Scientific Programming 16 (2008) 3
DOI 10.3233/SPR-2008-0245
IOS Press

Guest-Editorial

Biological data mining

Mohammed J. Zaki [a,*], Naren Ramakrishnan [b] and Srinivasan Parthasarathy [c]
[a] *Computer Science Department, Rensselaer Polytechnic Institute, Troy, NY 12180, USA*
[b] *Computer Science Department, Virginia Tech, Blacksburg, VA 24061, USA*
[c] *Department of Computer Science and Engineering, Ohio State University, Columbus, OH 43210, USA*

Biological data mining purports to extract useful knowledge from massive datasets gathered in biology, and in other related life sciences areas such as medicine and neuroscience. In the recent past, we have witnessed major transformations of these applied sciences into data-driven endeavors. For instance, the study of even a focused aspect of cellular activity, such as gene action, now benefits from multiple high-throughput data acquisition technologies such as microarrays, genome-wide deletion screens, and RNAi assays. Consequently, analysis and mining techniques, especially those that provide data reduction down to manageable quantities, have become a mainstay of these application domains.

This special issue presents novel research in biological data mining applications. The selected submissions went through two rounds of reviews by at least three reviewers. We are very grateful to the anonymous reviewers in helping us select the following papers for this special issue. The first paper, *Semi-supervised learning for classification of protein sequence data*, by Brian King and Chittibabu Guda, presents a comprehensive evaluation of semi-supervised techniques for classifying protein sequences into the various subcellular localization categories. They adopt variants of Bayesian text mining methods to formulate a generative model for protein classification. They show that their variant EM algorithms are highly effective, surpassing a transductive SVM approach. In the paper, *Discover gene specific local co-regulations from time-course gene expression data*, Ji Zhang, Qigang Gao and Hai Wang, present a genetic algorithm approach for finding co-regulated genes based on temporal expression data. They utilize progressive window based techniques, combined with efficient nearest neighbor look-up to further speed up the computations. The third paper, *Inferring neuronal network connectivity from spike data: A temporal data mining approach*, by Debprakash Patnaik, P.S. Sastry and K.P. Unnikrishnan, applies frequent episode discovery to the task of inferring the underlying neuronal connectivity patterns from temporal multi-neuronal spike data streams, which comprise of symbolic time-series data.

*Corresponding author. E-mail: zaki@cs.rpi.edu.