Algorithms for Computational Biology: Eighth Edition

Carlos Martín-Vide and Miguel A. Vega-Rodríguez

Index Terms—Algorithms, computational biology

This special section of *IEEE/ACM Transactions on Computational Biology and Bioinformatics* presents extended versions of some of the best papers accepted at the Eighth International Conference on Algorithms for Computational Biology, AlCoB 2021, held online due to the COVID-19 pandemic on November 9-11, 2021. The conference was organized by the Department of Computer Science at the University of Montana and the Institute for Research Development, Training and Advice - IRDTA, Brussels/London.

AlCoB 2021 was the eighth event in a series dedicated to promoting and displaying excellent research using string and graph algorithms and combinatorial optimization to deal with problems in biological sequence analysis, genome rearrangement, phylogeny reconstruction, and structure prediction.

Out of 22 submissions to the conference, 12 were accepted (which represents an acceptance rate of 55%). Among them, the authors of 8 papers were invited to submit to this special section. Each submission was reviewed by three experts and, based on their comments, the guest editors decided to accept all 7 papers for this special section (which represents an acceptance rate of about 32% out of the submissions to the conference).

Next, we briefly present the papers included in this special section.

In their paper "Reversal and Indel Distance with Intergenic Region Information", Alexsandro Oliveira Alexandrino, Klairton Lima Brito, Andre Rodrigues Oliveira, Ulisses Dias and Zanoni Dias introduce a new reversal distance variation that considers both gene order and intergenic region sizes for genomes with distinct gene content. Previously in the literature, the reversal distance problem incorporated intergenic regions only for genomes with the same set of genes. This paper advances the state-of-the-art by dealing with a more general model. A structure called Labeled Intergenic Breakpoint Graph is introduced, which is an adaptation of the well-known Breakpoint Graph structure. Using the Labeled Intergenic Breakpoint Graph, the authors show a lower bound for the distance and a 2.5-approximation algorithm. Also, they present experimental results on simulated data to evaluate the performance of these algorithms and experimental results on real data. These experiments on real data are used to compare distinct models that use reversals and to create a phylogenetic tree for the Cyanorak 2.1 dataset.

Klairton Lima Brito, Alexsandro Oliveira Alexandrino, Andre Rodrigues Oliveira, Ulisses Dias and Zanoni Dias, in their paper "Genome Rearrangement Distance with a Flexible Intergenic Regions Aspect", introduce a generalization of the genome rearrangement problems that consider both the gene order and the size of the intergenic regions. This generalized version adds more flexibility to the models by allowing the use of a range of values for the size of each intergenic region in the target genome rather than a single value. The authors investigate the case where the orientation of the genes is unknown and use a model composed exclusively of transposition events, which exchange two adjacent parts from the genome. Besides, they consider the case where the orientation of the genes is known and employ two models: one allows only reversals, which is an event that inverts a segment from the genome that also flips the orientation of the affected genes, and the other allows both reversals and transpositions. The authors show an NP-hardness proof and present an algorithm with a constant approximation factor for all three problems. Finally, they design simulated datasets of genomes for each problem and show practical results obtained by using the proposed algorithms.

Cherry-picking operations performed on phylogenetic networks, in an orderly way, as a cherry-picking sequence carry some information about the topology of the network. Because of this, ordered cherry-picking sequences between networks can be compared in order to determine isomorphism of the networks. In the paper "Defining Phylogenetic Network Distances Using Cherry Operations", Kaari Landry, Aivee Teodocio, Manuel Lafond and Olivier Tremblay-Savard extrapolate this use to the general comparison of differences between two networks, and define four novel phylogenetic network distances based on cherry-picking sequences. Three of the four distances are shown to be equal despite their different formulations. They show that computing these three distances is NP-hard, even when restricted to a tree and network. It is further shown that the fourth distance is also NP-hard to calculate. A diameter on the three equal distances is provided by showing an upper bound relative to the size of the input networks. The

- *Carlos Martín-Vide is with the Research Group on Mathematical Linguistics, Rovira i Virgili University, 43002 Tarragona, Spain. E-mail: carlos.martin@urv.cat.*
- *Miguel A. Vega-Rodríguez is with ARCO Research Group, University of Extremadura, 10003 Cáceres, Spain. E-mail: mavega@unex.es.*

Manuscript date of current version 5 June 2023.
(Corresponding author: Miguel A. Vega-Rodríguez.)
Digital Object Identifier no. 10.1109/TCBB.2022.3218808

authors show that the three equal distances can be computed in quadratic time on two trees and provide a dynamic programming algorithm that does so.

