Deep Sequencing Data Analysis

Bijoy K. Ghosh, Aniruddha Datta, and Ranadip Pal

1 Introduction

This special section consists of recent advances in Deep Sequencing Data Analysis for systems biology research. Deep sequencing technologies have been primarily applied to genomic sequencing but have recently been applied for transcriptomic profiling or mapping histone modifications. Deep Sequencing technology shows clear advantages over existing profiling technologies in terms of amount of sequence coverage, revealing new transcriptomic insights, measurement of expression of different transcript isoforms and accuracy of defining transcription level. However, being a relatively newer method for transcriptomic profiling, standardized approaches for analysis of deep sequencing expression data are still being developed. The analysis and application of deep sequencing data presents enormous challenges in the areas of machine learning, signal processing, systems theory and statistics. The emphasis of the special issue is on the latest computational challenges and finding rigorous and novel engineering approaches to tackle structural and functional systems biology problems using deep sequencing technologies.

The special section is comprised of five papers that present recent advances in the area of deep sequencing analysis.

The paper "Detecting Multivariate Gene Interactions in RNASeq Data Using Optimal Bayesian Classification" by Jason M. Knight, Ivan Ivanov, Karen Triff, Robert S. Chapkin, and Edward R. Dougherty considers the problem of detecting multivariate genetic interactions from RNA-seq data through a Bayesian computational framework comprised of a hierarchical statistical model of the RNA-Seq processing pipeline and the corresponding optimal Bayesian classifier. Markov Chain Monte Carlo sampling and Monte Carlo integration is utilized to compute quantities where no analytical formulation exists. The performance of the approach is illustrated on an expression dataset from a dietary intervention study where gene pairs are identified that have low classification error but not identified as differentially expressed. An open source version of the package to perform optimal Bayesian classification is available at http://bit.ly/obc_package.

A pipeline to assess de novo transcriptome assemblies is considered in the paper "Examining De Novo Transcriptome Assemblies via a Quality Assessment Pipeline" by Noushin Ghaffari, Osama A. Arshad, Hyundoo Jeong, John Thiltges, Michael F. Criscitiello, Byung-Jun Yoon, Aniruddha Datta, and Charles D. Johnson. A workflow of multiple quality check measurements is considered that in combination provide a clear evaluation of the assembly performance. The authors present novel transcriptome assemblies and functional annotations for Pacific whiteleg shrimp (Li-topenaeus vannamei), a mariculture species with great national and international interest, and no solid transcriptome/genome reference. The investigations showed that assessing the quality of an assembly purely based on the assembler's statistical measurements can be mislead-ing; and thus a hybrid approach consisting of statistical quality checks and further biological-based evaluations can be beneficial.

In the paper "Computational Prediction of Pathogenic Network Modules in *Fusarium verticillioides*," Mansuck Kim, Huan Zhang, Charles Woloshuk, Won-Bo Shim, and Byung-Jun Yoon performed a comparative analysis of wild type and loss-of-virulence mutant F. verticillioides co-expression networks to identify subnetwork modules that are associated with its pathogenicity. F. verticillioides co-expression networks were constructed from RNA-Seq data and network search identified subnetwork modules that are differentially activated between the wild type and mutant F. verticillioides, which considerably differ in terms of pathogenic potentials. The anlysis identified four potential pathogenicity-associated subnetwork modules, each of which consists of interacting genes with coordinated expression patterns, but whose activation level is significantly different in the wild type and the mutant. The predicted modules were comprised of functionally coherent genes and were topologically cohesive. Furthermore, they contained several orthologs of known pathogenic genes in other fungi, which may play important roles in the fungal pathogenesis.

Optimal Fault detection and diagnosis for transcriptional circuits observed through next generation sequencing data is consided in the paper "Optimal Fault Detection and Diagnosis in Transcriptional Circuits Using Next-Generation Sequencing" by Arghavan Bahadorinejad and Ulisses M. Braga-Neto. The fault detection consists of an innovations filter followed by a fault certification step, and requires no knowledge about the system faults. The innovations filter

- *B. K. Ghosh is with the Department of Mathematics and StatisticsTexas Tech University, 1108 Memorial Circle, Lubbock, TX 79409. E-mail: bijoy.ghosh@ttu.edu.*
- *A. Datta is with the Department of Electrical and Computer Engineering, Texas A&M University, 188 Bizzell Street, College Station, TX 77843. E-mail: datta@ece.tamu.edu.*
- *R. Pal is with the Department of Electrical and Computer Engineering, Texas Tech University, 1012 Boston Ave, Lubbock, TX 79409. E-mail: ranadip.pal@ttu.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.2018.2805179

uses the optimal Boolean state estimator, called the Boolean Kalman Filter (BKF). Performance is assessed by means of false detection and misdiagnosis rates, as well as average times until correct detection and diagnosis. The efficacy of the proposed methodology is demonstrated via numerical experiments using a p53-MDM2 negative feedback loop Boolean network with stuck-at faults that model mutation events commonly found in cancer.

