ISSN (Online): 2091-2609
DOI Prefix: 10.3126/ijasbt
International Journal of Applied Sciences and Biotechnology
(IJASBT)

Mini Review

BASICS OF BIOINFORMATICS IN BIOLOGICAL RESEARCH

Ashwini Kamble[1*] and Rajesh Khairkar[2]

[1]Department of Biochemistry, Mahatma Gandhi Institute of medical sciences, Sevagram, wardha. Maharashtra, India.
[2]Department of CE/IT, NUVA College of Engineering, Kalmeshwar Nagpur, India

Corresponding author's email: dr.ashwinipravin@gmail.com

Abstract

The concept of laboratory rat is giving way to the computer mouse arose after the famous handshake between Clinton-Blair for the completion of the human genome in April 2003.Bioinformatics is defined as the application of computational techniques to understand and organize the information associated with biological macromolecules.

There is availability of large databases of genomic information which has enabled research efforts for discovering methods for diagnosis and treatment of human diseases using DNA microarrays and proteomics experiments. But there are various problems while doing this like it's always challenging to develop proper and sophisticated analysis method which can properly use genomic data bases considering its and heterogeneity of the data.

The main purpose of this first paper is to explore and explain Bioinformatics in a more scientific way, and try highlighting applications of bioinformatics in the medical sector.

Keywords: Bioinformatics; Microarrays; Proteomics.

Introduction

Since the birth of the Bioinformatics in the 1980s the field has been rapidly expanding, keeping pace with the expansion of genome sequence data. Bioinformatics is a field of conceptualizing biology in terms of molecules and applying informatics techniques for understanding as well as organizing the information of these molecules. In other words it is defined as the application of computational techniques to understand and organize the information associated with biological macromolecules (Pandey and Divyasheesh, 2016).

Bioinformatics intends to use information technology for biological purpose. This can be explained in simple terms as life can be an information technology and the genes which determine organism's physiology is the most basic as digital storehouse of information. Traditionally, bioinformatics has had a structural orientation mainly because of its use in Rational Drug Design (RDD) and Structure-based drug design (SBDD). SBDD and RDD both use different computational methods for discovering new compounds with good selectivity, efficacy and safety. Take for example the concept of laboratory rat which is giving way to the computer mouse arose after the famous handshake between Clinton-Blair for the completion of the human genome in April 2003 (Ouzounis, 2012). We can say in many countries wet lab experiments and use of bioinformatics goes hand in hand in clinical and biological researches (Daisuke and Troy, 2006).

Literature is replete with large databases of genomic information which has enabled research efforts for discovering methods for diagnosis and treatment of human diseases using DNA microarrays and proteomics experiments. But there are various problems while doing this like it's always challenging to develop proper and sophisticated analysis method which can properly use genomic data bases considering its applications and heterogeneity of the data.

The main purpose of this first paper is to explore and explain Bioinformatics in a more scientific way, and try highlighting applications of bioinformatics in the medical sector like oncological research.

Basics of Bioinformatics Tools

The bioinformatics have numerous applications broadly defined as,

1. organization of data in such a way that it allows researchers to access existing information and to submit new information they have produced, eg DNA Data Bank of Japan (National Institute of Genetics),
2. developing most appropriate tools and resources for analysis of data (such as FASTA (Pearson and Lipman 1988) and PSI- BLAST (Altschul *et al*., 1997) and
3. Using these tools to interpret the result in biological manner.
4. So with the help of bioinformatics, anyone can now perform global analyses of all the available data for uncovering the common principles that apply across many systems and also can highlight novel features.

Tools for Systemic Collection and Organisation of Biological Data

Biological databases are meant for this purpose**.** Biological databases are libraries of life sciences information, collected from scientific experiments, published literature, high-throughput experiment technology, and computational analysis (Attwood *et al*., 2011).

