Guest Editors Introduction to the Special Section on Software and Databases

Dong Xu, Kun Huang, and Jeanette Schmidt

SOFTWARE tools and information systems in bioinformatics and computational biology are playing more and more important roles in biology and medical research. This special section consists of a selection of papers focuing on software and databases that are central in bioinformatics and computational biology. Following a rigorous review process, 11 papers were selected for publication. These papers cover a broad range of topics, including computational genomics and transcriptomics, analysis of biological networks and interactions, drug design, biomedical signal/image analysis, biomedical text mining and ontologies, biological data mining, visualization and integration, and high performance computing application in bioinformatics.

Mapping reads to the reference genome in next-generation sequencing is a critical step in effectively using the data. In "ResSeq: Enhancing Short-Read Sequencing Alignment By Rescuing Error-Containing Reads", Weixing Feng, Peichao Sang, Deyuan Lian, Yansheng Dong, Fengfei Song, Meng Li, Bo He, Fenglin Cao, and Yunlong Liu developed an open-source tool to retrieve additional reliably aligned reads (reads with more than a pre-defined number of mismatches) using a Bayesian-based approach. In this method, they first retrieve the sequence context around the mismatched nucleotides within the already aligned reads. Then, using the derived pattern, they evaluate the remaining (typically discarded) reads with more than the allowed number of mismatches, and calculate a score that represents the probability of a specific alignment being correct. This strategy improves alignment sensitivity.

Microhomology-mediated break-induced replication (MMBIR) is one of the mechanisms that cause genomic destabilization that may lead to cancer. In "MMBIRFinder: A Tool to Detect Microhomology-Mediated Break-Induced Replication", Matthew W. Segar, Cynthia J. Sakofsky, Anna Malkova, and Yunlong Liu developed MMBIRFinder, a method that detects template-switching events associated with MMBIR from whole-genome sequencing data based on a half-read alignment approach to identify potential regions of interest. Clustering of these potential regions helps narrow the search space to regions with strong evidence. Subsequent local alignments identify the template-switching events with single-nucleotide accuracy. MMBIRFinder has demonstrated good performance using both simulated and real data. The study on real data from both normal breast tissues and breast tumor samples led to the identification of template-switching events residing in the promoter region of seven genes that have been implicated in breast cancer.

Compressing heterogeneous collections of trees represents an open challenge in computational phylogenetics. In "Heterogeneous Compression of Large Collections of Evolutionary Trees", Suzanne J. Matthews extended TreeZip's capacity in compressing homogeneous tree collections to compress heterogeneous collections of trees. The test results indicate that TreeZip averages 89.03 percent (72.69 percent) space savings on unweighted (weighted) collections of trees when the level of heterogeneity in a collection is moderate. Combining the TreeZip compressed file with general-purpose compression yields even more space savings.

Cluster analysis of biological networks is one of the most important approaches for identifying functional modules and predicting protein functions. In "ClusterViz: A Cytoscape APP for Cluster Analysis of Biological Network", Jianxin Wang, Jiancheng Zhong, Gang Chen, Min Li, Fang-xiang Wu, and Yi Pan developed a Cytoscape 3 APP ClusterViz for cluster analysis and visualization based on the framework of Open Services Gateway Initiative. Three commonly used clustering algorithms, FAG-EC, EAGLE and MCODE, are included in the current version, with the capacity to include more clusters.

In "Building Transcriptional Association Networks in Cytoscape with RegNetC", Isabel A. Nepomuceno-Chamorro, Alfonso Marquez-Chamorro, and Jesus S. Aguilar-Ruiz implemented a Regression Network plugin for Cytoscape (RegNetC) based on the RegNet algorithm for the inference of transcriptional association network from gene expression profiles. Unlike the correlation-based methods that analyze each pair of genes individually, RegNetC can detect the relationship between each gene and the remaining genes simultaneously. Model trees favours localized similarities over more global similarity, which is one of the major drawbacks of correlation-based methods.

In "iPFPi: A System for Improving Protein Function Prediction through Cumulative Iterations", Kamal Taha, Paul D. Yoo, and Mohammed Alzaabi propsed an

- *D. Xu is with the Department of Computer Science, Informatics Institute, and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211. E-mail: xudong@missouri.edu.*
- *K. Huang is with the Department of Biomedical Informatics and the Department of Computer Science and Engineering, the Ohio State University, Columbus, OH 43210. E-mail: kun.huang@osumc.edu.*
- *J. Schmidt is with Affymetrix Inc., 3420 Central Expressway, Santa Clara, CA 95051. E-mail: Jeanette_Schmidt@affymetrix.com.*

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.2015.2454931

annotation system called iPFPi that predicts the functions of un-annotated proteins based on the abstracts of biomedical literature associated with the target protein. The authors constructed a novel semantic similarity measure that takes into consideration several factors using a pairwise beats and looses procedure. Tests showed some improvement over some recent protein function prediction systems.

