Low-Dimensional Representation of Biological Sequence Data

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

ABSTRACT

Systems of interest in bioinformatics and computational biology tend to be large, complex, interdependent, and stochastic. As our ability to collect sequence data at finer resolutions improves, we can better understand and predict system behavior under different conditions. Machine learning algorithms are a powerful set of tools designed to help with this understanding. However, many of the most effective of these algorithms are not immediately amenable to application on symbolic data. It is often necessary to map biological symbols to real vectors before performing analysis or prediction using sequence data. This tutorial will cover several techniques for embedding sequence data. Common methods utilizing k -mer count and binary vector representations will be addressed along with state of the art methods based on neural networks, like BioVec, and graph embeddings, like Node2Vec and multilateration. This tutorial is in collaboration with M. Lladser. Slides, datasets, and code from the tutorial will be made freely available for future use on GitHub.

KEYWORDS

sequence data, neural networks, multilateration, graph embeddings, symbolic data science

ACM Reference Format:

Richard C. Tillquist. 2019. Low-Dimensional Representation of Biological Sequence Data. 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/10.1145/3307339.3343174>

1 SUMMARY

High-throughput sequencing technologies allow cellular reactions to be observed and quantified in different environments at the level of DNA, RNA, and proteins. Datasets produced using these tools are symbolic, consisting of strings of nucleic acids when DNA or RNA are being considered and of amino acids when proteins are the objects of interest.

Because of the volume of data these assays are capable of generating, machine learning techniques are often well suited to addressing important biological questions. Unfortunately, many of the most frequently used machine learning algorithms, including k -nearest neighbors (KNN) and support vector machines (SVM), require that

the data being processed be represented as numeric feature vectors residing in some real space $\mathbb{R}^d$. Thus, dealing with sequence information directly can be difficult. Oftentimes, numeric values related to e.g. biochemical properties of sequences of nucleotides and amino acids, like thermodynamic stability, structural flexibility, hydrophobicity, charge, and polarity, are used as features [2, 3]. Unfortunately, such features do not always capture everything of importance and it is frequently useful to include a more direct representation of sequence data as a source of information for learning [4, 6]. Techniques like KNN and SVM can manipulate sequence information more directly through an embedding of the data in real space or through the use of a kernel.

Here we will discuss several embedding methods applicable to biological sequence data. Common highly dimensional representation methods like those based on k -mer count vectors and binary representations will be covered briefly, while a greater focus will be placed on newer techniques using neural networks [1] and general graph embedding techniques [5, 7]. The overall goal of the tutorial is to demonstrate the variety of available embedding tools and to highlight the fact that the choice of embedding matters. Understanding how an embedding works can be important in interpreting and understanding downstream results.

ACKNOWLEDGMENTS

The materials for this tutorial have been partially funded by the NSF ISS BIGDATA grant No. 1836914.

REFERENCES

- [1] Ehsaneddin Asgari and Mohammad RK Mofrad. 2015. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLOS ONE* 10, 11 (2015), e0141287.
- [2] Joel R Bock and David A Gough. 2001. Predicting protein–protein interactions from primary structure. *Bioinformatics* 17, 5 (2001), 455–460.
- [3] CZ Cai, LY Han, Zhi Liang Ji, X Chen, and Yu Zong Chen. 2003. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Research* 31, 13 (2003), 3692–3697.
- [4] Jianlin Cheng, Arlo Randall, and Pierre Baldi. 2006. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins: Structure, Function, and Bioinformatics* 62, 4 (2006), 1125–1132.
- [5] Aditya Grover and Jure Leskovec. 2016. Node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 855–864.
- [6] Vanessa Isabell Jurtz, Alexander Rosenberg Johansen, Morten Nielsen, Jose Juan Almagro Armenteros, Henrik Nielsen, Casper Kaae Sønderby, Ole Winther, and Søren Kaae Sønderby. 2017. An introduction to deep learning on biological sequence data: examples and solutions. *Bioinformatics* 33, 22 (2017), 3685–3690.
- [7] Richard C. Tillquist and Manuel E. Lladser. 2019. Low-dimensional representation of genomic sequences. *Journal of Mathematical Biology* 79, 1 (2019), 1–29. <https://doi.org/10.1007/s00285-019-01348-1>

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/10.1145/3307339.3343174>