₹ 200 ISSN - 2249-555X
Volume : 1 Issue : 5 February 2012

Indian Journal Of Applied Research

Journal for All Subjects

www.ijar.in

Listed in International ISSN Directory, Paris.

ISSN - 2249-555X

Indian Journal of Applied Research

Journal for All Subjects

Editor,
Indian Journal Of Applied Research
8-A, Banans, Opp. SLU Girls College, New Congres Bhavan, Paldi,
Ahmedabad-380006**,** Gujarat, INDIA
Contact.: +91-9824097643 E-mail : editor@ijar.in

Index

Volume : 1 | Issue : 5 | February 2012 ISSN - 2249-555X

Research Paper Engineering

Finding The Nearest Neighbors In Biological Databases

* Er. Pankaj Bhambri ** Dr. O.P. Gupta
*** Er. Franky Goyal

* Assistant Prof., Department of Info. Tech., Guru Nanak Dev Engg. College, Ldh.

** Dy. Director, School of Info. Tech., Punjab Agriculture University, Ludhiana

*** Research Scholar, Department of Comp. Sci. & Engg., Guru Nanak Dev Engg. College, Ldh.

Keywords : Biological, Finding

Introduction

Introduction to Bio-informatics

The term bioinformatics was given by Pauline Hogeweg and Ben Hesper in 1978 for the study of informatics processes in biotic systems. Bioinformatics is the application of computer science and information technology to the field of biology and medicine. Bioinformatics is the science of managing, mining, and interpreting information from biological sequences and structures. Bioinformatics deals with algorithms, databases and information systems, web technologies, artificial intelligence and soft computing, information and computation theory, software engineering, data mining, image processing, modeling and simulation, signal processing, discrete mathematics, for generating new knowledge of biology and medicine, and improving & discovering new models of computation. It is an interdisciplinary research area, which may be defined as the interface between biological and computational sciences. Java, XML, Perl, C, C++, Python, R, SQL and MATLAB are the programming languages popularly used in this field.

Scope

Bioinformatics consists of two subfields:

- The development of computational tools and databases
- The application of these tools and databases in generating biological knowledge to better understand living systems.

These two subfields are complementary to each other. The tool development includes writing software for sequence, structural, and functional analysis, as well as the construction and curating of biological databases. These tools are used in three areas of genomic and molecular biological research: molecular sequence analysis, molecular structural analysis, and molecular functional analysis. The analyses of biological data often generate new problems and challenges that use to development of new and better computational tools.

DNA

DNA (deoxyribonucleic acid) is the genetic material. This is profoundly powerful statement to molecular biologists. It is the information stored in DNA that allows the organization of inanimate molecules into functioning of living cells and organisms that are able to regulate their internal chemical composition, growth, and reproduction. All plants, animals, bacteria and some viruses contain genetic information in the form of DNA. A characteristic of living organisms is that DNA is reproduced and passed on to the next generation. DNA contains the instructions for making proteins. It is composed of linear array of nucleotides, each of which has a base plus a deoxyribose sugar and a phosphate. There were two main types of bases, purines (R) and pyrimidines (Y). Each main type of base came in two flavours. Two types of purines, adenine (A) and guanine (G), and two types of pyrimidines, cytosine (C) and thymine (T).DNA is in the form of double helix.

Figure 1.1: Structure of DNA

DNA is the king of molecules.

The molecule forms a double helix in which two strands of DNA spiral about one other. Strands refer to two chains.

Problem Formulation

The major problem being faced by biologists and researchers is huge amount of raw data but with a lack of means to effectively use this data.

