Foreword

Costas S. Iliopoulos · Alessio Langiu

Published online: 20 May 2017
© Springer International Publishing 2017

Past decades have witnessed a growing interest in applying mathematical theories to deal with the huge collections of data that modern technologies produce at an increasing pace. This branch of Computer Science has been recently named Big Data and it gathers researchers from Bioinformatics, Machine Learning, Information Theory, Graph Theory, Data Base, Math and Algorithms fields.

This special Issue of Mathematics in Computer Science focuses on Algorithms for Big Data and it has meant to collect new research papers on methods, algorithms, techniques, codes, which can provide significant advances in analysing and managing collections of massive and fast changing data, especially data produced by the Next-Generation-Sequencing technology. Most of the papers in this issue are extended versions of selected papers presented at the 2nd International Conference on Algorithms for Big Data (ICABD), held in Palermo, Italy, 7–9 April 2014.

Many algorithms and data structures have been devised to efficiently solve bioinformatics problems concerning searching and modelling genomic data. These sequences are characterised by being massive and high-repetitive collections of nucleotides or amino acid sequences. The common data model used by the paper in this Volume is the string, i.e. a sequence of symbols, and, hence, a relevant part of these papers stem from the stringology research field with a special interest in bioinformatics.

Ben Nsira et al. On-line string matching in highly similar DNA sequences presents a novel on-line exact string matching solution for highly similar sequences, i.e. a collection of sequenced genomes of the same reference specie, as an extension of the classical Morris–Pratt algorithm, which may be of fundamental interest in those cases where building an index is not feasible. *Kim et al. On Representations of Ternary Order Relations in Numeric Strings* extends recent results on order-preserving matching to the ternary order relation. This kind of matching is used when the searched occurrences as not an exact copy of the pattern but something that have the same shape. It was meant for numerical series, such as the stock market, but it has been applied in comparing expression level from microarray and gene chip data. *Karkkainen et al. Engineering a Lightweight External Memory Suffix Array Construction Algorithm*

C. S. Iliopoulos · A. Langiu (✉)
King's College London, London, UK
e-mail: Alessio.Langiu@kcl.ac.uk

C. S. Iliopoulos
e-mail: csi@kcl.ac.uk

A. Langiu
ICAR-CNR, National Research Council, Palermo, Italy
e-mail: Alessio.Langiu@cnr.it

propose a theoretical result on building the suffix array, i.e. a full-text index, on external memory which also has an enormous practical impact in many alignment and comparison bioinformatics tools. *Gagie et al. Compressed Spaced Suffix Arrays* is a paper devoted to a compressed version of the spaced suffix array, which is commonly used in sequence similarity query where spaced seeds are used. *Tischler. Faster Average Case Low Memory Semi-External Construction of the Burrows–Wheeler Transform* paper contains an interesting advancement for the suffix array construction algorithm, which is intimately connected to the Burrows–Wheeler Transform computation algorithm, in the case of partial use of external memory. *Davoodi et al. On Succinct Representations of Binary Trees* investigate the succinctness of binary trees obtained by transformation from ordinal trees, pointing out whose transformation leads to a convenient tree which supports inorder numbering which, in turn, is of fundamental value in efficiently support the range-minimum query. *Gagie et al. Block Graphs in Practice* faces the problem whether the block graphs compression scheme for highly repetitive data collections, such as versioning data and genomic data, is practical, i.e. fast and effective as much as competitor solutions, and shows that such compression scheme, in their own implementation, provides better compression and faster random access than other methods. *Vlachakis et al. Advanced Protein Alignments based on Sequence, Structure and Hydropathy Profiles; The paradigm of the viral Polymerase enzyme* proposes a new bioinformatics tool for protein similarity search which takes into account the secondary structure of proteins more than their amino acid sequences. It is able to recover secondary structures from online repositories as well as compute its own prediction. *Daykin et al. Indeterminate String Factorizations and Degenerate Text Transformations* deals with indeterminate or degenerate strings, i.e. strings where in some positions have a set of letters instead of a single symbol. Since, for space efficiency, a long sequence, such as a genome, is usually factorised into Lyndon factors before building indexes on their factors, such as suffix arrays for pattern matching purpose, authors define Indeterminate Lyndon Words and establish their associated unique string factorization; then they introduce the novel Degenerate Burrows–Wheeler Transform which may apply the indeterminate Lyndon factorization. *Hasan et al. Palindromic Subsequence Automata and Longest Common Palindromic Subsequence* presents a novel weighted finite automata, which is a compact representation of all the palindromic subsequences of a string, and a solution to the Longest Common Palindromic Subsequence problem, which is generally used in predicting the secondary structure of viruses and proteins, especially in discovering binding structures such as the hairpins. *Ferdousa et al. Solving the Minimum Common String Partition Problem with the Help of Ants* presents a solution for the minimum common partition problem of two strings, which has application in Genome Comparison. Whilst this problem is known to be NP-hard, a combinatorial optimization approach has been used which exploits a metaheuristic technique, namely, the MAX–MIN ant system, to solve it.

We thank the managing editor, Prof. Dongming Wang, the editorial office at MCS, and all the anonymous referees which have done an excellent work in supporting the reviewing process, in cooperating to rich the high level of all the eleven papers presented in this Issue.