Journal of Computational Intelligence in Bioinformatics
ISSN: 0973-385X Volume 2 Number 3 (2009) pp. 159–165

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

Bioinformatics Tools in Genomic and Proteomic Research (Review)

***[1]J. Gaikwad Vishnu, [2]P. S. Deshmukh and [3]N. M. Patil**

1,3Department of Biotechnology, Tatyasaheb Kore Institute of Engineering and Technology, Warananagar, TAl: Panhala, Dist: Kolhapur.-4116113, Maharashtra, INDIA.
[2]Dept. of Biotechnology, MGM' college of computer science, Nanded.
**Corresponding Author E-mail: vishnu.gaikwad@gmail.com*

Abstract

Bioinformatics is the rapidly growing and developing field in computational science era. The major databases which are useful for life science research are NCBI, DDBJ, EMBL, TIGR, PDB, SWIEE-PROT and TrEMBL. These databases are public databases, conduct research in computational biology, and develop software tools for analyzing genome data.

The Basic Local Alignment Search Tool (BLAST) and (FASTA) Fast Approximation of Smith and Waterman Algorithm for comparing gene and protein sequences against others in public databases.

Sequence entries in the major genomic databases currently rise exponentially, because of that the gap between available, deposited sequence data and analysis by means of conventional molecular biology is rapidly increasing, it means that new approaches of genomic and proteomic analysis necessary.

In present review, we describe the basic bioinformatics tools and databases for genomic and proteomic research such as: Computational genomics, Genomic databases, Sequence alignment, Proteomic Databases, Gene prediction, Promoter prediction and for proteomic research such as: secondary structure prediction, molecular modeling. In such a way bioinformatics plays a wide role in gainful research in genomics and proteomics.

Key words: bioinformatics, databases, annotation, genomics, proteomics, sequence alignment.

Introduction

In 1972, Walter Fiers and his team at the Laboratory of Molecular Biology of the University of Ghent (Ghent, Belgium) were the first to determine the sequence of a gene: the gene for Bacteriophage MS2 coat protein [1]. In 1976, the team determined the complete nucleotide-sequence of bacteriophage MS2-RNA[2]. The first DNA-based genome to be sequenced in its entirety was that of bacteriophage Φ-X174; (5,368 bp), sequenced by Frederick Sanger in 1977 [3]. The first free-living organism to be sequenced was that of *Haemophilus influenzae* (1.8 Mb) in 1995, and since then genomes are being sequenced at a rapid pace. A rough draft of the human genome was completed by the Human Genome Project in early 2001, creating much fanfare.

In 2004, researchers from the International Human Genome Sequencing Consortium (IHGSC) of the HGP announced a new estimate of 20,000 to 25,000 genes in the human genome [4].The current release of GenBank contained 47 million sequences (There are approximately 85,759,586,764 bases in 82,853,685 sequence records in the traditional GenBank divisions and 108,635,736,141 bases in 27,439,206 sequence records in the WGS division as of February 2008.) with a current exponential increase of novel submissions [5].

The bioinformatics field is currently developing tools and approaches are being established for genomic and proteomic applications. These approaches have become excellent aids to answer each and every genome-related question. In this review, we are focusing on globally used bioinformatics tools that are easily accessible over the World Wide Web.

Computational genomics

Computational genomics is the study of deciphering biology from genome sequences using computational analysis [6], including both DNA and RNA. Computational genomics focuses on understanding the human genome, and more generally the principles of how DNA controls the biology of any species at the molecular level. With the current abundance of massive biological datasets, computational studies have become one of the most important means to biological discovery [7].

Genomic databases

The generated sequence data are stored in large genomic repositories, such as of the European Molecular Biology Laboratory (EMBL)/European Bioinformatics Institute (EBI) [8], ENSEMBL (http://www.ensembl.org), The Ensembl database project provides a bioinformatics framework to organize biology around the sequence of large genomes [9].

