Journal of Computational Intelligence in Bioinformatics
ISSN 0973-385X Volume 6, Number 1 (2013) pp. 17-26

http://www.ripublication.com/jcib.htm

Approaches of Biological Sequence Alignment with Comparison of Experimental Results

Niyati J. Buch and Mahesh H. Panchal

ME-CSE Student, Dept. of Computer Engineering,
Kalol Institute of Technology & Research Centre, Gujarat, India
Head & Associate Prof., Dept. of Computer Engineering,
Kalol Institute of Technology & Research Centre, Gujarat, India

Abstract

One of the major research areas in bioinformatics is sequence alignment. There are various optimal as well as heuristic techniques for alignment of biological sequences. Gap penalty and scoring schemes are used to compute the alignment score. Some experiments performed using BLAST are included for better understanding of the heuristic algorithm by comparing the results.

Key words-Bioinformatics; Sequence Alignment; Gap Penalty; Scoring Matrix; BLAST

Introduction

Bioinformatics uses the statistical analysis of protein sequences and structures to help annotate the genome, to understand their function, and to predict structures when only sequence information is available. And as a first step to this analysis, sequences need to be aligned. This can be either local subsequence alignment or global whole sequence alignment. To know if the alignment is good, alignment score is calculated using a scoring scheme i.e. scoring matrix for protein sequences and match-mismatch score for nucleotide sequences. Sometimes to align the sequences, gaps may need to be added to obtain a better alignment. There are various techniques for alignment, optimal and heuristic. Optimal methods are accurate but time consuming. And heuristic methods are faster but one has to compromise with accuracy. This paper details sequence alignment, scoring schemes and the result of some experiments performed using a tool based on heuristic local alignment algorithm.

Section II introduces the field of bioinformatics, various applications and challenges in the area. Section III defines what sequence alignment is, its types and types of gap penalty required during the alignment of biological sequences. In section IV, the types of scoring matrices and comparison between them are discussed. Section V gives an overview to heuristic techniques for sequence alignment esp. BLAST and in section VI, comparison among some experiments of local sequence alignment is shown.

Bioinformatics, Applications and Challenges

Bioinformatics is a branch of biological science which deals with the study of methods for storing, retrieving and analyzing biological data, such as nucleic acid (DNA/RNA) and protein sequence, their structure and functions and genetic interactions.

The primary goal of bioinformatics is to increase the understanding of biological processes. However, it also focuses on developing and applying computationally intensive techniques to achieve this goal. Bioinformatics involves the creation and advancement of databases, algorithms, computational and statistical techniques and theory to solve formal and practical problems arising from the management and analysis of biological data. Common activities in bioinformatics include mapping and analyzing DNA and protein sequences, aligning different DNA and protein sequences to compare them, and creating and viewing 3-D models of protein structures.

There are various applications [1] of bioinformatics. In forensic science, sequencing is used to identify particular individual and also to determine the paternity of the child. Sequencing can be used to predict physiological and behavioral traits. DNA sequencing enables to determine the genome sequence. The genes can be identified which are responsible for causing genetic diseases like Alzheimer's disease, cystic fibrosis and many other diseases caused by the disability of genes to function properly. Acquired diseases like cancers can also be detected by observing certain genes. In gene therapy [2], sequencing is used to identify the defected genes and replace them with the healthy ones. In agriculture, specific genes are used in some plants to increase their resistance against insects and pests and increase the productivity and nutritional value (Genetically Modified Food). Sequencing is also required in the procedure of cloning.

Protein structure prediction is one of the major challenges in bioinformatics. It is the prediction of the three-dimensional structure of a protein from its amino acid sequence. Homology searches is very important feature of bioinformatics application but accurately and optimally performing the same for a large database for various species and at different evolutionary hierarchy is quite an issue. Also one needs to do multiple alignments in order to detect regular patterns in families of proteins and to generate sequence genomes by superimposing nucleotides. Multiple alignments can be also used in phylogeny construction. Though multiple alignments can be used for many purposes but

implementing the same and reducing computational time for it is a major issue. Genomic sequence analysis and gene-finding is a one of key areas in bioinformatics that deals with about with a million or more nucleotide sequences and it is quite a challenging task to mine or extract data from it.

