International
Journal
Of Advanced Research in Engineering & Management (IJAREM)

BIOINFORMATICS

Khadri S.S
Guest lecturer
Govt Degree College, Sindhanur, Raichur (Dist) Karnataka, India

***Abstract*—** Bioinformatics is conceptualizing biology in terms of molecules (in the sense of physical-chemistry) and then applying, informatics, techniques (derived from disciplines such as applied math, CS, and statistics) to understand and organize the information associated with these molecules, on a large-scale. Bioinformatics is MISfor Molecular Biology Information. Bioinformatics is both an umbrella term for the body of biological studies that use computer programming as part of their methodology, as well as a reference to specific analysis "pipelines" that are repeatedly used, particularly in the fields of genetics and genomics. Common uses of bioinformatics include the identification of candidate genes and nucleotides (SNPs). Often, such identification is made with the aim of better understanding the genetic basis of disease, unique adaptations, desirable properties (esp. in agricultural species), or differences between populations. In a less formal way, bioinformatics also tries to understand the organisational principles within nucleic acid and protein sequences.

***Keywords*—** Genetic material; Mechanical motion/support; Control of growth/differentiation; General Types of Informatics in Bioinformatics; The Character of Molecular Biology Information: Redundancy and Multiplicity; Sequence alignment and Sequence database

I. INTRODUCTION

Bioinformatics is the fields of science in which biology, computer science, and information technology merge to form asingle discipline. It is the emerging field that deals with the application of computers to the collection, organization, analysis, manipulation, presentation, and sharing of biologic data to solve biological problems on the molecular level. Bioinformatics is the mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information.

The term bioinformatics was coined by paulien hogeweg in 1979 for the study of informatics processes in biotoc systems. The national center for biotechnology information defines bioinformatics as:" bioinformatics is the field of science in which biology, computer science and information technology merges into a single discipline. there are three important sub-disciplines' within bioinformatics: the development of new algorithms and statistics with which to assess relation ships among members of large data sets; the analysis and interpretation of various types of data including nucleotide and amino acid sequences, proteins domains, and proteins structures; and the development and implementation of tools that enable efficient access and management of different types of information."

II. SEQUENCE ANALYSIS

The sequences of different genes or proteins may be aligned side-by-side to measure their similarity. This alignment compares protein sequences containing WPP domains.

Since the Phage Φ-X174 was sequenced in 1977,the DNA sequences of thousands of organisms have been decoded and stored in databases. This sequence information is analyzed to determine genes that encode proteins, RNA genes, regulatory sequences, structural motifs, and repetitive sequences. A comparison of genes within a species or between different species can show similarities between protein functions, or

relations between species (the use of molecular systematic to construct phylogenetic trees). With the growing amount of data, it long ago became impractical to analyze DNA sequences manually. Today, computer programs such as BLAST are used daily to search sequences from more than 260 000 organisms, containing over 190 billion nucleotides. These programs can compensate for mutations (exchanged, deleted or inserted bases) in the DNA sequence, 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 (which was used, for example, by The Institute for Genomic Research to sequence the first bacterial genome, *Haemophilias' influenza*) does not produce entire chromosomes. Instead it generates the sequences of many thousands of small DNA fragments (ranging from 35 to 900 nucleotides long, depending on the sequencing technology). The ends of these fragments overlap and, when aligned properly by a genome assembly program, can be used to reconstruct the complete genome. Shotgun sequencing yields sequence data quickly, but the task of assembling the fragments can be quite complicated for larger genomes. For a genome as large as the human genome, it may take many days of CPU time on large-memory, multiprocessor computers to assemble the fragments, and the resulting assembly usually contains numerous gaps that must be filled in later. Shotgun sequencing is the method of choice for virtually all genomes sequenced today, and genome assembly algorithms are a critical area of bioinformatics research.

Another aspect of bioinformatics in sequence analysis is annotation. This involves computational gene finding to search for protein-coding genes, RNA genes, and other functional sequences within a genome. Not all of the nucleotides within a genome are part of genes. Within the genomes 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.