The paper "Sparse Triangular Decomposition for Computing Equilibria of Biological Dynamic Systems Based on Chordal Graphs", by Chenqi Mou and Wenwen Ju, aims at an efficient computation of equilibria of biological dynamic systems. This is achieved via applying a sparse triangular decomposition based on chordal graphs to exploit the inherent sparsity of such systems. In order to handle parametric biological dynamic systems, an extended theory of block chordal graphs is established, based on which a new algorithm of sparse triangular decomposition for parametric systems is proposed. The extensive experiments carried out with the methods presented by the authors confirm their computational efficiency.

Best match graphs (BMG) are a key intermediate in graph-based orthology detection for families of homologous genes. They are known to contain a large amount of information on the gene tree. In their paper "Best Match Graphs with Binary Trees", David Schaller, Manuela Geiss, Marc Hellmuth and Peter F. Stadler describe a near-cubic algorithm that determines whether a BMG is binary-explainable, i.e., whether it can be explained by a fully resolved gene tree and, if so, it constructs such a tree. Moreover, it is shown that all such binary trees explaining a given BMG are refinements of the uniquely defined binary-refinable tree. This tree is in general a substantial refinement of the least resolved tree of a BMG, which is also uniquely defined. The authors show also that the problem of editing an arbitrary vertex-colored graph to a binary-explainable BMG with a minimal number of edge insertions and deletions is NP-complete. Finally, an integer linear program formulation for this task is provided.

Yin Yao and Martin Frith, in their paper "Improved DNA-versus-Protein Homology Search for Protein Fossils", develop a sensitive method to detect ancient and highly-degraded protein fossils in DNA sequences. Protein fossils are DNA segments that are descended from protein-coding DNA, but are no longer protein-coding. They can tell us about past evolution of genomes, transposable elements, and viruses (paleovirology) from ancient viral insertions into host genomes. The authors work out a statistical method to find regions of decayed homology between DNA sequences and a database of protein sequences. This method found a diverse ecosystem of transposable elements that inhabited our aquatic ancestors in the Paleozoic Era.

A classical formulation for multiple sequence alignment called the "maximum weight trace" (MWT) problem was proposed in 1993 by John Kececioglu. Paul Zaharias, Vladimir Smirnov and Tandy Warnow, in their paper "Large-Scale Multiple Sequence Alignment and the Maximum Weight Trace Alignment Merging Problem", show how the MWT problem can be powerfully extended to the problem of merging a set of disjoint alignments. Both MWT and its extension to alignment merging (MWT-AM) are NP-hard optimization problems, and so heuristics are needed. Over the last 15 years, alignment merging has become a basic algorithmic step in designing methods for large-scale multiple sequence alignment using divide-and-conquer. In these methods, a dataset is divided into disjoint subsets, alignments are computed on the disjoint subsets, and then they are merged together. While several well established methods use this kind of approach, the most recent of these methods, MAGUS, has the best alignment accuracy and can scale to datasets with more than 100000 sequences. In this paper, the authors explore the Graph Clustering Merger technique used in MAGUS for merging disjoint alignments, and show that it is an excellent heuristic for the MWT-AM problem. This paper explains why MAGUS works so well in practice, and ties this performance to a new version of a classical problem in bioinformatics.

We thank you the authors for their contributions, the reviewers for their valuable work, and the editorial team of the journal for their professional support and collaboration.

Carlos Martín-Vide is full professor with Rovira i Virgili University, Tarragona. His research interests include automata and language theory, molecular computing, theoretical computer science and mathematical and computational linguistics. He is (co)author of more than 300 papers. He has been involved in the definition, operation and monitoring of several European funding initiatives in support of fundamental research in mathematics and computer science.

Miguel A. Vega-Rodríguez received the PhD degree in computer engineering from the University of Extremadura, Spain, in 2003. He is a full professor (Catedrático de Universidad) of computer architecture with the Department of Computer and Communications Technologies, University of Extremadura. He has authored or co-authored more than 740 publications including journal papers (more than 180 JCR-indexed journal papers), book chapters, and peer-reviewed conference proceedings, for which he got several awards (such as best paper awards). He has edited more than 20 special issues of JCR-indexed journals. His main research interests include parallel and distributed computing, multiobjective optimization, evolutionary and bio-inspired computation, bioinformatics, and reconfigurable and embedded computing.