Archaea Taxonomic Classification

Jorge Miguel Silva[1]
jorge.miguel.ferreira.silva@ua.pt
Diogo Pratas[1]
pratas@ua.pt
Tânia Caetano[2]
tcaetano@ua.pt
Sérgio Matos[1]
aleixomatos@ua.pt

[1] DETI/IEETA
University of Aveiro
Portugal

[2] CESAM and Department of Biology
University of Aveiro
Portugal

[3] Department of Virology
University of Helsinki
Finland

Abstract

Archaea are a domain of single-celled organisms that live in almost every environment and play significant environmental roles, such as carbon fixation and nitrogen cycling. However, their classification is difficult because most have not been isolated in a laboratory and detected only by their gene sequences in environmental samples. Moreover, archaeal genomes are characterized by significant dissimilarity. This manuscript provides an automatic classification methodology by applying an ensemble method using a combination of reference-free compression measures with GC-content and length. Notably, the results show that we can automatically and accurately distinguish between Archaea genomes at different taxonomic levels.

1 Introduction

Archaea are a domain of single-celled organisms that lack a nucleus. Their cells have unique properties which are distinct from both bacteria and eukaryota domains. Archaea and bacteria are generally similar in size and shape. However, despite the morphological similarities to bacteria, Archaea have genes and metabolic pathways more closely related to eukaryotes, prominently for the enzymes involved in transcription and translation. In addition, other aspects of archaeal biochemistry are unique, such as their reliance on ether lipids in their cell membranes. Furthermore, Archaea are characterized for having a significant genomic inter-dissimilarity.

Despite being firstly detected living in extreme environments such as hot springs and salt lakes with no other organisms, they live in almost every environment. In the human microbiome, they are essential in the gut, mouth, and skin. Furthermore, they play significant environmental roles, such as carbon fixation, nitrogen cycling, organic compound turnover, and maintaining microbial symbiotic and syntrophic communities.

Currently, Archaea are further divided into multiple recognized phyla. However, classification is difficult because most have not been isolated in a laboratory and detected only by their gene sequences in environmental samples. Studying a DNA sequence's complexity (or quantity of information) may help solve this classification problem. As such, this manuscript proposes an Archaea genomic taxonomic classification tool. Specifically, it performs classification without resorting directly to the sequence of the reference genomes. Instead, it uses an ensemble of three predictors, namely normalized compression and two simple property characteristics, for probabilistic classification of DNA sequences.

It is counter-intuitive to think that it is possible to classify a genome recurring only to how much it can be compressed, its length, and the percentage of Guanine and Cytosine. For example, to determine its phylum, order, class or genus. Furthermore, this manuscript shows that it is not only possible but that it can be done automatically with high accuracy, using a small and diverse dataset recurring to alignment-free approaches [11]. The complete study can be fully replicated using the repository https://github.com/jorgeMFS/Archaea.

2 Methods

2.1 Database

The Archaea NCBI database is minimal when compared to other domains of life. The dataset comprises 216 complete reference genomes retrieved from the NCBI database (link) on 30 September 2021. In addition, the taxonomic description was also retrieved from the NCBI database and manually corrected to classify different taxonomic levels correctly. This mapping is available in the project to simplify future usage and replication.

2.2 Normalized Compression (NC)

An efficient compressor, $C(x)$, provides an upper bound approximation for the Kolmogorov complexity ($K(x)$), where $K(x) < C(x) \leq |x|$ ($|x|$ is the length of string x in the appropriate scale). Usually, an efficient data compressor is a program that approximates both probabilistic and algorithmic sources using affordable computational resources (time and memory). Although the algorithmic nature may be more complex to model, data compressors can have embedded sub-programs to handle this nature. The normalized version, known as the Normalized Compression (NC), is defined by

$$\mathrm{NC}(x) = \frac{C(x)}{|x| \log_2 |\mathrm{A}|}, \tag{1}$$

where $C(x)$ is the compressed size of x in bits, $|\mathrm{A}|$ the number of different elements in x (size of the alphabet). Given the normalization, the NC enables to compare the proportions of information contained in the strings independently from their sizes [7]. If the compressor is efficient, then it can approximate the quantity of probabilistic-algorithmic information in data using affordable computational resources. In our work, to determine the NC, we made use of the state-of-the-art DNA sequence compressor: GeCo3 [10].

