Editorial

Rasiah Loganantharaj

Bioinformatics Research Lab,
University of Louisiana at Lafayette,
Lafayette 70504, Louisiana
E-mail: logan@cacs.louisiana.edu

Biographical notes: Rasiah (Raja) Loganantharaj is a Professor at the Center for Advanced Computer Studies, University of Louisiana in Lafayette. He is currently working for the National Institute of Environmental Health Sciences, at the Research Triangle Park, NC as a bioinformatics scientist. His research investigation includes; micro RNA target prediction algorithms, next generation sequence assembly and analysis such as DNA-seq, RNA-seq and Methyl-seq. He had worked on other aspects of bioinformatics including micro array analysis, gene ontology, and interaction networks. He has published papers in bioinformatics, artificial intelligence, and computer science.

As we have started to generate a huge amount of data rapidly with the advancement of next generation sequencing technology at lower cost along with high throughput expression data set, a paradigm shift is started to take place from hypothesis generation and verification to data exploration for discovering knowledge. This special issue on Exploration and Exploitation of Data in Bioinformatics consisting of extended version of seven selected papers presented at the 7th Annual Biotechnology and Bioinformatics Symposium (BIOT-2010) held at the University of Louisiana at Lafayette on October 14–15, 2010. It featured three keynote presentations addressing trends and the state of art in bioinformatics and biotechnology, 15 podium presentations, and 10 poster presentations. All the submitted original contributions were reviewed by at least two technical program committee members who are expert in the area. The average number of reviews per contribution was three. The seven papers are broadly categorised into three groups namely

- algorithms and methods for clustering data from next generation sequencing and high throughput expression so as to facilitate down stream analysis
- combining diverse modalities or features so as to improve the effectiveness of classification and hence the prediction accuracy
- heuristics and search.

In the paper entitled 'A new algorithm for quantifying binding site pattern similarity with applications for Next Generation Sequencing' from the first group, the authors had introduced a new novel algorithm for gapped dynamic alignment of Position Frequency Matrices (PFMs) called PfmSim. Using such similarity measures, transcription factors can be clustered based on their PFMs which provides insight into gene regulation such as

structurally related transcription factors that compete for binding site. Further PfmSim can be used for similarity ranking of database PFMs to identify motifs derived from ChIP-seq peaks after *de novo* motif discovery. They have demonstrated the effectiveness of their algorithm by comparing it with a similar algorithm using simulated data sets.

Unlike many other clustering algorithms, the one introduced in the paper entitled 'An effective graph-based clustering technique to identify coherent patterns from gene expression data' automatically captures the coherent patterns in gene expression and clusters them without any parameters. The authors have clustered four different expression data sets from public domain and demonstrated the utility of their algorithm by comparing the results against several other popular clustering algorithms.

The papers from group 2 clearly demonstrate that that fidelity of signals improves when combining signals from diverse modalities. To facilitate such integration, the authors of paper entitled 'Large margin classifiers and Random Forests for integrated biological prediction' have proposed a large margin random forests classification approach based on random forests proximity. They have used random forests proximity kernel or its derivative kernels to obtain large margin classifiers. They demonstrated the effectiveness of their approach by comparing against some popular classifiers on four biological datasets.

Moulavi et al. have proposed an approach based on combining gene expression with interaction network to improve the feature selection in a paper entitled 'Combining gene expression and interaction network data to improve kidney lesion score prediction'. The general practice of predicting lesion scores from microarray expressions from patients' renal biopsies were error prone due to high dimensionality and intrinsic noisy nature of this data. They have explored different types of feature selection methods beyond statistical approach and have shown that the one based on topological features of interaction network works better than the biological feature selection method for prediction.

The following three papers in the last group fall into heuristics and search. In the paper entitled 'Detecting molecular selection on single amino acid replacements' the author had explored amino acid properties with variable window size to computationally screen Single Nucleotide Polymorphisms (SNPs) for adaptive changes and had presented some promising results. This approach is in contrast to nucleotide-based methods, which fail to detect positive selection associated with adaptive SNPs.

In the paper entitled 'Computational analysis of adaptive antigenic mutations of the human influenza haemagglutinin for vaccine strain selection' have analysed haemagglutinin (HA) sequence so as to identify significant antigenic selection sites and to link these sites and their respective amino acids to the structural changes over time. They have identified four and eight positively selected sites for H1N1 and H3N2 respectively. They also have explored some strains containing a significant number of selection sites, which make them reasonable candidates for a vaccine.

Duax et al. explored retrieval of ribosomal proteins in prokaryotic species from a search vector composed of some combination of a small subset from 20 amino acids in a paper entitled 'Evolution of bacterial ribosomal protein L1'. Using three dimensional structural information, they have developed search vectors composed primary of Gly, Ala, Arg, and Pro residues (GARP) distributed across the entire protein sequence that retrieve 98% of each of the ribosomal proteins in prokaryotic species with no false hits.

We are grateful to Dr. Yi Pan, the Editor-in-Chief of the IJBRA journal, for his continuous support of the BIOT series, encouragements, and for his willingness to dedicate a special issue of IJBRA to BIOT 2010. We also thank the staffs at IJBRA for being patient and diligent in bring this issue in time. We hope that you will enjoy reading these manuscripts as much as the conference attendees enjoyed hearing their presentations at the symposium.