Bioinformatics: Homology Detection and Protein Families
Computation in the Area of New Research

Manjeet Kaur Bedi¹, Kendar Pratap² and Raj Kumari³

¹ Assistant Professor,
Govt. National P.G.College, Sirsa,Haryana (INDIA)

² Lecturer,
Govt. National P.G.College, Sirsa,Haryana (INDIA)

³ Lecturer,
Govt. National P.G.College, Sirsa,Haryana (INDIA)

Abstract
It is the name given to these mathematical and computing approaches used to glean understanding of biological processes. It is the application of information technology to the field of molecular biology. Bioinformatics entails the creation and advancement of databases, algorithms, computational and statistical techniques, and theory to solve formal and practical problems arising from the management and analysis of biological data. Over the past few decades rapid developments in genomic and other molecular research technologies combined developments in information technologies have combined to produce a tremendous amount of information related to molecular biology. Common activities in Bioinformatics include mapping and analyzing DNA and protein sequences, aligning different DNA and protein sequences to compare them and creating and viewing 3-D models of protein structures. The primary goal of bioinformatics is to increase our understanding of biological processes. What sets it apart from other approaches, however, is its focus on developing and applying computationally intensive techniques (e.g., data mining, and machine learning algorithms) to achieve this goal. Major research efforts in the field include sequence alignment, gene finding, genome assembly, protein structure alignment, protein structure prediction, prediction of gene expression and protein-protein interactions, and the modeling of evolution.

Keywords: statistical, DNA, encode polypeptides, CPU time, Eco System, Micro Aray

1. Introduction
Bioinformatics is that branch of life science which deals with the study of application of information technology to the field of molecular biology. Bioinformatics was applied in the creation and maintenance of a database to store biological information at the beginning of the "genomic revolution", 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.

In order to study how normal cellular activities are altered in different disease states, the biological data must be combined to form a comprehensive picture of these activities. 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: a) the development and implementation of tools that enable efficient access to, and use and management of, various types of information, b) 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.
2. Research Areas in Bioinformatics

a. Shotgun Sequence Analysis

Today, computer programs are used to search the genome of thousands of organisms, containing billions of nucleotides. These programs would compensate for mutations (exchanged, deleted or inserted bases) in the DNA sequence, in order to identify sequences that are related, but not identical. A variant of this sequence alignment is used in the sequencing process itself. The so-called shotgun sequencing technique does not give a sequential list of nucleotides, but instead the sequences of thousands of small DNA fragments (each about 600-800 nucleotides long). The ends of these fragments overlap and, when aligned in the right way, make up the complete genome. Shotgun sequencing yields sequence data quickly, but the task of assembling the fragments can be quite complicated for larger genomes. DNA sequences of hundreds of organisms have been decoded and stored in databases. The information is analyzed to determine genes that encode polypeptides, as well as regulatory sequences. A comparison of gene homologs within a species or between different species can show similarities between protein functions, or relations between species (the use of molecular systematics to construct phylogenetic trees). With the growing amount of data, it long ago became impractical to analyze DNA sequences manually. In the case of the Human Genome Project, it took several days of CPU time (on one hundred Pentium III desktop machines clustered specifically for the purpose) to assemble the fragments. Shotgun sequencing is the method of choice for virtually all genomes sequenced today, and genome assembly algorithms are a critical area of bioinformatics research.

b. Automatic Search:

Another aspect of bioinformatics in sequence analysis is the automatic search for genes and regulatory sequences within a genome. Not all of the nucleotides within a genome are genes. Within the genome of higher organisms, large parts of the DNA do not serve any obvious purpose. This so-called junk DNA may, however, contain unrecognized functional elements. Bioinformatics helps to bridge the gap between genome and proteome projects—for example, in the use of DNA sequences for protein identification.

3. Gene Expression Techniques in Problem

The expression of many genes can be determined by measuring mRNA levels with multiple techniques including microarrays, expressed cDNA sequence tag (EST) sequencing, serial analysis of gene expression (SAGE) tag sequencing, massively parallel signature sequencing (MPSS), or various applications of multiplexed in-situ hybridization. All of these techniques are extremely noise-prone and/or subject to bias in the biological measurement, and a major research area in computational biology involves developing statistical tools to separate signal from noise in high-throughput gene expression studies. Such studies are often used to determine the genes implicated in a disorder; one might compare microarray data from cancerous epithelial cells to data from non-cancerous cells to determine the transcripts that are up-regulated and down-regulated in a particular population of cancer cells.

4. Genome Annotation a New Research

Annotation is the process of marking the genes and other biological features in a DNA sequence. The first genome annotation software system was designed in 1995. Sequenced and analyzed the first genome of a free-living organism to be decoded eg. bacterium Haemophilus influenzae. A software system to find the genes (places in the DNA sequence that encode a protein), the transfer RNA, and other features, and to make initial assignments of function to those genes. Most current genome annotation systems work similarly, but the programs available for analysis of genomic DNA are constantly changing and improving.

