BMC Genomics

INTRODUCTION Open Access

Genomic signatures and gene networking: challenges and promises

Ke Zhang[1*], Mehdi Pirooznia[2], Hamid R Arabnia[3], Jack Y Yang[4], Liangjiang Wang[5,6], Zuojie Luo[7], Youping Deng[8,9,10*]

From BIOCOMP 2010. The 2010 International Conference on Bioinformatics and Computational Biology Las Vegas, NV, USA. 12-15 July 2010

Abstract

This is an editorial report of the supplement to BMC Genomics that includes 15 papers selected from the BIOCOMP'10 - The 2010 International Conference on Bioinformatics & Computational Biology as well as other sources with a focus on genomics studies.

BIOCOMP'10 was held on July 12-15 in Las Vegas, Nevada. The congress covered a large variety of research areas, and genomics was one of the major focuses because of the fast development in this field. We set out to launch a supplement to BMC Genomics with manuscripts selected from this congress and invited submissions. With a rigorous peer review process, we selected 15 manuscripts that showed work in cutting-edge genomics fields and proposed innovative methodology. We hope this supplement presents the current computational and statistical challenges faced in genomics studies, and shows the enormous promises and opportunities in the genomic future.

Although the high throughput technology has made continuous progress during the last decade in terms of size, cost and signal quality, it remains challenging to deduce reliable predictive signatures from genomic data due to a small sample size and a large number of variables. Much effort of this supplement is devoted to developing predictive models for various types of genomic data, including mRNA, microRNA, and genome DNA. One of the important applications for RNA microarray data is to identify differentially expressed genes that can lead to gene signatures for predicting disease status and drug response. Mao *et al.* [1] investigated the differential gene expression between African Americans and Caucasians in white blood cells expression profiles for both type 2 diabetes patients and healthy people. The newly identified gene markers implicate the genetic basis for distinct risks of type 2 diabetes between these two populations. For microarray data, Due to the high cost of GeneChip, microarray experiments often have low sample sizes that present challenging for statistical analysis. In view of the difficulties for applying the common-used microarray analysis methods such as t-test, SAM, and FDR for small sample size experiment, Chen and his colleagues [2] proposed a model-based information sharing method (MBIS) that enhances the power of statistical test by utilizing information shared among genes. Next-generation sequencing technology enables the quantification of gene expression in the species whose gene chips are not available in market. Chen and his colleagues [3] used the 454 pyrosequencing technology to perform the transcriptome sequencing for an important herb medicine, the root of *Panax notoginseng*. This work discovered more than 20K unique transcripts and around 900 putative transcription factors.

When using gene signature for classification of disease phenotypes, it is critical to determine a subset of genes that is reliable across various studies and that provides high predictive power for the disease status. Liu *et al.* [4] have developed a gene selection algorithm, Recursive Feature Addition, that combines supervised learning method and statistical similarity measures. The gene

* Correspondence: ke.zhang@med.und.edu; youping_deng@rush.edu
[1]Department of Pathology, Bioinformatics Core, School of Medicine and Health Sciences, University of North Dakota, Grand Forks, ND 58201, USA
[8]Department of Internal Medicine, Rush University Medical Center, Chicago, IL 60612, USA
Full list of author information is available at the end of the article

signature was further optimized via a novel algorithm, Lagging Prediction Peephole Optimization. On the other hand, Shi and his colleagues [5] aimed to minimize the number of genes in a gene signature while maintaining its predictive power. They proposed a method called Minimize Feature's Size that makes use of similarity analyses between different endpoints and at multiple levels such as probe, gene, and GO. Both manuscripts validated their methods by comparing with various gene signature algorithms using benchmark microarray data.

