Bioinformatics – A Comprehensive Review

Abstract

The world is flooded with data today and managing these data in a scientific manner is under the perview of information science. Bioinformatics is a branch of this information science which handles all those data that are generated in everyday research in molecular biology. Bioinformatics not only stores data, but with the myriad of computational tools, it also allows you to decipher new ideas for future research. Bioinformatics as a discipline encompasses wide range of subjects like structural biology, genomics, evolutionary biology and also medical science.

This review gives a basic idea about methodology and application of the subject along with how the subject evolved to this stage. It has been created to cater this knowledge to the undergraduate students.

Keywords: BLAST, DDBJ, EMBL, Data Retrieval, Phylogeny, Global and Local Alignment

Sanjeev Pandey
Assistant Professor,
Deptt. of Botany,
Banwarilal Bhalotia College,
Asansol, West Bengal

Introduction

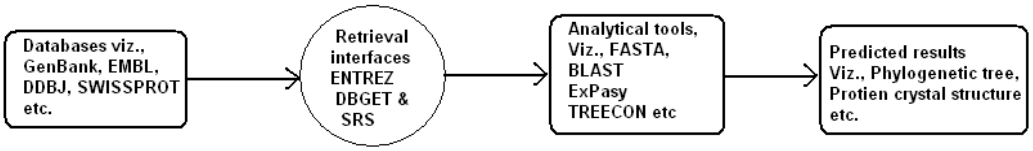
In the present era of genomics and proteomics, high throughput data are being generated at a phenomenal rate (Reichhardt, 1999). As of 15 August 2014, GenBank repository of nucleic acid sequences contained 174,108,750 reported entries of sequences (GenBank release note) and as of 19th March, 2014 the UniProtB/Swiss-Prot database (Apweiler *et al*. 2010) of protein sequences contained 542,782 reported entries of sequences. On an average, these databanks are doubling almost every year. Bioinformatics is that branch of information science which is capable of managing these data in a more productive way. It is a new branch of science which uses computational approach to answer biological questions on the basis of the available nucleotide and protein sequences that are generated in everyday research. It is a field of science in which biology, computer science and information technologies merge into a single discipline. According to National Center for Biotechnology Informations (NCBI), there are three important sub-disciplines within bioinformatics – the development of new algorithms and statistics which can be used to assess relationship among members of large data sets; the analysis and interpretation of various types of data including nucleotide and amino acids sequences, protein domains and protein structure and the development and implementation of tools that enable efficient access and management of different types of information.

History and Development

Bioinformatics as a cross-disciplinary field began its journey in 1960s with the effort of Margaret O. Dayhoff. Walter M. Fitch, Russell F. Doolittle and others and has matured as a fully developed discipline since then (Sabu M. Thampi). The first biological database in this regard was constructed a few years after the first protein sequence began to become available. The first protein sequence reported was bovine insulin in 1956. Nearly a decade later, the first nucleotide sequence was reported in terms of yeast alanine tRNA with its 77 bases. In 1965, it was Margaret Oakley Dayhoff who published the initial edition of *Atlas of Protein Sequence and Structure*, the first comprehensive, computerized and publicly available collection of protein sequences (Dayhoff, M. O. 1979). This became the first database and a model for many other molecular databases that evolved later on. After that it was PDB (Protein Data Bank) which appeared in 1972 with a collection of ten X-ray crystallographic protein structures. Later on a more advanced database for protein sequences, called the SWISSPROT, was created in the year 1987. After the formation of databases, tools became available to search sequences in the databases. It started with a dynamic algorithmic tool that matched two sequences at a very slow rate but with accuracy. This was soon replaced by FASTA in 1988 which had a relatively greater speed than the dynamic algorithm. This was followed by the BLAST algorithm in 1990 which was even faster than

FASTA. The later editions of databases and tools which developed time to time have been discussed separately in this comprehensive review.

In bioinformatics, the databases, the data retrieval systems and the analytical tools of different types operate in concert to produce a meaningful result as has been shown in the figure – 1.