Ensembl is a joint project between EMBL - EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes. Ensembl is primarily funded by the Wellcome Trust. This site provides free access to all the data and software from the Ensembl project.

GenBank is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences. A new release is made every two months.

GenBank is part of the International Nucleotide Sequence Database Collaboration, The National Center for Biotechnology Information (NCBI, GenBank) database [10] and the DNA Database of Japan (DDBJ) [11]. These three main repositories work in close collaboration and exchanging their sequence information daily with updating.

Entry in the GenBank includes description of the sequence, the scientific name and taxonomy of the source organism, and a table of features that identifies coding regions (CDs regions). Bibliographic references with a link to the Medline unique identifier for all published sequences are included along.

Approximately 31 million entries within GenBank are of human origin. Especially the biological model organisms of the mouse, *Escherichia coli,* rat, fish, fly, frog, *saccharomyces cerevisiae* and worm have been deposited in these genomic databases. **wFleaBase: the *Daphnia* genome database:** wFleaBase is a project of the *Daphnia* Genomics Consortium [12]and is designed to be a resource where users can search and retrieve sequence data for genes of ecological importance, or find putative genes modulating traits of interest based on their homologies to functionally characterized genes in other model organisms. Therefore, wFleaBase is an organized repository of *Daphnia* specific sequences with standard bioinformatic tools to facilitate gene discovery. This function includes BLAST analyses and links to gene reports for other eukaryotic genomic models via euGenes [13]. However, for most of these other model species, characterized genes are ineluctably biased toward those sets whose phenotypic effects are observed in the benign settings of a laboratory. With the benefit of these genomic databases, we have the huge data on our tip of finger.

Sequence alignment

Computational approaches to sequence alignment generally fall into two categories: global alignments and local alignments. Calculating a global alignment is a form of global optimization that "forces" the alignment to span the entire length of all query sequences. By contrast, local alignments identify regions of similarity within long sequences that are often widely divergent overall. Local alignments are often preferable, but can be more difficult to calculate because of the additional challenge of identifying the regions of similarity. A variety of computational algorithms have been applied to the sequence alignment problem, including slow but formally optimizing methods like dynamic programming, and efficient, but not as thorough heuristic algorithms or probabilistic methods designed for large-scale database search.

Pairwise sequence alignment methods are used to find the best-matching piecewise (local) or global alignments of two query sequences. Pairwise alignments can only be used between two sequences at a time, but they are efficient to calculate and are often used for methods that do not require extreme precision (such as searching a database for sequences with high homology to a query). The three primary methods of producing pairwise alignments are dot-matrix methods, dynamic programming, and word methods [14].

Many sequence visualization programs also use color to display information about the properties of the individual sequence elements; in DNA and RNA sequences, this equates to assigning each nucleotide its own color. In protein alignments, color is

often used to indicate amino acid properties to aid in judging the conservation of a given amino acid substitution. For multiple sequences the last row in each column is often the consensus sequence determined by the alignment; the consensus sequence is also often represented in graphical format with a sequence logo in which the size of each nucleotide or amino acid letter corresponds to its degree of conservation [15].

Sequence comparison and alignment programs are Basic Local Alignment Search Tool, or BLAST, is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of different proteins or the nucleotides of DNA sequences. A *BLAST search* enables a researcher to compare a query sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold. The BLAST program was designed by Eugene Myers, Stephen Altschul, Warren Gish, David J. Lipman and Webb Miller at the NIH and was published in J. Mol. Biol. in 1990 [16].

As well as the FASTA, the FASTA package is available from fasta.bioch.virginia.edu [17], [18] and ClustalW [19], [20] algorithms.

Gene prediction

Gene finding typically refers to the area of computational biology that is concerned with algorithmically identifying stretches of sequence, usually genomic DNA, that are biologically functional. This especially includes protein-coding genes, but may also include other functional elements such as RNA genes and regulatory regions. Gene finding is one of the first and most important steps in understanding the genome of a species once it has been sequenced.

