Data-driven Variable-length Segmentation of Biological Sequences: Applications in Metagenomics and Proteomics

Ehsaneddin Asgari[1 2] **Philipp C. Münch**[2] **Till R. Lesker**[2] **Alice C. McHardy**[* 2] **Mohammad R.K. Mofrad**[* 1 3]

Abstract

In this paper, we propose a data-driven segmentation approach for dividing biological sequences into frequent variable-length sub-sequences inspired by Byte-Pair Encoding (BPE) text compression algorithm. In contrast to the recent use of BPE in natural language processing for vocabulary size reduction, we used this idea to increase the size of symbols in the biological sequences replacing the k-mer representations. We investigate the use of this segmentation in 16S rRNA gene processing (Asgari et al., 2019b) and show that this representation can improve the performance of biomarker detection in 16S rRNA processing. Furthermore, we extend the BPE to perform a probabilistic segmentation of protein sequences and show that it can be used for the task of motif discovery and protein sequence embedding (Asgari et al., 2019a).

1. Introduction

Bioinformatics and natural language processing (NLP) are research areas that have greatly benefited from each other since their beginnings and there have been always methodological exchanges between them. Levenshtein distance (Levenshtein, 1966) and Smith–Waterman (Waterman et al., 1976) algorithms for calculating string or sequence distances, the use of formal languages for expressing biological sequences (Searls, 1993; 2002), training language model-based embeddings for biological sequences (Asgari & Mofrad, 2015), and using state-of-the-art neural named entity recognition architecture (Lample et al., 2016) for secondary structure prediction (Johansen et al., 2017; Asgari et al., 2019c) are some instances of such influences. Similar to the complex syntax and semantic structures of natural languages, certain biophysical and biochemical grammars dictate the formation of biological sequences. This assumption has motivated a line of research in bioinformatics to develop and adopt language processing methods to gain a deeper understanding of how functions and information are encoded within biological sequences (Yandell & Majoros, 2002; Searls, 2002; Asgari & Mofrad, 2015; Asgari, 2019). However, one of the apparent differences between biological sequences and many natural languages is that biological sequences (DNA, RNA, and proteins) often do not contain clear segmentation boundaries, unlike the existence of tokenizable words in many natural languages (Adel et al., 2017). This uncertainty in the segmentation of sequences has made overlapping k-mers one of the most popular representations in machine learning for all areas of bioinformatics research, including proteomics (Grabherr et al., 2011; Asgari & Mofrad, 2015), genomics (Jolma et al., 2013; Alipanahi et al., 2015), epigenomics (Awazu, 2016; Giancarlo et al., 2015), and metagenomics (Wood & Salzberg, 2014; Asgari et al., 2018). However, it is unrealistic to assume that fixed-length k-mers are units of biological sequences and that more meaningful units need to be introduced. This means that although choosing a fixed k value for sequence k-mers simplifies the problem of segmentation, it is an unrealistic assumption to assume that all important part of the sequences have the same length and we need to relax this assumption.

Although in some sequence-labeling tasks (e.g. secondary structure prediction or binding site prediction) sequences are implicitly divided into variable-length segments as the final output, methods to segment sequences into variable-length meaningful units as inputs of downstream machine learning tasks are needed. Here, we propose a segmentation approach for dividing biological sequences into frequent variable-length sub-sequences inspired by byte pair encoding (BPE) algorithm, which is a text compression algorithm introduced in 1994 (Gage, 1994) that has been also used for compressed pattern matching in genomics (Chen et al., 2004) as well. Recently, BPE became a popular word segmentation method in machine translation in NLP for vocabulary size reduction, which also allows for open-vocabulary neural machine translation (Sennrich et al., 2016). In contrast to the use of BPE in NLP for vocabulary size reduction, we used this idea to increase the size of symbols from 4 nucleotides (in DNA and RNA sequences) or 20 amino

[*] Equal contribution [1] Department of Bioengineering, University of California, Berkeley, CA 94720, USA [2] Computational Biology of Infection Research, Helmholtz Centre for Infection Research, BS 38124, Germany [3] Molecular Biophysics and Integrated Bioimaging, Lawrence Berkeley National Lab, Berkeley, CA 94720, USA. Correspondence to: Ehsaneddin Asgari <asgari@berkeley.edu>, Alice C. McHardy <Alice.McHardy@helmholtz-hzi.de>, Mohammad R.K. Mofrad <mofrad@berkeley.edu>.