Figure, 1 – A Work-Flow of Bioinformatics Study

The Databases

As mentioned earlier databases are repositories where data are submitted and stored in the form of nucleotide or amino acid sequences. There are three types of databases –

Primary Databases

A primary or archival dataset is one which represents experimental results (with some interpretation) but are never a curated review. These databases store data in their raw form that have been obtained and submitted directly by some researcher by sequencing some known/unknown gene or polypeptide. The primary databases may contain data which might have some error or redundancy. GenBank, EMBL, DDBJ and SWISSPROT are primary databases. ***GenBank*** (Benson *et al.* 1997) is a nucleotide sequence database. It is maintained by NCBI (National Center for Biotechnology Information) which is a part of National Institute of Health (NIH), a federal agency of the US Government. The ***EMBL*** (Baker *et al.* 2000) (European Molecular Biology Laboratory) nucleotide sequence database is maintained by the European Bioinformatics Institute (EBI) in Hinxton, Cambridge, UK. DNA Databank of Japan (***DDBJ***) is also a nucleotide sequence database which began in a collaboration with GenBank and EMBL (Tateno *et al*. 1997). It is run by National Institute of Genetics. All these nucleic acid based databases share data among themselves and also with the user.

SWISS-PROT on the otherhand is a protein sequence database. It was created in 1986 by Amos Bairoch at Swiss Institute of Bioinformatics (SIB) and subsequently developed by Rolf Apweiler at European Bioinformatics Institute (EBI) (Bairoch and Apweiler 1996; Séverine Altairac, 2006). TrEMBL (translated EMBL) is another database in same format. TrEMBL was created to provide automated annotations for those proteins which are not in SWISS-PROT. It was also created by SIB and EBI themselves. Another similar database created by PIR (Protein Information Resource) is PSD (Protein sequence databse, designated as PIR-PSD) (Wu *et al*. 2003). To pool overlapping resource and expertise of these three databases (Swiss-prot, TrEMBL and PIR-PSD) another database known as **UniProt** was created in 2003 (Apweiler *et al*. 2004).

Secondary Databases

The secondary databases are curated databases that add value to what is already present in primary database. Examples of secondary databases include, PDB (Protein Data Bank) and PROSITE etc. The protein data bank (***PDB***) is a repository of three dimensional structure of proteins and nucleic acids (Bernstein *et al*. 1977). The PDB was established in 1971 by Dr. Walker Hamilton, at the suggestion of American Crystallography Association (ACA) (Phillips, D. C. 1971). The data which are obtained through X-ray diffraction or NMR spectroscopy are submitted to this data bank. The ***PROSITE*** is another protein database that consists of entries describing the protein families, domains and functional sites as well as amino acid patterns and profile in them (Hulo *et al*. 2006).

Composite Databases

When primary databases are combined with secondary databases and filtered they form the non-redundant composite databases. ***SCOP*** (Structural Classification of Protein) is an example of composite database. It is a curated repository which stores 3-D structures of proteins that have been classified on the basis of their similarities and evolutionary relationships (Murzin *et al*. 1995).

Data Retrieval

The biological data is widely distributed across the world wide web (www) and is available to any learned worker. There are some sequence retrieval programs that can extract data from the database and can feed into the analytical tools and serve as interface between the two. The most popularly used retrieval tool is '**Entrez'**. It is a www-based data retrieval tool developed by the NCBI, which can be used to search for informations in 11 integrated NCBI databases (http://www/ncbi.nlm.nih.gov/Entrez/). ***DBGET*** is another data retrieval tool maintained by Kyoto University and the University of Tokyo (http://www.genome.ad.jp/dbget/dbget2.html). It covers more than 20 databases and is closely associated with KEGG. ***SRS*** (Sequence Retrieval System) is yet another retrieval tool like Entrez and DBGET (http://srs/ebi/ac/uk/). It was developed by EBI (European Bioinformatics Institute) that integrates over 80 molecular biology databases.

