ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 9, November 2012

Bioinformatics: A revolutionary way of using Computer technology for evidence based medicine

Onkar S Kemkar[1] and Dr P B Dahikar[2]

PCD ICSR, VMV College Campus, Wardhaman Nagar, Nagpur – 440008, India[1]

Kamla Nehru Mahavidyalaya, Sakkardara Square, Nagpur – 440009, India[2]

***ABSTRACT*: A flood of data means that many of the challenges in biology are now challenges in computing. Bioinformatics, the application of computational techniques to analyze the information associated with bimolecular on a large-scale, has now firmly established itself as a discipline in molecular biology, and encompasses a wide range of subject areas from structural biology, genomics to gene expression studies. In this review we provide an introduction and overview of the current state of the field. We discuss the main principles that underpin bioinformatics analyses, look at the types of biological information and databases that are commonly used, and finally examine some of the studies that are being conducted, particularly with reference to transcription regulatory systems.**

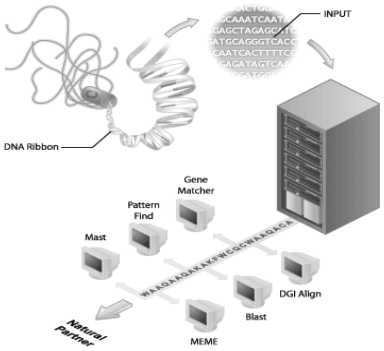
Keywords: Healthcare system, Bioinformatics, Biological data

I INTRODUCTION

Bioinformatics is the combination of biology and information technology. The discipline encompasses any computational tools and methods used to manage, analyze and manipulate large sets of biological data. Essentially, bioinformatics has three components:

- The creation of databases allowing the storage and management of large biological data sets.
- The development of algorithms and statistics to determine relationships among members of large data sets.
- The use of these tools for the analysis and interpretation of various types of biological data, including DNA, RNA and protein sequences, protein structures, gene expression profiles, and biochemical pathways[1,2].

The term bioinformatics first came into use in the 1990s and was originally synonymous with the management and analysis of DNA, RNA and protein sequence data. Computational tools for sequence analysis had been available since the 1960s, but this was a minority interest until advances in sequencing technology led to a rapid expansion in the number of stored sequences in databases such as GenBank.



Now, the term has expanded to incorporate many other types of biological data, for example protein structures, gene expression profiles and protein interactions[3]. Each of these areas requires its own set of databases, algorithms and statistical methods. Bioinformatics is largely, although not exclusively, a computer-based discipline. Computers are important in bioinformatics for two reasons:

First, many bioinformatics problems require the same task to be repeated millions of times. For example, comparing a new sequence to every other sequence stored in a database or comparing a group of sequences systematically to determine evolutionary relationships. In such cases, the ability of computers to process information and test alternative solutions rapidly is indispensable. Second, computers are required for their problem-solving power [4, 5]. Typical problems that might be addressed using bioinformatics could include solving the folding pathways of protein given its amino acid sequence, or deducing a biochemical pathway given a collection of RNA expression profiles. Computers can help with such problems, but it is important to note that expert input and robust original data are also required

Over the past few decades, major advances in the field of molecular biology, coupled with advances in genomic technologies, have led to an explosive growth in the biological information generated by the scientific community [6, 7, 8]. This deluge of genomic information has, in turn, led to an absolute requirement for computerized databases to store, organize, and index the data and for specialized tools to view and analyze the data.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 9, November 2012

II WHAT IS A BIOLOGICAL DATABASE?

A **biological database** is a large, organized body of persistent data, usually associated with computerized software designed to update, query, and retrieve components of the data stored within the system. A simple database might be a single file containing many records, each of which includes the same set of information [9]. For example, a record associated with a nucleotide sequence database typically contains information such as contact name, the input sequence with a description of the type of molecule, the scientific name of the source organism from which it was isolated, and often, literature citations associated with the sequence[10].

