ISSN:0975-9646

Sequence Alignment in DNA Using Smith Waterman and Needleman Algorithms

M.P Sudha
School of Computing Sciences, Vel's University, Pallavaram, Chennai-600 117

P.Sripriya
Assistant Professer, School of Computing Sciences, Vel's University, Pallavaram, Chennai-600 117

Abstract-Algorithm and scoring parameters Eg "best" Two methods for searching protein and DNA Evolution of protein and DNA sequence is done using database. 1. Local comparison i) Ignoring difference-outside most similar region ii) Find similarity between two sequence 2. Gobal Comparison. More appropriate when homology has been established when Building evolutionary trees comparison methods are preferred for functionally Conserved non homologous domains. Avoiding high similarity scores with unrealed sequences is more important as calculating related sequences while searching protein sequences databases. Thus comparison algorithm scoring matrix And Gap penalty are not most effective.

INTRODUCTION

Cells are fundamental working units of every living system and All the instructions which direct contained in the DNA (deoxyribonucleic acid). DNA consists of chemical and physical components. It is a side-by side arrangement (e.g., ATTCCGGA).genome is organism's complete set of DNA. Which vary in size: smallest genome consists of 600,000 DNA base pairs, human and mouse genomes consists 3 billion .DNA in the human genome arranged into 24 distinct. Chromosomes—physically separate molecules range from about 50 million to 250 million. major chromosomal abnormalities, including missing or extra copies or gross breaks and rejoining (translocations), detected by microscopic examination. Each chromosome contains many genes, consists of 2% human genome rest of noncoding regions.Human genome contain 30,000 genes. Which perform major functions of cellular structures. Proteins are large, complex molecules of subunits called amino acids. Chemical properties that distinguish the 20 different amino acids cause the protein chains to fold into three-dimensional structures that define functions in the cell. constellation of proteins in a cell called proteome. The dynamic proteome changes from minute to minute . Protein's chemistry and behavior are specified by gene sequence ,number and identities. Studies to explore protein structure and activities, known as proteomics, This focus research on molecular basis of health and disease.

What is Bio-Informatics?

Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned.

- ✓ Study of biological information.
- ✓ Interface of biology and computers.
- ✓ Computational molecular biology.
- ✓ Includes genomic Sub fields: DNA informatics, protein informatics, proteomics.

Comparison of sequences

The most fundamental operation in protein informatics is finding the best alignment between a query sequence and one or more additional sequences

Once candidate homologs have been identified, they can be evaluated using statistical methods and structural and biological information.

The correspondence between two aligned sequences can be expressed in a similarity score and/or viewed graphically, e.g., dot plots, alignments, motifs or patterns.

SCORING SYSTEMS

PAM matrices

Using many sets of 2 aligned sequences, for each amino acid pair Ai, Aj, count the # of times Ai aligns with Aj and divide that number by the total # of amino

acid pairs in all of the alignments, resulting in the frequency, f(i,j)

• Let fi and fj, respectively, denote the frequencies at which Ai and Aj appear in the sets of sequences

• Then the (i,j) entry for the ideal PAM matrix is

$$\log \frac{f(i,j)}{f(i)\,f(j)}$$

BLOSUM (BLOcks SUbstitution Matrices)

- Many sequences from aligned families are used to generate the matrices
- Sequences identical at >X% are eliminated to avoid bias from proteins over-represented in the database
- Specific matrices refer to these clustering cut-offs, i.e., BLOSUM62 reflects observed substitutions between segments <62% identical
- In analogy to PAM matrices, a log-odds matrix is calculated from the frequencies A ij of observing residue i in one cluster aligned against residue j in another cluster

Properties of Sequence Alignment

DNAShould use evolution sensitive measure of similarity

Should allow for alignment on exons => searching for local alignment as opposed to global alignment

Proteins Should allow for mutations => evolution sensitive measure of similarity
Many proteins do not display global patterns of similarity, but instead appear to be built from functional modules => searching for local alignment as opposed to global alignment

The Smith – Waterman Algorithm

- Smith-Waterman searching method:
- Compare query to each sequence in database
- Do full Smith-Waterman pair wise comparisons
- Use search results to generate statistics
- A more sensitive approach to searching
- Much slower than BLAST or FASTA.
- Use dynamic programming.

Smith Waterman Nucleotide & Peptide Variants

SWN

Smith Waterman nucleotide (SWN) compares nucleic acid sequences. Paracel's implementation allows the user to specif arbitrary match/mismatch matrix so that SWN may be used for both contextual and evolutionary comparisons. The matrix need not be symmetric to permit modeling directional substitutions.

SWP

Smith Waterman peptide (SWP) compares peptide sequences. Generally SWP is used for homology analysis and one of the evolutionary matrices, e.g., BLOSUM, is used. Unlike BLASTP.SWP does not restrict the value of permitted gap penalties.

