Low-Dimensional Representation of Genomic Sequences

Richard C. Tillquist
richard.tillquist@colorado.edu
University of Colorado
Boulder, Colorado

Manuel E. Lladser
manuel.lladser@colorado.edu
University of Colorado
Boulder, Colorado

ABSTRACT

The symbolic nature of biological sequence data greatly complicates its analysis. As many of the most powerful, widely used analysis techniques work exclusively in the context of Euclidean spaces, methods of embedding DNA, RNA, and protein sequences in such spaces are of critical importance. Here we examine the utility of multilateration as the foundation for representing biological sequences in Euclidean space. With respect to discrete metric spaces, this technique is closely related to the concept of metric dimension from graph theory. The goal is to discover a subset of vertices of minimum size such that all vertices in the graph may be uniquely identified based on distances to the vertices in this set. Multilateration is analogous to trilateration, the process of identifying points in the plane using distances to three non-colinear points. Interpreting the space of all k -mers as a Hamming graph, we are able to find such sets efficiently. Resulting sequence representations tend to be more compact than traditional binary or k -mer count vectors and, unlike Multidimensional Scaling (MDS) and Node2Vec, they apply over all k -mers and do not need to be recomputed when new data is encountered. To test the efficacy and practicality of multilateration we classify DNA 20-mers centered at intron-exon boundaries in the *Drosophila melanogaster* genome using features derived from binary and k -mer count vectors as well as MDS, Node2Vec, and multilateration. The performance of multilateration-based features is competitive with other techniques and allows long genomic sequences to be embedded efficiently. This highlight showcases the key findings in “Low-dimensional representation of genomic sequences” [J Math Biol. 2019 Jul;79(1):1-29].

KEYWORDS

feature extraction, graph embeddings, Hamming graph, k -mers, metric dimension, multilateration, resolving set, symbolic data science

ACM Reference Format:

Richard C. Tillquist and Manuel E. Lladser. 2019. Low-Dimensional Representation of Genomic Sequences. In *10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (ACM-BCB '19)*, September 7–10, 2019, Niagara Falls, NY, USA. ACM, New York, NY, USA, 1 page. <https://doi.org/10.1145/3307339.3342620>

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

ACM-BCB '19, September 7–10, 2019, Niagara Falls, NY, USA

© 2019 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-6666-3/19/09.

<https://doi.org/10.1145/3307339.3342620>

ACKNOWLEDGMENTS

This research was partially funded by the NSF IGERT Grant 1144807, and NSF IIS Grant 1836914. The authors acknowledge the BioFrontiers Computing Core at the University of Colorado-Boulder for providing High-Performance Computing resources (funded by National Institutes of Health 1S10OD012300), supported by BioFrontiers IT group.