For researchers to benefit from the data stored in a database, two additional requirements must be met:

- easy access to the information
- a method for extracting only that information needed to answer a specific biological question

Bioinformatics is the application of computer technology to the management of biological information. Computers are used to gather, store, analyze and integrate biological and genetic information which can then be applied to gene-based drug discovery and development

Biology in the 21st century is being transformed from a purely lab-based science to an information science as well.

III WHAT IS BIOINFORMATICS?

Bioinformatics is the field of science in which biology, computer science, and information technology merge to form 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. At the beginning of the "genomic revolution", a bioinformatics concern was the creation and maintenance of a database to store biological information, such as nucleotide and amino acid sequences. Development of this type of database involved not only design issues but the development of complex interfaces whereby researchers could both access existing data as well as submit new or revised data[11,12].

Ultimately, however, all of this information must be combined to form a comprehensive picture of normal cellular activities so that researchers may study how these activities are altered in different disease states. Therefore, the field of bioinformatics has evolved such that the most pressing task now involves the analysis and interpretation of various types of data, including nucleotide and amino acid sequences, protein domains, and protein structures. The actual process of analyzing and interpreting data is referred to as **computational biology**. Important sub-disciplines within bioinformatics and computational biology include:

- The development and implementation of tools that enable efficient access to, and use and management of, various types of information
- The development of new algorithms (mathematical formulas) and statistics with which to assess relationships among members of large data sets, such as methods to locate a gene within a sequence, predict protein structure and/or function, and cluster protein sequences into families of related sequences [13].

IV WHY IS BIOINFORMATICS SO IMPORTANT?

The rationale for applying computational approaches to facilitate the understanding of various biological processes includes:

- A more global perspective in experimental design.
- The ability to capitalize on the emerging technology of **database-mining** - the process by which testable hypotheses are generated regarding the function or structure of a gene or protein of interest by identifying similar sequences in better characterized organisms .

Although a human disease may not be found in exactly the same form in animals, there may be sufficient data for an animal model that allow researchers to make inferences about the process in humans.

V EVOLUTIONARY BIOLOGY

New insight into the molecular basis of a disease may come from investigating the function of homolog's of a disease gene in model organisms. In this case, **homology** refers to two genes sharing a common evolutionary history. Scientists also use the term homology, or homologous, to simply mean similar, regardless of the evolutionary relationship.

Equally exciting is the potential for uncovering evolutionary relationships and patterns between different forms of life. With the aid of nucleotide and protein sequences, it should be possible to find the ancestral ties between different organisms. Thus far, experience has taught us that closely related organisms have similar sequences and that more distantly related organisms have more dissimilar sequences. Proteins that show significant sequence conservation, indicating a clear evolutionary relationship, are said to be from the same **protein family**. By studying **protein folds** (distinct protein building blocks) and families, scientists are able to reconstruct the evolutionary relationship between two species and to estimate the time of divergence between two organisms since they last shared a common ancestor.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 9, November 2012

VI PROTEIN MODELING

The process of evolution has resulted in the production of DNA sequences that encode proteins with specific functions. In the absence of a protein structure that has been determined by X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy, researchers can try to predict the three-dimensional structure using **protein or molecular modeling**. This method uses experimentally determined protein structures **(templates)** to predict the structure of another protein that has a similar amino acid sequence **(target)**.

Although molecular modeling may not be as accurate at determining a protein's structure as experimental methods, it is still extremely helpful in proposing and testing various biological hypotheses. Molecular modeling also provides a starting point for researchers wishing to confirm a structure through X-ray crystallography and NMR spectroscopy. Because the different genome projects are producing more sequences and because novel protein folds and families are being determined, protein modeling will become an increasingly important tool for scientists working to understand normal and disease-related processes in living organisms.