Smith Waterman Frame Variants

Paracel accelerates three per-character, frame shift-tolerant and Smith Waterman style algorithms. In each of these algorithms, at each character position the score is determined by evaluating whether to stay in the current reading frame and accepting a match/mismatch score or an amino acid insertion/deletion(indel) or to jump to another reading frame and incur a frame shift penalty along with a match/mismatch score. This contrasts to the equivalent BLAST search types in which six static protein translations corresponding to three forward frames and three reverse frames are used in the comparison. Paracel's frame search variants tolerate frame shifts that are most often the result of sequencing errors and produce longer meaningful alignments then can be produced by BLAST.

SWX

Paracel's frame search compares nucleic acid query sequences to protein data. This search is used to find putative homologous proteins for newly sequenced ESTs, RNAs, and cDNAs. An independently adjustable frame shift penalty may be set to reflect the overall quality of the nucleic acid sequences. Additionally, this algorithm uses protein scoring matrices that can be chosen to reflect the evolutionary distance between the nucleic acid sequences and the organisms represented in the protein database. An affined gap penalty is generally used to model evolutionary variations.

TSWN

Searching a peptide sequence against nucleic acid coding regions is performed with Paracel's reverse frame algorithm. This comparison allows a user to annotate unknown peptide sequences by comparing them to databases of nucleic acid coding regions or to locate putative genes with known proteins. An independently adjustable frame shift penalty is available to model the possibility that a sequencing error in the nucleic acid data has occurred. Protein scoring matrices are used along with affined gap penalties to model evolutionary variations. Double affined gap penalties may be used to evaluate gene structure.

TSWX

Lastly, Parcel others a double frame nucleic acid to nucleic acid comparison at the protein level. This search allows for frame shifts at each character of both nucleic acid sequences. This search is useful for comparing homologous coding regions that are sufficiently separated by evolution to have differing codon usuage.

The Smith-Waterman algorithm does not impose any additional restrictions on the model of sequence evolution used in database searching. The Smith Waterman algorithm places no restriction on the alignment it reports other than that it have a positive score in terms of the similarity table used to score the alignment (17). Biologically, this means that the weights or scores assigned to replacements that occur more frequently than expected at random must outweigh those assigned to the replacements that occur less often than expected at random. In other words the preponderance of the evidence is in favor of the two aligned sequence sections being homologous (although the preponderance may not be great enough to justify inferring that the sequences are homologous). Both BLAST (20) and FASTA (19) place additional restrictions on the alignments that they report in order to speed up their operation. **Because of this Smith Waterman is more sensitive than either BLAST or FASTA**

Smith-waterman is mathematically rigorous, it is guaranteed to find the best scoring alignment between the pair of sequences being compared.

It does this by constructing a two dimensional table of partial alignment scores. The tables, as show below has one dimension or axis for each sequence. Each cell in the table contains the score for the best partial alignment that terminates with the pair of sequence residues (one from each sequence) that correspond to that cell in the table. That best scoring partial alignment will be extended to subsequent cells in the table only when it is the prior cell that results in the best scoring partial alignment for the subsequent cell. **In this way all possible alignments are considered until they are proven inferior to a competing alignment that also involves aligning at least one of the same pairs of sequence residues.** The final alignment is thus the best the best scoring alignment possible.

Because of this mathematical rigor and lack of restrictions the Smith Waterman algorithm is more sensitive than either BLAST or FASTA. This additional sensitivity comes at the price of being a very much slower way to search a sequence database than are either BLAST or FASTA (4). Because of this Smith-Waterman is most often run on either a supercomputer or sometimes special purpose hardware is purchased. The examples shown here with a 470 amino acids query sequence searching 89,912 sequences with 28,507,787 amino acids took between 20 and 25 minutes on a single processor of the Cray C-90 at the Pittsburgh Supercomputing Center.

Similarly Scores: DNA PAM 47, Match = 5, Mismatch =-4;

DNA sequence comparison is the most powerful tool available today for inferring structure and function from sequence because of the constrains of protein evolution-a protein fold into a functional structure. DNA sequence similarity can routinely be used to infer relationships between DNA s that last shared a common ancestor 1-2:5 billion years ago. Our ability to identify distantly related proteins has improved over the past five years with the development of accurate statistical estimates, which have provided better normalization methods, and with the use of optimized scoring methods, and with the use of optimized scoring parameters. In using sequence similarity to infer homology, one should remember

Open Cap = 0, Extend Cap=-7

Needleman-Wunsch algorithm

The Needleman-Wunsch algorithm consists of three steps:

1. Initialisation of the score matrix

2. Calculation of scores and filling the traceback matrix

3. Deducing the alignment from the traceback matrix

The traceback

Traceback = the process of deduction of the best alignment from the traceback matrix.

There are three possible moves: diagonally (toward the top-left corner of the matrix), up, or left.

The traceback is completed when the first, top-left cell of the matrix is reached (”done” cell).