But for creating these database(s) it requires some raw materials.

a) Source of information for databases

Raw DNA sequences, protein sequences, macromolecular structures, genome sequences, and other whole genome data forms sources.

GenBank (R) is a place where comprehensive database that contains publicly available nucleotide sequences, nearly for more than 240 000 named organisms, obtained primarily through submissions from individual laboratories and batch submissions from large-scale sequencing projects (Benson *et al*., 2013).

This database is produced and maintained by the National Center for Biotechnology Information (NCBI) as part of the International Nucleotide Sequence Database Collaboration (INSDC).GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis (Benson *et al*., 2013).

Scientific researchers have stressed on genome sequencing and revealed that genomes consist of baseletters, ranging from 1.6 million bases in Haemophilus influenza to 3 billion in humans (Luscombe *et al*., 2001) scientist now have reached to the stage where measurement of expression levels of almost every gene in a given cell on a whole-genome level is possible although public availability of such data is still limited. Interestingly this measurement is carried out under different environmental conditions, different stages of the cell cycle and different cell types in multi-cellular organisms (Luscombe, 2001).

Apart from this primary nucleotide database there are variety of other area in which databases have prepared like protein sequence databases, proteomic databases, Protein structure, Protein model ,RNA databases, Carbohydrate structure ,Protein-protein and other molecular interactions, Signal transduction pathway databases, Metabolic pathway and Protein Function, Gene expression databases (mostly Microarray data) etc.

b) Classification of databases

Biological databases can be classified into primary, secondary and composite databases.

Primary Database(s) are those which contain information of the sequence or structure alone. For e.g. Gen bank and DDBJ for genome sequence.

Secondary Database(s) are those which contain derived information from primary databases. They contains information like the conserved sequence, signature sequence and active site residues of the protein families etc. some of the databases like SCOP developed at Cambridge university, CATH developed at university college of London, eMOTIF at standford etc. are created and hosted by individual researchers at their individual laboratories.

Composite Database(s) includes variety of primary database sources which obviates the need to search multiple resources. The NCBI i.e. national centre for Biotechnology information, hosts nucleotides and protein databases in their large high available arrays of computer servers, provide free access to the persons involved in research.

Now we can discuss about some of the well-known databases.

1. Nucleotide and Genome sequences

The GenBank (Benson, 2000) EMBL (Baker, 2000)and DDBJ (Okayama, 1998) databases contain DNA sequences for individual genes that encode protein and RNA products the biggest excitement currently lies with the availability of complete genome sequences for different organisms. They have uniform data format (but not identical) and exchange on daily basis.

The composite protein sequence database, the Entrez nucleotide database (Schuler, 1996)compiles sequence data from these primary databases. they not only provide the raw nucleotide sequence, but they store information in detail regarding all chromosomes in an organism, detailed views of single chromosomes marking coding and non-coding regions, list of completed genomes, and single genes. Adding to this at each level there are graphical presentations, precomputed analyses and links to other sections of Entrez (Luscombe *et al*., 2001)

Another database, COGs database classifies proteins encoded in 21 completed genomes on the basis of sequence similarity.^{21}

The essential function of these databases is to predict the function of proteins which are uncharacterized by their homology to characterized proteins, in addition to identify phylogenetic patterns of protein occurrence (Tatusov, 1997).

2. *Protein sequence databases*

Protein sequence databases are even categorized into primary, secondary and composite databases. Primaryprotein sequence databases contain more than 300,000 protein sequences. SWISS-PROT (Bairoch and Apweiler 2000) and PIR International (Mc Garvey, 2000) acts as primary as well as secondary databases they acts as repositories as well as describe the proteins' functions, its domain structure and post-translational modifications. Composite databases like OWL (Bleasby 1994) and the NRDB (Bleasby and Wootton 1990). Compile and filter sequence data from different primary databases to produce more complete databases than the individual databases.