Sequence Alignment

A sequence is a set of objects which is listed in a specific order, one after another. In bioinformatics, a sequence under consideration is either a protein sequence or nucleotide sequence (DNA/RNA) [3]. A protein sequence is represented by a string of letters coding for the 20 different types of amino acid. The full names of the amino acids are rarely given; instead, 3-letter or 1-letter abbreviations are usually recorded for conciseness. Nucleic acids (DNA/RNA) consist of a chain of linked units called nucleotides. A, C, G, and T represents the four nucleotide bases of a DNA strand - Adenine, Cytosine, Guanine and Thymine [4].

Before two sequences are compared, a sequence alignment needs to be produced i.e. to find the optimal alignment between the two sequences. A sequence alignment is a way of arranging the sequences of DNA, RNA or protein to identify regions of similarity that may be a consequence of functional, structural or evolutionary relationships between the sequences [3]. Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix. Gaps are inserted between the residues so that identical or similar characters are aligned in successive columns.

The purpose of alignment is to highlight similarity between the sequences [5]. Similar sequences may share a common ancestral sequence. And due to common ancestry, similar sequences have similar functionality. Alignment of two residues implies that those residues are performing similar roles in the two different proteins. This allows the information known about specific residues in one sequence to be potentially transferred to the residues aligned in the other sequence [6].

Global Alignment [7]

Global Alignment aims to align as many characters in each sequence as possible. It is most useful when the sequences in the query set are similar and of roughly equal size. An optimal alignment is computed that is required to extend from the start of the given sequences to their ends. Optimal global alignment can be obtained by using Needleman-Wunsch Algorithm.

Two sequences are inputs to the Needleman-Wunsch Algorithm and scoring scheme is chosen in accordance to the type of sequence. If it is a protein sequence then scoring matrix either PAM or BLOSUM is used and if it is a nucleotide sequence then match score and mismatch score is decided. Also, the gap penalty score either linear (constant) or affine gap penalty is decided. Generate similarity matrix using the algorithm. Start from the right lower corner of similarity matrix and backtrack from this position until end of matrix

is encountered to find global alignment.

E.g.: Sequences FTFTALILLAVAV and FTALLLAAV can be globally aligned as:

```
F T F T A L I L L A V A V
|     | | |   | | |   | |
F - - T A L - L L A - A V
```

Local Alignment [7]

Local Alignment focuses on segments of sequence with the highest density of matches. Local alignment seeks an alignment that is highest scoring among all alignments between an arbitrary section of the first sequence and an arbitrary section of the second sequence. It is more useful for dissimilar sequences that may contain regions of similarity or for similar sequence motifs (patterns) within their larger sequence context. Optimal global alignment can be obtained by using Smith-Waterman Algorithm.

Input two sequences to Smith-Waterman Algorithm and generate the similarity matrix. Required parameters are similar to those used in Needleman-Wunsch Algorithm. Find the maximum value from similarity matrix and backtrack from this position until zero or end of matrix is encountered to find local alignment.

E.g.: Sequences FTFTALILLAVAV and FTALLLAAV can be locally aligned as:

```
F T F T A L I L L - A V A V
    | | | |   | |   | |
- - F T A L - L L A A V - -
```

Gap Penalty [8]

A gap is a space introduced into an alignment to compensate for insertions and deletions in one sequence relative to another. Gaps are needed to be introduced during alignment to obtain optimal alignment score. Gap penalty is a pre-defined negative score for each gap in the sequence. Gap penalty contributes to the overall score of alignments, and therefore, the size of the gap penalty relative to the entries in the similarity matrix affects the alignment that is finally selected. Selecting a higher gap penalty will cause less favorable characters to be aligned, to avoid creating as many gaps.