2.3 Other Measures

The two other measures used to perform Archaea taxonomic classification are the GC-Content (GC) and the length of the genome $|x|$.

GC-Content (GC) represents the proportion of guanine (G) and cytosine (C) bases out the quaternary alphabet $\{A,C,G,T/U\}$. This includes thymine (T) in DNA and uracil (U) in RNA. The GC percentage is given by the number of cytosine (C) and guanine (G) bases in an Archaea genome x with length $|x|$ according to

$$\mathcal{GC}(x) = \frac{100}{|x|} \sum_{i=1}^{|x|} \mathcal{N}(x_i || x_i \in \Xi), \tag{2}$$

where x_i is each symbol of x (assuming causal order), Ξ is a subset of the genomic alphabet containing the symbols $\{G,C\}$ and $\mathcal{N}$ the program that counts the numbers of symbols in Ξ.

GC-content is variable between different organisms. In addition, the GC-content value correlates with the organism's life-history traits, genome size [9], and GC-biased gene conversion [3]. As such, this measure is useful to perform Archaea classification. Furthermore, an organism with a genome high in GC-content is rich in energy and more prone to mutation. Thus, over time, a species tends to decrease its GC-content to become more stable, giving us further information regarding Archaea characterization.

For comparison of the obtained results, we assessed the outcomes obtained using a random classifier. For that purpose, for each task, we determined the probability of a random sequence being correctly classified (p_{hit}) as

$$p_{hit} = \sum_{i=0}^{n} [p(c_i) * p_{correct}(c_i)], \tag{3}$$

where $p(c_i)$ is the probability of each class, determined as

Table 1: Results obtained for Archaea taxonomic classification task regarding the phylum, class, order, family, and genus. The features used were the genome's sequence length (SL), the GC-content (GC) and the Normalized Compression (NC) values. The classifiers used were Linear Discriminant Analysis (LDA), Gaussian Naive Bayes (GNB), K-Nearest Neighbors (KNN), Support Vector Machine (SVM) and XGBoost classifier (XGB). The performance was measured using the accuracy (ACC) and the Weighted F1-score (F1-score). The probability of a sequence being correctly classified using a random classifier was determined (p_{hit}).

Data Characteristics			**Random**	**GNB**$_{SQ+GC+NC}$		**SVM**$_{SQ+GC+NC}$		**KNN**$_{SQ+GC+NC}$		**LDA**$_{SQ+GC+NC}$		**XGB**$_{NC}$		**XGB**$_{SQ+GC+NC}$	
Classification	**N. Classes**	**Samples**	p_{hit}	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**
Phylum	4	215	25.00	58.14	0.56	68.60	0.56	69.77	0.65	66.28	0.58	77.91	0.76	**80.23**	**0.80**
Class	9	168	11.11	63.24	0.57	61.76	0.52	63.24	0.59	64.71	0.63	66.18	0.65	**69.12**	**0.68**
Order	25	242	4.00	26.80	0.17	28.87	0.16	28.87	0.19	36.08	0.3	34.02	0.34	**41.24**	**0.43**
Family	39	219	2.56	27.27	0.22	29.55	0.15	32.95	0.18	36.36	0.32	35.23	0.34	**39.77**	**0.42**
Genus	97	213	1.03	12.79	0.09	17.44	0.05	15.12	0.04	27.91	0.23	19.77	0.14	**29.07**	**0.24**

$$p(c_i) = \frac{|samples_{class}|}{|samples_{total}|}.$$

On the other hand, $p_{correct}(c_i)$ is the probability of that class being correctly classified. In the case of a random classifier,

$$p_{correct}(c_i) = \frac{1}{|classes|}.$$

3 Results

In this section, we performed five different classification tasks for each Archaea sequence from the dataset. Specifically, the sequences were classified regarding their phylum, class, order, family, and genus.

We applied 5 types of classifiers: Linear Discriminant Analysis (LDA) [6], Gaussian Naive Bayes (GNB) [8], K-Nearest Neighbors (KNN) [4], Support Vector Machine (SVM) [2] and XGBoost classifier (XGB)[1]. To select the best performing method we computed the Accuracy and the Weighted F1-score.