Genome annotation

In the context of genomics, annotation is the process of marking the genes and other biological features in a DNA sequence. This process needs to be automated because most genomes are too large to annotate by hand, not to mention the desire to annotate as many genomes as possible, as the rate of sequencing has ceased to pose a bottleneck. Annotation is made possible by the fact that genes have recognisable start and stop regions, although the exact sequence found in these regions can vary between genes.

The first genome annotation software system was designed in 1995 by Owen White, who was part of the team at The Institute for Genomic Research that sequenced and analyzed the first genome of a free-living organism to be decoded, the bacterium *Haemophilias influenza*. White built a software system to find the genes (fragments of genomic sequence that encode proteins), the transfer RNAs, 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, such as the Gene Mark program trained and used to find protein-coding genes in *Haemophilias influenza*, are constantly changing and improving.

Computational evolutionary biology

Evolutionary biology is the study of the origin and descent of species, as well as their change over time. Informatics has assisted evolutionary biologists by enabling 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 not necessarily related.

Comparative genomics

The core of comparative genome analysis is the establishment of the correspondence between genes (morphology analysis) or other genomic features in different organisms. It is these intergenomic 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 end symbiosis, 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.
Many of these studies are based on the homology detection and protein families' computation.

III. 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.

Open-source bioinformatics software

Many free and open-source software tools have existed and continued to grow since the 1980s. The combination of a continued need for new algorithms for the analysis of emerging types of biological readouts, the potential for innovative *in silicone* experiments, and freely available open code bases have helped to create opportunities for all research groups to contribute to both bioinformatics and the range of open-source software available, regardless of their funding arrangements. The open source tools often act as incubators of ideas, or community-supported plug-ins in commercial applications. They may also provide *defector* standards and shared object models for assisting with the challenge of bio information integration.

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 derive from the fact that end users do not have to deal with software and database maintenance overheads.
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.

Bioinformatics workflow management systems

A Bioinformatics workflow management system is a specialized form of a workflow management system designed specifically to compose and execute a series of computational or data manipulation steps, or a workflow, in a Bioinformatics application. Such systems are designed to

provide an easy-to-use environment for individual application scientists themselves to create their own workflows

provide interactive tools for the scientists enabling them to execute their workflows and view their results in real-time

Simplify the process of sharing and reusing workflows between the scientists.

enable scientists to track the provenance of the workflow execution results and the workflow creation steps.

IV. STRUCTURAL BIOINFORMATICS

Protein Structure Basics

Starting from this chapter and continuing through the next three chapters, we introduce

The basics of protein structural bioinformatics. Proteins perform most essential biological and chemical functions in a cell. They play important roles in structural, enzymatic, transport, and regulatory functions. The protein functions are strictly determined by their structures. Therefore, protein structural bioinformatics is an essential element of bioinformatics. This chapter covers some basics of protein structures and associated databases, preparing the reader for discussions of more advanced topics of protein structural bioinformatics.

HIERARCHY

Protein structures can be organized into four levels of hierarchies with increasing

Complexity. These levels are primary structure, secondary structure, tertiary structure,

And quaternary structure. A linear amino acid sequence of a protein is the primary

Structure. This is the simplest level with amino acid residues linked together through