The secondary databases help the user determine whether a new sequence belongs to a known protein family. The most popular databases in this are PROSITE. It is one of the most popular database of short sequence patterns and profiles that characterize biologically significant sites in proteins. PRINTS on the other hand, expand on this concept and provide an essence of protein fingerprints – groups of conserved motifs that characterize a protein family. Finally, Pfam-A is another database which comprises accurate manually compiled alignments on the other hand Pfam-B is an automated clustering of the whole SWISS-PROT database (Bateman *et al.*, 2000). These different secondary databases have recently been incorporated into a single resource named InterPro (Attwood *et al.*, 1999).

3. *Structural databases*

These are databases of the macromolecular structures. The Protein Data Bank, PDB provides a primary archive of all 3D structures for macromolecules such as proteins, RNA, DNA and various complexes (Bernstein, F.C.1977; Berman, H.M., 2000).The problem with individual Protein Data Bank is that information regarding entries can be difficult to extract. This problem has overcome by PDBsum (Laskowski, 1997). PDBsum has capability of providing a separate Web page for every structure in the protein databases and helps in detailed structural analyses, schematic diagrams and data on interactions between different molecules in a given entry (Luscombe, 2001) CATH (Pearl, 2000), SCOP (Lo Conte, 2000) and FSSPb (Holm and Sander. 1998) databases are the three major databases which classify proteins by structure to identify structural and evolutionary relationships. Similarly, there are various other databases which focus on particular types of macromolecules for ex. Nucleic Acids Database, NDB(Berman,1992) for structures related to nucleic acids, the HIV protease database (Vondrasek,1997) for HIV-1, HIV-2 and SIV protease structures and their complexes, and ReLiBase for receptor-ligand complexes (Hendlich, 1998).

Table 1: Databases and their bioinformatics' sources

Database	**Bioinformatics sources**
Protein sequence (primary)	SWISS-PROT PIR-International
Protein sequence (composite)	OWL NRDB
Protein sequence (secondary)	PROSITE PRINTS Pfam
Macromolecular structures	Protein Data Bank (PDB) Nucleic Acids Database (NDB) HIV Protease Database ReLiBase PDBsum CATH SCOP FSSP
Nucleotide sequences	GenBank EMBL DDBJ
Genome sequences Entrez genomes GeneCensus COGs	Entrez genomes GeneCensus COGs
Integrated databases	InterPro Sequence retrieval system (SRS) Entrez

Data Integration

Data integration is most important step in the field of bioinformatics. Because individual data does not carry much significance until it combines with the other information available regarding that structure. In other words it is the way of putting individual pieces of information in context with respect to other data. Data integration, as it looks like, however is not always straightforward to access as there are differences in nomenclature and file formats.

There are several methods to overcome this problem

1) Can be solved to some extent by providing cross-references
2) At a more advanced level, there have been efforts to integrate access across several data sources.
3) SRS is the Sequence Retrieval System (Etzold *et al*., 1996) which allows databases to be indexed to each other.
4) Entrez facility (Schuler *et al*., 1996) which provides similar gateways to DNA and protein sequences, genome mapping data, 3D macromolecular structures and the PubMed bibliographic database.

So in this way a search for any specific gene in either database will allow smooth transitions to the genome it comes from, the protein sequence it encodes, its structure, bibliographic reference and equivalent entries for all related genes.

Use(S) Of Integrated Data

So after integrating available information, integrated data is to be utilized in different areas.

As depicted in the Table 2, data source formed can be utilized for different purpose(s) using bioinformatics' techniques. For example genomics data can be used relating specific genes to diseases, in metabolic pathways for characterization of protein content etc. by using bioinformatics methods. Likewise protein sequence data can be utilized for multiple sequence alignments algorithms, sequence comparison algorithms, Identification of conserved sequence motifs etc.

Conclusions

Bioinformatics methods have become indispensable to biological investigations. In this review we have tried to provide baseline information regarding role of bioinformatics in the biomedical research. In our next article we will try to cover role of bioinformatics in specific medical conditions.

