Novel Approaches to Marker Gene Representation Learning Using Trained Tokenizers and Jointly Trained Transformer Models

1st Alexander Manuele
Computer Science)
Dalhousie University
Halifax, Canada
al366886@dal.ca

2nd Robert G. Beiko
Computer Science
Dalhousie University
Halifax, Canada
beiko@cs.dal.ca

***Abstract*—Next-generation DNA sequencing technologies have made DNA sequence data far more widely available, opening new avenues of research. Analysis of marker gene data sets has many shortcomings, including sparsity, high cardinality, and intra-study dependencies during feature engineering. We present two novel approaches to feature representation of DNA marker gene data, first showing that trained tokenizers can replace traditional sliding-window based segmentation techniques, then proposing a training scheme to learn dense-vector representations of DNA sequences using transformer language models. We demonstrate that our representations match or exceed previously published approaches while providing fixed-length, low cardinality representations.**

***Index Terms*—representation learning, DNA, transformer, language modelling, microbiome**

I. Introduction

A. Marker gene analysis

The ribosomal rRNA 16S gene serves as a marker gene for analysis of prokaryotic microbes. The microbial diversity of an environment (i.e. the microbiome) is often characterized via high-throughput sequencing of the 16S gene, using differences in the gene sequences to determine the species present in a sample and their abundances [1]. It is increasingly common to apply machine learning techniques to these data, either to characterize the individual sequences or to characterize the collection of sequences that makes up a sample as a whole. Modelling individual DNA sequences is usually done by chunking a sequence into overlapping segments of length k, called k-mers. Sequences are then classified by counting the k-mers that appear per sequence and modelling the count vectors [2] [3]. Similarly, collections of sequences that comprise environmental samples can be classified by creating count vectors of the denoised or clustered unique sequences present in the sample [4] [5] [6] [1]. Each of these representations have severe shortcomings, however: k-mer count modelling produces high-dimensional, sparse data which provide no context or distance information. Microbiome sample count-vectors perform well but their cardinality relies on the unique organisms identified in a group of samples, preventing the use of trained models across studies.

B. Previous work

Some researchers have sought to address the shortcomings of k-mer count representations for sequences, showing that neural language models can be trained to create dense-vector representations of k-mers [7] [8] [9]. Specifically for 16S marker genes, [8] went further, showing that the dense-vector representations of the k-mers in a sequence can be aggregated to form dense-vector representations of the full sequence.

C. Contributions

We investigate two trained tokenization strategies for tokenizing 16S marker genes into non-overlapping tokens, byte-pair encoding (BPE) and unigram language modelling, and show that we can achieve significant compression of sequences without loss of performance in down-stream tasks. Having shown that we can use these strategies to effectively represent 16S sequences, we train a transformer encoder to model BPE encoded 16S sequences using a masked-language modelling task. We subsequently fine-tune the pre-trained transformer to produce fixed-length, dense-vector representations of 16S sequences which outperform the previously published method in all metrics tested. We show that our dense-vector sequence representations can be used for classification, clustering, and nearest-sequence lookup. Finally, we show that we can encode all the sequences comprising a microbial community sample and aggregate them to produce fixed-length dense-vector representations of samples without any loss in classification accuracy vs canonical count-based classification methods.

II. Trained Marker Gene Tokenizers Compress Sequences Without Loss of Performance in Downstream Tasks

We investigate BPE and unigram trained tokenizers as an alternative to k-mer tokenization for 16S DNA sequences. We find that using BPE or unigram tokens results in an approximately 7 fold reduction in sequence length due to the use of non-overlapping tokens and the collapse of conserved DNA regions into single tokens (Fig 1).

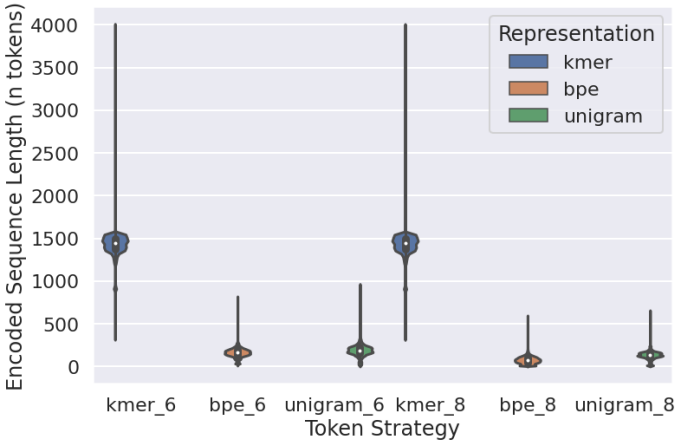

Fig. 1. Sequence lengths in n tokens of 410,078 tokenized 16S DNA sequences using k-mer (blue), byte-pair encoding (orange), and unigram language modelling (green) with vocabulary sizes of 4^6 and 4^8.