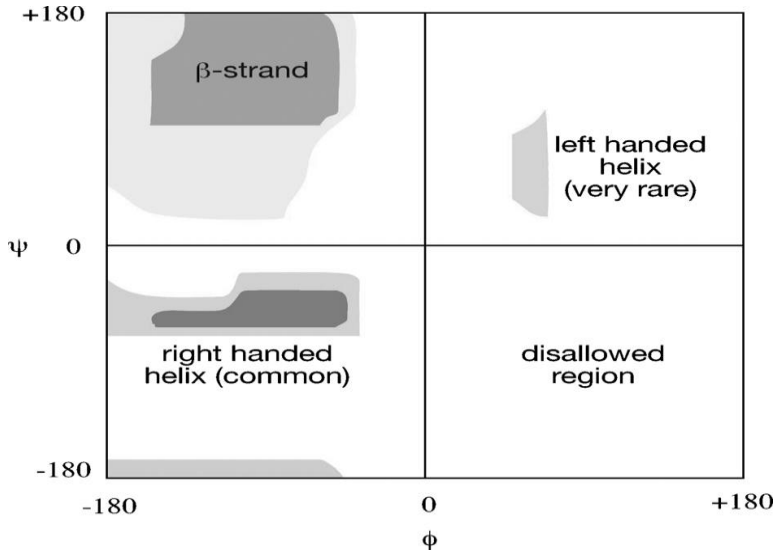

Figure 1: A Rama chandran plot with allowed values of φ and ψ in shaded areas. Regions favored by α-helices and β-strands (to be explained) are indicated.

Peptide bonds. The next level up is the secondary structure, defined as the local conformation of a peptide chain. The secondary structure is characterized by highly regular and repeated arrangement of amino acid residues stabilized by hydrogen bonds between main chain atoms of the C=O group and the NH group of different residues. The level above the secondary structure is the tertiary structure, which is the three dimensional arrangement of various secondary structural elements and connecting regions. The tertiary structure can be described as the complete three-dimensional assembly of all amino acids of a single polypeptide chain. Beyond the tertiary structure is the quaternary structure, which refers to the association of several polypeptide chains into a protein complex, which is maintained by non covalent interactions. In such a complex, individual polypeptide chains are called *mono mer*s or *subunits.* Intermediate between secondary and tertiary structures, a level of super secondary structure is often used, which is defined as two or three secondary structural elements forming a unique functional domain, a recurring structural pattern conserved in evolution.

Stabilizing Forces

Protein structures from secondary to quaternary are maintained by non covalent forces. These include electrostatic interactions, vander Waals forces, and hydrogen bonding. Electrostatic interactions are a significant stabilizing force in a protein structure.

They occur when excess negative charges in one region are neutralized by positive Charges in another region. The result is the formation of salt bridges between oppositely charged residues. The electrostatic interactions can function within a relatively long range (15 Å).

Hydrogen bonds are a particular type of electrostatic interactions similar to dipole– dipole interactions involving hydrogen from one residue and oxygen from another. Hydrogen bonds can occur between main chain atoms as well as side chain atoms.

Hydrogen from the hydrogen bond donor group such as the N–H group is slightly positively charged, whereas oxygen from the hydrogen bond acceptor group such as the C=O group is slightly negatively charged. When they come within a close distance (<3 Å), a partial bond is formed between them, resulting in a hydrogen bond. Hydrogen bonding patterns are a dominant factor in determining different types of protein secondary structures.

Van der Waals forces also contribute to the overall protein stability. These forces are instantaneous interactions between atoms when they become transient dipoles. A transient dipole can induce another transient dipole nearby. The dipoles of the two atoms can be reversed a moment later. The oscillating dipoles result in an attractive force. The van der Waals interactions are weaker than electrostatic and hydrogen bonds and thus only have a secondary effect on the protein structure.

In addition to these common stabilizing forces, disulfide bridges, which are covalent bonds between the sulfur atoms of the cysteine residue, are also important in maintaining some protein structures. For certain types of proteins that contain metal ions as prosthetic groups, non covalent interactions between amino acid residues and the metal ions may play an important structural role.

Starting from this chapter and continuing through the next three chapters, we introduce The basics of protein structural bioinformatics. Proteins perform most essential biological and chemical functions in a cell. They play important roles in structural, enzymatic, transport, and regulatory functions. The protein functions are strictly determined by their structures. Therefore, protein structural bioinformatics is an essential element of bioinformatics. This chapter covers some basics of protein structures and associated databases, preparing the reader for discussions of more advanced topics of protein structural bioinformatics.

