Selected Articles from the 2012 IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS 2012)

Yufei Huang, Yidong Chen, and Xiaoning Qian

1 Introduction

The 2012 IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS 2012) was held in Washington DC from December 2nd to 4th. As in the past GENSIPS workshops, GENSIPS 2012 provided a forum for signal processing researchers, bioinformaticians, computational biologists, biostatisticians, and biomedical researchers to exchange ideas and discuss the challenges confronting bioinformatics and computational systems biology communities due to the high dimensionality and variability of modern high-throughput biomedical data as well as high complexity of genomics and proteomics. The theme of GENSIPS 2012 is *Data Mining and Modeling Methods* in genomics emphasizing the applications of signal processing and statistics in next-generation sequencing (NGS) and cancer systems biology. GENSIPS 2012 featured prominent plenary speakers including Dr. John Quackenbush from the Department of Biostatistics at Harvard University, Dr. Eric P. Hoffman from the Children's National Medical Center at George Washington University, and Dr. Jinghui Zhang from the St. Jude Children's Research Hospital.

2 Articles

This special section contains seven selected articles, which are significantly extended versions based on seven favorably reviewed papers in the GENSIPS 2012 conference proceeding. In GENSIPS 2012, each submitted article to the conference was reviewed by a minimum of two reviewers and these seven invited papers to this special section were top-ranked among more than sixty submissions. The extended journal versions were then further reviewed according to rigorous peer-review criteria.

In these accepted articles, advanced probabilistic models, graph algorithms, as well as novel optimization methods have been implemented for integrative analysis and effective visualization of diverse "omics" data [1], [2], [3], [7], network-based biomarker discovery for disease prognosis [1], [3], [5], high-throughput sequencing data analysis [4], [6], and drug sentivity prediction [2].

In [1], Gregory et al. have presented a novel statistical framework for integrative analysis of multi-platform genomics data based on decompositions of large numbers of platform-specific features into smaller numbers of latent features. The proposed framework aims to discover how diverse molecular features interact both within and between platforms based on matched patient samples. Principal components, partial least squares, non-negative matrix factorization, and sparse counterparts of each have been implemented to define the latent features and then the derived latent features are integrated in a predictive model for clinical outcomes by Bayesian model averaging. The performance comparison of these decompositions with respect to clinical outcome prediction on real and simulated data has demonstrated that the principal component decompositions achieved the best performance. Furthermore, the latent feature interactions have been shown to preserve interactions between the original features in a way that aids prediction and enables the selection of outcome-related features. The methods have been further applied to a glioblastoma multiforme dataset from The Cancer Genome Atlas to predict patient survival time by integrating gene expression, microRNA, copy number and methylation data. The selected prognostic genes indeed have known associations with glioblastoma.

Another integrative analysis of diverse "omics" data is presented by Berlow et al. [2] for the predictive modeling of tumor sensitivity to anti-cancer drugs. In this paper, the authors propose a new approach for drug sensitivity prediction based on integrated functional and genomic characterizations, which may further provide insights into personalized tumor proliferation. The proposed model is inferred from data on cell viability for a training set of molecularly targeted drugs and available genomic characterizations. The modeling approach when applied to data from the Cancer Cell Line Encyclopedia shows a significant gain in prediction accuracy as compared to elastic net and random forest techniques based on genomic characterizations. The high prediction accuracy of the framework based on functional data alone has been validated on experimental data from a 60 targeted drug screen applied to a mouse Embryonal Rhabdomyosarcoma cell culture. The authors also show that the

- *Y. Huang is with the Department of Electrical and Computer Engineering, University of Texas at San Antonio, San Antonio, TX 78249. E-mail: yufei.huang@utsa.edu.*
- *Y. Chen is with the Department of Epidemiology & Biostatistics, University of Texas Health Science Center at San Antonio, San Antonio, TX 78229. E-mail: cheny8@uthscsa.edu.*
- *X. Qian is with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 78743. E-mail: xqian@ece.tamu.edu.*

For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2014.2353218

accuracy of tumor sensitivity prediction to targeted drugs can be considerably improved by incorporating functional and genomic characterizations in modeling.

Tian et al. [3] introduce a novel integration of network biology and imaging to study cancer phenotypes and responses to treatments at molecular systems level by intergrating clinical measurements from *in vivo* imaging. In this paper, Differential Dependence Network (DDN) analysis is used to detect and visualize statistically significant topological rewiring in molecular networks between two phenotypic conditions, and *in vivo* magnetic resonance imaging (MRI) is used to more accurately define phenotypic sample groups for such differential analysis. The effectiveness of DDN is demonstrated by simulation and real data analysis. In the ND2-SmoA1 medulloblastoma mouse model treatment study by the FDA-approved antineoplastic agent, arsenic trioxide (ATO), the authors combine both MRI and Reverse Phase Protein Microarray (RPPM) data to assess tumor responses to ATO. Specifically, Kaplan-Meier survival and MRI-based tumor growth analyses have been used to establish the effectiveness of treatment with DDN analysis of the RPMA data further revealing newly identified rewiring "hubs" of biological networks triggered by ATO at the systems level.

