Pattern Recognition in Bioinformatics: An Introduction

J.C. Rajapakse[1,4,5], L. Wong[2], and R. Acharya[3]

[1] BioInformatics Research Center, Nanyang Technological University, Singapore
[2] National University of Singapore, Singapore
[3] Computer Science and Engineering, The Penn State University, USA
[4] Singapore-MIT Alliance, N2 50 Nanyang Avenue, Singapore
[5] Biological Engineering Division, Massachusetts Institute of Technology, USA
asjagath@ntu.edu.sg

The information stored in DNA, a chain of four nucleotides (A, T, G, and C), is first converted to mRNA through the process of transcription and then converted to the functional form of life, proteins, through the process of translation. Only about 5% of the genome contains useful patterns of nucleotides, or genes, that code for proteins. The initiation of translation or transcription process is determined by the presence of specific patterns of DNA or RNA, or motifs. Research on detecting specific patterns of DNA sequences such as genes, protein coding regions, promoters, etc., leads to uncover functional aspects of cells. Comparative genomics focus on comparisons across the genomes to find conserved patterns over the evolution, which possess some functional significance. Construction of evolutionary trees is useful to know how genome and proteome are evolved over all species by ways of a complete library of motifs and genes.

A protein's functionality or its interaction with another protein is mainly determined by its 3-D structure and the surface pattern. Prediction of protein's 3-D structure from its 1-D amino-acid sequence remains an open problem in structural genomics; protein-protein interactions determine all essential functions in living cells. Computational modeling and visualization tools of 3-D structures of proteins help biologists to infer cellular activities.

The challenge in functional genomics is to analyze gene expression data accumulated by microarray techniques to discover the clusters of co-regulated genes and thereby gene regulatory networks, leading to the understanding of regulatory mechanisms of genes and pathways. Molecular imaging provides techniques for *in vivo* sensing and imaging of molecular events, which measure biological processes in living organism at the molecular and cellular level. The techniques to fuse and integrate different kinds of information derived from different life science data are yet to be explored.

The knowledge in databases of biomedicine and phenotypes, combined with genotypes, is increasingly unmanageable by traditional text-based methods. Advanced data mining techniques, where the use of ontologies for constructing precise descriptors of medical concepts and procedures, are required in the field of medical informatics. The increasing amount of biological literature are posing new challenges in the field of text mining which techniques could find pathways and interaction networks from pure mining of literature.

J.C. Rajapakse, L. Wong, and R. Acharya (Eds.): PRIB 2006, LNBI 4146, pp. 1 – 3, 2006.
© Springer-Verlag Berlin Heidelberg 2006

Finding a particular structure of a sequence or surface pattern of a protein, that has a specific biological function or is involved in interactions with other molecule, is a fundamental question which could be addressed by pattern recognition algorithms. Further, pattern recognition has already shown promise in the following areas of bioinformatics:

- Computational genomics and comparative genomics
- Gene expression analysis and functional genomics
- Alignment of sequences: DNA, protein, structures, etc.
- Phylogenic analysis of species, sequences, structures
- Structural genomics and proteomics
- Functional and molecular imaging
- Data mining, data integration, and visualization
- Information fusion such as combining sequences, expressions, texts, etc.
- Pathway analysis, gene regulatory networks, etc.
- Disease modeling
- Medical informatics

Statistical, fuzzy, and neural network clustering techniques have been successfully applied to gene expression data analysis. Graph-based pattern recognition techniques have found applications in recognition of motifs, gene regulatory networks, and protein-protein interactions [1, 2, 3]. Support vector machines and information theory based approaches are increasingly used in feature selection or gene selection [4, 5]. Markov models and hidden Markov models are becoming popular in sequence alignments and gene or RNA structure finding [6, 7]. Statistical and neural network based predictors have found signals in genomic sequences and protein structures [2, 4, 8, 9]. As underpinnings of life sciences data are becoming clearer, pattern recognition algorithms would find more and more useful and relevant in solving computational biology and bioinformatics problems.

References

[1] E Eskin, PA Pevzner (2002), "Finding composite regulatory patterns in DNA sequences", Bioinformatics, 18:S354-S363.

[2] MN. Nguyen and JC. Rajapakse (2005), “Two-stage support vector regression approach for predicting accessible surface areas of amino acids,” *PROTIENS: Structure, Function, and Bioinformatics*, 63: 542-550.

[3] Min Zou, Suzanne D. Conzen (2005), "A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data", *Bioinformatics,* 21:71-79.

[4] Haifeng Li, Tao Jiang (2005), "A class of edit kernels for SVMs to predict translation initiation sites in eukaryotic mRNAs", *Journal of Computational Biology*, 12(6):702-718.

[5] Guo-Liang Li, Tze-Yun Leong (2005), “Feature selection for the prediction of translation initiation sites”, *Genomics Proteomics Bioinformatic,* 3(2):73-83.

[6] WH Majoros, M Pertea, SL Salzberg (2005), "Efficient implementation of a generalized pair hidden Markov model for comparative gene finding", *Bioinformatics*, 21(9):1782-1788.

[7] Dustin E. Schones, Pavel Sumazin, Michael Q. Zhang (2005), "Similarity of position frequency matrices for transcription factor binding sites", *Bioinformatics,* 21:307-313.

[8] Te-Ming Chen, Chung-Chin Lu, and Wen-Hsiung Li (2005), "Prediction of splice sites with dependency graphs and their expanded bayesian networks", *Bioinformatics*, 21: 471-482.

[9] Gideon Dror, Rotem Sorek, Ron Shamir (2005), "Accurate identification of alternatively spliced exons using support vector machine", *Bioinformatics*, 21:897-901.