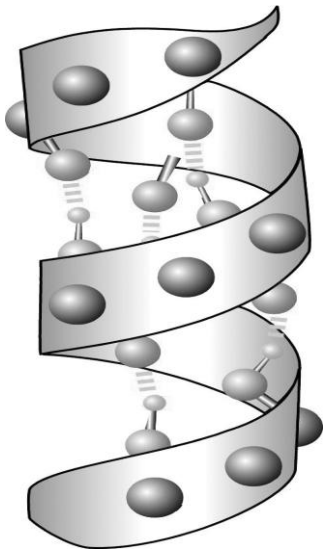

Figure 2: A ribbon diagram of a α-helix with main chain atoms (as grey balls) had shown. Hydrogen bonds between the carbonyl oxygen (red) and the amino hydrogen (green) of two residues are shown in yellow dashed lines (see color plate section).

V. CONCLUSIONS

Databases are fundamental to modern biological research, especially to genomic studies. The goal of a biological database is two fold: information retrieval and knowledge discovery. Electronic databases can be constructed either as flat files, relational, or object oriented. Flat files are simple text files and lack any form of organization to facilitate information retrieval by computers. Relational databases organize data as tables and search information among tables with shared features. Object-oriented databases organize data as objects and associate the objects according to hierarchical relationships. Biological databases encompass all three types. Based on their content, biological databases are divided into primary, secondary, and specialized databases. Primary databases simply archive sequence or structure information; secondary databases include further analysis on the sequences or structures. Specialized databases cater to a particular research interest. Biological databases need to be interconnected so that entries in one database can be cross-linked to related entries in another database. NCBI databases accessible through Entrez are among the most Integrated databases. Effective information retrieval involves the use of Boolean operators.

REFERENCES

[1] Aebersold, R., andMann, M. 2003.Mass spectrometry-based proteomics. *Nature* 422:198–207.

[2] Cutler, P. 2003. Protein arrays: The current state-of-the-art. *Proteomics* 3:3–18.

[3] Donnes, P., andHoglund, A. 2004. Predicting protein subcellular localization: past, present, and future. *Genomics Proteomics Bioinformatics* 2:209–15.

[4] Droit, A., Poirier, G. G., and Hunter, J.M. 2005. Experimental and bioinformatic approaches for interrogating protein-protein interactions to determine protein function. *J.Mol. Endocrinol.*34:263–80.

[5] Eisenhaber, F., Eisenhaber,B., andMaurer-Stroh, S. 2003. "Prediction of post-translational modifications from amino acid sequence: Problems, pitfalls, and methodological hints." In *Bioinformatics and Genomes: Current Perspectives*, edited by M. A. Andrade, 81–105.Wymondham, UK: Horizon Scientific Press. Emanuelsson, O. 2002. Predicting protein subcellular localisation from amino acid sequence.

[6] Huynen, M. A., Snel, B.,Mering, C., and Bork, P. 2003. Function prediction and protein networks. *Curr. Opin. Cell Biol*. 15:191–8.

[7] Mann, M., and Jensen, O. N. 2003. Proteomic analysis of post-translational modifications. *Nature Biotechnol*. 21:255–61.

[8] Nakai, K. 2000. Protein sorting signals and prediction of subcellular localization. *Adv. Protein Chem*. 54:277–344.

[9] Nakai, K. 2001. Review: Prediction of in vivo fates of proteins in the era of genomics and proteomics. *J. Struct. Biol.* 134:103–16.

[10] Phizicky, E., Bastiaens, P. I. H., Zhu, H., Snyder, M., and Fields, S. 2003. Protein analysis on a proteomic scale. *Nature* 422:208–15.

[11] Sadygov, R. G., Cociorva, D., and Yates, J. R. III. 2004. Large-scale database searching using tandem mass spectra: Looking up the answer in the back of the book. *Nat. Methods* 1:195–202.

[12] Tyers, M., andMann, M. 2003. From genomics to proteomics. *Nature* 422:193–7.

[13] Valencia, A., and Pazos, F. 2002. Computational methods for the prediction of protein interactions. *Curr. Opin. Struct. Biol*. 12:368–73.

[14] Valencia, A., and Pazos, F. 2003. Prediction of protein-protein interactions from evolutionary information.*Methods Biochem. Anal.* 44:411–26.