© PHOTODISC

Computational Techniques and Pattern Recognition

Pattern Discovery in Bioinformatics

BY JAGATH C. RAJAPAKSE

With the recent advances of high-throughput sequencing, microarray technology, and high-content screening, biologists are able to gather data about molecular events at an unprecedented rate. Yet, biologists today are facing an uphill task in making inferences from their data because of the lack of necessary techniques and tools for analyzing such data. The field of bioinformatics has evolved for two reasons: for the creation and maintenance of biological databases and for the extraction of knowledge underlying life sciences data to unravel the mysteries of biological phenomena and disease.

Bioinformatics data come in various forms such as biological sequences, molecular structures, gene and protein expressions, molecular networks, cellular images, and literature. A major aspect of discovering biological knowledge is to search, predict, and model specific patterns of the data that are likely to be associated with an important biological phenomenon or another set of data. One of the major discoveries in bioinformatics is that specific patterns in our genome and proteome are able to decipher our characters and how prone we are for certain diseases. To date, pattern recognition algorithms have been successfully applied or catered to address a wide range of bioinformatics problems. This issue highlights a few such applications that were selected from the presentations at the Third International Association for Pattern Recognition (IAPR) International Conference on Pattern Recognition in Bioinformatics (PRIB 2007), Singapore.

The genome comprises of all the DNA molecules of a cell or an organism and carries the blueprint of the function of the cell or organism. The information stored in DNA—a macromolecule of four nucleotides (A, T, G, and C)—is first transcribed to mRNA and then translated to proteins. During translation, patterns of three nucleotides, referred to as codons, are converted to one of the 20 amino acids. Though DNA and proteins form complex three-dimensional (3-D) patterns, they are often represented and stored as one-dimensional (1-D) sequences of nucleotides or amino acids, respectively. Only about 5% of the human genome contains useful patterns of nucleotides, or genes, that code for proteins. However, many believe that the rest of the genome, often referred to as junk, may still contain undiscovered sequence segments of functional significance.

The article by Rajapakse et al. describes a comparative genomics approach to detect conserved segments in noncoding regions of vertebrates and an implementation on a computing grid. The DNA segments preserved across several species are likely to have important functions for their existence over the evolution.

Differences across individual genomes of the same species are known as polymorphisms. In recent years, single nucleotide polymorphisms (SNPs) have been widely investigated as genetic markers for disease susceptibility of individuals. The patterns of statistically linked multiple SNPs on the same gene, referred to as haplotypes, are useful for the identification of genes or gene–gene interactions involved in complex diseases. The article by Assawamakin et al. presents a nonparametric classification technique for identifying genetic markers by finding a mapping between the inferred haplotypes and disease/control status.

Proteins are the functional forms of cells, and protein–DNA and protein–protein interactions are responsible for the majority of biological functions. A protein's functions and interactions with other proteins are primarily determined by its 3-D structure. Computational biologists often handle the prediction of protein structures from amino acid sequences by first predicting the protein secondary structure (PSS) and thereby the 3-D structure (see Figure 1). The PSS represents the backbone structure of the protein and patterns seen in PSS are broadly classified into α-helices, β-stands, and coils. Today, bioinformatics approaches to PSS prediction from amino acid sequences have reached about 80%, and molecular dynamic approaches to visualize protein folding and interactions have only been possible for small peptides of the order of tens of amino acids. Therefore, the predictions of proteins' 3-D structure and protein–protein interactions remain as grand challenges in bioinformatics.

Because of the difficulty in determining the proteins' 3-D structure, the functions of proteins are alternatively predicted through classification or homology finding. Proteins with similar characteristics such as physicochemical properties, codon usage, structural motifs, etc. are likely to have similar functions. Gupta et al. propose wavelet-variant features derived from seven physicochemical properties of amino acids as novel features for a support vector machine classifier for functional annotation of G-protein coupled receptors. A

Digital Object Identifier 10.1109/MEMB.2009.933570

0739-5175/09/$25.00©2009IEEE

protein domain is a part of protein sequence or a building block, which can function, evolve, and exist independently from the rest of the protein. Lin et al. propose a mountain clustering method to find a library of structural building blocks for the construction of 3-D structures of proteins.

Microarrays allow measurements of expressions of thousands of genes simultaneously. One major objective of the analysis of gene expressions is to identify which genes are responsible for specific conditions. For example, in cancer, genes responsible for different cancers or different stages of a particular cancer are of high importance. This is achieved by the classification of patterns of gene expressions gathered across tissue samples under different conditions. There are two challenges for microarray data analysis: 1) only hundreds of samples are gathered in a microarray study compared with the thousands of genes measured and 2) only a few genes are responsible for a particular condition or disease stage. This has made gene selection (also referred to as feature selection or variable selection) a highly investigated topic in bioinformatics. The article by Ooi et al. describes how a gene's relevancy and redundancy is compromised by accurately estimating their priority in a filter-based approach of feature selection. Wrapper and embedded methods explicitly incorporate the selection criterion into the classification algorithm.