Some Bioinformatics Tools

Fasta

Pronounced "fast A" and stands for "FAST-ALL' is a sequence alignment software which is used for aligning sequences (proteins or nucleic acids) globally or locally. It was developed for the first time by Lipman and Pearson in 1988 to align protein sequences, but now it is also used to align nucleotide and translated DNA to protein sequences as well (Pearson and Lipman 1988). The FASTA format could be availed from fasta.bioch.virginia.edu. FASTA program follows a largely heuristic approach to speed up sequence execution. It is at least five times faster than the earlier existing algorithm. The basic idea of FASTA is to add a fast prescreen step to locate the highly matching segments between two sequences,

and then extend these matching segments to local alignments using more rigorous algorithms such as Smith-Waterman. The search speed and selectivity are controlled by a word size parameter called "*ktup*". By default, for protein sequence, *ktup* value is 2 while it is 6 for DNA comparison. Lesser the *ktup* value more sensitive will be the alignment but at the same time it will be slower.

Blast

BLAST (**B**asic **L**ocal **A**lignment **S**earch **T**ool) is more advanced algorithmic software for comparing nucleotide or amino acid sequences. It was developed by Altschul *et al*. in 1990. BLAST is ten times faster than FASTA. It is a set of search programs designed for the Windows platform and is used to perform fast similarity searches regardless of whether the query is for protein or DNA. Depending on the type of sequences to compare, there are different types of blast programs:

1. blastp compares an amino acid query sequence against a protein sequence database
2. blastn compares a nucleotide query sequence against a nucleotide sequence database
3. blastx compares a nucleotide query sequence translated in all reading frames against a protein sequence database
4. tblastn compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames
5. tblastx compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

Clustalw

It is a fully automated sequence alignment tool for DNA and protein sequences (http://www.clustal.org/clustal2/). It returns the best match over a total length of input sequences, be it a protein or a nucleic acid. This software is very helpful in multiple alignment for generating phylogenetic tree.

RasMol

It is a powerful research tool to display the structure of DNA, proteins, and smaller molecules (www.RasMol.org and www.OpenRasMol. org). Protein Explorer, a derivative of RasMol, is an easier to use program.

Applications of Bioinformatics

Today bioinformatics has become inevitable in molecular biological research. In addition to identifying an unknown sequence and its function, bioinformatics helps us to understand evolutionary relationship, protein structure and function, reading ORFs and their annotations, designing new drug against a functional domain of protein and many more things. Some of these aspects can be discussed here as follows.

Finding Structure and Function of Genes and Proteins

Bioinformatics facilitates us with huge amount of data which can be compared with one another to draw some inferences. One of such objective is to compare the nucleotide and amino acid sequence among various groups of organisms and draw a line of evolutionary relationship among them. The most common method in this regard is *sequence alignment,* which provides an explicit mapping between residues of two or more sequences (Mount, 2004). There are two types of sequence alignments –

1. Pair wise sequence alignments and
2. Multiple sequence alignments

Pair Wise Alignment

In pairwise sequence alignment, two sequences are compared at a time in pairs mainly to find out homology between them. Homology is quantitated by quantifying the level of matches. A mismatch represents a mutational event that might have occurred at some point of time in course of evolution. Similarly a gap in the alignment indicates either insertion or deletion in either of the sequences. The pair-wise sequence similarity analyses commonly uses two types of *dynamic programming algoriths* – *Needelman-Wunsch algorithm* (Needleman et al 1970) and *Smith-Waterman algorithm* (Smith et al 1981). Both are almost identical but, the main difference is that, Needelman-Wunsch algorithm finds *global similarity* between sequences while Smith-Waterman algorithm finds *local similarity*. A global similarity is that which starts from the left end of a sequence and covers the entire length of the sequence. Local similarity covers only a small part of the sequence. A local similarity is more acceptable because most of the biological sequences are often not similar over their entire length. For example, a gene which consists of exons and introns will show homology in the exon regions i.e., in local regions only while the introns will differ markedly. Similarly a protein sequence shows homology in certain domains and not in the entire structure.

Introduction of gaps in an alignment is a common practice. To achieve maximum alignment, gaps have to be introduced. Presence of many gaps means as many insertions or deletion. But insertion and deletion are relatively rare types of mutations and hence too many gaps in an alignment do not make biological sense.