Presented at the ICML 2020 Workshop on Computational Biology (WCB). Copyright 2020 by the author(s).

acids (in protein sequences) or to a large set of variable-length frequent sub-sequences, which are potentially meaningful in bioinformatics tasks. In the present article, we introduce applications of this data-driven segmentation (i) in metagenomics for an accurate biomarker detection from 16S rRNA gene sequences, and (ii) in proteomics for discriminative motif discovery and variable-length embedding of protein sequences.

2. Nucleotide-pair encoding (NPE) for 16S rRNA gene processing

Identifying distinctive taxa for microbiome-related diseases is considered key to the establishment of diagnosis and therapy options in precision medicine and imposes high demands on the accuracy of microbiome analysis techniques. We propose an alignment- and reference- free subsequence based 16S rRNA data analysis, as a new paradigm for microbiome phenotype and biomarker detection. Our method, called DiTaxa, substitutes standard OTU-clustering by segmenting 16S rRNA reads into the most frequent variable-length subsequences called Nucleotide-pair encoding (NPE). We compared the performance of DiTaxa to the state-of-the-art methods in phenotype and biomarker detection, using human-associated 16S rRNA samples for periodontal disease, rheumatoid arthritis, and inflammatory bowel diseases, as well as a synthetic benchmark dataset. DiTaxa performed competitively to the k-mer based state-of-the-art approach in phenotype prediction while outperforming the OTU-based state-of-the-art approach in finding biomarkers in both resolution and coverage evaluated over known links from literature and synthetic benchmark datasets (Asgari et al., 2019b). The overview of DiTaxa pipeline and a sample result on periodontal disease is shown in Figure 1.

3. Peptide-pair encoding (PPE) for protein sequence processing

We present peptide-pair encoding (PPE), a general-purpose probabilistic segmentation of protein sequences into commonly occurring variable-length sub-sequences. We modify the BPE text compression algorithm by adding a sampling framework allowing for multiple ways of segmenting a sequence. PPE segmentation steps can be learned over a large set of protein sequences (Swiss-Prot) or even a domain-specific dataset and then applied to a set of unseen sequences. This representation can be widely used as the input to any downstream machine learning tasks in protein bioinformatics. In particular, here, we introduce this representation through protein motif discovery and protein sequence embedding. (i) DiMotif: we present DiMotif as an alignment-free discriminative motif discovery method and evaluate the method for finding protein motifs in three different settings: (1) comparison of DiMotif with two existing approaches on 20 distinct motif discovery problems which are experimentally verified, (2) classification-based approach for the motifs extracted for integrins, integrin-binding proteins, and biofilm formation, and (3) in sequence pattern searching for nuclear localization signal. The DiMotif, in general, obtained high recall scores, while having a comparable F1 score with other methods in the discovery of experimentally verified motifs. Having high recall suggests that the DiMotif can be used for short-list creation for further experimental investigations on motifs. In the classification-based evaluation, the extracted motifs could reliably detect the integrins, integrin-binding, and biofilm formation-related proteins on a reserved set of sequences with high F1 scores. (ii) ProtVecX: we extend k-mer based protein vector (ProtVec) embedding to variable-length protein embedding using PPE sub-sequences. We show that the new method of embedding can marginally outperform ProtVec in enzyme prediction as well as toxin prediction tasks. In addition, we conclude that the embeddings are beneficial in protein classification tasks when they are combined with raw k-mer frequency features (Asgari et al., 2019a).