Another objective of microarray experiments is to identify coregulated genes, which requires collection of gene expressions of a sample at different times or experimental conditions; for instance, to study the progression of an infection, gene expressions are gathered at several time points spanning across hours or days after the host is infected by the pathogen. The clustering of expression profiles is able to find the genes that have similar behavioral patterns across experiments or time and thereby identifies genes belonging to the same biological pathway. There is an increased interest in biclustering techniques, because the clustering of expression profiles simultaneously across the samples and the genes infers the genes and underlying pathways responsible for a specific condition. Figure 2 illustrates an application of hierarchical biclustering on a high-content microscopic screening of tissues to study the effect of a drug at different concentrations. Biclustering of morphological features of cells and concentrations reveals the concentrations at which cell morphologies undergo similar and major changes.

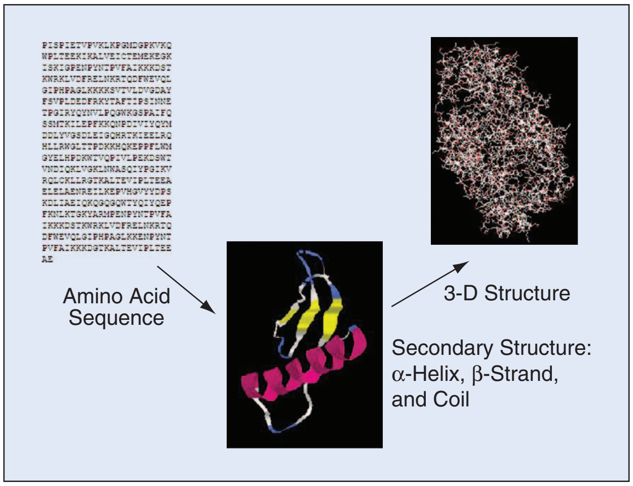

Fig. 1. The bioinformatics approach to prediction of protein structure involves predicting the secondary structure first and thereby the 3-D structure of the protein.

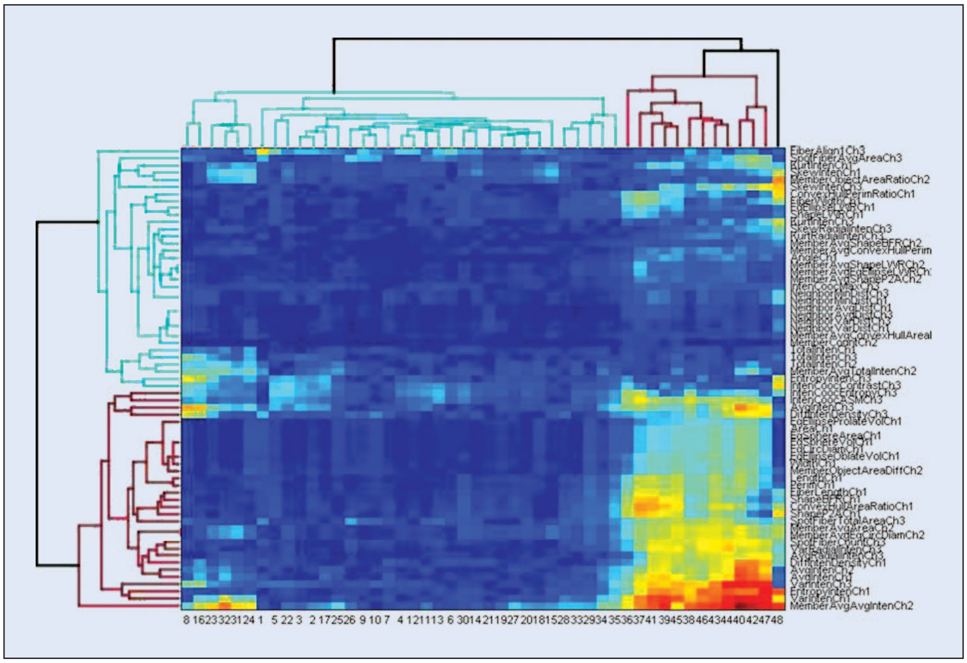

Fig. 2. Biclustering of morphological features of cells against the concentration of the drug: the rows represent morphological features and columns drug concentrations. The color codes indicate the statistically significant changes of morphological features relative to untreated cells.

Building gene regulatory networks (GRNs) has become possible, for example, by modeling with ordinary differential equations (ODEs) or Bayesian networks when microarray data are gathered over a large number of time points, but remains as an unchallenged problem when only short time series of gene expressions are available. Gathering gene expression data at a high spatial resolution, such as in a multicellular environment of a tissue sample, is of high practical and clinical relevance. Hohm and Zitzler describe an ODE-based model of a GRN and then propose an evolutionary approach to adjust its parameters to fit into a multicell environment.

