Normalized Compression Distance for DNA Classification

Gavin L.A. Hearne
Drexel University
Philadelphia, PA, USA

Mohammad S. Refahi
Drexel University
Philadelphia, PA, USA

Haozhe (Neil) Duan
Drexel University
Philadelphia, PA, USA

James R. Brown
Drexel University
Philadelphia, PA, USA

Gail L. Rosen
Drexel University
Philadelphia, PA, USA

ABSTRACT

The increasingly common use of next-generation sequencing has enabled greater access to large-scale (meta-)genomic datasets than ever before. The resulting deluge of data has made the quest for efficient DNA sequence classification methods an urgent challenge for downstream analyses. Traditional sequence alignment-based methods for DNA sequence classification struggle when presented with increasingly large volumes of sequence data due to the computational complexity of alignment. Subsequently, there is a need for methods capable of sequence identification without alignment. Normalized compression distance (NCD) has demonstrated capabilities in the field of text classification as a low-resource alternative to deep neural networks by leveraging compression algorithms to approximate Kolmogorov information distance. In an effort to apply this technique toward genomics tasks akin to tools such as Many-against-Many sequence searching (MMseqs) and Kraken2, we have explored the use of a *gzip*-based NCD towards both gene labeling of ORFs (open reading frames) and taxonomic classification of short reads. This demonstrates the efficacy of NCD in diverse multitask classification, and we further explore the capacity for NCD to classify larger libraries of metagenomic reads.

CCS CONCEPTS

• **Applied computing** → **Computational genomics**.

KEYWORDS

Metagenomics, Bioinformatics, Compression Distance, Microbiome

ACM Reference Format:

Gavin L.A. Hearne, Mohammad S. Refahi, Haozhe (Neil) Duan, James R. Brown, and Gail L. Rosen. 2024. Normalized Compression Distance for DNA Classification. In *15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '24), November 22–25, 2024, Shenzhen, China.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3698587.3701490

1 METHODS AND RESULTS

Our framework (Figure 1) utilises an implementation of NCD-*gzip* [1] designed to precalculate the compressed lengths of training sequences to reduce computational complexity at inference [2], and is supplemented by features to sub-sample training database elements to facilitate metagenomic analysis. This framework is evaluated on a pre-built 6-class human gene classification, as well as custom databases designed for full prokaryotic gene classification and metagenomic read taxonomy classification tasks.

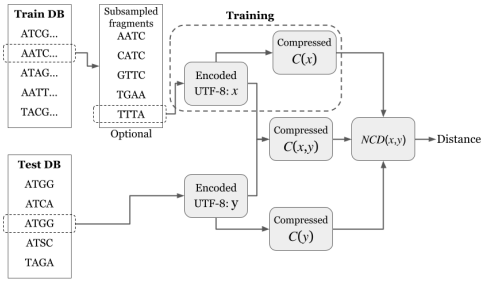

Figure 1: NCD process for a single comparison.

NCD-*gzip* demonstrates exceptional performance in human gene classification (0.883 macro F1). For prokaryotic gene classification (Figure 2) and metagenomics, it rivals state of the art classifiers such as Kraken2 and MMseqs2 on the superkingdom level, and unlike taxonomic classifiers, it can also identify gene label, albeit with less performance than alignment. However, the computational requirements for this method can be prohibitive, and scaling to large datasets can result in slow inference times.

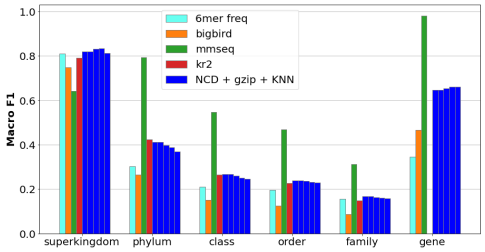

Figure 2: Macro-averaged F1 prokaryotic gene classification. Repeated NCD bars are (from left to right) k-values 1 -> 5.

ACKNOWLEDGMENTS

This work is supported by National Science Foundation under Grant Number #1936791 , #1919691 and #2107108. We thank the University Research Computing Facility for their paid services. We thank Dr. Bahrad Sokhansanj for discussion about related methods.

REFERENCES

[1] Zhiying Jiang, Matthew Yang, Mikhail Tsirlin, Raphael Tang, Yiqin Dai, and Jimmy Lin. 2023. "Low-Resource" Text Classification: A Parameter-Free Classification Method with Compressors. In *Findings of the Association for Computational Linguistics: ACL 2023.* 6810–6828.

[2] Ken Schutte. 2023. Bad numbers in the "gzip beats BERT" paper? https://kenschutte.com/gzip-knn-paper/

BCB '24, November 22–25, 2024, Shenzhen, China

ACM ISBN 979-8-4007-1302-6/24/11
https://doi.org/10.1145/3698587.3701490