Advanced gene finders for both prokaryotic and eukaryotic genomes typically use complex probabilistic models, such as Hidden Markov Models, in order to combine information from a variety of different signal and content measurements. The GLIMMER system is a widely used and highly accurate gene finder for prokaryotes. Gene Mark is another popular approach. Eukaryotic *ab initio* gene finders, by comparison, have achieved only limited success; notable examples are the GENSCAN and geneid programs. A few programs like CONTRAST also use machine learning approaches like support vector machines for successful gene prediction. The SNAP gene finder is HMM-based like Genscan and attempts be more adaptable to different organisms, addressing problems related to using a gene finder on a genome sequence that it was not trained against [21].

Promoter prediction

In biology, a promoter is a regulatory region of DNA generally located upstream (towards the 5' region of the sense strand) of a gene that allows transcription of the gene, essential to the regulation of a gene's expression. As the transcription initiation point, which represents the extreme 5' end of the mRNA, may not be known accurately.

Developments in promoter prediction like PromoterScan [22] has been viewed as one of the first promoter prediction algorithms with acceptably high specificity.

Recently, PromoterInspector[23]and Dragon Promoter Finder [24] made further progress in specificity and sensitivity of promoter prediction algorithms.

PromoterScan [22] Predicts Promoter regions based on scoring homologies with putative eukaryotic Pol II promoter sequences.

Secondary Structure Prediction and protein modelling:

Prediction of secondary structure is facilitated by using bioinformatics tools such as GORIV, to predict secondary structures of proteins with known structure (We may access them at Swissprot or directly at PDB) [25], also the PredictProtein server where the PHD-method is implemented.[26].

For protein modeling swiss-model server is very useful for the same. It is a fully automated protein structure homology-modeling server, accessible via the ExPASy web server, or from the program DeepView (Swiss Pdb-Viewer), Automated mode, Alignment Mode and Project Mode. The purpose of this server is to make Protein Modelling accessible to all biochemists and molecular biologists World Wide. There are several methods [27].

Functional discovery

Proteomics also seeks to understand the interactions between proteins based on structure information and how these interactions help to form metabolic networks.

Examples of the proteomic tools used to identify protein–protein interactions are yeast forward and reverse hybrid systems, developed almost ten years ago. It was shown that DNA binding and activating functions of yeast transcription factors were located on two different domains.

By fusing an unrelated protein (protein A) to the activation domain and a second protein (protein B) to the binding domain, one is able to examine whether or not these proteins interact to restore transcriptional activity in yeast cells. Transcription of a reporter gene is used to identify yeast cells with restored activity. This system is known as the yeast two-hybrid system[28]. The power of the two-hybrid system is clearly demonstrated by publication of a complete protein–protein interaction map of *S. cerevisiae* by CuraGen Inc. (New Haven, CT, USA)[29].

Future Scope

With the rapidly emergence and vast development of this field, it has the bright perspectives in upcoming decades.

References

[1] Min Jou W, Haegeman G, Ysebaert M, Fiers W., Ncleotide sequence of the gene coding for the bacuteriophage MS2 coat protein, Nature. 1972 May 12;237(5350):82-8

[2] Fiers W et al., Complete nucleotide-sequence of bacteriophage MS2-RNA - primary and secondary structure of replicase gene, Nature, 260, 500-507, 1976

[3] Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M., Nucleotide sequence of bacteriophage phi X174 DNA, Nature. 1977 Feb 24;265(5596):687-95

[4] IHGSC (2004). "Finishing the euchromatic sequence of the human genome." *Nature* 431: 931–945.doi:10.1038/nature03001.

[5] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J and Wheeler DL: GenBank: update. Nucleic Acids Res 32: D23-D26, 2004.

