Editorial

Ion Mandoiu*

Computer Science & Engineering Department,
University of Connecticut,
Storrs, CT 06269, USA
Email: ion@engr.uconn.edu
*Corresponding author

Mihai Pop

Computer Science Department,
University of Maryland,
College Park, MD 20742, USA
Email: mpop@umiacs.umd.edu

Sanguthevar Rajasekaran

Computer Science & Engineering Department,
University of Connecticut,
Storrs, CT 06269, USA
Email: rajasek@engr.uconn.edu

John L. Spouge

Computational Biology Branch,
National Center for Biotechnology Information,
Bethesda, MD 20894, USA
Email: spouge@nih.gov

Biographical notes: Ion Mandoiu is an Associate Professor in the Computer Science and Engineering Department at the University of Connecticut, Storrs. He received his PhD in Computer Science from Georgia Institute of Technology. His current research focuses on scalable algorithms for high-throughput sequencing data analysis.

Mihai Pop is an Associate Professor of Computer Science and Interim Director of the Center for Bioinformatics and Computational Biology at the University of Maryland, College Park. He received his PhD in Computer Science from Johns Hopkins University, USA and his current research focuses on genome and metagenome assembly, sequence alignment, and metagenomic association statistics.

Sanguthevar Rajasekaran is the UTC Chair Professor of Computer Science and Engineering and the Director of Booth Engineering Center for Advanced Technology (BECAT) at the University of Connecticut, Storrs. He received his PhD degree in Computer Science from Harvard University. His research

interests include bioinformatics, parallel algorithms, data mining, randomised computing, and combinatorial optimisation. He is a Fellow of the IEEE and AAAS.

John L. Spouge is a Senior Investigator at the National Center for Biotechnology Information. He received his MD from the University of British Columbia, Canada, and his DPhil in Applied Mathematics from Oxford University. His current research focuses on sequence alignment statistics and coalescent theory, applied in particular to virology and DNA barcodes.

This special issue includes a selection of papers presented at the 2nd IEEE International Conference on Computational Advances in Bio and medical Sciences (ICCABS), held in Las Vegas, Nevada on 23–25 February 2012. Computational techniques are revolutionising the way in which research is conducted in science and engineering. Unsurpassed advances have been made in myriads of application domains. This is particularly true in the areas of biology, medicine, and drug discovery. Even though a number of conferences exist today in the general area of bioinformatics, they focus on computational biology to a large extent. ICCABS has the goal of bringing together scientists in all the three areas and hence serving as a platform for bridging the research efforts in these areas.

In 2012 the technical programme of ICCABS included 25 extended abstracts selected by the Programme Committee from a number of 55 submissions received in response to the call for papers. The programme also included 8 invited talks and a poster session, and featured keynote talks by two distinguished speakers. The authors of 11 extended abstracts were invited to submit full versions of their manuscripts and were selected for publication in this special issue following a rigorous review process.

In the following we provide a synopsis of selected papers, which cover a broad range of topics, from efficient algorithms for sequence alignment to phylogenetics and stochastic modeling of biological networks.

- Sequence alignment is a fundamental problem in bioinformatics. In local pairwise sequence alignment, two sequences of lengths m and n are given and the problem is to find their best local alignment. The running time of the Smith–Waterman algorithm for local alignment on a single-core CPU is unacceptably long when m and n are very large. The paper by Li, Ranka, and Sahni develops a single-GPU parallelisation of the Smith-Waterman algorithm, demonstrating speed-up by an order of magnitude over previous GPU algorithms. Furthermore, the memory required is at least one order of magnitude lower than that required by previous GPU implementations.
- In computational biology, finding approximate patterns (‘motifs’) in sequences helps biologists understand the function of the sequences. The paper by Bandyopadhyay et al. examines the NP-hard problem of (l, d)-motif search, also known as the Planted Motif Search, where motifs planted in sequences may have at most d mismatches relative to some ‘ideal motif’ of length l. The authors develop the PMS6 algorithm, which runs twice as fast as an implementation of the fastest known algorithm for exhaustive motif search.