Machine learning algorithms are key instruments for many bioinformatics applications, including prediction of gene-functions based on available biomolecular annotations. In "Software Suite for Gene and Protein Annotation Prediction and Similarity Search", Davide Chicco and Marco Masseroli described a gene function prediciton software suite using Latent Semantic Indexing (LSI), which takes advantages of both inferred and available annotations to search for semantically similar genes. The suite consists of three components: BioAnnotationPredictor for predicting new gene-functions based on singular value decomposition of available annotations, SimilBio for discovering similarities between genes via LSI, and a new web service SemSim in the bio search computing framework.

Adverse drug reaction (ADR) is a common clinical problem and is also one of the major factors leading to failure of new drug development. In "Rapid Assessment of Adverse Drug Reactions by Statistical Solution of Gene Association Network", Yan-Ping Xiang, Ke Liu, Xian-Ying Cheng, Cheng Cheng, Fang Gong, Jian-Bo Pan, and Zhi -Liang Ji proposed a novel naıve Bayesian model for rapid assessment of clinical ADRs with frequency estimation. This model covered 611 US FDA approved drugs, 14,251 genes, and 1,254 distinct ADR terms. It achieved an average detection rate of 99.86 and 99.73 percent in identifying known ADRs in internal test data set and external case analyses, respectively.

In "Quantitative Measurement of Split of the Second Heart Sound (S2)", Shovan Barma, Bo-Wei Chen, Ka Lok Man, and Jhing-Fa Wang proposed a quantitative measurement of split of the second heart sound (S2) based on nonstationary signal decomposition to deal with overlaps and energy modeling of the subcomponents of S2. HVD method is used to decompose the S2 into a number of components. A2s and P2s are localized using smoothed pseudo Wigner-Ville distribution followed by reassignment method. The method measures the split efficiently based on some tests.

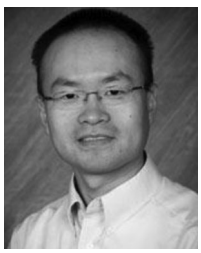
Post-acquisition denoising is an important step for ensuring the quality of any quantitative measurement from MRI. In "An Optimized LMMSE Based Method for 3D MRI Denoising", Hosein M. Golshan and Reza P.R. Hasanzadeh developed a new filtering method based on the linear minimum mean square error (LMMSE) estimation. This method employs the self-similarity property of the MRI data to restore the noise-less signal. It takes into account the structural characteristics of images and the Bayesian mean square error (Bmse) of the estimator to address the denoising problem. Experimental results demonstrated the effectiveness of the proposed method in comparison with related state-of-the-art methods.

Bioconductor provides over 700 software packages for large-scale genomic data analytics based on the R-programming language. However, executing these packages in cloud infrastructure is often challenging. In "RBioCloud: A Light-Weight Framework for Bioconductor and R-based Jobs on the Cloud", Blesson Varghese, Ishan Patel, and Adam Barker designed and developed an open-source light-weight framework called 'RBioCloud' for executing R-scripts using Bioconductor packages in cloud environment. RBioCloud offers a set of simple command-line tools for managing the cloud resources, the data and the execution of the job. The feasibility of RBioCloud is demonstrated using three biological test cases.

The authors would like to thank the authors for their contributions and the reviewers for volunteering their time to review the submissions. They would also like to thank the editor-in-chief, Dr. Ying Xu, for his support of this issue.

Dong Xu
Kun Huang
Jeanette Schmidt
Guest Editors

Dong Xu received the PhD degree from the University of Illinois, Urbana-Champaign, in 1995 and did two years of postdoctoral work at the US National Cancer Institute. He is the James C. Dowell professor and chair of the Computer Science Department, with appointments in the Christopher S. Bond Life Sciences Center and the Informatics Institute at the University of Missouri-Columbia. He was a staff scientist at Oak Ridge National Laboratory until 2003 before joining the University of Missouri. His research includes protein structure prediction, high-throughput biological data analyses, in silico studies of plants, microbes and cancers, and mobile App development for healthcare. He has published more than 240 papers. He received the 2001 R&D 100 Award, 2003 Federal Laboratory Consortium's Award of Excellence in Technology Transfer, and 2010 Outstanding Achievement Award from International Society of Intelligent Biological Medicine. He is an editor-in-chief of the *International Journal of Functional Informatics and Personalised Medicine* and an associate editor-in-chief of the *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.



Kun Huang received the BS degree in biological sciences and the BE degree in computer science from Tsinghua University in 1996, and the MS degrees in physiology, electrical engineering, and mathematics all from the University of Illinois at Urbana-Champaign (UIUC). He then received the PhD degree in electrical and computer engineering also from UIUC in 2004 with a focus on computer vision and machine learning. Currently, he is an associate professor in the Department of Biomedical Informatics, The Ohio State University (OSU). His research interests include bioinformatics, computational biology, bioimage informatics, and machine learning. He has coauthored more than 140 papers.

Jeannette Schmidt received the PhD degree in applied mathematics and computer science from Weizmann Institute of Science in 1986. She was the executive director in the Center for Biomedical Computing at Stanford University and is currently the vice president for informatics at Affymetrix, Inc.