Bioinformatics covers a wide range of subject areas including structural biology, genomics and gene expression studies etc. as we have seen Bioinformatics principle approach is to compare and group the data according to biologically meaningful similarities and then, based on this, analysing one type of data to infer and understand the observations for another type of data. This helps us to understand the biological information in large scale dimensions both in depth and breadth.

So in total it enables us to examine individual systems in detail, to compare them with those that are related to find out similar principals in them and also distinguish some features which are unique to some systems.

Table 2: Data sources in bioinformatics and subject areas that utilize this data.

DATA SOURCE	RESEARCH AREAS
Genomes	1) Phylogenetic analysis 2) Linkage analysis relating specific genes to diseases 3) characterization of protein content metabolic pathways 4) Characterization of repeats 5) Structural assignments to genes
Raw DNA sequence	1) Identification of introns and exons 2) Separating coding and non-coding regions 3) Forensic analysis 4) Gene product prediction
Protein sequence	1) Multiple sequence alignments algorithms 2) Sequence comparison algorithms 3) Identification of conserved sequence motifs
Gene expression	1) Mapping expression data to sequence, structural and biochemical data 2) Correlating expression patterns
Macromolecular structure	1) Protein geometry measurements 2) Secondary, tertiary structure prediction 3) 3D structural alignment algorithms 4) Surface and volume shape calculations 5) Intermolecular interactions

References

Altschul SF, Madden TL, Schaffer AA, Zhan J, Zhang Z, Miller W and et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. **25**(17):3389-3402. DOI: 10.1093/nar/25.17.3389

Attwood TK, Flower DR, Lewis AP, Mabey JE, Morgan SR, Scordis P and et al. (1999). PRINTS prepares for the new millennium. *Nucleic Acids Res.* **27**(1): 220-225. DOI: 10.1093/nar/27.1.220

Attwood TK, Gisel A, Eriksson, NE, Bongcam-Rudloff E (2011) Concepts, Historical Milestones and the Central Place of Bioinformatics in Modern Biology: A European Perspective". *Bioinformatics - Trends and Methodologies*. InTech. Retrieved 8 Jan 2012. DOI: 10.5772/23535

Baker W, Van den Broek, A, Camon E, Hingamp P, Sterk P, Stoesser G and et al. (2000) The EMBL nucleotide sequence database. *Nucleic Acids Res* 28(1):19-23. DOI: 10.1093/nar/28.1.19

Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL (2000) The Pfam protein families database. *Nucleic Acids Res*. **28**(1): 263-266. DOI: 10.1093/nar/28.1.263

Benson DA, Karsch-Mizrachi I, Lipman, DJ, Ostell J, Rapp BA and Wheeler DL (2000) GenBank. *Nucleic Acids Res*. **28**(1): 15-18. DOI: 10.1093/nar/28.1.15

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2013) GenBank. *Nucleic Acids Res.* **41**(Database issue): D36-42. DOI: 10.1093/nar/gks1195. DOI: 10.1093/nar/gks1195

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H and et al. (2000) The Protein Data Bank. *Nucleic Acids Res*. **28**(1): 235-242. DOI: 10.1093/nar/28.1.235

Berman HM, Olson WK, Beveridge DL, Westbrook J, Gelbin A, Demeny T and et al.(1992).The Nucleic Acid Database. A comprehensive relational database of three dimensionalstructures of nucleic acids. *Biophysics Journal* **63**(3): 751-759. DOI: 10.1016/S0006-3495(92)81649-1

Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, Rodgers JR and et al. (1977) The Protein Data Bank. A computer-based archival file for macromolecular structures. *European Journal of Biochemistry* **80**(2): 319-24. DOI: 10.1111/j.1432-1033.1977.tb11885.x

Bleasby AJ and Wootton JC (1990). Construction of validated, non-redundant composite protein sequence databases. *Protein Engineering.* **3**(3): 153-159. DOI: 10.1093/protein/3.3.153

Bleasby AJ, Akrigg D, Attwood TK.(1994). OWL—a non-redundant composite protein sequence database. *Nucleic Acids Res.* **22**(17): 3574-3577.