The Steps of Protein Modeling

- Identify the proteins with known three-dimensional structures that are related to the target sequence
- Align the related three-dimensional structures with the target sequence and determine those structures that will be used as templates

VII GENOME MAPPING

Genomic maps serve as a scaffold for orienting sequence information. A few years ago, a researcher wanting to localize a gene, or nucleotide sequence, was forced to manually map the genomic region of interest, a time-consuming and often painstaking process. Today, thanks to new technologies and the influx of sequence data, a number of high-quality, genome-wide maps are available to the scientific community for use in their research. Computerized maps make gene hunting faster, cheaper, and more practical for almost any scientist. In a nutshell, cientists would first use a genetic map to assign a gene to a relatively small area of a chromosome. They would then use a physical map to examine the region of interest close up, to determine a gene's precise location. In light of these advances, a researcher's burden has shifted from mapping a genome or genomic region of interest to navigating a vast number of Web sites and databases.

Map Viewer: A Tool for Visualizing Whole Genomes or Single Chromosomes

Map Viewer is a tool that allows a user to view an organism's complete genome, integrated maps for each chromosome (when available), and/or sequence data for a genomic region of interest. When using Map Viewer, a researcher has the option of selecting either a "Whole-Genome View" or a "Chromosome or Map View". The Genome View displays a schematic for all of an organism’s chromosomes, whereas the Map View shows one or more detailed maps for a single chromosome. If more than one map exists for a chromosome, Map Viewer allows a display of these maps simultaneously.

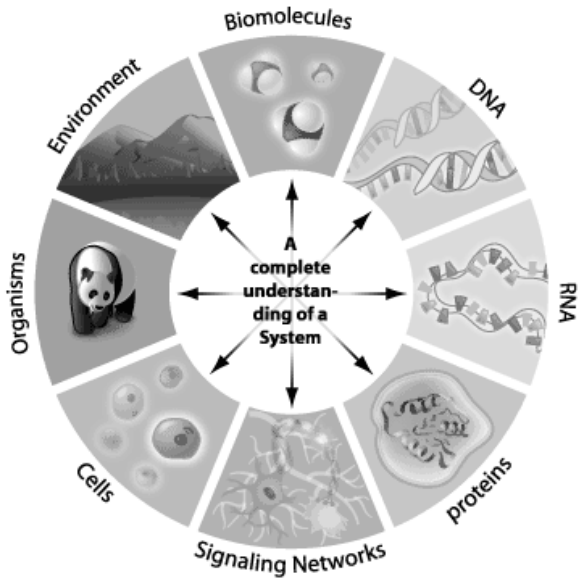
Using Map Viewer, researchers can find answers to questions such as:

- Where does a particular gene exist within an organism's genome?
- Which genes are located on a particular chromosome and in what order?
- What is the corresponding sequence data for a gene that exists in a particular chromosomal region?
- What is the distance between two genes?

The rapidly emerging field of bioinformatics promises to lead to advances in understanding basic biological processes and, in turn, advances in the diagnosis, treatment, and prevention of many genetic diseases. Bioinformatics has transformed the discipline of biology from a purely lab-based science to an information science as well. Increasingly, biological studies begin with a scientist conducting vast numbers of database and Web site searches to formulate specific hypotheses or to design large-scale experiments. The implications behind this change, for both science and medicine, are staggering.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 9, November 2012



1)

VIII GENETIC MAPS

Types of Landmarks Found on a Genetic Map

Genetic maps use landmarks called genetic markers to guide researchers on their gene hunt.

Just like interstate maps have cities and towns that serve as landmarks, genetic maps have landmarks known as **genetic markers**, or "markers" for short. The term marker is used very broadly to describe any observable variation that results from an alteration, or mutation, at a single genetic locus. A marker may be used as one landmark on a map if, in most cases, that stretch of DNA is inherited from parent to child according to the standard rules of inheritance. Markers can be within genes that code for a noticeable physical characteristic such as eye color, or a not so noticeable trait such as a disease. **DNA-based reagents** can also serve as markers. These types of markers are found within the non-coding regions of genes and are used to detect unique regions on a chromosome. DNA markers are especially useful for generating genetic maps when there are occasional, predictable mutations that occur during **meiosis**—the formation of gametes such as egg and sperm—that, over many generations, lead to a high degree of variability in the DNA content of the marker from individual to individual.