[6] Koonin EV (2001) Computational Genomics, National Center for Biotechnology Information, National Library of Medicine, NIH (PubMed ID: 11267880)

[7] Computational Genomics and Proteomics at MIT

[8] Hubbard T, And rews D, Caccamo M, Cameron G, Chen Y, Clamp M, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, Down T, Durbin R, Fernandez-Suarez XM, Gilbert J, Hammond M, Herrero J, Hotz H, Howe K, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Keenan S, Kokocinsci F, London D, Longden I, McVicker G, Melsopp C, Meidl P, Potter S, Proctor G, Rae M, Rios D, Schuster M, Searle S, Severin J, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Trevanion S, Ureta-Vidal A, Vogel J, White S, Woodwark C and Birney E: Ensembl 2005. Nucleic Acids Res 33: D447-D453, 2005.

[9] T. Hubbard, D. barker, E Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. doen, R. Durbim, E. Eyras, J. Gilbert, M. Hammond, L. Huminiecki, A. Kasprzyk, H. Lehvaslaiho, P. Lijnzaad, C. Melsopp, E. Mongin, R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E Stupka, A. Uretqa-Vidal, I. Vastrik and M. Clamp(2002), nucleic Acid Res., Vol 30, No.1

[10] Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Church DM, Di Cuccio M, Edgar R, Federhen S, Helmberg W, Kenton DL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, Ostell J, Pontius JU, Pruitt KD, Schuler GD, Schriml LM, Sequeira E, Sherry ST, Sirotkin K, Starchenko G, Suzek TO, Tatusov R, Tatusova TA, Wagner L and Yaschenko E: Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 33: D39-D45, 2005.

[11] Tateno Y, Imanishi T, Miyazaki S, Fukami-Kobayashi K, Saitou N, Sugawara H and Gojobori T: DNA Data Bank of Japan (DDBJ) for genome scale research in life science. Nucleic Acids Res 30: 27-30, 2002.

[12] DGC: The Daphnia Genomics Consortium[http://daphnia.cgb.indiana.edu/]

[13] euGenes: A eukaryote organism genome information service [http://eugenes.org/]

[14] Mount DM. (2004). *Bioinformatics: Sequence and Genome Analysis 2nd ed.*. Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY.. ISBN 0-87969-608-7.Schneider TD, Stephens RM (1990). "Sequence logos: a new way to display consensus sequences". *Nucleic Acids Res* **18**: 6097–6100. doi:10.1093/nar/18.20.6097. PMID 2172928.Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). "Basic local alignment search tool". *J Mol Biol* **215** (3): 403410.doi:10.1006/jmbi.1990.9999.PMID2231712.http://wwwmath.mit.edu/~lippert/18.417/papers/altschuletal1990.pdf.

[17] Pearson WR: Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol 132: 185-219, 2000.

[18] Pearson WR and Lipman DJ: Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85: 2444-2448, 1988.

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

[20] Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG and Thompson JD: Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 31: 3497-3500, 2003.

[21] Korf I. (2004-05-14). "Gene finding in novel genomes". *BMC Bioinformatics* 5: 59-67. doi:10.1186/1471-2105-5-59. PMID 15144565.

[22] Prestridge DS: Predicting Pol II promoter sequences using transcription factor binding sites. J Mol Biol 249: 923-932, 1995.

[23] Scherf M, Klingenhoff A and Werner T: Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. J Mol Biol 297: 599-606, 2000.

[24] Bajic VB, Seah SH, Chong A, Zhang G, Koh JL and Brusic V: Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoters. Bioinformatics 18: 198-199, 2002.

[25] R.F. Doolittle Ed., Garnier J, Gibrat J-F, Robson B. GOR IV GOR secondary structure prediction method version IV Methods in Enzymology., vol 266, 540-553 , 1996

[26] B Rost, G Yachdav and J Liu (2004) The PredictProtein Server. Nucleic Acids Research 32(Web Server issue):W321-W326.

[27] Arnold K., Bordoli L., Kopp J., and Schwede T. (2006). The SWISS-MODEL Workspace: A web-based environment for protein structure homology modelling. Bioinformatics, 22,195-201.

[28] Vidal, M. and Legrain, P., *Nucleic Acids Res.*, 1999, **27**, 919–929.

[29] Uetz, P. *et al*., *Nature*, 2000, **403**, 623–627.