition of a sequence, each event contains only a single item. Thus, the derived BIDE algorithm mines only frequent closed single-item sequences. Figure 3 shows the lexicographic frequent tree built from the running example. In figure 3, each node contains a frequent sequence and its corresponding support, and the sequences in the dotted ellipses are non-closed ones. Figure 4 shows the system flow diagram for genetic algorithm with random solutions.

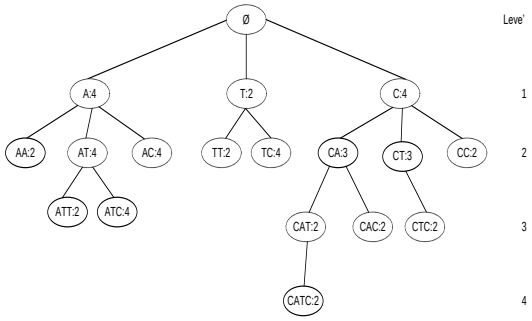

Figure 3: The lexicographic frequent sequence tree

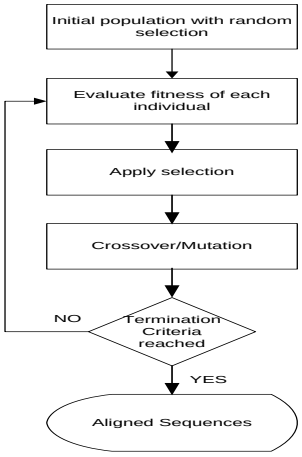

Figure 4: Conventional genetic algorithm

4. Conclusion

Pairwise sequence alignment is an extremely useful tool for DNA and protein sequence analysis. A primary reason for this continued interest in protein sequence alignment is the centrality of comparative sequence analysis in modern computational biology: accurate alignments from the basis of many bioinformatics studies, and advances in alignment methodology can confer sweeping benefits in a wide variety of application domains. Proposed algorithm and dynamic programming algorithm are used to align protein sequences. And calculate the alignment scores which result from these algorithms. After comparing the calculated scores, the final result shows the optimal alignment. The proposed algorithm implements the spaces less than the dynamic programming algorithm. The score of proposed algorithm is greater than the dynamic programming algorithm. So, the proposed algorithm in sequence alignment method gives the better result. Table 2 shows the results for the comparison of the alignments between proposed algorithm and dynamic programming approach (SMS). In multiple sequence alignment GA with BIDE algorithm is used. By using BIDE algorithm, GA will

reduce the next obtained generation of populations, time complexity and space complexity.

References

[1] Alexander Chan , “An Analysis of Pairwise Sequence Alignment Algorithm Complexities: Needleman-Wunsch, Smith-Waterman, FASTA, BLAST and Gaped BLAST”, 5075504 ,Biochemistry 218 ,Final Project Clark KL, Negations as failure, in Gallaire H, Winker J (eds.), *Logic and Data Bases*, Plenum Press, New York, pp. 293–306, 1973.

[2] B. Al-Lazikani , J. Jung, Z. Xiang and B. Honig*, “Protein Structure Prediction”, Department of Biochemistry and Molecular Biophysics, Howard Hughes Medical Institute, Columbia University.

[3] M.S. Rosenberg, “Sequence Alignment”, Concepts and History, Arizona State University.

[4] M.S. Rosenberg, “Sequence Alignment”, Concepts and History, Arizona State University.

[5] P.Johri, “Atomic Level Sequence Analysis-A Review”, Amity Institute of Biotechnology, Amity University Lucknow, Uttar Pradesh, India, Vol. 2, No. 4 (2013): 173-179.

[6] S.R.Shenoy and B.Jayaram, “Proteins: Sequence to Structure and Function- Current Status” , Department of Chemistry & Supercomputing Facility for Bioinformatics and Computational Biology, Indian Institute of Technology Delhi, HauzKhas, New Delhi 110016 , India.

[7] T.F.Smith and M.S.Waterman, "Identification of common molecular subsequence," J. Molecular Biology, vol. 147, pp. 195-197, 1981.

[8] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," J. Molecular Biology, vol. 48, pp. 443-453, 1970.

[9] V.Anitha and B.Poorna, “DNA Sequence Matching using Boolean Algebra”, 2010 International Conference on Advances in Computer Engineering.

[10] S.S. Bandyopadhyay, S.S. Paul, A.Konar, “Improved Algorithms for DNA Sequence Alignment and Revision of Scoring Matrix”, Proceedings of International Conference on Intelligent Sensing and Information Processing, 2005 .

[11] Pin-Teng Chang, Lung-Ting Hung, Kuo-Ping Lin, Chih-sheng Lin, Kuo-Chen Hung, “Protein Sequence Alignment Based on Fuzzy Arithmetic and Genetic Algorithm”, 2006 IEEE International Conference on Fuzzy Systems.

[12] E. Loekito, J. Bailey, and J. Pei, “A Binary Decision Diagram Based Approach for Mining Frequent Subsequences,” Knowl. Inf. Syst., vol. 24, no. 2, pp. 235–268, Sep. 2010.

[13] K. Pavel and P. Vladimir, “Efficient Motif Finding Algorithms for Large-Alphabet Inputs,” BMC Bioinformatics 2010, 11(Suppl 8) :S1, doi: 10.1186/1471-2105-11-S8-S1.

[14] M. Piipari, T. A. Down, and T. J. P. Hubbard, “Large-Scale Gene Regulatory Motif Discovery with NestedMICA,” … Pattern Discov., vol. 7, p. 1, 2011.

[15] F. Hadzic, T. Dillon, and H. Tan, Mining of Data with Complex Structures. 2011, p. 348

[16] H. Chen-Ming, C. Chien-Yu, and L. Baw-Jhiune, “WildSpan: mining structured motifs from protein sequences,” Algorithms Mol. Biol., vol. 6, no. 1, p. 6, 2011.

[17] G. Chen and Q. Zhou, “Heterogeneity in DNA multiple alignments: modeling, inference, and applications in motif finding,” Biometrics, vol. 66, no. 3, pp. 694–704, 2010.

[18] Gibson TJ Thompson JD, Higgins DG. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res, 22:4673–4680, 1994.

[19] K. Chellapilla and G. B. Fogel.Multiple sequence alignment using evolutionary programming. In Proceedings of the 1999 Congress on Evolutionary Computation (CEC’99), pages 445–452, 1999.

[20] E. A. O’Brien C. Notredame and D. G. Higgins. “RAGA:RNA sequence alignment by genetic algorithm”. Nucleic Acids Research, 25:4570–4580, 1997.

[21] J.Wang and J.Han “BIDE: Efficient Mining of Frequent Closed Sequences”Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, U.S.A.

[22] K.Karadimitriou and D. H. Kraft “Genetic Algorithms and The Multiple Sequence Alignment Problem In Biology”, Department of Computer Science, Louisiana State University, Baton Rouge, Louisiana 70803.