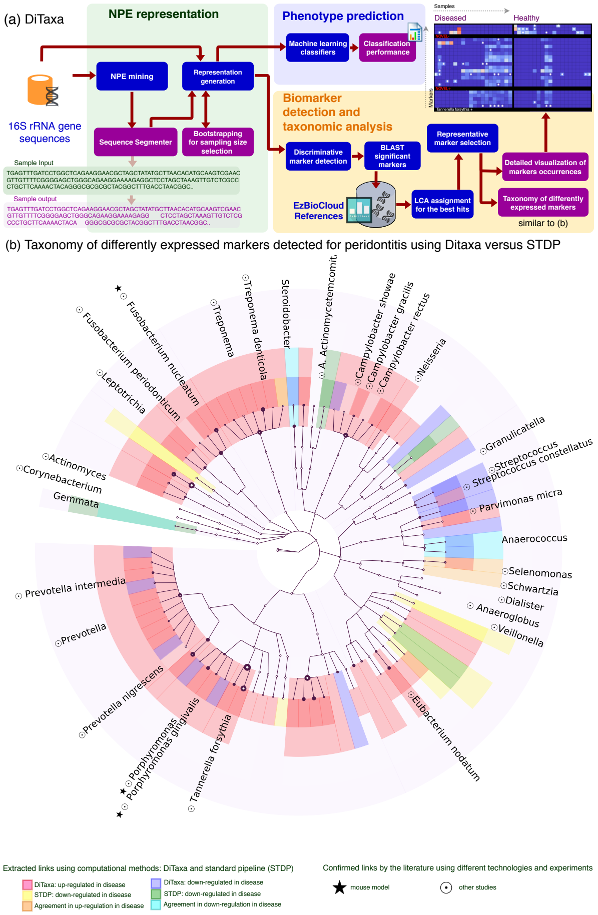
(a) DiTaxa
NPE representation
16S rRNA gene sequences
NPE mining
Representation generation
Sequence Segmenter
Bootstrapping for sampling size selection
Sample Input
TGAGTTTGATCCTGGCTCAGAAGGAACGCTAGCTATATGCTTAACACATGCAAGTCGAAC
GTTGTTTTCGGGGAGCTGGGCAGAAGGAAAAGAGGCTCCTAGCTAAAGTTGTCTCGCC
CTGCTTCAAAACTACAGGGCGCGCGCTACGGCTTTGACCTAACGGC..
Sample output
TGAGTTTGATCCTGGCTCAGAAGGAACGCTAGCTATATGCTTAACACATGCAAGTCGAAC
GTTGTTTTCGGGGAGCTGGGCAGAAGGAAAAGAGG CTCCTAGCTAAAGTTGTCTCG
CCCTGCTTCAAAACTACA GGGCGCGCGCTACGGCTTTGACCTAACGGC..
Phenotype prediction
Machine learning classifiers
Classification performance
Samples
Diseased
Healthy
Markers
NOVEL-
NOVEL+
Tannerella forsythia +
Biomarker detection and taxonomic analysis
Discriminative marker detection
BLAST significant markers
EzBioCloud References
LCA assignment for the best hits
Representative marker selection
Detailed visualization of markers occurrences
Taxonomy of differently expressed markers
similar to (b)
(b) Taxonomy of differently expressed markers detected for peridontitis using Ditaxa versus STDP
★ ⊙ Fusobacterium nucleatum
⊙ Fusobacterium periodonticum
⊙ Treponema
⊙ Treponema denticola
Steroidobacter
⊙ A. Actinomycetemcomit.
⊙ Campylobacter showae
⊙ Campylobacter gracilis
⊙ Campylobacter rectus
⊙ Neisseria
⊙ Leptotrichia
⊙ Granulicatella
⊙ Streptococcus
⊙ Streptococcus constellatus
⊙ Actinomyces
⊙ Corynebacterium
Gemmata
⊙ Parvimonas micra
Anaerococcus
⊙ Selenomonas
⊙ Schwartzia
⊙ Dialister
⊙ Anaeroglobus
⊙ Veillonella
⊙ Prevotella intermedia
⊙ Prevotella
⊙ Prevotella nigrescens
★ ⊙ Porphyromonas
★ ⊙ Porphyromonas gingivalis
⊙ Tannerella forsythia
⊙ Eubacterium nodatum
Extracted links using computational methods: DiTaxa and standard pipeline (STDP)
🟥 DiTaxa: up-regulated in disease
🟨 STDP: down-regulated in disease
🟧 Agreement in up-regulation in disease
🟦 DiTaxa: down-regulated in disease
🟩 STDP: down-regulated in disease
🟦 Agreement in down-regulation in disease
Confirmed links by the literature using different technologies and experiments
★ mouse model
⊙ other studies

Figure 1. (a) Computational workflow of DiTaxa, DiTaxa has three main components: (i) NPE representation, (ii) Phenotype prediction, (ii) Biomarker detection and taxonomic analysis. The purple boxes denote the outputs of the approach. (b) Sample output: Taxonomy of differently expressed markers for samples from patients with periodontal disease versus healthy subjects: comparison of DiTaxa and standard pipeline (STDP).