The final paper in this section "MeTDiff: A Novel Differential RNA Methylation Analysis for MeRIP-Seq Data" by Xiaodong Cui, Lin Zhang, Jia Meng, Manjeet K. Rao, Yidong Chen, and Yufei Huang presents a computational tool MetDiff for predicting differential N6-Methyladenosine (m6A) methylation sites from Methylated RNA immunoprecipitation sequencing (MeRIP-Seq) data. Compared with the existing algorithm exomePeak, the advantages of MeTDiff are that it explicitly models the reads variation in data and devises a power likelihood ratio test for differential methylation site prediction. Comprehensive evaluation of MeTDiff's performance using both simulated and real datasets showed that MeTDiff is much more robust and achieved much higher sensitivity and specificity over exomePeak.

Bijoy Ghosh
Aniruddha Datta
Ranadip Pal
Guest Editors

Bijoy K. Ghosh received the PhD degree in engineering sciences from the Decision and Control Group of the Division of Applied Sciences, Harvard University, Cambridge, Massachusetts, in 1983. From 1983 to 2007, he was with the Department of Electrical and Systems Engineering, Washington University, St. Louis, Missouri, where he was a professor and director of the Center for BioCybernetics and Intelligent Systems. Currently, he is the Dick and Martha Brooks Regents professor of Mathematics and Statistics at Texas Tech University, Lubbock, Texas. He received the D. P. Eckmann Award in 1988 from the American Automatic Control Council, the Japan Society for the Promotion of Sciences Invitation Fellowship in 1997, the Chinese Academy of Sciences Invitation Fellowship in 2016, and the Indian Institute of Technology, Kharagpur, distinguished visiting professorship, in 2016. He became a fellow of the IEEE in 2000, a fellow of the International Federation on Automatic Control, in 2014, and a fellow of South Asia Institute of Science and Engineering in 2016. He has held visiting positions at the Tokyo Institute of Technology, Osaka University, and Tokyo Denki University, Japan, University of Padova in Italy, Royal Institute of Technology and Institut Mittag-Leffler, Stockholm, Sweden, Yale University, Technical University of Munich, Germany, Chinese Academy of Sciences, China, and Indian Institute of Technology, Kharagpur, India. His current research interest include the study of human head-eye mechanics, cyberphysical systems and control problems in rehabilitation engineering.

Aniruddha Datta received the BTech degree in electrical engineering from the Indian Institute of Technology, Kharagpur, in 1985, the MSEE. degree from Southern Illinois University, Carbondale, in 1987, and the MS (applied mathematics) and PhD degrees from the University of Southern California, in 1991. In August 1991, he joined the Department of Electrical and Computer Engineering, Texas A&M University where he is currently the J. W. Runyon, Jr. '35 professor II and director for the Center for Bioinformatics and Genomic Systems Engineering (CBGSE). His areas of interest include adaptive control, robust control, PID control, and genomic signal processing. He has authored or coauthored five books and more than 180 journal and conference papers on these topics. He is a fellow of the IEEE, has served as an associate editor of the *IEEE Transactions on Automatic Control* (2001-2003), the *IEEE Transactions on Systems, Man and Cybernetics-Part B* (2005-2006), the *IEEE Transactions on Biomedical Engineering* (2013-2015) and is currently serving as an associate editor of the *EURASIP Journal on Bioinformatics and Systems Biology*, *the ACM/IEEE Transactions on Computational Biology and Bioinformatics, the IEEE Journal of Biomedical and Health Informatics,* and *IEEE Access*.

Ranadip Pal received the BTech degree in electronics and electrical communication engineering from the Indian Institute of Technology, Kharagpur, India, in 2002, and the MS and PhD degrees in electrical engineering from Texas A & M University, College Station, in 2004 and 2007, respectively. Since August 2007, he has been with Texas Tech University where he is currently an associate professor in the Electrical and Computer Engineering Department. His research areas are genomic signal processing, stochastic modeling and control, machine learning and computational biology. He is the author of more than 80 peer reviewed articles including publications in high impact journals such as *Nature Medicine* and *Cancer Cell* and author of a book entitled *Predictive Modeling of Drug Sensitivity*. He received the NSF CAREER Award, 2010, President's excellence in Teaching Award, 2012; Whitacre Research Award, 2014, and the Chancellor's Council Distinguished Research Award, 2016.