Bioinformatics techniques facilitate not only the discovery of novel drugs but also the effective delivery of therapy and vaccines. Motakis et al. propose a regression model to relate expression profiles of genes and gene pairs to different types of breast cancer. The authors find a synergetic effect on patient's survival time and identify gene markers associated with low- and high-risk subtypes of breast cancer. The genes identified are supported by an analysis of information from

One of the major discoveries in bioinformatics is that specific patterns in our genome and proteome are able to decipher our characters and how prone we are for certain diseases.

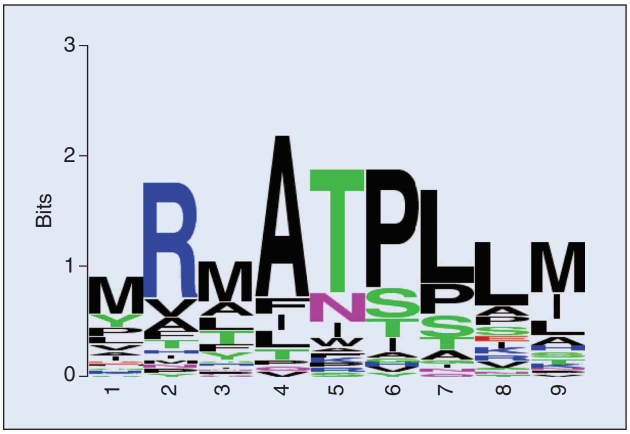

Fig. 3. The logo of a 9-mer motif binding peptides to MHC molecules.

gene ontology. The article by Pennisi et al. proposes a multiobjective evolutionary algorithm to find the optimal vaccination pattern for cancer immunization.

The binding of DNA or protein molecules takes place at the specific segments or motifs of sequences and structures. The knowledge of a binding site or motif of two molecules enables one to facilitate or inhibit the binding of the molecules. The article by Rajapakse and Lin describes an evolutionary algorithm for finding a consensus motif by combining already-known experimental motifs and a method for characterizing the motifs based on the physicochemical properties of amino acids. The method is applied to characterize peptide-binding motifs to major histocompatibility complex (MHC) molecules and may be useful for epitope-based drug discovery. Figure 3 illustrates a peptide-binding motif to an MHC molecule with a logo that indicates the conservation of specific amino acids at different sites. A motif is often characterized by a few strongly preserved sites for specific amino acids.

Bioinformatics solutions to the problems in systems biology are emerging, where the interest is to handle molecular pathways comprising of networks of molecules, genes, or proteins. This requires gathering of data from diverse sources, such as gene and protein expressions, and synergistic fusion of multiple modalities. Identification of core subnetworks by graph mining provides insights into important sets of molecules for the functioning of a pathway. This enables further streamlining of genes that are to be tested in wet labs or in clinical setting. Detection of bipartites, cliques, and bicliques of interaction and regulatory networks by using pattern recognition techniques enables making inferences on binding sites of protein complexes and on regulatory mechanisms of genes. Text-mining techniques also play a significant role in building protein–protein interaction networks, as only a single or a few interactions of protein interactions are usually discovered in a single experiment.

The functional information of organisms is stored as specific patterns of their genomes and proteomes, and the discovery of these patterns is the core to the finding of solutions to many bioinformatics problems. Efforts to perform traditional biological and clinical experiments or different stages of drug discovery pipelines, in silico or by computational means with bioinformatics tools, have saved millions of dollars and are increasingly pursued today. Therefore, computational techniques and pattern recognition in bioinformatics will play an essential and significant role in future life sciences and medicine.

Jagath C. Rajapakse received his M.Sc. and Ph.D. degrees in electrical and computer engineering from the University of Buffalo (USA). He was a visiting fellow at the National Institute of Mental Health (USA) and a visiting scientist at the Max Planck Institute of Brain and Cognitive Sciences (Germany). Currently, he is a professor of computer engineering and the director of BioInformatics Research Center, Nanyang Technological University (NTU), Singapore. He is a visiting professor to the Department of Biological Engineering, Massachusetts Institute of Technology (MIT). His research interests include neuroinformatics, bioinformatics, modeling brain connectivity through functional brain imaging, and building pathways from gene and protein expressions obtained by microarrays and high-content microscopic imaging. He has authored or coauthored more than 225 peer-reviewed journal and conference papers and book chapters and was listed among the most cited scientists. He serves as an associate editor of *IEEE Transactions on Medical Imaging, IEEE Transactions on Computational Biology and Bioinformatics*, and *IEEE Transactions on Neural Networks* and in editorial boards of several other journals.

Address for Correspondence: Jagath C. Rajapakse, Bioinformatics Research Center, Nanyang Technological University, Singapore 639798. E-mail: asjagath@ntu.edu.sg.