Sequence Alignment

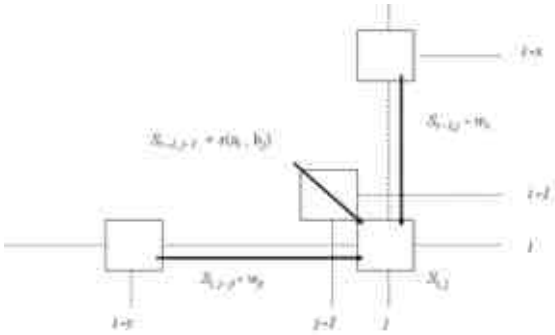
In bioinformatics, a sequence alignment is a way of arranging the primary sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationship between the sequences. Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix. Gaps are inserted between the residues so that residues with identical or similar characters are aligned in successive columns. In protein sequence alignment, the degree of similarity between amino acids occupying a particular position in the sequence can be interpreted as a rough measure. The absence of substitutions, or the presence of only very conservation substitutions in a particular region of the sequence, suggest that this region has structural or functional importance. Although DNA and RNA nucleotide bases are more similar to each other than amino acids, the conservation of base pairs can indicate a similar functional or structural role. Sequence alignment can be used for non- biological sequences, such as those present in natural language or in financial data.

Global and Local Alignment

Global alignments, which attempt to align every residue in every sequence, are most useful when the sequences in the query set and of roughly equal size. A general alignment technique is called the Needleman-Wunsch algorithm and is based on dynamic programming. Local alignments are more useful for dissimilar sequences that are suspected to contain regions of similarity or similar sequence motifs within their larger sequence context. The Smith-Waterman algorithm is a general local alignment method also based on dynamic programming. With sufficiently similar sequences, there is no difference between local and global alignments. The work is focused on protein sequence alignment. The optimal alignment between two sequences will be computed. The optimal alignment is the one which gives the maximum alignment score for the input sequences. It is also possible to obtain multiple optimal alignments for the two sequences.

Implementation

Figure 2: Formal Description of dynamic programming

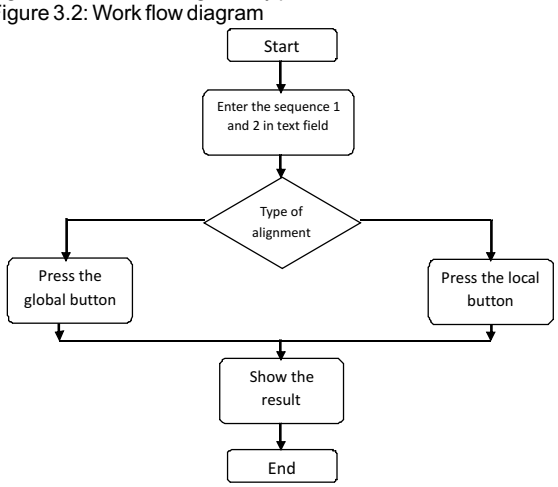


This diagram indicates the moves that are possible to reach a certain position (*i,j*) starting from the previous row and column at position (*i -1, j-1*) or from any position in the same row or column. Diagonal move with no gap penalties or move from any other position from column *j* or row *i*, with a gap penalty that depends on the size of the gap. For two sequences a = a_1, a_2, ..a_i and b = b_1, b_2, ..b_j, where S_{ij} = S (a_1,…a_i, b_1,…b_j) then

$$S_{ij} = \max \{S_{i-1,j-1} + s(a_i b_j),$$
$$\max (S_{i-x,j} - w_x),$$
$$x \geq 1$$
$$\max (S_{i,j-y} - w_x),$$
$$Y \geq 1$$
$$\}$$

where S_{ij} is the score at position at *i* in sequence a and *j* in sequence b, $s(a_i b_j)$ is score for aligning the character at positions *i* and *j*, w_x is the penalty for a gap of length *x* in sequence a, and w_x is the penalty for a gap of length *y* in sequence b. Note that S_{ij} is a type of running best score as the algorithm moves through every position in the matrix.

Figure 3.2: Work flow diagram



Results And Discussion

Pairwise Sequence Alignment

The purpose of sequence alignment algorithms is to detect evolutionary, and thus structural and functional, relations among sequences. The successful sequence comparison would allow us to infer the biological properties of new sequences from data accumulated on related genes. In trying to characterize a newly discovered sequence (DNA, protein) the first step is always to check whether there are similar sequences in known databases and their annotation.

Scoring Matrices

The different scoring matrices used for homology prediction are PAM and BLOSUM.