In [4], Lu et al. propose a novel framework of Minor Allele Frequency (MAF)-based logistic principal component analysis (MLPCA) to analyze high-density Single Nucleotide Polymorphism (SNP) data. As SNP data is categorical, the authors develop a new probabilitic model considering the categorical nature and derive aggregated statistics by explicitly modeling the correlation between rare variant SNP data. The derived aggregated statistics by MLPCA can then be tested as a surrogate variable in regression models to detect gene-environment interaction from rare variants for association analyses with a given trait. Derived MLPCA-based methods aggregate rare variants by an optimal linear combination of the best SNP subset and thus could capture the best combined effect from individual rare variants belonging to the corresponding gene. Based on both simulation data set as well as Genetic Analysis Workshop 17 (GAW17) data, the authors have evaluated and compared the power of MLPCA-based methods with four existing collapsing methods in gene-environment interaction association analysis. The experimental results have demonstrated that MLPCA achieves higher statistical power than those existing methods on two forms of genotype representations, and it can be further improved by introducing the appropriate sparsity penalty.

A network-based model is proposed by Sajaddi et al. [5] to identify prognostic biomarkers by considering the interaction effects as well as the individual effects of candidate risk factors. In this new network-based framework, a node represents a candidate risk factor, and individual and pairwise interaction effects are quantified as node and edge weights respectively. Biomarker identication is then formulated as a Maximum Weighted Multiple Clique Problem (MWMCP) that searches for a collection of cliques whose total weights over both nodes and edges are maximized. As a result, the identified cliques have the highest predictive power with the most synergistic interactions among them. The authors develop an analytical algorithm based on column generation to achieve high quality solutions as well as a fast heuristic algorithm for large-scale networks. Experimental results with both randomly generated networks and constructed interaction networks from type 1 diabetes and breast cancer datasets have shown that the proposed methods can effectively identify critical biomarkers for better prediction accuracy.

Detection and annotation of SNPs have been among the central topics in modern genomics research, as SNPs are believed to play important roles on the manifestation of phenotypic events, such as disease susceptibility. A Bayesian approach, BM-SNP, is derived in [6] to identify SNPs based on the posterior inference using next-generation sequencing data. In particular, BM-SNP computes the posterior probability of nucleotide variation at each covered genomic position using the contents and frequency of the mapped short reads. The position with a high posterior probability of nucleotide variation is flagged as a potential SNP. The analysis by BM-SNP on two cell-line NGS data shows a high ratio of overlap (>95%) with the dbSNP database. Compared with MAQ, BM-SNP identifies more SNPs that are in dbSNP, with higher quality. The SNPs that are called only by BM-SNP but not in dbSNP may serve as new discoveries. The proposed BM-SNP method integrates information from multiple aspects of NGS data, and therefore achieves high detection power. BM-SNP is fast, capable of processing whole genome data at 20-fold average coverage in a short amount of time.

Flow cytometry is a widely used technology for simultaneously measuring multiple proteins at the single-cell level. A typical flow cytometry experiment collects measurements for a large number of cells, in the order of hundred thousand or higher. Analysis of such data often aims to cluster cells into subpopulations with distinct phenotypes. Currently, the most widely used analysis method in the flow cytometry community is manual gating, a process that clusters cells based on visual inspection of user-defined biaxial plots, which is highly subjective. Automated clustering algorithms have been proposed to improve gating. However, completely removing the manual component can be challenging. Instead of aiming for automation, Qiu [7] proposes a novel visualization techinque to facilitate manual gating. The proposed method views a flow cytometry dataset as a high-dimensional point cloud of cells, derives the skeleton of the cloud, and unfolds the skeleton to generate 2D visualizations, similar to unfolding an origami back to a piece of paper. The proposed visualization has been tested on real data and quantitative comparison for visualization is performed with principal component analysis and multidimensional scaling.

3 Competing Interests

The authors declare that they have no competing interests.

4. Authors' Contributions

All authors served as guest editors for the special section, with YH serving as the Lead Editor. All authors helped write this editorial.

ACKNOWLEDGMENTS

This special section would not have been possible without the support of the contributing authors, reviewers, and program committee members of IEEE GENSIPS 2012, especially Drs. Ranadip Pal and Joseph Yue Wang. The authors would also like to thank Ms. Joyce Arnold for her kind help throughout the process.

REFERENCES

[1] K. B. Gregory, A. A. Momin, K. R. Coombes, and V. Baladandayuthapani, "Latent feature decompositions for integrative analysis of multi-platform genomic data," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 11, no. 6, pp. 984–994, Nov. 2014.

[2] N. Berlow, S. Haider, Q. Wan, M. Geltzeiler, L. E. Davis, C. Keller, and P. Pal, "An integrated approach to anti-cancer drug sensitivity prediction," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 11, no. 6, pp. 995–1008, Nov. 2014.

