Entropy-based Kullback-Leibler Taxonomic Classification of Biological Sequences

Viorel Munteanu, Nicolae Drabcinski, Dumitru Ciorbă, Viorel Bostan
Technical University of Moldova, viorel.munteanu@lt.utm.md,
nicolae.drabcinski1@lt.utm.md, dumitru.ciorba@fcim.utm.md,
viorel.bostan@adm.utm.md, ORCID: 0000-0002-4133-5945, 0009-0008-4381-836X,
0000-0002-3157-5072, 0000-0002-2422-3538

Keywords: alignment-free method, Kullback-Leibler divergence, phylogeny

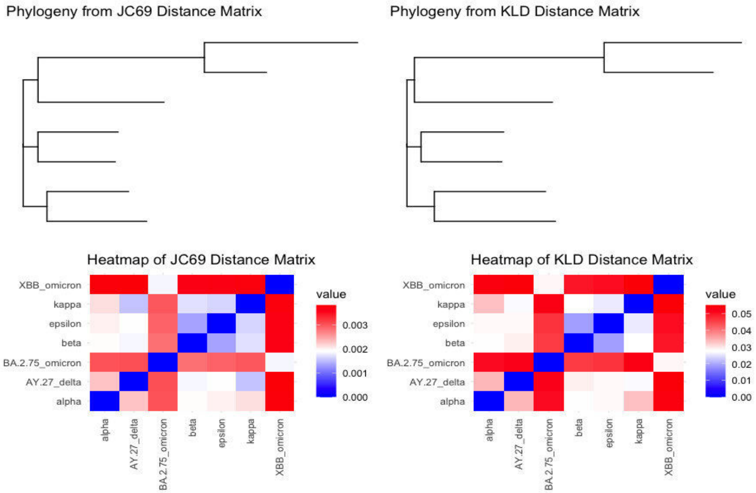
Abstract. Accurate classification of biological sequences is fundamental for understanding their functional, structural, and evolutionary significance. Traditional alignment-based methods often face challenges when applied to large, highly diverse datasets, especially when sequences have low identity or are distantly related [1]. Alignment-free methods, an established category in computational biology, have emerged as powerful alternatives to traditional alignment approaches, offering solutions for these challenges. Here we present an efficient alignment-free method for sequence similarity measure and taxonomic classification that relies on *k*-mer frequency distribution using Kullback-Leibler (KL) divergence between two probability distributions [2]:

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) log\left(\frac{P(x)}{Q(x)}\right), \tag{1}$$

where $P(x)$ and $Q(x)$represents the probability of observing *k*-mer x in the first and second sequence, respectively. This measure is asymmetric, $D_{KL}(P||Q) \neq D_{KL}(Q||P)$, correspondingly don't satisfy the proprietis of a true distance metric, such as symmetry and the triangle inequality. To account for this asymmetry, we compute the symmetric KL divergence, which averages the KL divergence in both directions:

$$D_{KL}^{S} = (P,Q) = \frac{1}{2}(D_{KL}(P||Q) + D_{KL}(Q||P)) \tag{2}$$

Our preliminary results show that the D_{KL}^{S}-based method for sequence comparison and taxonomic classification performs with high accuracy, closely matching traditional alignment-based methods (Fig.1)[3].



Phylogeny from JC69 Distance Matrix

Phylogeny from KLD Distance Matrix

Heatmap of JC69 Distance Matrix

Heatmap of KLD Distance Matrix

Fig 1. Both the traditional JC69 (left) and the D_{KL}^{S} metric (right) produce consistent phylogenetic trees and similar distances (heatmaps) across both methods (here *k*-mer length is 10bp).

Acknowledgments. This work was partially supported by the State Program LIFETECH No. 020404 at the Technical University of Moldova.

References

[1] Zielezinski, Andrzej, Susana Vinga, Jonas Almeida, and Wojciech M. Karlowski. 2017. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biology* 18: 186. https://doi.org/10.1186/s13059-017-1319-7.

[2] Kullback, S., and R. A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics* 22. Institute of Mathematical Statistics: 79–86.

[3] Bioinformatics-Lab-TUM/TCS_LLR. 2024. Bioinformatics Lab, TUM. https://github.com/Bioinformatics-Lab-TUM/TCS_LLR/tree/main