Furthermore, we performed classification using three different features: the Normalized Compression (NC), GC-content (GC), and sequence length (SL). These three features were fed to all the classifiers, and the accuracy and weighted F1-score were measured to determine which classifier was best suited for this task.

Table 1 depicts the accuracy and weighted F1-score values obtained for each classifier. For all classification tasks, the best performing classifier was the XGBoost classifier. Regarding the features used, despite the NC feature being the most relevant, combining it with the GC-Content and Sequence Length improved the accuracy and F1-score result. This improvement increased when the number of classes was higher. Overall, there is a decrease in accuracy and F1-score when there is an increase in the number of classes. Specifically, we obtained the best performance in the phylum classification of the Archaea (accuracy - 80.23%, F1-score - 0.80) and our lowest performance in genus classification (accuracy - 29.07%, F1-score - 0.24). This decrease is mainly because the average number of samples per class decreases as the number of classes increases. As such, many classes lack a valid number of samples to be accurately classified. Moreover, part of the classification inaccuracies can be explained by possible errors in the assembly process of the original sequence or eventual sub-sequence contamination of parts of the genomes. Other inaccuracies could be due to several genomes being reconstructed using older methods that have been improved since then [5].

As far as we know, this is the first attempt at performing this type of taxonomic classification using reference-free methods. As such, for comparison purposes, we assessed the outcomes obtained using a random classifier. Specifically, for each task, we determined the probability of a random sequence being correctly classified (p_{hit}). Overall, there is a vast improvement relative to the random classifier, showing the importance of the features used in the classification process. The results are particularly encouraging given the small sample size and the many classification labels of the dataset. We conclude that these classification results show that this metric can be utilized for taxonomic classification, particularly if more sequence samples are added to the public dataset.

4 Conclusion

This manuscript evaluates the capability of using complexity measures to perform Archaea classification at different taxonomic levels. For this purpose, we used the NC, the GC-content, and length of the genome sequence. The best results were obtained using all mentioned features in the XGBoost classifier. Notably, the results showed that we can automatically and accurately distinguish between Archaea genomes at different taxonomic levels. As far as we are aware, this is the first study where reference-free classification of the Archaea at different taxonomic levels is performed. As such, we compared our obtained results with a random classifier. As a result, we extensively outperform a random classifier, proving these measures' efficiency in performing this type of classification. However, the results obtained showed a decrease in accuracy when approaching the lowest taxonomic levels due to an increase in the number of classes and a decrease in the number of samples per class. As such, when future entries are added to the database, accuracy may significantly increase in the lowest taxonomic levels. Future work involves the addition of other experts regarding the proteomes of Archaea.

Overall, this manuscript shows that the efficient approximation of the Kolmogorov complexities of Archaea sequences as measures of complexity have a profound impact on genomes identification and classification.

References

[1] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2.

[2] Nello Cristianini, John Shawe-Taylor, et al. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[3] Laurent Duret and Nicolas Galtier. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual review of genomics and human genetics*, 10:285–311, 2009.

[4] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. Knn model-based approach in classification. *Springer Berlin Heidelberg*, pages 986–996, 2003. doi: 10.1007/978-3-540-39964-3_62.

[5] Jennifer Lu and Steven L Salzberg. Removing contaminants from databases of draft genomes. *PLoS computational biology*, 14(6): e1006277, 2018.

[6] Geoffrey J McLachlan. *Discriminant analysis and statistical pattern recognition*, volume 544. John Wiley & Sons, 2004.

[7] Diogo Pratas and Armando J Pinho. On the approximation of the Kolmogorov complexity for DNA sequences. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 259–266. Springer, 2017.

[8] Irina Rish et al. An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001.

[9] Jonathan Romiguier, Vincent Ranwez, Emmanuel JP Douzery, and Nicolas Galtier. Contrasting gc-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome research*, 20(8):1001–1009, 2010.

[10] Milton Silva, Diogo Pratas, and Armando J Pinho. Efficient DNA sequence compression with neural networks. *GigaScience*, 9(11), 11 2020. ISSN 2047-217X. giaa119.

[11] et al. Zielezinski, Andrzej. Benchmarking of alignment-free sequence comparison methods. *Genome Biology*, 20(144), 2019. doi: 10.1186/s13059-019-1755-7.