Dynamic programming algorithms get around this problem by using gap penalties. A simple scoring system contains a positive additive contribution of 1 for each matching pair and a gap penalty of 1 is subtracted for each gap. For example in the following pair of sequences, there are 16 matches and one gap, thus the total alignment score is 15;

Seq1: AATTGATTGCGCATTTAAAGGG
Seq2: AACTGA- - - CGATTCTTAAGGG

A most complex form of gap penalties is known as ***affine***. It has both constant and proportional contributors. It is represented by the formula, 'A+Bl', where A is the constant penalty, is called the **gap opening penalty** and is applied to gap of any length. The constant B is called the **gap-extension penalty** and *l* is the length of the gap. In this affine penalty system, opening a gap is more strongly penalized, but once a gap is opened it should cost less to extend it.

The pair-wise sequence similarity searches are used very commonly to predict gene or protein functions. The underlying theory is that similar sequences are likely to be homologous and therefore would have similar functions. Thus, it helps to determine the orthologs and paralogs. When two homolgous genes in different species have the same function, they are known as **orthologs**; whereas when two genes in the same or different species have

different functions they are known as **paralogs (**Vallender, 2009).

Multiple Alignment

Unlike pair-wise alignments, they involve more than two sequences to be aligned and it carries more information than the pair-wise alignments do. Multiple alignments are a key to the prediction of protein secondary structure, residue accessibility, function and identification of residues important for specificity and to draw evolutionary relationship. Multiple alignments also provide the basis for the most sensitive sequence searching algorithm (Chenna *et al*. 2003).

Automatic alignment program, such as CLUSTAL W (Thompson *et al*. 1994) gives good quality multiple alignments for sequences that have greater similarities. This, along with most current programs, uses the method of progressive alignment (Feng and Doolittle. 1987). The first step in multiple alignment is to retrieve data for comparison. It is followed by editing these sequences with respect to their length and similarity. This is followed by similarity assessment among the set of sequences by comparing them pair-wise with randomization. The first pair of sequences has maximum similarity and serves as guide tree. Using this guide tree other sequences are progressively aligned according to their percent of similarities. This system of alignment is very effective but it suffers from the problem that the alignment errors made early in the process can never be rectified. However, with the help of an alignment visualization tools, such as, ALSCRIPT/JalView etc., it is possible to remove sequences manually that are disrupting the alignment. The problem of progressive alignment has been removed in programs like T-Coffee (Notredame *et al*. 2000), MultAlin (Corpet, 1988), PRRP (Gotoh, 1996) and DIALIGN (Morgenstern *et al*. 1998) algorithms. The significance of the alignment is finally assessed using “Monte Carlo” test of significance.

Phylogenetic Studies

Phylogenetics, also known as phylogenetic systematics, studies evolutionary relatedness among various groups of organisms. Phylogentic relationship can be represented with the help of tree like structre called phylegenetic tree/ dendrogram/ cladogram. **Phylogenetic trees** are genealogical trees which are built up with information gained from the comparison of either nucleic acid sequences or amino acid sequences of a conserved gene or protein respectively. As the ribosomal RNA has a common function in all domains of life, they have been conserved in the course of evolution and serve as important molecule for such comparison. Among them, the 16S rDNA in case of prokaryotes and 18S rDNA in case of eukaryotes have opitimum length and hence have been mostly exploited for constructing such phylogenetic trees. Amino acid sequences of proteins such as Cyt c are also conserved and can be used for phylogenetic analysis. On the contrary, proteins like β-amylase or haemoglobin etc are unsuitable for phylogenetic consideration, because they don't occur throughout the living matter.

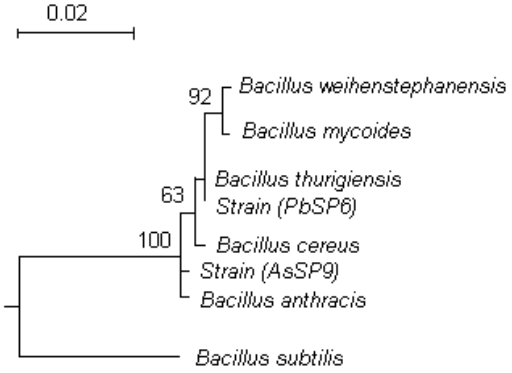
For making a phylogenetic tree, the multiple alignment file becomes the input for a phylogenetic analysis program. Based on the level of similarity or distance among organisms with respect to their sequences, an inference is drawn to construct a meaningful phylogram. There are three major methods employed in this regard:

1. distance matrix or Neighbour joining (Saitou and Nei, 1987),
2. maximum parsimony (Farris *et al*. 1970), and
3. maximum likelihood (Felsentein, 1981).