The advocacy of personalized medicine in complicated diseases such as cancer and neural defects has made it increasingly important to identify genomic signatures that are associated with clinical outcomes. Several manuscripts of this supplement address questions in this aspect using various types of data. Zhao *et al*. [6] investigated a number of models for predicating cancer overall survival using gene expression profile from microarray data and found that the maximum predictive power of each model is limited by the correlation between endpoint and gene expression. Instead of looking at mRNA expression level, Zhang and his colleagues [7] focused on identifying DNA copy number variation that is correlated with cancer over survival. They developed a novel and efficient algorithm using a hidden Markov model to take into account the correlation between markers in SNP array. The algorithm classified glioma samples with distinct overall survival time. In the manuscript by Wang and his colleagues [8], they moved further to associate single nucleotide variations with single amino acid polymorphisms (SAPs) that can be used for predicting disease risk. They have validated their results using public datasets such as 1000 Genome Project and Genetic Analysis Workshop (GAW17). Protein-protein interaction (PPI) networks Protein functions were utilized by Huang and Chen [9] to predict drug cardiotoxicity. They proposed a systems biology framework to predict adverse drug reactions (ADR) using supervised learning methods such as support vector machine. This framework has a potential large impact to pharmaceutical industry for ADR is one of the major reasons for drug withdrawals in clinical trials.

Nowadays researchers are interested in not only what individual genes are activated, but also how genes interact with each other. Modelling gene networking presents a high challenge for bioinformatics because of incomplete information of gene functions and gene-gene interactions. About half of the manuscripts in this supplement are addressing questions regarding gene networking. Li *et al*. [10] developed a modified version of dynamic Bayesian Network for time-series microarray data, and showed that the proposed method provided an enhanced accuracy for predicting gene regulatory network structure. Wang and his colleagues [11] applied network analysis to protein-protein interaction data from the STRING database and identified a number of proteins that are associated with proteases Malaria parasite. These results illustrated the diverse functions of protease and implicated novel targets of drug design for Malaria.

DNA- or RNA-binding proteins play a critical role in gene regulatory networking. Liu and his colleagues [12] integrated the RNA sequence and secondary structures to to identify the consensus sequence of protein-RNA binding sites. This novel model-based approach, called RNA-MotifModeler, demonstrated a number of statistical advantages when being applied to the RNA-binding protein SRSF1. The effect of epigenetic modification on gene regulation has been widely investigated. In this supplement, a manuscript by Zhao and his colleagues [13] looked into the combined regulation of epigenetic modification and miRNA in mediating gene networking. They conducted a genome-wide study and showed that DNA methylation and miRNA function are complementary to each other for gene regulation. This finding would advance our predictive models for gene regulatory networks by incorporating the epigenetic and miRNA factors.

Some of the authors devoted their efforts to traditional bioinformatics areas. Many alignment algorithm for DNA or protein sequences were proposed in 1980s, nonetheless, multiple sequence alignment is still a challenging question because of its computing intensity, which manifests with the advancement of the next-generation sequencing. Nguyen, Pan and Nong [14] provided a solution for multiple sequence alignment by combining the pair-wise dynamic programming algorithm with parallel computing approach using R-Mesh. This new method achieved computing time at $O(m)$, where m is the number of sequences. Bio-imaging analysis is another active field of bioinformatics. Tang and his colleagues [15] proposed a robust method to reduce the specles that present obstacles for ultrasound image post-processing. They used a detail preserving anisotropic diffusion filter and showed that their method prohibit over-diffusion and preserved the important structure information.

We are proud of the high quality of the manuscripts contained within this issue. We hope they would guide current genomics research and indicate the trend for future study.

Acknowledgements

This supplement will not be possible without the support of the International Society of Intelligent Biological Medicine (ISIBM).

Author details

[1]Department of Pathology, Bioinformatics Core, School of Medicine and Health Sciences, University of North Dakota, Grand Forks, ND 58201, USA.

[2]School of Medicine, Johns Hopkins University, Baltimore, MD 21287, USA. [3]Department of Computer Science, University of Georgia, Athens, Georgia 30602-7404, USA. [4]Department of Radiation Oncology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA. [5]Department of Genetics and Biochemistry, Clemson University, Clemson, SC 29634, USA. [6]J.C. Self Research Institute of Human Genetics, Greenwood Genetic Center, Greenwood, SC 29646, USA. [7]Office of the President of The First Hospital, Guangxi Medical University, Nanning, Guangxi, China. [8]Department of Internal Medicine, Rush University Medical Center, Chicago, IL 60612, USA. [9]Rush University Cancer Center, Rush University Medical Center, Chicago, IL 60612, USA. [10]Department of Biochemistry, Rush University Medical Center, Chicago, IL 60612, USA.