We classify 16S sequences using k-mer, BPE, and unigram token count vectors. We find that BPE and unigram token representations are more sensitive to machine learning model classification but that there is no significant difference in the best performance from each representation, showing that the reduction of sequence length from trained tokenizers is not a trade-off.

III. Trained tokens can replace k-mers for representation learning

We recreate the 16S DNA sequence embedding described by [8]. Their model uses a two step process; first, they train a word embedding algorithm to project DNA k-mers to continuous d-dimensional vector space. Then, they produce sequence level d-dimensional vector representations by using a normalized averaging procedure of the k-mer vectors from k-mers found in a sequence.

We repeat the procedure described by [8] using k-mers, BPE tokens, and unigram tokens as the word units for projection to d-dimensional vector space. We note that the reduced sequence length from BPE and unigram tokenization strategies results in ≈ 10 fold reduction in training time of the word embedding model. We perform taxonomic classification and clustering using each feature representation and find no significant difference in performance between token strategies in any metric. We conclude that the reduction in training time coupled with the homeostatic performance metrics suggests that BPE and unigram tokenization are preferable for representation learning tasks to k-mer tokenization.

IV. Transformer language models can be fine tuned to produce high quality 16S gene embeddings

Bidirectional transformer architectures (also known as transformer encoders) can be pre-trained using large, un-annotated data sets using unsupervised training objectives and subsequently fine-tuned for several applications [10] [11]. One such fine-tuning application is the ability to train the model to produce dense-vector representations of text sequences by training the model using sequence pairs annotated with a similarity score [12].

Using a large corpus of BPE tokenized 16S DNA sequences, we pre-train a bidirectional transformer model using the masked-language model training objective [13]. Subsequently, we create a corpus of annotated sequence pairs, annotating pairs of 16S sequences with pairwise alignment scores using VSEARCH [14]. We add a pooling layer to the pre-trained transformer model, thusly configuring it to produced dense-vector outputs from input sequences. We then fine-tune the model to minimize the mean squared error between cosine similarity of sequence embeddings and pairwise alignment score of the corresponding sequences.

This fine tuning process results in a model which creates high-quality fixed-length dense-vector representations of variable length 16S sequence inputs. We show that these sequence embeddings can be used for fast approximate nearest-sequence lookup, finding that nearest-vector lookup using cosine similarity scores overlap significantly with nearest-sequence lookup using the *de facto* gold standard biological sequence search algorithm, BLAST [15].

We compare our sequence embedding method against the method published by [8]. We find that our embeddings produce higher quality clusters when clustering sequences to different ranks in taxonomic hierarchy and have a much higher Spearman rank correlation coefficient between cosine similarities and pairwise alignment, both metrics which were reported in [8]. We find that both our embedding method and the previously published method classify 16S sequences with excellent performance. We note that unlike the previously published method, which relies on statistics calculated over a corpus of sequences, our embedding method produces high quality dense-vector representations using only the information contained in an individual sequence.

V. Trained sequence embeddings can be aggregated to form fixed-length representations of microbiome data

Microbiome data samples are typically represented as count vectors with entries corresponding to the number of sequence units identified in the sample. The canonical method for performing machine learning classification of these data is to use log-ratio transformations of these count vectors as features [16] [4]. The cardinality of these count vectors is determined by the study from which the data originated, as the counts reference all sequence units identified across a study [4]. We address this shortcoming by using aggregations of our dense-vector sequence representations to produce sample level dense-vector representations. We propose that microbiome samples can be represented as dense-vectors by simply calculating weighted averages of the embeddings of representative sequences in a sample, using centered log-ratio transformed sequence counts as the weights.

We acquire data sets from five microbiome studies with binary classification objectives. For each data set, we train a random forest classifier to predict the classification target using canonical methods (i.e. centered log-ratio transformation of the sequence count vectors) and our proposed feature representation method of aggregated sequence embeddings. Our method provides the advantages of fixed-length representations and significantly reduced feature cardinality. We find that in four of the five data sets, our method does not perform significantly differently than the canonical method and that in the fifth data set our method performs significantly better (table I).

TABLE I
CLASSIFICATION OF DISEASE STATE FROM STOOL MICROBIOTA SAMPLES

Study	ROC (Abundance CLR)	ROC (Embedding)	*p*-value
IBD	0.770 +/- 0.134	0.753 +/- 0.202	0.843
HIV	0.957 +/- 0.055	0.930 +/- 0.058	0.475
CRC1	0.754 +/- 0.143	0.606 +/- 0.207	0.230
ASD	0.581 +/- 0.058	0.755 +/- 0.027	0.001
CRC2	0.456 +/- 0.099	0.511 +/- 0.125	0.466

VI. CONCLUSION

To our knowledge, we are the first researchers to investigate the properties and viability of byte-pair encoding or unigram language modelling for tokenizing DNA sequences. We have shown that either representation can replace *k*-mer representation in classification and language modelling tasks, providing researchers with the advantages of reduced sequence length and controllable vocabulary size.

We extend the state of the art of DNA marker gene embedding, creating 16S sequence embeddings which outperform the previously published method in several metrics. Our method performs as well or better than the previously published method without the need to calculate data set-wide statistics for embedding calculations.