Daisuke Kihara YDY, Troy H. (2006). Bioinformatics resources for cancer research with an emphasis on gene function and structure prediction tools. *Cancer Informatics* **2**: 25-35.

Etzold T, Ulyanov A and Argos P (1996). SRS: information retrieval system for molecularbiology data banks. *Methods Enzymol*. **266**: 114-128. DOI: 10.1016/S0076-6879(96)66010-8

Hendlich M (1998) Databases for protein-ligand complexes. *Acta Crys.t D.* **54**(1): 1178-1182. DOI : 10.1107/S0907444998007124

Holm L and Sander C (1998) Touring protein fold space with Dali/FSSP. *Nucleic Acids Res*. **26**(1): 316-319. DOI: 10.1093/nar/26.1.316

Laskowski RA, Hutchinson EG, Michie AD, Wallace AC, Jones ML and Thornton JM (1997) PDBsum: a Web-based database of summaries and analyses of all PDB structures. *Trends in Biomedical Science* **22**(12): 488-490. DOI: 10.1016/S0968-0004(97)01140-7

Lo Conte L, Ailey B, Hubbard TJ., Brenner SE, Murzin AG and Chothia C (2000) SCOP: a structural classification of proteins database. *Nucleic Acids Res*. **28**(1): 257-259. DOI: 10.1093/nar/28.1.257

Luscombe NM, Greenbaum D and Gerstein M (2001).What is bioinformatics? An introduction and overview. Yearbook of Medical Informatics.

Mc Garvey PB, Huang H, Barker WC, Orcutt BC, Garavelli JS, Srinivasarao GY and et al.(2000) PIR: a new resource for bioinformatics. *Bioinformatics* **16**(3): 290-291. DOI: 10.1093/bioinformatics/16.3.290

Okayama T, Tamura T, Gojobori T, Tateno Y, Ikeo K, Miyazaki S and et al. (1998) Formal design and implementation of an improved DDBJ DNA database with a new schema and object-oriented library. *Bioinformatics* **14**(6): 472-8. DOI: 10.1093/bioinformatics/14.6.472

Ouzounis CA (2012) Rise and demise of bioinformatics? Promise and progress. *PLoS Comput. Biol.* **8**: e1002487. DOI: 10.1371/journal.pcbi.1002487

Pandey AS and Divyasheesh V (2016) Applications of Bioinformatics in Medical Renovation and Research. *International Journal of Advanced Research in Computer Science and Software Engineering* **6**(3): 56-58.

Pearl FM, Lee D, Bray JE, Sillitoe I, Todd AE, Harrison AP and et al. (2000) Assigning genomic sequences to CATH. *Nucleic Acids Res.* **28**(1): 277-282. DOI: 10.1093/nar/28.1.277

Pearson WR and Lipman DJ (1988) Improved tools for biological sequence comparison. *Proceedings of National Academy of Sciences U S A* **85**(8): 2444-2448. DOI: 10.1073/pnas.85.8.2444

Schuler GD, Epstein JA, Ohkawa H, Kans JA (1996) Entrez: molecular biology database and retrieval system. *Methods Enzymol*. **266**: 141-62. DOI: 10.1016/S0076-6879(96)66012-1

Tatusov RL, Koonin EV and Lipman DJ (1997) A genomic perspective on protein families. *Science*. **278**(5338): 631-637. DOI: 10.1126/science.278.5338.631

Vondrasek J and Wlodawer A (1997) Database of HIV proteinase structures. Trends in Biochemical Science **22**(5): 183. DOI: 10.1016/s0968-0004(97)01024-4