One of the commonly applied distance matrix methods is Neighbour-joining (Nj) method. It uses the number of nucleotide or amnio acid substitutions between sequences as a distance between a pair of sequence. Sequences with most substitutions are distanly related. Maximum parsimony method implies the idea that closely related sequences will have less chance that they bear substitutions. Thus it is opposite to distance matrix method. Maximum parsimony method is simple but statistically inconsistent. Maximum likelyhood method is most popular statistical method, however, being very slow, it has not been implemented on the internet as widely as other methods.

One of the most common phylogenetic program for building phylogenetic tree is PHYLIP (PHYLogeny Inference Package) (Fitch and Margoliash, 1967; Felsenstein, 1991). This program is freely available in the Internet. Another such phylogenetic program which uses protein sequence and structure is ExPaSy (Expert Protein Analysis System) (Gasteiger *et al*. 2003). In addition to these, there are PAUP and TREECON (Van de Peer and Wachter 1994) which are also very popular phylogenetic tree construction programs. PAUP stands for Phylogenetic analysis using parsimony (Swofford et al 1996) and is one of the most sophisticated parsimony programs available.

In a phylogenetic tree, the nodes are given with a bootstrap value (Felsenstein, 1985; Barzilay and Lee, 2002). Bootstrapping is basically a method of evaluating the reproducibility of the tree. It is the proportion of bootstrap replicates that support the monophyly of the clade. Bootstrap method was invented in 1979 by Efron (Efron, 1979) and was introduced as tree evaluation method by Felsentein (1985). A phylogenetic tree prepared using TREECON software has been shown in the figure – 2 for understanding (Pandey *et al*. 2011).



Figure, 2 – A Phylogenetic Tree Developed Through TREECON Software

Prediction of Structure of Proteins

Prediction of protein structure from sequence is one of the most challenging tasks in today's

computational biology. Although most information of 3-dimensional structure is encoded in the amino acid sequence it is still unknown which information controls the process of protein folding. Among millions of possible folding products, proteins take up one working, native structure. Since it is very difficult and expensive to evaluate structures by methods like X-ray diffraction or NMR spectroscopy, there is a big need for the unfailing prediction of 3-dimensional structures of proteins from sequence data (Figure – 3). Today there are methods which are able to give a quite reliable result from available sequence data (Sali and Blundell, 1993).

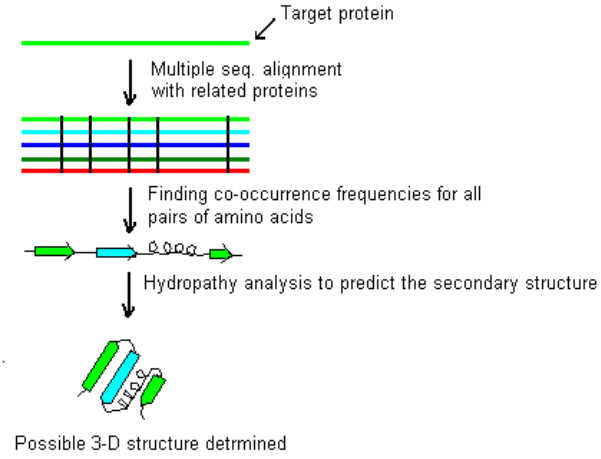


Figure 3 – Steps for Protein Structure Prediction

Rational Drug Designing

If the predicted crystalline structure of the protein is available, algorithms such Autodock, DOCK, GRAMM, FlexX etc (Brocklehurst et al 1999) can be used to identify potential interacting ligands to be used as drug, therefore allowing rational drug design.