Finally, we show that our sequence embeddings can be used to create fixed-length dense-vector representations of microbiome data samples. The ability to represent microbiome data with a fixed-length representation that is not dependant on study level meta data and statistics is an essential development towards the creation of production-ready host phenotype prediction models.

REFERENCES

[1] R. Knight, A. Vrbanac, B. C. Taylor, A. Aksenov, C. Callewaert, J. Debelius, A. Gonzalez, T. Kosciolek, L.-I. McCall, D. McDonald, A. V. Melnik, J. T. Morton, J. Navas, R. A. Quinn, J. G. Sanders, A. D. Swafford, L. R. Thompson, A. Tripathi, Z. Z. Xu, J. R. Zaneveld, Q. Zhu, J. G. Caporaso, and P. C. Dorrestein, "Best practices for analysing microbiomes," *Nature Reviews Microbiology*, vol. 16, no. 7, pp. 410–422, May 2018. [Online]. Available: https://doi.org/10.1038/s41579-018-0029-9

[2] J. S. Johnson, D. J. Spakowicz, B.-Y. Hong, L. M. Petersen, P. Demkowicz, L. Chen, S. R. Leopold, B. M. Hanson, H. O. Agresta, M. Gerstein, E. Sodergren, and G. M. Weinstock, "Evaluation of 16s rRNA gene sequencing for species and strain-level microbiome analysis," *Nature Communications*, vol. 10, no. 1, Nov. 2019. [Online]. Available: https://doi.org/10.1038/s41467-019-13036-1

[3] N. A. Bokulich, B. D. Kaehler, J. R. Rideout, M. Dillon, E. Bolyen, R. Knight, G. A. Huttley, and J. G. Caporaso, "Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin," *Microbiome*, vol. 6, no. 1, May 2018. [Online]. Available: https://doi.org/10.1186/s40168-018-0470-z

[4] L. J. Marcos-Zambrano, K. Karaduzovic-Hadziabdic, T. L. Turukalo, P. Przymus, V. Trajkovik, O. Aasmets, M. Berland, A. Gruca, J. Hasic, K. Hron, T. Klammsteiner, M. Kolev, L. Lahti, M. B. Lopes, V. Moreno, I. Naskinova, E. Org, I. Paciência, G. Papoutsoglou, R. Shigdel, B. Stres, B. Vilne, M. Yousef, E. Zdravevski, I. Tsamardinos, E. C. de Santa Pau, M. J. Claesson, I. Moreno-Indias, and J. Truu, "Applications of machine learning in human microbiome studies: A review on feature selection, biomarker identification, disease prediction and treatment," *Frontiers in Microbiology*, vol. 12, Feb. 2021. [Online]. Available: https://doi.org/10.3389/fmicb.2021.634511

[5] G. B. Gloor and G. Reid, "Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data," *Canadian Journal of Microbiology*, vol. 62, no. 8, pp. 692–703, Aug. 2016. [Online]. Available: https://doi.org/10.1139/cjm-2015-0821

[6] A. Statnikov, M. Henaff, V. Narendra, K. Konganti, Z. Li, L. Yang, Z. Pei, M. J. Blaser, C. F. Aliferis, and A. V. Alekseyenko, "A comprehensive evaluation of multicategory classification methods for microbiomic data," *Microbiome*, vol. 1, no. 1, Apr. 2013. [Online]. Available: https://doi.org/10.1186/2049-2618-1-11

[7] P. Ng, "dna2vec: Consistent vector representations of variable-length k-mers," 2017.

[8] S. Woloszynek, Z. Zhao, J. Chen, and G. L. Rosen, "16s rrna sequence embeddings: Meaningful numeric feature representations of nucleotide sequences that are convenient for downstream analyses," *PLOS Computational Biology*, vol. 15, no. 2, pp. 1–25, 02 2019. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1006721

[9] R. Menegaux and J.-P. Vert, "Continuous embeddings of dna sequencing reads and application to metagenomics," *Journal of Computational Biology*, vol. 26, no. 6, pp. 509–518, 2019, pMID: 30785347. [Online]. Available: https://doi.org/10.1089/cmb.2018.0174

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[11] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

[12] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *CoRR*, vol. abs/1908.10084, 2019. [Online]. Available: http://arxiv.org/abs/1908.10084

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[14] T. Rognes, T. Flouri, B. Nichols, C. Quince, and F. Mahé, "VSEARCH: a versatile open source tool for metagenomics," *PeerJ*, vol. 4, p. e2584, Oct. 2016. [Online]. Available: https://doi.org/10.7717/peerj.2584

[15] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, Oct. 1990. [Online]. Available: https://doi.org/10.1016/s0022-2836(05)80360-2

[16] G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue, "Microbiome datasets are compositional: And this is not optional," *Frontiers in Microbiology*, vol. 8, Nov. 2017. [Online]. Available: https://doi.org/10.3389/fmicb.2017.02224