Published: 23 December 2011

References

1. Mao J, Ai J, Zhou X, Shenwu M, Ong M, Blue M, Washington JT, Wang XN, Deng Y: **"Transcriptomic profiles of peripheral white blood cells in type II diabetes and racial differences in expression profiles,".** *BMC Genomics* 2011, **12(Suppl 5)**:S12.
2. Chen Z, Liu Q, McGee M, Kong M, Huang X, Deng Y, Scheuermann RH: **"A gene selection method for GeneChip array data with small sample sizes,".** *BMC Genomics* 2011, **12(Suppl 5)**:S7.
3. Luo H, Sun C, Sun Y, Wu Q, Li Y, Song J, Niu Y, Cheng X, Xu H, Li C, Liu J, Steinmetz A, Chen S: **"Analysis of the transcriptome of Panax notoginseng root uncovers putative triterpene saponin-biosynthetic genes and genetic markers,".** *BMC Genomics* 2011, **12(Suppl 5)**:S5.
4. Liu Q, Sung AH, Chen Z, Chen L, Liu J, Huang X, Qiao M, Wang Z, Deng Y: **"Gene selection and classification for cancer microarray data based on machine learning and similarity measures,".** *BMC Genomics* 2011, **12(Suppl 5)**:S1.
5. Chang C, Wang J, Zhao C, Fostel J, Tong W, Bushel PR, Deng Y, Pusztai L, Symmans WF, Shi T: **"Maximizing biomarker discovery by minimizing gene signatures,".** *BMC Genomics* 2011, **12(Suppl 5)**:S6.
6. Zhao C, Shi L, Tong W, Shaughnessy JDJ, Oberthuer A, Pusztai L, Deng Y, Symmans FW, Shi T: **"Maximum predictive power of the microarray-based models for clinical outcomes is limited by correlation between endpoint and gene expression profile,".** *BMC Genomics* 2011, **12(Suppl 5)**:S3.
7. Zhang K, Yang Y, Devanarayan V, Xie L, Deng Y, Sens D: **"A hidden Markov model-based algorithm for identifying tumor subtype using array CGH data,".** *BMC Genomics* 2011, **12(Suppl 5)**:S10.
8. Hu Y, Liu Y, Jung J, Dunker KA, Wang Y: **"Changes in predicted protein disorder tendency may contribute to disease risk,".** *BMC Genomics* 2011, **12(Suppl 5)**:S2.
9. Huang L-C, Chen JY: **"Predicting drug's cardiotoxicity with a systems biology approach,".** *BMC Genomics* 2011, **12(Suppl 5)**:S11.
10. Li H, Wang N, Gong P, Perkins EJ, Zhang C: **"Learning the structure of transition gene regulatory networks from time series gene expression data,".** *BMC Genomics* 2011, **12(Suppl 5)**:S13.
11. Lilburn TG, Cai H, Zhou Z, Wang Y: **"Protease-associated cellular networks in malaria parasite Plasmodium falciparum,".** *BMC Genomics* 2011, **12(Suppl 5)**:S9.
12. Wang X, Juan L, Lv J, Wang K, Sanford J, Liu Y: **"Predicting sequence and structural specificities of RNA binding regions recognized by splicing factor SRSF1,".** *BMC Genomics* 2011, **12(Suppl 5)**:S8.
13. Su Z, Xia J, Zhao Z: **"Functional complementation between transcriptional methylation regulation and post-transcriptional microRNA regulation in the human genome,".** *BMC Genomics* 2011, **12(Suppl 5)**:S15.
14. Nguyen KD, Pan Y, Nong G: **"Parallel progressive multiple sequence alignment on reconfigurable mesh,".** *BMC Genomics* 2011, **12(Suppl 5)**:S4.
15. Liu X, Liu J, Xu X, Chun L, Tang J, Deng Y: **"A robust detail preserving anisotropic diffusion for speckle reduction in ultrasound images,".** *BMC Genomics* 2011, **12(Suppl 5)**:S14.

doi:10.1186/1471-2164-12-S5-I1
Cite this article as: Zhang *et al.*: Genomic signatures and gene networking: challenges and promises. *BMC Genomics* 2011 **12**(Suppl 5):I1.