Commonly Used DNA Markers

- RFLPs, or restriction fragment length polymorphisms, VNTRs, or variable number of tandem repeat polymorphisms
- Microsatellite polymorphisms
- SNPs, or single nucleotide polymorphisms

From Linkage Analysis to Genetic Mapping

IX PHYSICAL MAPS

Types of Physical Maps and What They Measure

Physical maps can be divided into three general types: **chromosomal** or **cytogenetic maps**, **radiation hybrid (RH) maps**, and **sequence maps**. The different types of maps vary in their degree of **resolution**, that is, the ability to measure the separation of elements that are close together. The higher the resolution, the better the picture.

The lowest-resolution physical map is the chromosomal or **cytogenetic map**, which is based on the distinctive banding patterns observed by light microscopy of stained chromosomes. As with genetic linkage mapping, chromosomal mapping can be used to locate genetic markers defined by traits observable only in whole organisms. Because chromosomal maps are based on estimates of physical distance, they are considered to be physical maps. Yet, the number of base pairs within a band can only be estimated.

RH maps and sequence maps, on the other hand, are more detailed. RH maps are similar to linkage maps in that they show estimates of distance between genetic and physical markers, but that is where the similarity ends. **RH maps** are able to provide more precise information regarding the distance between markers than can a linkage map.

The physical map that provides the most detail is the sequence map. **Sequence maps** show genetic markers, as well as the sequence between the markers, measured in base pairs.

X CONCLUSION

The future of bioinformatics is integration. For example, integration of a wide variety of data sources such as clinical and genomic data will allow us to use disease symptoms to predict genetic mutations and vice versa. The integration of GIS data, such as maps, weather systems, with crop health and genotype data, will allow us to predict successful outcomes of agriculture experiments. Another future area of research in bioinformatics is large-scale comparative genomics. For example, the development of tools that can do 10-way comparisons of genomes will push forward the discovery rate in this field of bioinformatics. Along these lines, the modeling and visualization of full networks of complex systems could be used in the future to predict how the system (or cell) reacts to a drug for example. A technical set of challenges faces bioinformatics and is being addressed by faster computers, technological advances in disk storage space, and increased bandwidth. Finally, a key research question for the future of bioinformatics will be how to computationally compare complex biological observations, such as gene expression patterns and protein networks. Bioinformatics is about converting biological observations to a model that a computer will understand. This is a very challenging task since biology can be very complex.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 9, November 2012

REFERENCES

[1] Anderson, James G, and Kenneth W. Goodman. Ethics and Information
[2] Technology: A Case-Based Approach to a Health Care System in Transition.
[3] Health Informatics. New York: Springer, 2002. BA Call Number: 174.2 A5451 (B4)
[4] Bourne, Philip E., and Helge Weissig, eds. Structural Bioinformatics. Methods of Biochemical Analysis 44. Hoboken, NJ: Wiley-Liss, 2003.BA Call Number: 572.8733 (B1)
[5] Bremer, Eric G, eds. Knowledge Discovery in Life Science Literature: PAKDD 2006
[6] International Workshop, KDLL 2006, Singapore, April 9, 2006: Proceedings.
[7] Lecture Notes in Computer Science 3886. Lecture Notes in Bioinformatics. Berlin:Springer, 2006.BA Call Number: 006.3 P1111 (B4)
[8] Campbell, A. Malcolm, and Laurie J. Heyer. Discovering Genomics, Proteomics, and Bioinformatics. San Francisco: Benjamin Cummings, 2003. BA Call Number: 572.86 (B4 -- Closed Stacks)
[9] Dwyer, Rex A. Genomic Perl: From Bioinformatics Basics to Working Code. Cambridge, UK: Cambridge University Press, 2003. BA Call Number: 572.80285 D9935 (B1)
[10] Algorithms for Molecular Biology (AMB). BioMed Central. 2006-2008.
[11] BMC Bioinformatics. BioMed Central. 2000-2008. Cancer Informatics. Libertas Academica. 2005-2008.
[12] Chem-Bio Informatics Journal (CBI). Chem-Bio Informatics Society. 2001-2007.
[13] Computational Biology and Chemistry. Elsevier Science. 2003-2008. Source: ScienceDirect (Database)