Guest Editorial: Pattern Recognition in Bioinformatics

Elena Marchiori, Alioune Ngom, and Raj Acharya

DEVELOPMENT and application of pattern recognition techniques in the field of bioinformatics is of utmost importance for gaining new insights about phenomena in life sciences through the analysis of biological data. In this special section, three research manuscripts in their significantly extended form were selected from the papers presented at the Eighth IAPR International Conference on Pattern Recognition in Bioinformatics (PRIB 2013), which was held in Nice, France. These papers tackle three core problems in bioinformatics using different pattern recognition techniques.

In "Fast Entropic Profiler. An Information Theoretic Approach for the Discovery of Patterns in Genomes", Comin et al. introduce a fast algorithm for pattern discovery in genomes, named Fast Entropic Profiler (FastEP). This algorithm is based on a local entropy function that captures the importance of a region with respect to the whole genome. FastEP has a linear time and linear space complexity. Additionally, the authors also propose an efficient alternative to data normalization leading to a simpler implementation as well as a faster execution of FastEP. Computational experiments and results show that FastEP is suitable for large genomes and for the discovery of patterns with unbounded length.

In "A Segmentation-based Method to Extract Structural and Evolutionary Features for Protein Fold Recognition", Dehzangi et al. introduce a segmentation-based feature extraction technique providing local evolutionary information embedded within a position specific scoring matrix (PSSM) and structural information embedded within a predicted secondary structure of proteins using SPINE-X. By applying a support vector machine (SVM) on the extracted features, the authors enhance the protein fold prediction accuracy by 7.4 percent over the best results reported in the literature. They report 73.8 percent prediction accuracy for a data set consisting of proteins with less than 25 percent sequence similarity rates and 80.7 percent prediction accuracy for a data set with proteins belonging to 110 folds with less than 40 percent sequence similarity rates. The authors investigate the relation between the number of folds and the number of features being used and show that the number of features should be increased to get better protein fold prediction results when the number of folds is relatively large.

In "Outlier Analysis and Top Scoring Pairs for Integrated Data Analysis and Biomarker Discovery", Ochs et al. investigate pathway deregulation, a key driver of carcinogenesis, with proteins in signaling pathways serving as primary targets for drug development. They introduce a novel approach that identifies pathways of interest by integrating outlier analysis within and across molecular data types with gene set analysis. The authors use the results to seed the top-scoring pair algorithm to identify robust biomarkers associated with pathway deregulation. Application of this methodology to pediatric acute myeloid leukemia (AML) data results in the identification of biomarkers in primary AML tumors. The authors demonstrate the robustness of their approach with an independent primary tumor data set, and show that the identified biomarkers also function well in relapsed pediatric AML tumors.

Elena Marchiori
Alioune Ngom
Raj Acharya
Guest Editors

Elena Marchiori received the MSc degree in mathematics and the PhD degree in computer science from the University of Padua, Italy. After receiving the PhD degree, she was employed in particular at the Centre for Mathematics and Computer Science in Amsterdam and at the Leiden Institute of Advanced Computer Science. Since January 2008, she has been an associate professor at the Radboud University Nijmegen. She is a member of the Machine Learning Group and head of the Section Intelligent Systems of the Institute for Computing and Information Science. She has published more than 100 scientific papers on methods, applications, and tools in computer science and machine learning for biomedical informatics. Her current research interests in bioinformatics include comparative and integrative analysis of biological networks, clustering and feature selection, in particular for computational diagnosis and biomarker detection. She is involved in various multidisciplinary research projects where she collaborates with domain experts from life sciences and medicine.

- *E. Marchiori is with the Department of Computer Science, Faculty of Sciences, Radboud University, Nijmegen, Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands. E-mail: elenam@cs.ru.nl.*
- *A. Ngom is with the School of Computer Science, University of Windsor, 401 Sunset Avenue, N9B 3P4, Windsor, Ontario, Canada. E-mail: angom@uwindsor.ca.*
- *R. Acharya is with the Department of Computer Science and Engineering, College of Engineering, Penn State University, 342B IST Building, University Park, PA 16802. E-mail: acharya@cse.psu.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.2014.2315668

Alioune Ngom is a full professor at the University of Windsor, Ontario, Canada. Prior to joining the University of Windsor, he held the position of an assistant professor in the Department of Mathematics and Computer Science at Lakehead University, Thunder Bay, Ontario, Canada, from 1998 to 2000. During his short stay at Lakehead University, he co-founded Genesis Genomics Inc. (now, Mitomics Inc), in 1999; a biotechnology company specializing in the analysis of the mitochondrial genome and the identification and design of mtDNA biomarkers for the early detection of cancer. His main research interests include but are not limited to computational intelligence and machine learning methods and their applications in computational biology and bioinformatics problems such as microarray analysis, protein analysis, oligonucleotide selection, bioimage analysis, and protein interaction networks and gene regulatory network analysis.

Raj Acharya is a professor and head of the Department of Computer Science and Engineering at Penn State University. He was a research scientist at General Electric (Thomson) CSF Laboratory, Paris, France. His main research focus is in the areas of data mining, bioinformatics/genomic signal processing, and network sciences & engineering. He is an associate editor of the *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. He is the founding chair of IAPR Technical Committee on Pattern Recognition in Bioinformatics. He is on the steering committee of the International Conference on Pattern Recognition in Bioinformatics. He and his work have been profiled in publications such as *Businessweek*, *Mathematics Calendar*, *Drug Discovery*, *Diagnostic Imaging*, *Environnews*, *EurekAlert*, *Science Daily*, and *The Scientist* among others. He is working as part of a multidisciplinary international team on a US NSF-United Nations Digital Government Surveillance project involving Hot Spot Bio-Geo Informatics. He has recruited 18 new tenure-track faculty members. All of the eligible junior faculty members have been awarded US National Science Foundation (NSF) Career grants. Three junior faculty members were awarded the prestigious US NSF PECASE award by President Obama. During his tenure, Penn State CS research expenditure has moved from 64th (2001) in the 8th (2011) in the nation. CS citations ranks 5th in the world. Recently, the CS Department at Penn State was awarded a $35.5 Million US Army Network Sciences CTA, a $48 Million CRA (US Army), and a $10 Million US NSF Expedition award. He is a fellow of the IEEE and AIMBE.