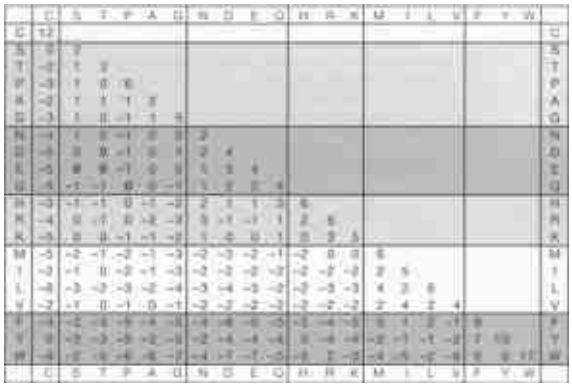
Point Accepted Mutations (PAM)

PAM series is the first major amino acid substitution series of scoring matrices. The PAM matrices (also called Dayhoff PAM matrices) were first constructed by Margaret Dayhoff, who compiled alignments of seventy-one groups of very closely related protein sequences. Because of the use of very closely related homologs, the observed mutations were not expected to significantly change the common function of the proteins. Thus, the observed amino acid mutations are considered to be accepted by natural selection.

An ideal PAM matrix can be obtained as follows:

- Collect and align many pairs of homologous sequences that are known to be n PAM units diverged
- Calculate the number of times amino acid a_i aligns opposite a_j, and divide by the total number of pairs in the aligned data, that gives us $p(a_i a_j)$
- Calculate the number of times a_i appears in both sequences and divide by their total length; this gives us $p(a_i)$ and then $p(a_j)$ for a_j
- There are many different PAMs, which are representing different evolutionary scenarios. PAM 250 represents a level of 250% of changes expected in 2500 MY
- PAM is more suitable for studying quite distant proteins, BLOSUM is for more conserved proteins of domains

Table 4 PAM250 Substitution Matrix



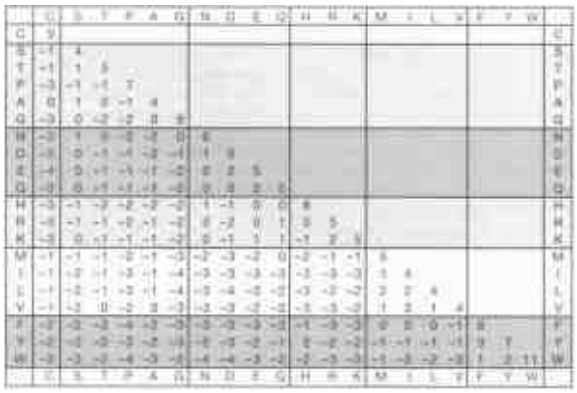
Amino acids are grouped according to to the chemistry of the side group: (C) sulfhydryl, (STPAG)-small hydrophilic, (NDEQ) acid, acid amide and hydrophilic, (HRK) basic, (MILV) small hydrophobic, and (FYW) aromatic. Log odds values: +10 means that ancestor probability is greater, 0 means that the probabilities are equal, -4 means that the change is random. Thus the probability of alignment YY/YY is 10+10=20, whereas YY/TP is 3-5=-8, a rare and unexpected between homologous sequences.

BLOSUM Series of matrices

BLOSUM matrices were an alternative to PAM matrices. Their construction was based on a larger collection of over 2000 ungapped, aligned amino acid sequence blocks. Blocks are a short contiguous interval in a multiple alignment of amino acid sequences.

These matrices are derived from BLOCKS. It de-emphasis of pairs of rows that are too similar; for example if two rows that appear in a given block are identical, then one is removed;

Table 5 BLOSUM62 Substitution Matrix



Ideology of BLOSUM is similar but it is calculated from a very different and much larger set of proteins, which are much more similar and create blocks of proteins with a similar pattern.

Conclusion And Future Scope

Conclusion

The model is constructed for predicting different proteins sequence alignments along with the alignment scores. The protein alignment is useful in predicting the homologous sequences and thus useful for assigning the class for unknown protein. There are different sequence formats available from which plain text format is utilized. Global and local alignments are predicted along with the alignment scores. Needleman Wunsch algorithm is used for the global alignment problem and Smith Waterman algorithm is used for the local alignment problem. There are different applications of global and local alignments. The global alignment accesses the complete amino acid sequences without skipping any gap whereas the local alignment accesses the partial alignment skipping the initial and ending gap penalties. Scoring matrix assists in computing the alignment score. It consists of mutation scores for different amino acid alignments. The model is user friendly which provides different options to the user. The overall computation is also shown in the matrix along with the final sequence alignment.