Functional Genomics

The 'functional genomics' deal with high throughput functional annotations of the whole set of genes, present in a genome (Hieter and Boguski 1997). With the help of **ORF readers** and **ESTs (Expressed Sequence Tags),** number of possible genes in a genome are determined. Under the canopy of **pharmacogenomics' (**Jain, 1999) functional genomics has a massive impact on pharmaceutical industry as well, as it provides a fast track identification and validation of target protein. Functional genomics also helps in designing chips for **microarray (**Pease *et al.* 1994; Cho *et al.* 1998) which can later be implemented to read the expression profile of a particular cell, tissue or organism.

References

1. Altschul S. F., Gish W., Miller W, Myer E. W. and D. J. Lipman. 1990. Basic Local Alignment Search Tool, *Journal of Mol. Bio.*, 215 (3): 403-410.
2. Apweiler R., Martin M.J., O'Donovan C., Magrane M., Alam-Faruque Y., Antunes R., Barrell D., Bely B., Bingley M. and D. Binns. 2010. UniProt Consortium The Universal Protein Resource (UniProt). *Nucleic Acids Res*. 2010;38: D142–D148.
3. Apweiler, R., Bairoch, A.. and C. H. Wu. 2004. "Protein sequence databases". *Current Opinion in Chemical Biology* **8** (1): 76–80.
4. Bairoch, A. and R. Apweiler, 1996. "The SWISS-PROT protein sequence data bank and its new supplement TREMBL". *Nucleic Acids Research* 24 (1): 21–25.
5. Baker, W., van den Broek, A., Camon, E., Hingamp, P., Sterk, P., Stoesser, G., and M. A. Tuli. 2000. The EMBL nucleotide sequence database, *Nucleic Acids Research*. 28: 19-23.
6. Barzilay, R. and L. Lee. 2002. "Bootstrapping Lexical Choice via Multiple-Sequence Alignment" (PDF). *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* 10: 164–171.
7. Benson, D. A., Karsch-Mizrachi, I, Lipman, D. J., Ostell, J., Rapp, B. A. and D. L. Wheeler. 1997. GenBank. *Nucleic Acids Research*, 25: 1-6.
8. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. and M. Tasumi, 1977. *J. Mol. Biol.* 112: 535±542.
9. Brocklehurst,, S. M., Hardman, C. M. and S. J. T. Johnston. 1999. Creating integrated computer systems for target discovery and drug discovery. In: *Pharmainformatics: A Trends Guide*. M. Owen (Ed.). New York, Elsevier Science Ltd: 12-15.
10. Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G. and J.D. Thompson. 2003. "Multiple sequence alignment with the Clustal series of programs". *Nucleic Acids Res* 31 (13): 3497–500
11. Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J. and R. W. Davis. 1998. *Mol. Cell* 2: 65–73.
12. Corpet, F. 1988. Multiple sequence alignment with hierarchical clustering, *Nucleic. Acids Res.* 16(22): 10881-90.
13. Dayhoff, M. O. 1979. Atlas of Protein Sequence and Structure, Volume 5, Supplement 3, 1978. *National Biomedical Research Foundation*, Washington, D.C.
14. Efron, B. 1979. Bootstrap Methods: Another look at the Jackknife. *The Annals of Statistics,* Vol. 7, No. 1: 1-26.
15. Vallender Eric J. 2009. Bioinformatic approaches to identifying orthologs and assessing evolutionary relationships. *Methods*. 49(1): 50–55.
16. Lewitter, F. 1998. Text-based database searching. *Trends Guide to Bioinformatics*, : 3-5
17. Farris J. S., Kluge, A.G. and M. J. Eckhardt 1970. A numerical approach to phylogene tic systematics. *Syst. Zool*. 19: 172 - 191.
18. Felsentein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach, *Journal of Molecular Evolution.* 17(6): 368-376
19. Felsenstein, J. 1985. Con®dence limits on phylogenies: An approach using the bootstrap. *Evolution* 39: 783 - 791.