[3] Y. Tian, S. S. Wang, Z. Zhang, O. C. Rodriguez, E. Petricoin III, I.-M. Shih, D. Chan, M. Avantaggiati, G. Yu, S. Ye, R. Clarke, C. Wang, B. Zhang, Y. Wang, and C. Albanese, "Integration of network biology and imaging to study cancer phenotypes and responses," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 11, no. 6, pp. 1009–1019, Nov. 2014.

[4] M. Lu, H.-S. Lee, D. Hadley, J. Z. Huang, X. Qian, "Logistic principal component analysis for rare variants in gene-environment interaction analysis," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 11, no. 6, pp. 1020–1028, Nov. 2014.

[5] S. J. Sajjadi, X. Qian, B. Zeng, A. A. Adl, "Network-based methods to identify highly discriminating subsets of biomarkers," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 11, no. 6, pp. 1029–1037, Nov. 2014.

[6] Y. Xu, X. Zheng, Y. Yuan, M. R. Estecio, J.-P. Issa, P. Qiu, Y. Ji, and S. Liang, "BM-SNP: A Bayesian model for SNP calling using high throughput sequencing data," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 11, no. 6, pp. 1038–1044, Nov. 2014.

[7] P. Qiu, "Unfold high-dimensional clouds for exhaustive gating of flow cytometry data," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 11, no. 6, pp. 1045–1051, Nov. 2014.

Yufei Huang (M'02) received his Ph.D. degree in electrical engineering from the State University of New York at Stony Brook in 2001. Since 2002, he has been with the Department of Electrical and Computer Engineering at the University of Texas at San Antonio (UTSA), where he is now a Professor. He has been a visiting professor at the Center of Bioinformatics, Harvard Center for Neurodegeneration & Repair. He is now also an adjunct professor of the Greehey Children's Cancer Institute and Dept. of Epidemiology and Biostatistics at the University of Texas Health Science Center at San Antonio. Dr. Huang's current research interests include uncovering the functions of mRNA methylation using high throughput sequencing technologies, microRNA functions and target identification, brain-machine-interaction using EEG data, and deep learning algorithms and application. He was a recipient of US National Science Foundation (NSF) Early CAREER Award in 2005, Best Paper Award of 2006 Artificial Neural Networks in Engineering Conference, and 2007 Best Paper Award of *IEEE Signal Processing Magazine*. He is an Associate Editor of the *IEEE Transactions on Signal Processing*, *BMC Systems Biology*, and *EURASIP Journal on Bioinformatics and Computational Biology*.

Yidong Chen received his B.S. and M.S. degrees in Electrical Engineering from Fudan University, Shanghai, China, and the Ph.D. degree in Imaging Science from the Rochester Institute of Technology, Rochester, NY. From 1986 to 1988, he worked in the Department of Electronic Engineering of Fudan University as an assistant professor. During 1988 to 1989, he was a visiting scholar in the Department of Computer Engineering, Rochester Institute of Technology. From 1995 to 1996, he worked at the Hewlett Packard Company as a research engineer, specializing in digital and color image processing. Since 1996, he has worked on the microarray technology development effort at the National Human Genome Research Institute (NHGRI), US National Institutes of Health (NIH), Bethesda, MD, as a special expert, staff scientist, and later associate investigator in the field of microarray image analysis, statistical data analysis, and bioinformatics. From 2006-2008, he worked in the Genetics Branch at National Cancer Institute (NCI) as a staff scientist. During the 12-year period with NHGRI and NCI, he worked with biologists closely and applied statistical methods to various cancer research projects, and contributed to about numerous peer-reviewed publications in methods of microarray analysis. Currently, he is a professor in the Department of Epidemiology and Biostatistics at the University of Texas Health Science Center at San Antonio (UTHSCSA), and the director of the Computational Biology and Bioinformatics (CBBI) at Greehey Children's Cancer Research Institute (GCCRI) of UTHSCSA. His research interests include bioinformatics methods in next-generation sequencing technologies and analysis methods, gene and microRNA expression analysis, DNA mutation analysis, and long noncoding RNA profiling. His research interests also include integrative genomic data integration, genetic data visualization and management, and genetic network modeling in translational cancer research.

Xiaoning Qian (S'01–M'07) received the PhD degree in Electrical Engineering from Yale University, New Haven, CT, in 2005. Currently, he is an assistant professor in the Department of Electrical & Computer Engineering, Texas A&M University, College Station, TX. He also is a courtesy assistant professor in the Department of Computer Science & Engineering and the Department of Pediatrics at the University of South Florida, Tampa FL, in which he spent four years before joining Texas A&M. He was with the Bioinformatics Training Program at Texas A&M University, sponsored by the National Cancer Institute (NCI). His current research interests include computational network biology, genomic signal processing, and biomedical image analysis.