5. Computational biology

It is defined as biology that defined by computer. In our paper computational evolutionary biology is refer as the origin and descent of species, as well as their change over time. Informatics has assisted evolutionary biologists in several key ways; it has enabled researchers to:

Trace the evolution of a large number of organisms by measuring changes in their DNA, rather than through physical taxonomy or physiological observations alone,

More recently, compare entire genomes, which permits the study of more complex evolutionary events, such as gene duplication, horizontal gene transfer, and the prediction of factors important in bacterial speciation,

Build complex computational models of populations to predict the outcome of the system over time

Track and share information on an increasingly large number of species and organisms
Future work endeavours to reconstruct the now more complex tree of life. The area of research within computer science that uses genetic algorithms is sometimes confused with computational evolutionary biology, but the two areas are unrelated.

6. Modeling biological systems

Systems biology involves the use of computer simulations of cellular subsystems (such as the networks of metabolites and enzymes which comprise metabolism, signal transduction pathways and gene regulatory networks) to both analyze and visualize the complex connections of these cellular processes. Artificial life or virtual evolution attempts to understand evolutionary processes via the computer simulation of simple (artificial) life forms.

7. Biodiversity of an Ecosystem

Biodiversity of an ecosystem might be defined as the total genomic complement of a particular environment, from all of the species present, whether it is a biofilm in an abandoned mine, a drop of sea water, a scoop of soil, or the entire biosphere of the planet Earth. One very exciting potential of this field is that entire DNA sequences, or genomes of endangered species can be preserved, allowing the results of Nature’s genetic experiment to be remembered in silico, and possibly reused in the future, even if that species is eventually lost.

Databases are used to collect the species names, descriptions, distributions, genetic information, status and size of populations, habitat needs, and how each organism interacts with other species. Specialized software programs are used to find, visualize, and analyze the information, and most importantly, communicate it to other people. Computer simulations model such things as population dynamics, or calculate the cumulative genetic health of a breeding pool (in agriculture) or endangered population (in conservation).

8. Bioinformatic Techniques

a. Analysis of regulation

In our research the bioinformatics techniques have been applied to explore various steps in this process. For example, promoter analysis involves the identification and study of sequence motifs in the DNA surrounding the coding region of a gene. Regulation is the complex or chestration of events starting with an extracellular signal such as a hormone and leading to an increase or decrease in the activity of one or more proteins. These motifs influence the extent to which that region is transcribed into mRNA. Expression data can be used to infer gene regulation: one might compare microarray data from a wide variety of states of an organism to form hypotheses about the genes involved in each state. In a single-cell organism, one might compare stages of the cell cycle, along with various stress conditions (heat shock, starvation, etc.). One can then apply clustering algorithms to that expression data to determine which genes are co-expressed. For example, the upstream regions (promoters) of co-expressed genes can be searched for over-represented regulatory elements.

b. Problems of Protein Microarray

Protein microarrays and high throughput (HT) mass spectrometry (MS) can provide a snapshot of the proteins present in a biological sample. Bioinformatics is very much involved in making sense of protein microarray and HT MS data; the former approach faces similar problems as with microarrays targeted at mRNA, the latter involves the problem of matching large amounts of mass data against predicted masses from protein sequence databases, and the complicated statistical analysis of samples where multiple, but incomplete peptides from each protein are detected.

c. Analysis of mutations in cancer

In cancer, the genomes of affected cells are rearranged in complex or even unpredictable ways. Massive sequencing efforts are used to identify previously unknown point mutations in a variety of genes in cancer. Bioinformaticians continue to produce specialized automated systems to manage the sheer volume of sequence data produced, and they create new algorithms and software to compare the sequencing results to the growing collection of human genome sequences and germline polymorphisms. New physical detection technology are employed, such as oligonucleotide microarrays to identify chromosomal gains and losses (called comparative genomic hybridization), and single nucleotide polymorphism arrays to detect known point mutations. These detection methods simultaneously measure several hundred thousand sites throughout the genome, and when used in high-throughput to measure thousands of samples, generate terabytes of data per experiment. Again the massive amounts and new types of data generate new opportunities for bioinformaticians. The data is often found to contain considerable variability, or noise, another type of data that requires novel informatics development is the analysis of lesions found to be recurrent among many tumors.
9. Application of Bioinformatics

Important application of bioinformatics is protein structure prediction. The amino acid sequence of a protein, the so-called primary structure, can be easily determined from the sequence on the gene that codes for it. In the vast majority of cases, this primary structure uniquely determines a structure in its native environment. (Of course, there are exceptions, such as the bovine spongiform encephalopathy - aka Mad Cow Disease - prion.) Knowledge of this structure is vital in understanding the function of the protein. For lack of better terms, structural information is usually classified as one of secondary, tertiary and quaternary structure. A viable general solution to such predictions remains an open problem. As of now, most efforts have been directed towards heuristics that work most of the time.

