THE past two decades have witnessed rapid technological advances in biological data collection. These advances in biotechnology enabled interrogation of cellular systems at various levels, leading to generation and collection of large-scale biological data (mostly in public databases) at an exponential rate. The explosion of biological data is leading to a paradigm shift in research methods in life sciences from hypothesis-driven to data-driven research. In the last decade, sophisticated algorithms for knowledge discovery and data mining have demonstrated great promise in extracting novel biological information from complex, heterogeneous, and very high-dimensional biological data sets.

The International Workshop on Data Mining in Bioinformatics (BIOKDD), held in conjunction with the ACM Conference on Knowledge Discovery and Data Mining (KDD) for 11 years, has successfully established a tradition of providing a platform for the presentation and discussion of advances in data mining techniques that primarily target biological data. In 2012, the workshop was held in Beijing, China, and BIOKDD continued the tradition of bringing together data mining researchers and life scientists, emphasizing novel problems with various types of biological data. This special section features extended versions of three papers that were presented in BIKODD.

In “Biological Sequence Classification with Multivariate String Kernels,” Pavel P. Kuksa presents a framework for extending string kernel-based approaches in classifying biological sequences. Observing that one-dimensional string representations of biological sequences may not robustly capture the relationships among distantly related sequences, Kuska develops a method that uses multi-dimensional representations of sequences, also taking into account the physical and chemical properties of protein chains. Experiments on three different protein sequence classification tasks show that Kuska’s multivariate kernels provide a 15-20 percent improvement over one-dimensional kernels.

In “Text Categorization of Biomedical Data Sets Using Graph Kernels and a Controlled Vocabulary,” Said Bleik, Meenakshi Mishra, Jun Huan, and Min Song also focus on the development of kernels to represent biomedical data in innovative ways that enable extracting hidden relationships. They use a graph kernel to represent biomedical articles, which enables them to take into account the semantic relationships among biomedical concepts while categorizing biomedical text. Experiments on a rich collection of biomedical articles show that the use of graph-based kernels provides considerable performance improvement over common text-based classifiers.

Finally, in “Chemical Name Extraction Based on Automatic Training Data Generation and Rich Feature Set,” Su Yan, W. Scott Spangler, and Ying Chen tackle the problem of automatically extracting chemical names from text. For this purpose, they use randomly generated text data as training data to enable selection of useful text features. Their results show that the resulting method delivers better or comparable results compared to state-of-the-art methods, but with less human effort. The authors also observe that structural and semantic components of chemical names follow a Zipfian distribution, as in many natural languages.

As guest editors of this special section, we would like to thank the contributing authors, BIKODD program committee, the reviewers who reviewed the journal version of the papers for TCBB, and the TCBB editorial staff for their invaluable contributions.

Tamer Kahveci
Saeed Salem
Mehmet Koyuturk
Guest Editors
Tamer Kahveci received the BS degree in computer engineering and information science from Bilkent University, Turkey, in 1997, and the PhD degree in computer science from the University of California at Santa Barbara in 2004. He is currently an associate professor in the Computer and Information Science and Engineering Department at the University of Florida. He has worked on indexing sequence and protein structure databases, sequence alignment, and computational analysis of biological networks. Dr. Kahveci received the ORAU Ralph E. Powe Jr. Faculty Enhancement Award in 2006, the US National Science Foundation Career Award in 2008, the CSB (Computational Systems Biology) best paper award in 2008, and the ACM-BCB (Bioinformatics and Computational Biology) best student paper award in 2010.

Saeed Salem received the PhD degree in computer science in 2009 from Rensselaer Polytechnic Institute, Troy, New York. He is currently an assistant professor in the Department of Computer Science at North Dakota State University, Fargo. His research interests are in data mining, bioinformatics, and social networks analysis. Dr. Salem is a co-recipient of the PAKDD best paper award in 2009. He serves on the program committees of several data mining and bioinformatics conferences, including ICDM, KDD, CIKM, SDM, and BIBM.

Mehmet Koyutürk received the BS and MS degrees in electrical and electronics engineering and computer engineering in 1998 and 2000, respectively, from Bilkent University, Ankara, Turkey. He received the PhD degree in computer science from Purdue University, West Lafayette, Indiana, in 2006. He is currently an associate professor of electrical engineering and computer science at Case Western Reserve University, Cleveland, Ohio. His research interests include the analysis and integration of high-throughput biological data sets, biological networks, and genomic sequences. He received the US National Science Foundation CAREER Award in 2010 and the best paper award in EVOBIO 2010. He serves as an associate editor for the IEEE/ACM Transactions on Computational Biology and Bioinformatics and on the program committees of several bioinformatics conferences, including RECOMB, ISMB, ACM-BCB, and EVOBIO.