There are two methods to calculate gap penalty.

- Constant gap penalty: It has only one parameter (d) which is a penalty per unit length of gap.
- Affine gap penalty: It uses a gap opening penalty (G = 10 to 15) and a gap extension penalty (L = 1 to 2). It there is a gap of length n, then affine gap penalty is G + L * (n).

Suppose values are G = 10, L = 1, one continuous gap of size 8 will result in an affine gap penalty of 1 * (10 + 7 * (1)) = 17. While four separate gaps of size 2 will result in an affine gap penalty of 4 * (10 + 1 * (1)) = 44.

Hence, few large gaps are better than many small gaps. The value of L is always smaller as compared to G so as to encourage gap extension rather than a gap introduction.

Scoring Matrices [9] [10]

Substitution matrices (also known as scoring matrices) are used to assign individual scores to the aligned sequence positions. A substitution matrix defines values for all possible pairs of residues. It describes the rate at which one character in a sequence changes to other character.

PAM (Point Accepted Mutation) [11]

PAM units are used to measure the amount of evolutionary distance between two amino acid sequences. One PAM unit is an evolutionary time period over which 1% of the amino acids in a sequence are expected to undergo accepted mutations. For any specific pair (Ai, Aj) of amino acids, the (i, j) entry in the PAM-N matrix reflects the frequency at which Ai is expected to replace Aj in two sequences that are N PAM units diverged. Example of PAM matrices are PAM 250 and PAM 120.

BLOSUM (BLOcks of Amino Acid SUbstitution Matrix) [12]

PAM is derived from global alignments of proteins, while BLOSUM comes from alignments of shorter sequences or blocks (A block is a short contiguous interval in a multiple alignment of amino acid sequences.) of sequences that match each other at some defined level of similarity. The BLOSUM method incorporates much more data into its matrices, and is therefore more accurate. The matrix built from blocks with no more than x% of similarity is called BLOSUM-x. E.g. the matrix built using sequences with no more than 62% similarity is called BLOSUM-62. The some of BLOSUM matrices are BLOSUM 45, BLOSUM 62 and BLOSUM 80.

TABLE 1: COMPARISON BETWEEN PAM AND BLOSUM MATRIX

PAM	BLOSUM
Used for global alignment	Used for local alignment
Based on explicit evolutionary model	Based on experiments
Covers a smaller dataset than BLOSUM	Covers a larger dataset than PAM
Tries to match entire sequence	Focuses on highly mutable regions
Higher number means more evolutionary distance	Higher number means higher similarity

Tools for Sequence Alignment

When dynamic programming techniques are used to compute similarity between two sequences, the time and space cost will be O (n * m) which is not feasible for larger sequences. So, heuristic or approximate algorithms like FASTA [13] [14] and BLAST [15] were developed to speed up the process while attempting to keep as much sensitivity as possible. Tools for both these algorithms, FASTA [16] and BLAST [17] are available online and free for academic use.

BLAST (Basic Local Alignment Search Tool) [17] provides the following programs:

TABLE 2: VARIOUS PROGRAMS IN BLAST

Program	Query	Database
BLASTP	Protein	Protein
BLASTN	DNA	DNA
BLASTX	Translated DNA	Protein
TBLASTN	Protein	Translated DNA
TBLASTX	Translated DNA	Translated DNA

To begin working with BLAST tool, first choose a program depending on what type of query one has (a protein sequence, a DNA sequence or a translated DNA sequence) and with what database one wants to search in. Various databases are available online for both protein sequences and DNA sequences.

Once the program is chosen, enter the query sequence. BLAST takes FASTA format, bare sequence or sequence identifiers as input query. Also one can upload a query file (text file containing queries formatted in FASTA format). If the sequence is too long or one need to search for only a segment then range of segment can also be provided.

Next one needs to choose the search set or the database for searching the query sequence. BLAST provides with a default option but user can change as

per requirement. Also user can choose the program i.e. specific algorithm using which the alignment would be done.