One of the key ideas in bioinformatics is the notion of homology. In the genomic branch of bioinformatics, homology is used to predict the function of a gene: if the sequence of gene A, whose function is known, is homologous to the sequence of gene B, whose function is unknown, one could infer that B may share A's function. In the structural branch of bioinformatics, homology is used to determine which parts of a protein are important in structure formation and interaction with other proteins. In a technique called homology modeling, this information is used to predict the structure of a protein once the structure of a homologous protein is known. This currently remains the only way to predict protein structures reliably.

One example of this is the similar protein homology between hemoglobin in humans and the hemoglobin in legumes (leghemoglobin). Both serve the same purpose of transporting oxygen in the organism. Though both of these proteins have completely different amino acid sequences, their protein structures are virtually identical, which reflects their near identical purposes.

10. Analysis of genome in Research

The core of comparative genome analysis is the establishment of the correspondence between genes (orthology analysis) or other genomic features in different organisms. It is these intergenic maps that make it possible to trace the evolutionary processes responsible for the divergence of two genomes. A multitude of evolutionary events acting at various organizational levels shape genome evolution. At the lowest level, point mutations affect individual nucleotides. At a higher level, large chromosomal segments undergo duplication, lateral transfer, inversion, transposition, deletion and insertion. Ultimately, whole genomes are involved in processes of hybridization, polyploidization and endosymbiosis, often leading to rapid speciation. The complexity of genome evolution poses many exciting challenges to developers of mathematical models and algorithms, who have recourse to a spectra of algorithmic, statistical and mathematical techniques, ranging from exact, heuristics, fixed parameter and approximation algorithms for problems based on parsimony models to Markov Chain Monte Carlo algorithms for Bayesian analysis of problems based on probabilistic models.

Computational Technologies as High-throughput image analysis

Computational technologies are used to accelerate or fully automate the processing, quantification and analysis of large amounts of high-information-content biomedical imagery. Modern image analysis systems augment an observer's ability to make measurements from a large or complex set of images, by improving accuracy, objectivity, or speed. A fully developed analysis system may completely replace the observer. Although these systems are not unique to biomedical imagery, biomedical imaging is becoming more important for both diagnostics and research. Some examples are:

- high-throughput and high-fidelity quantification and sub-cellular localization (high-content screening, cytohistopathology)
- morphometrics
- clinical image analysis and visualization
- determining the real-time air-flow patterns in breathing lungs of living animals
- quantifying occlusion size in real-time imagery from the development of and recovery during arterial injury
- making behavioral observations from extended video recordings of laboratory animals
- infrared measurements for metabolic activity determination
- inferring clone overlaps in DNA mapping, e.g. the Sulston score

Protein-protein docking

In the last two decades, tens of thousands of protein three-dimensional structures have been determined by X-ray crystallography and Protein nuclear magnetic resonance
spectroscopy (protein NMR). One central question for the biological scientist is whether it is practical to predict possible protein-protein interactions only based on these 3D shapes, without doing protein-protein interaction experiments. A variety of methods have been developed to tackle the Protein-protein docking problem, though it seems that there is still much place to work on in this field.

**Software and tools**

Software tools for bioinformatics range from simple command-line tools, to more complex graphical programs and standalone web-services available from various bioinformatics companies or public institutions. The computational biology tool best-known among biologists is probably BLAST, an algorithm for determining the similarity of arbitrary sequences against other sequences, possibly from curated databases of protein or DNA sequences. The NCBI provides a popular web-based implementation that searches their databases. BLAST is one of a number of generally available programs for doing sequence alignment.

Commonly used software tools and technologies in this field include C++, XML, C, MySQL, SQL, Java, and Microsoft Excel.

**Web services in bioinformatics**

SOAP and REST-based interfaces have been developed for a wide variety of bioinformatics applications allowing an application running on one computer in one part of the world to use algorithms, data and computing resources on servers in other parts of the world. The main advantages lay in the end user not having to deal with software and database maintenance overheads[1] Basic bioinformatics services are classified by the EBI into three categories: SSS (Sequence Search Services), MSA (Multiple Sequence Alignment) and BSA (Biological Sequence Analysis). The availability of these service-oriented bioinformatics resources demonstrate the applicability of web based bioinformatics solutions, and range from a collection of standalone tools with a common data format under a single, standalone or web-based interface, to integrative, distributed and extensible bioinformatics workflow management systems.

**References**