Future Scope

Following improvements regarding the developed model of bioinformatics can be made:

- The model can be extended to implementing Gene expression.
- Protein structure prediction can be incorporated to further enhance the model.
- Database can be maintained for the known protein sequences and the search option can be provided to the user.
- Different scoring matrices can be included.

Screen Shots

Figure 6 : Specifying the type of alignment that operate on the specified sequences

Figure 6.2: PAM Scores for the sequences

Figure 6.3: BLOSUM Scores for the sequences

REFERENCES

[1] Bergeron, B. (2003), "Bioinformatics Computing", Pearson Education, pp. 110-160. | [2] Xiong, J. (2006), "Essential Bioinformatics", Texas A&M University, pp. 31-47. | [3] Krane, D. and Raymer, M. (2006), "Fundamental Concepts of Bioinformatics", Pearson Education Publishers. | [4] Li, J.J and Huang, De-S. (2005), "Characterizing Human Gene Splice Sites Using Evolved Regular Expressions", Proceedings of International joint Conference on Neural Networks, Montreal, Canada, July 31 August 4, 2005, pp. 493-498. | [5] Luscombe, N.M., Greenbaum, D., Gerstein, M. (2001), "What is Bioinformatics? A proposed definition and Overview of the field", Luscombe group publications, pp 346-358. | [6] Mount D. W. (2004), "Bioinformatics: Sequence and Genome Analysis" 2nd ed. Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY. ISBN 0-87969-6087. | [7] Altschul, F., S., Madden, L., T., Schaffer, A., A., Zhang, J., Zhang, Z., Miller, W. and Lipman, J., D. (2003), "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs," Nucleic Acids Research, vol. 25, no. 17, pp. 3389-3402. | [8] Allcock, W., Bresnahan, J., Kettimuthu, R., Link, M., Dumitrescu, C., Raicu, I., Foster, I. (2005), "The globus striped gridFTP framework and server, in: Proceedings of Super Computing 2005", SC05. | [9] Arndt, P., Burge, C., Hwa, T. (2003), "DNA sequence evolution with neighbor-dependent mutation" Journal of Computational Biology 10: 313_322. | [10] Needleman, S.B. & Wunsch, C. D. (1970), "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins", J.Mol . Biol., vol. 48, pp. 443-453. | [11] Smith, T. F. & Waterman, M. S. (1981), "Identification of common molecular subsequences", J. Mol. Biol., vol. 147, pp 195-197. | [12] Pongor, S. and Landsman, D. (1999), "Bioinformatics and the developing world", Biotechnology-and-Development-Monitor, No. 40, pp. 10-13. | [13] Rastogi, S. C., Mendiratta, N. and Rastogi, P. (2005), "Bioinformatics Methods and Applications", Microsoft White pp. 10. | [14] Zhang Y, Skolnick J. (2005), "The protein structure prediction problem could be solved using the current PDB library", Proc. Natl Acad Sci USA 102: 1029-34. | [15] Mabrouk, M., S., Hamdy, M., Mamdouh, M., Aboelfotoh, M. and Kadah, Y., M., (2006), "BIOINFTool: Bioinformatics and sequence data analysis in molecular biology using Matlab", proc. cairo international biomedical engineering conference. | [16] Li, Z., Zhang, S., Wang, Y., Zhang, Sun X. and Chen, L. (2007), "Alignment of molecular networks by integer quadratic programming" Vol. 23 ,no. 13 2007, pp. 16311639. | [17] Joshi, Y. and Vadhiyar, S. (2009), "Analysis of DNA sequence transformations on grids" J. Parallel Distrib. Comput. 69, pp. 80-90.

Sara Publishing Academy
Indian Journal Of Applied Research
Journal for All Subjects

Editor,
Indian Journal Of Applied Research
8-A, Banans, Opp. SLU Girls College,
New Congres Bhavan, Paldi, Ahmedabad-380006.
Contact.: +91-9824097643 E-mail : editor@ijar.in

Printed at Unique Offset, Novatsing Rupam Estate, Opp. Abhay Estate, Tavdipura, Shahibaug, Ahmedabad