algorithms

MDPI

Editorial

Special Issue: Algorithms in Bioinformatics

Christina Boucher

Department of Computer and Information Science and Engineering, Herbert Wertheim College of Engineering, University of Florida, Gainesville, FL 32611-0570, USA; christinaboucher@ufl.edu

In the past decade, there has been an effort to sequence and compare a large number of individual genomes of a given species, resulting in a large number of (reference) genomes of various species being made publicly available. For example, there is now data from the 1000 Genome Project, the 100K Genome Project, the 1001 Arabidopsis Genomes project, the Rice Genome Annotation Project, and the Bird 10,000 Genomes (B10K) Project are now public. Accompanying this endeavor is the first complete, gapless human reference genome, referred to as Telomere-to-Telomere (T2T), and genome in a bottle (GIAB) data. For these and other projects, the data generated are not limited to just short read data, as many different data types can be now generated in a high-throughput manner, including long reads, ultra-long reads, optical maps, mass spectrometry data, etc.

Given all the new and existing datasets, it is important to study how they can be studied alone or in combination. Datasets are now large enough that they surpass our ability to analyze them simultaneously. For example, standard read alignment methods can comfortably align sequence reads to less than 10 human genomes, but thousands publicly exist. Therefore, new (pangenome) methods are now emerging, including the VG, Giraffe, and Moni methods. Yet, this is only one area: read alignment, which aligns to a population of genomes. Fundamental to these pangenome approaches are novel data compression schemes. There are countless other areas that need to be advanced via algorithm design and analysis.

Brejová and Královič [1] tackle a problem in reconciling phylogenetic trees. Phylogenetic trees act as a high-level model of the evolutionary history of a set of species. The goal of reconciliation is to map nodes of a gene tree to the corresponding points in a species tree that represent the same points in a species' evolutionary history. Here, the authors focus on a specific variant of reconciliation that aims to compute all reconciliations of two unrooted trees and present a linear-time algorithm for solving this variant, which is a significant improvement compared to previous methods.

Abedin et al. [2] present a survey on the problem of finding all shortest unique substring queries in a given input string. Let S be a string of length n and $S[i,j]$ be the substring starting at position i and ending at position j. Then, $S[i,j]$ is called a repeat if it occurs more than once in S; otherwise, it is considered unique. The goal is then to find all unique substrings in S. The problem is both theoretically interesting since it intersects both data structures and algorithms and is also practically interesting since it can be applied in a variety of different scenarios, not only in bioinformatics but also in information retrieval and data mining.

In this same Special Issue, Bannai et al. [3] extend recent results regarding finding shortest unique substrings to obtain new time-space tradeoffs for this problem. In addition, they give an efficient algorithm for the generalization of finding k-mismatch shortest unique substrings.

The other contributions in this Special Issue contribute to the immensely growing field of machine learning algorithms in bioinformatics. Chen et al. [4] implement a convolution neural network (CNN) accelerator based on a field-programmable gate array (FPGA). CNNs have numerous applications in bioinformatics but require extensive computational resources, making it challenging to deploy a CNN model on a low-power device. The contribution of Chen helps ameliorate this limitation.

check for updates

Citation: Boucher, C. Special Issue: Algorithms in Bioinformatics. *Algorithms* **2023**, *16*, 21. https://doi.org/10.3390/a16010021

Received: 15 December 2022
Accepted: 29 December 2022
Published: 30 December 2022

In [5], the author studies how to identify genes that play a role in the growth of hepatocellular carcinoma (HCC) in Hepatitis C virus (HCV) patients using machine learning techniques. This work could potentially have significant impact on the development of new, more accurate antiviral treatment. HCV is described as one of the most dangerous viruses worldwide and is the foremost cause of HCC. Hence, this research could be impactful to public health.

Lastly, Cumbo et al. [6] give the details for a new, in-memory, cognitive-based, hyperdimensional supervised machine learning algorithm for the classification of tumor verses non-tumor samples based on DNA methylation data. Their results contribute to the growing evidence connecting tumor growth to DNA methylation, a genetic modification that regulates the functions of genes.

In summary, this Special Issue demonstrates the breadth of algorithms in bioinformatics—from hardware optimization of machine learning algorithms, to the application of machine learning, to finding disease-causing genetic variants, to stringology problems, and, lastly, to problems in phylogenetics. We hope the community continues to grown in its diversity and breadth.

Acknowledgments: Christina Boucher would like to thank all researchers who submitted their work, the invited expert reviewers.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Brejová, B.; Královič, R. A Linear-Time Algorithm for the Isometric Reconciliation of Unrooted Trees. *Algorithms* **2020**, *13*, 225. [CrossRef]
2. Abedin, P.; Külekci, M.O.; Thankachan, S.V. A survey on shortest unique substring queries. *Algorithms* **2020**, *13*, 224. [CrossRef]
3. Bannai, H.; Gagie, T.; Hoppenworth, G.; Puglisi, S.J.; Russo, L.M. More Time-Space Tradeoffs for Finding a Shortest Unique Substring. *Algorithms* **2020**, *13*, 234. [CrossRef]
4. Chen, C.; Li, Z.; Zhang, Y.; Zhang, S.; Hou, J.; Zhang, H. Low-Power FPGA Implementation of Convolution Neural Network Accelerator for Pulse Waveform Classification. *Algorithms* **2020**, *13*, 213. [CrossRef]
5. Abdel Samee, N.M. Classical and deep learning paradigms for detection and validation of key genes of risky outcomes of HCV. *Algorithms* **2020**, *13*, 73. [CrossRef]
6. Cumbo, F.; Cappelli, E.; Weitschek, E. A brain-inspired hyperdimensional computing approach for classifying massive dna methylation data of cancer. *Algorithms* **2020**, *13*, 233. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.