Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2009, Article ID 582697, 2 pages
doi:10.1155/2009/582697

Editorial

Computational Data Mining in Cancer Bioinformatics and Cancer Epidemiology

Zhenqiu Liu,[1] Dechang Chen,[2] Xuewen Chen,[3] and Haomiao Jia[4]

[1] *Division of Biostatistics and Bioinformatics, The University of Maryland Greenebaum Cancer Center, USA*
[2] *Department of Preventive Medicine and Biometrics, Uniformed Services University of the Health Sciences, USA*
[3] *Department of Electrical Engineering and Computer Science, The University of Kansas, USA*
[4] *Mailman School of Public Health, Columbia University, USA*

Correspondence should be addressed to Zhenqiu Liu, zliu@umm.edu

Received 18 August 2009; Accepted 18 August 2009

High throughput technologies such as microarray have produced huge amount of genomic and proteomic data in public domain. Many survey and clinical outcome data such as SEER data are also available. A long list of links to large health-related data sets can be found at the website http://www.ehdp.com/. All of these databases have different temporal and spatial assumptions for example, different frequencies of collection, different spatial resolution (by state, by county, by zip-code, and by square kilometer), and so forth. How to mine these data together and extract useful information is really a challenging task. This special issue brings together researchers from different disciplines and encourages collaborative research on cancer related computational data mining. The objectives of this special issue are intended to address two challenging issues. One is how to identify and evaluate biomarkers (features, risk/protector) factors. The other is to develop new or adapt existing algorithms to analyze data from different sources.

This special issue published 11 articles, which may be classified into three groups: (1) those concerned with problems with gene selection and predictions, (2) those developed methods for network construction and system biology with multi source genomic data, and (3) those related to medical informatics and methodology research.

The first group covers methods in gene selection and prediction, a fundamental problem in biomedical research. These studies open new avenues for identifying complex disease genes and biomarkers for disease diagnosis and for assessing drug efficacy and toxicity. For examples, the l_1 penalized methods can be efficiently implemented with different classifiers for gene identification and model prediction. In one article, Huang and Wu propose a novel method for cancer diagnosis using gene expression data by casting the classification problem as finding sparse representations of test samples with respect to training samples. The sparse representation is computed by the l_1-regularized least square method. The proposed method is more efficient than SVMs as it has no need of model selection. Receiver Operating Characteristic (ROC) analysis is a common tool for assessing the performance of biomarkers and prediction models. It gained much popularity in biomedicine. Liu et al., in another article, propose a novel method through regularized F-measure maximization. The proposed method assigns different costs to positive and negative samples and does simultaneous feature selection and prediction with l_1 penalty. This method is useful especially when data set is highly unbalanced or the labels for negative (positive) samples are missing, which is very common in biomedical research.

Also in the first group, an article by Jrad develops a multiclass cancer diagnosis with class-selective rejection scheme for gene selection. It gives a general formulation of the problem and proposes a possible solution based on ν-1-SVM coupled with its regularization path. The proposed classifier minimizes any asymmetric loss function and consists of rejecting some patients from one, some, or all classes in order to ensure a higher reliability while reducing time and expense costs. Another article on human cancer prediction by Martín-Merino et al. incorporates in

the ν-SVM algorithm a linear combination of non-Euclidean dissimilarities. The weights of the combination are learnt in a Hyper Reproducing Kernel Hilbert Space (HRKHS) using a Semidefinite Programming algorithm. This approach allows us to incorporate a smoothing term that penalizes the complexity of the family of distances and avoids overfitting. This method is more robust than the traditional support vector machines (SVMs). Another methodology article by Hua et al. proposes a Bayesian cut fitting to describe features in response to the skeletal age. Their method cannot only capture the entire pattern of feature variation but also carry the local properties regarding the skeletal age.

The second group includes genomic networks and system biology with multi source genomic data. In a biological system, genes perform different molecular functions and regulate various biological processes via interactions with other genes thus forming a variety of complex networks. Article by Han et al. proposes an integrative method based on the bootstrapping K-S test to evaluate a large number of microarray datasets generated from 21 different types of cancer in order to identify gene pairs that have different relationships in normal versus cancer tissues. The significant alteration of gene relations can greatly extend our understanding of the molecular mechanisms of human cancer. In another article, Spencer et al. utilize data mining methods based on machine learning to build a predictive model of lung injury by retrospective analysis of treatment planning archives. In addition, biomarkers for this model are extracted from a prospective clinical trial that collects blood serum samples at multiple time points. They utilize a 3-way proteomics methodology to screen for differentially expressed proteins that are related to RP. They present their proteomic methodology to investigate predictive biomarkers of RP that could eliminate informational gaps in the retrospective physical model. Article by Wang et al. constructs a single gene network based on linear programming and an integrated analysis of the significant function cluster using Kappa statistics and fuzzy heuristic clustering. Finally, in their article, Loganantharaj and Chung introduce an integrating protein-to-protein interaction information, pathway information with array expression data set to identify a set of "important" genes and potential signal transduction networks that help to target and reverse the oncogenic phenotype induced by tumor antigen such as integrin a6b4.

The third group comprises two articles, which cover advances in medical informatics and spatial and temporal data analyses. In their article, Chen et al. develop a prognostic system of cancer patients with ensemble clustering and SEER database. This system can be used to predict an outcome or a survival rate of cancer patients with more accuracy. The article by Song et al. proposes a method of classifying temporal gene expression curves in which individual expression trajectory is modeled as longitudinal data with a changeable variance and covariance structure. The method, mainly based on generalized mixed model, is illustrated by a dense temporal gene expression data in bacteria. The power and time points of measurements are also characterized via the longitudinal mixed model. Even if the method is developed for temporal gene expression data, it may be generally applicable to other spatial and temporal data analyses.

Zhenqiu Liu
Dechang Chen
Xuewen Chen
Haomiao Jia