The Needleman-Wunsch algorithm published in 1970, provides a method of finding the optimal global alignment of two sequences by maximizing the number of amino acid matches and minimizing the number of gaps necessary to align the two sequences. Because the Needleman-Wunsch algorithm finds the optimal alignment of the entire sequence of both proteins, it is a global alignment technique, and cannot be used to find local regions of high similarity.

In pairwise sequence alignment algorithms, a scoring function, *F*, must exist such that different scores can be assigned to different alignments of two proteins relative to the number of gaps and number of matches in the alignment. Thus, the alignment with the largest score must be the optimal alignment. In this scoring function, let *m* be the score for two residues matching, *s* is the penalty for mismatches, and *g* is the penalty for inserting a gap. The Needleman-Wunsch algorithm realizes that the score of aligning the entire proteins is the same as the sum of the scores of two subsequences of the proteins,

$$F(x1:M, y1:N) = F(x1:i, y1:j) + F(xi+1:M, yj+1:N)$$

where M is the length of sequence x, N is the length of sequence y, and $1<i<M$ and $1<j<N$. From this, we can see that the optimal score of two partial sequences is the sum of score of residue *i* in sequence x and residue *j* in sequence y, and the maximum score aligning the rest of the sequences. There are three possibilities

- *xi* and *yj* are the same so $F(i, j) = s(i, j) + F(i-1, j-1)$

*$s(i, j)=m$ if $xi=yj$; $s(i, j)=-s$ otherwise

- *xi* aligns to a gap so $F(i, j) = -d + F(i-1\ j)$
- *yi* aligns to a gap so $F(i, j) = -d + F(i, j-1)$

the Needleman-Wunsch algorithm essentially creates a matrix in which the horizontal and vertical axes each correspond to one of the protein sequences. Each amino acid in the protein sequence is assigned to a row or column starting at the N-terminus. For every cell (i, j) where i is the row and j is the column, if the residue i is the same as residue j, the score m is entered into the matrix. In this case, let $m=1$, and $s=d=0$

These weighted scores can affect the final alignment of the two protein sequences and the biological relevance of the alignment, but will not affect the time or space complexity of the algorithm because the number of operations will not change. This alignment is limited, however, because it can only align entire proteins. A different algorithm was developed to create local alignments backtracking process to keep record for values to be calculated in each iteration on parallel machines. This algorithm has been implemented on Grid using Alchemi Framework [8]. All the matrices in parallel version of Needleman-Wunsch algorithm are places in global memory space so that all available processors can access them at the same time to perform initialization and other calculations. By developing Needleman’s parallel algorithm we have Tahir Naveed reduced the calculation time from O(NxM) to O(N+M) by using two CPU to make a parallel computation[2].

CONCLUSION

Always compare DNA sequences if the genes encode DNAs.While most sequences that share statistically significant are homologous, many distantly related homologous sequences do not share significant homology. (Low complexity regions can display significant similarity in the absence of homology). Homologous sequences are usually similar over an entire sequence or domainThe Needleman-Wunsch algorithm aligns a pair of sequences over their entire lengths while the Smith-waterman algorithm finds the best matching regions in the same pair of Similarity searching techniques can be improved either by increasing the ability of a method to recognize distantly related sequences - increased sensitivity-or by lowering scores for unrelated sequences-increased selectivity. Since there are generally 1000-times more unrelated than related

sequences in a sequence database, improvements that reduce the scores of unrelated sequences can have dramatic effects Sequences. Putting the zero in the recursion is saying that if the partial alignment score becomes negative during the calculations we want to ignore that as well as ignore the preceding calculations and start over from a neutral score. Thus the best scoring regions do not have to overcome the effects of surrounding regions of low similarity in order to achieve a high score. The best scoring alignment is the alignment that ends at the cell in the table with the highest score.

REFERENCES

1. Of URFs and ORFs : A Primer on How to Analyze Derived Amino Acid Sequences, R.F. Doolittle Pages 3-36
2. Computer Methods for Macromolccular Sequence Analysis Methods in Enzymology, Vol 266 (1996) "Effective Protein Sequence Comparison" chapter W.R. Pearson Pages 227-258
3. Computer Methods for Macromolccular Sequence Analysis Methods in Enzymology, Vol 266 (1996) "Local Alignment Statistcs" chapter S.F. Altschul and W.Gish pages 460-480
4. Evolution of protein molecules T.H. Jukes and C.R. Cantor ,In mammalian Protein Metabolism (1969) 21-132
5. Atlas of Protein Sequence and Structure, M.O. Dayhoff 1978, Vol 5, Supp. 3. Natl. Biomed. Res. Found, Washington DC
6. Of URFs and ORFs - A Primer on How to Analyze Derived Amino Acid Sequences Ressell F. Doolittle 1987
7. Molecular Evolution. Wen - Hsiung Li 1997