BLAST also provides with default setting of algorithm parameters: General Parameters like maximum target sequences, expected threshold, word size (the length of the seed that initiates an alignment), etc. and Scoring Parameters like scoring matrix and gap costs. Also parameters specific to the algorithm may need to be set. But user can change these parameters as per requirement. With each option help is also provided that describes what each option or parameter does.

After all the options and parameters are set according to requirement, user just needs to click on BLAST button at the end of the webpage and wait until the tool does the search and obtains the results of the algorithm.

EXPERIMENTS IN BLAST

For these experiments, max target query was taken as 100. The option to automatically adjust parameters for short input sequences was chosen. Expected threshold was taken as 10. For protein query, word size was taken 3 and for nucleotide query word size was taken 11. Other specifically chosen parameters are given in the TABLE 3.

One of the best 100 results given by BLAST as output is shown in the TABLE 4 with values of various output parameters.

TABLE 3: INPUT PARAMETERS VALUES FOR EXPERIMENTS

Parameters	Experiment 1	Experiment 2	Experiment 3	Experiment 4	Experiment 5
Program	Standard Protein BLAST	Standard Protein BLAST	Standard Protein BLAST	Standard Nucleotide BLAST	Standard Nucleotide BLAST
Input query	ALYLVCGERGFFYTPKTR	MVLSPADKTNVKAAWGKHGSAGAEALERMFLSFPTTKTYFPHFDLSHGSAQ	MNVTSLGELEKACVTIP	CTCGTAACCAACCGAGAGAGA	ACGTCGTAGTTCCAGTC
Database	Non-redundant protein sequences(nr)	Non-redundant protein sequences(nr)	Non-redundant protein sequences (nr)	Human genomic plus transcript	Human genomic plus transcript
Algorithm	blastp	blastp	blastp	Blastn	blastn

Scoring Scheme	BLOSUM 62 matrix	BLOSUM62 matrix	BLOSUM 62 matrix	Match: 2 Mismatch: -3	Match: 1 Mismatch: -1
Gap Costs	Existence: 11 Extension: 1	Existence: 10 Extension: 1	Existence: 12 Extension: 1	Existence: 5 Extension: 2	Existence: 2 Extension: 1

TABLE 4: OUTPUT PARAMETERS VALUES OBTAINED FROM EXPERIMENTS

Results	Experiment 1	Experiment 2	Experiment 3	Experiment 4	Experiment 5
Accession	AEG19452.1	AAN04486.1	AAG35266.1 AF215934_1	NT_032977.9	NT_005612.16
Description	insulin [Homo sapiens]	hemoglobin alpha-2 [Homo sapiens]	Smad2 [Schistosoma mansoni]	Homo sapiens chromosome 1 genomic contig, GRCh37.p5 Primary Assembly	Homo sapiens chromosome 3 genomic contig, GRCh37.p5 Primary Assembly
Raw score	140	238	60	15	14
Bit score	62.2	88.8	28.6	30.2	28.2
E value	1e-11	2e-21	37	19	38
Max Ident	100%	87%	48%	100%	100%
Identities	18/18 (100%)	48/55 (87%)	12/25 (48%)	15/15 (100%)	14/14 (100%)
Positives	18/18 (100%)	49/55 (89%)	12/25 (48%)	-	-
Gaps	0/18 (0%)	4/55 (7%)	9/25 (36%)	0/15 (0%)	0/14 (0%)

Experiment 1, 2 and 3 were done using program blastp and database as non-redundant protein sequences (nr). Scoring scheme used was BLOSUM62 matrix. For experiment 1, input query of 18 characters was given and gap costs were taken as 11 as gap existence penalty and 1 as gap extension penalty. As output, the input query matched with insulin [Homo sapiens] with 100% ident and 0% gap. For experiment 2, input query of 55 characters was given and gap costs were taken as 10 as gap existence penalty and 1 as gap extension penalty. As output, the input query matched with hemoglobin alpha-2 [Homo sapiens] with 87% ident and 7% gap. For experiment 3, input query of

25 characters was given and gap costs were taken as 12 as gap existence penalty and 1 as gap extension penalty. As output, the input query matched with hemoglobin alpha-2 [Homo sapiens] with 48% ident and 36% gap.