20. Felsenstein, J. 1991. PHYLIP (phy logeny inference package) 3.4 manual. University of Washington, Seattle.
21. Feng, D.F. and R. F. Doolittle. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees, *J. Mol. Evol.* 25(4): 351-60.
22. Fitch, W. M., and E. Margoliash. 1967. Construction of phylogenetic trees. *Science* 155: 279–284
23. Gaasterland, T. 1998. Structural genomics: Bioinformatics in the driver's seat. Nature *Biotechnology*, 16: 625-627.
24. Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D. and A. Bairoch. 2003. "ExPASy: The proteomics server for in-depth protein knowledge and analysis". *Nucleic Acids Research.* 31 (13): 3784–8.
25. GenBank release notes. NCBI.
26. Gotoh, O. 1996. Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *J Mol Biol*. 264(4): 823-38.
27. Hieter, P. and M. Boguski. 1997. Functional Genomics: It's all how you read it. *Science*. 278: 601-602.
28. Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., De Castro, E., Langendijk-Genevaux, P. S., Pagni, M. & Sigrist, C. J. 2006. The PROSITE database. *Nucleic Acids Res*, 34: D227-230.
29. Jain, K. K. 1999. Strategies and technologies in functional genomics. *Drug Discovery Today*, 4: 50-53.
30. Morgenstern, B., Frech K., Dress, A. and T. Werner. 1998. DIALIGN: finding local similarities by multiple sequence alignment. *Bioinformatics*. 14(3):290-4.
31. Mount, D. M. 2004. *Bioinformatics: Sequence and Genome Analysis* (2nd ed.). Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY. ISBN 0-87969-608-7.
32. Murzin A. G., Brenner S. E., Hubbard T. and C. Chothia. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540
33. Needleman, Saul B. and Wunsch, Christian D. 1970. "A general method applicable to the search for similarities in the amino acid sequence of two proteins". *Journal of Molecular Biology* 48 (3): 443–53.
34. Notredame, C., Higgins, D.G., Heringa, J. 2000 T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol.* 8;302(1): 205-17.
35. Pandey, S., Saha, P., Biswas, S. and T. Maiti. 2011. Characterization of two metal resistant Bacilus strains isolated from slag disposal site at Burnpur, India. *J. Environ. Biol.* 32, 773-779.
36. Pease, A. C., Solas, D., Sullivan, E. J., Cronin, M. T., Holmes, C. P. & Fodor, S. P. A. 1994. *Proc. Natl. Acad. Sci.* USA 91: 5022–5026
37. Phillips, D. C. (1971). Cold Spring Harbor Symp. *Quant. Biol*: 589±592
38. Reichhardt, T. 1999. It's sink or swim as a tidal wave of data approaches. Nature, 399 (6736): 517-20.
39. Sabu M. Thampi, Dept of CSE, LBS College of Engineering, Kasaragod, Kerala-671542
40. Saitou, N. and M. Nei 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* 4(4): 406-25.
41. Sali, A. and T. L. Blundell.1993. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234, 779–815 .
42. Séverine Altairac, 2006. "Naissance d'une banque de données: Interview du prof. Amos Bairoch".*Protéines à la Une*, ISSN 1660-9824.
43. Smith, T. F. and M. S. Waterman. 1981."Identification of Common Molecular Subsequences" (PDF). *Journal of Molecular Biology*147: 195–197.
44. Swofford, D.L., Olsen, G.J., Waddell, P.J., and D. M. Hillis. 1996. Phylogenetic Inference. In *Molecular systematics*, 2nd edition, chap. 5; 407-514. Sinauer and Associates, Sunderland, Massachusetts
45. Tateno, Y., Miyazaki, S., Ota, M., Sugawara, H. and T. Gojobori. 1997. DNA Databank of Japan (DDBJ) in collaboration with mass sequencing teams, *Nucleic Acids Research*, 28: P24-26.
46. Thompson, J.D., Higgins, D.G. and T.J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* Nov 11;22(22): 4673-80.
47. Van de Peer, Y. and Y. De Wachter. 1994. TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. Comput. Applic. Biosci. 10, 569-70.
48. W. R Pearson and D. J. Lipman. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* USA Vol. 85: 2444-2448.
49. Wu, C. H., Yeh, L. S., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., Hu, Z., Kourtesis, P., Ledley, R. S., Suzek, B. E., Vinayaka, C. R., Zhang, J. and W. C. Barker. 2003. "The Protein Information Resource". *Nucleic Acids Research.* 31 (1): 345–347