References

Adel, H., Asgari, E., and Schütze, H. Overview of character-based models for natural language processing. In *International Conference on Computational Linguistics and Intelligent Text Processing*, pp. 3–16. Springer, 2017.

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nat. Biotechnol.*, 33(8):831–838, 2015.

Asgari, E. *Life Language Processing: Deep Learning-based Language-agnostic Processing of Proteomics, Genomics/Metagenomics, and Human Languages*. PhD thesis, UC Berkeley, 2019.

Asgari, E. and Mofrad, M. R. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PloS One*, 10(11):e0141287, 2015.

Asgari, E., Garakani, K., McHardy, A. C., and Mofrad, M. R. K. Micropheno: predicting environments and host phenotypes from 16s rrna gene sequencing using a k-mer based representation of shallow sub-samples. *Bioinformatics*, 34(13):i32–i42, 2018. doi: 10.1093/bioinformatics/bty296.

Asgari, E., McHardy, A. C., and Mofrad, M. R. Probabilistic variable-length segmentation of protein sequences for discriminative motif discovery (dimotif) and sequence embedding (protvecx). *Scientific reports*, 9(1):3577, 2019a.

Asgari, E., Münch, P. C., Lesker, T. R., McHardy, A. C., and Mofrad, M. R. K. DiTaxa: nucleotide-pair encoding of 16S rRNA for host phenotype and biomarker detection. *Bioinformatics*, 11 2019b. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty954. URL https://doi.org/10.1093/bioinformatics/bty954.

Asgari, E., Poerner, N., McHardy, A., and Mofrad, M. Deepprime2sec: Deep learning for protein secondary structure prediction from the primary sequences. *bioRxiv*, pp. 705426, 2019c.

Awazu, A. Prediction of nucleosome positioning by the incorporation of frequencies and distributions of three different nucleotide segment lengths into a general pseudo k-tuple nucleotide composition. *Bioinformatics*, 33(1):42–48, 2016.

Chen, L., Lu, S., and Ram, J. Compressed pattern matching in dna sequences. In *Computational Systems Bioinformatics Conference, 2004. CSB 2004. Proceedings. 2004 IEEE*, pp. 62–68. IEEE, 2004.

Gage, P. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38, 1994.

Giancarlo, R., Rombo, S. E., and Utro, F. Epigenomic k-mer dictionaries: shedding light on how sequence composition influences in vivo nucleosome positioning. *Bioinformatics*, 31 (18):2939–2946, 2015.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. Full-length transcriptome assembly from rna-seq data without a reference genome. *Nat. Biotechnol.*, 29(7):644–652, 2011.

Johansen, A. R., Sønderby, C. K., Sønderby, S. K., and Winther, O. Deep recurrent conditional random field network for protein secondary prediction. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 73–78. ACM, 2017.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., et al. Dna-binding specificities of human transcription factors. *Cell*, 152(1-2):327–339, 2013.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. Neural architectures for named entity recognition. In *Conference of the North American Chapter of the Association for Computational Linguistics / Human Language Technologies*, 2016.

Levenshtein, V. I. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet Physics Doklady*, volume 10, pp. 707–710, 1966.

Searls, D. B. The computational linguistics of biological sequences. *Artificial intelligence and molecular biology*, 2: 47–120, 1993.

Searls, D. B. The language of genes. *Nature*, 420(6912):211, 2002.

Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1715–1725. Association for Computational Linguistics, 2016. doi: 10.18653/v1/P16-1162. URL http://www.aclweb.org/anthology/P16-1162.

Waterman, M. S., Smith, T. F., and Beyer, W. A. Some biological sequence metrics. *Adv. Math. (NY)*, 20(3):367–387, 1976.

Wood, D. E. and Salzberg, S. L. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.*, 15(3):R46, 2014.

Yandell, M. D. and Majoros, W. H. Genomics and natural language processing. *Nat. Rev. Genet.*, 3(8):601, 2002.