Experiment 4 and 5 were done using program blastn and database as Human genomic plus transcript. For experiment 4, input query of 15 characters was given, scoring scheme taken was match = 2 and mismatch = -3 and gap costs were taken as 5 as gap existence penalty and 2 as gap extension penalty. As output, the input query matched with Homo sapiens chromosome 1 genomic contig, GRCh37.p5 Primary Assembly with 100% ident and 0% gap. For experiment 5, input query of 14 characters was given, scoring scheme taken was match = 1 and mismatch = -1 and gap costs were taken as 2 as gap existence penalty and 1 as gap extension penalty. As output, the input query matched with Homo sapiens chromosome 3 genomic contig, GRCh37.p5 Primary Assembly with 100% ident and 0% gap.

CONCLUSION

Sequence alignment is an important aspect of bioinformatics. There are various techniques that use scoring schemes and gap penalty score to find alignment score. Dynamic programming methods for global and local alignment always give an optimal score but require more computational time. Heuristic methods are faster than dynamic programming methods but may not always give optimal outputs.

FUTURE WORK

In bioinformatics, matching a subsequence is frequently more useful than matching the whole sequence, i.e. optimal local alignment. Also, multiple sequence alignment can be used to generate phylogenetic trees and discover homologous regions for protein and nucleotide sequences. But optimal techniques require more computational time and heuristic techniques do not give accuracy. To overcome these limitations, a proposed approach is to use a parallel environment to compute multiple sequence alignment based on dynamic programming techniques. Such an implementation will speed-up the computation and will always give an optimal alignment score.

REFERENCES

[1] Application of Bioinformatics http://www.biotecharticles.com/Genetics-Article/DNA-Sequencing-Method-Benefits-and-Applications-248.html

[2] Gene Therapy http://www.ebi.ac.uk/2can/bioinformatics/bioinf_realworld_1.html

[3] Mount, David W. Bioinformatics: Sequence and Genome Analysis CSHL Press, 2004

[4] Claverie, J. M. and C. Notredame, Bioinformatics for Dummies, John Wiley & Sons, 2006
[5] Purpose of alignment. http://www.nd.edu/~gmadey/bio05/ClassNotes/sa.pdf
[6] Andreas D. Baxevanis, B. F. Francis Ouellette Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, John Wiley & Sons, 2004
[7] Types of Alignment. Kun-Mao Chao : Sequence Alignment (2005)
[8] Jonathan Pevsner, Bioinformatics and Functional Genomics, Wiley-Blackwell, 2009
[9] Kun-Mao Chao, Louxin Zhang Sequence Comparison: Theory and Methods, Springer, 2008
[10] Scoring Matrices http://www.cs.tau.ac.il/~rshamir/algmb/98/scribe/html/lec03/node9.html
[11] PAM. cs124.cs.ucdavis.edu/lectures/scoringmatrices.pdf
[12] BLOSUM. S. Henikoff and J. G. Henikoff Amino acid substitution matrices from protein blocks, Proceedings of the National Academy of Science USA, 89(22):10915–10919, November 1992.
[13] Lipman, D J; Pearson, W R Rapid and sensitive protein similarity searches, Science 227: 1435–41 (1985)
[14] Pearson, W R; Lipman, D J, Improved tools for biological sequence comparison, Proceedings of the National Academy of Sciences of the United States of America 85 : 2444–8 (1988)
[15] Altschul, S; Gish, W; Miller, W; Myers, E; Lipman, D Basic local alignment search tool, Journal of Molecular Biology 215 (3): 403–410 (1990)
[16] http://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml
[17] http://blast.ncbi.nlm.nih.gov/Blast.cgi