- The paper by Spouge et al. deals with the problem of finding subsequences of unusual composition in biological sequences. The authors present a generalisation of the Ruzzo-Tompa algorithm for finding subsequences of unusual composition. This generalisation enables one to find subsequences with greatest total gapped scores. To illustrate the usefulness of the generalisation, the problem of finding repeats is considered.
- Identification of closely related, ecologically distinct populations of bacteria would have benefits for microbiologists working in many fields, including systematics, epidemiology, and biotechnology. The paper by Francisco et al. tested several algorithms for demarcating 'ecotypes' of bacteria on both simulated bacterial sequences and *Bacillus* strains isolated from Death Valley, which is known to contain multiple bacterial ecotypes. The authors conclude that the Ecotype Simulation algorithm performs significantly better than the other algorithms tested, but that it is presently too slow and will require acceleration to be useful in the routine analysis of environmental DNA samples.
- The use of an outgroup is the most common strategy for rooting phylogenetic trees. The results of the paper by Ackerman et al. concern the effect that an outgroup can have on the topology of phylogentic trees created by different distance-based algorithms. The authors show that for hierarchical algorithms such as UPGMA and a class of bisecting algorithms including bisecting k-means the topology of the ingroup is not affected when adding sufficiently distant outgroups. In contrast, for the widely-used neighbour joining algorithm the authors show that the topology of the ingroup can be affected by an arbitrarily distant outlier even when distances within the ingroup are additive.
- Massive sequencing of genomes requires a corresponding scaling of genomic annotation. The paper by Thrasher et al. describes a framework enabling labs of various sizes to parallelise genomic annotation without forcing them to invest in a particular type of batch system. The authors demonstrate results on *Caenorhabditis japonica* and *Anopheles gambiae* PEST genomes within the Amazon EC2 cloud computing framework, a framework that can run bioinformatics tools on clusters, grids, and clouds, even during early stages of development.
- The paper by Zhang et al. describes a novel machine learning framework integrating several standard techniques in text-mining, feature selection, and network analysis to find genomic features associated with binary microbial traits based on whole genome sequence data. The proposed method is used to identify clusters of orthologs (COGs) and Pfam domain families associated with traits such as sporulation, Gram stain, motility and oxygen requirement. Cross-validation experiments show that the proposed methods select features that yield excellent classification accuracy.
- Non-coding RNA elements in the 3' untranslated regions participate in post-transcriptional regulation of genes, affecting their stability, translation efficiency, and subcellular localisation. The paper by Zhong et al. describes a clustering pipeline for RNA structures, using a similarity measure that compensates for the effect of length. The pipeline uses graphical cliques to cluster the structures, improving performance relative to a traditional hierarchical clustering algorithm. The results are validated with several known families of RNA structures.

- Several non-coding RNAs (ncRNAs) fold into alternate native structures. The computational prediction of an ncRNA's alternate structures can be done with the analysis of the ncRNA's energy landscape. In prior work, Li, Zhong, and Zhang have developed an algorithm using this approach. In this paper they improve the prediction accuracy of their prior algorithm by incorporating structural conservation information.
- Biochemical and biomedical systems can be studied with stochastic models. One of the challenges in this approach lies in discovering values of underlying parameters from experimentally observed facts. In their paper, Hussain et al. present a new parameter discovery algorithm based on Wald's sequential probability ratio test and statistical model checking. As a case study, they perform a massively parallel in silico validation of artificial pancreata.
- One of the difficulties in employing stochastic differential equations (SDEs) to study rare events lies in the limited availability of analytic methods for SDEs. As a result, stochastic simulations are commonly used to estimate the probability of a rare event. But these simulations are costly. Ghosh et al. introduce a new algorithm to quantify the likelihood of rare events in SDE models. They use their algorithm to investigate the likelihood of irregularities in cell size and time between cell divisions.

We would like to thank the Programme Committee members and external reviewers for volunteering their time to review the manuscripts submitted to the conference and the special issue. We would also like to thank the Editor-in-Chief, Professor Yi Pan, for providing us with the opportunity to showcase some of the exciting research presented at ICCABS in the *International Journal of Bioinformatics Research and Applications*. Last but not least we would like to thank all authors; the conference could not continue to thrive without their high-quality contributions.