Old School vs. New School: Evaluating DNA Sequence Classification on Random Forests, BLAST Searches, and Foundation Models.

Eric Nunes
eenunes@umass.edu

Thomas Potts
tpotts@umass.edu

Abstract

In biology, nucleotide sequence classification is an important task because it can be used to identify functionalities and properties of proteins. In this paper, we evaluate two non-LLM baselines against biological foundation models (FM) on a classification task from the Diverse Genomic Embedding Benchmark (DGEB). We built two classifier models: a k-nearest neighbors (kNN) classifier using BLAST searches as a similarity metric, and a random forest (RF) model based on normalized k-mer counts. We then compare the macro F1 scores to the results published in the DGEB paper, evaluating whether traditional machine learning approaches can outperform LLM-based methods. Our findings support the hypothesis to a moderate degree - the BLAST k-NN classifier achieved an F1 score of 0.629, outperforming the top foundation model by 24.3%; on the other hand, the RF models achieved an F1 score of 0.485, a 13.8% improvement from the worst FM but a 4.2% decrease from the top LLM-based model. Our research suggests that further development in the biological foundation models is needed to catch up with the performance of single-task, rudimentary ML algorithms in DNA sequence classification.
Keywords: Computational Biology, Bioinformatics, Foundation Models, Random Forests, BLAST, Classification

Github Repository: https://github.com/eric27n/compsci-690u-project
Presentation Slides/Poster: https://tinyurl.com/compsci-690u-nunes-potts

1. Introduction

Foundation models (FM) are a form of large-language model (LLM) that has the ability to generalize in a topic. In the mid-2020s, there is both intrigue and skepticism regarding the use of biologically-based FMs, sparking debate between newer LLM-approaches and traditional ML algorithms across many tasks in bioinformatics [2][3]. In 2024, West-Roberts et al. introduced DGEB as a benchmark to evaluate FMs across different tasks and datasets in bioinformatics.[5] One of the nucleotide-based baselines was a multi-label classification task on the Minimum Information about a Biosynthetic Gene cluster (MIBiG) dataset [4]. The Diverse Genomic Embedding Benchmark only applied large protein and genomic language models to classify the MIBiG entries, and so simpler, non-NLP based models have not been evaluated for this task. Thus, we have decided to implement two methods using traditional computational biology and simpler machine learning methods and compare their results to the MIBiG benchmark.

In this paper, we evaluated the correctness of a simple multi-label classification task on nucleic-acid sequences. The first model classifies using BLAST, a traditional computational biology method, and the second model classifies using Random Forest, an ensemble machine learning method. The MIBiG database supports both amino acid and nucleotide sequence modalities, but for this paper, we apply the methods solely on the nucleotide sequence database. We compared the results of a random forest implementation, a BLAST-enhanced k-nearest neighbors algorithm, and genomic foundation models attempting to classify sequences of DNA into six categories.

The dataset being used to evaluate the classifications is a modified version of the MIBiG protein dataset, altered and published by West-Roberts et al in their paper. Note that there are two datasets under the MIBiG name used in the paper, one classifying DNA sequences (used by nucleic acid (NA)-based foundation models) and the other protein sequences (used by amino acid (AA)-based foundation models); for this paper, we plan to evaluate solely on the nucleotide sequences because the NA language models struggled with the DNA classification task more than the AA models with the protein classification. The dataset contains 6 different classes across 1,763 training entries and 441 testing entries. While the multi-class classification tasks were done reasonably well by the AA models, the NA models fared rather poorly, with the best NA model achieving an F1 score of 0.506.

We seek to examine the performance of two traditional ML algorithms against a more novel NLP-based approach. We hypothesize that if given enough information on the samples, then both the random forest and BLAST k-NN algorithms outperform the foundation models due to being more tailored in solving one task, rather than a foundation model that can generalize well across multiple tasks though not perform better at a specific task. In addition, the novelty and complexity of these foundation models could affect performance, as can be seen regarding the NT model, and how the largest parameter model (2.5b) fared slightly worse than simpler models like the 100m or 250m parameter models.

2. Method

For the BLAST-based classifier, we utilize the local tool, BLAST+, provided by the National Library of Medicine (NLM) [1]. As the MIBiG database is not included in the NLM's online list of BLAST databases, we must first create a BLAST database of the MIBiG nucleotide dataset. As for splitting the data into a train and test set, we evaluate on the split data provided by the DGEB. For classification, our algorithm follows a similar structure to the k-nearest neighbors method, such that the "nearest" instances are the most similar sequences returned using BLAST. For each test instance, the algorithm applies blastn on the reference database, finds the labels of the top k hits, and predicts the majority label. The hyperparameter for this BLAST implementation is simply k. Macro F1 score is the primary evaluation metric for this task, as that was used in the benchmark.

For the Random Forest classifier, we use the RandomForestClassifier class provided by scikit-learn for its ubiquitousness. In addition, we use the RandomizedSearchCV class to apply both cross validation and hyperparameter tuning to find the most optimal set of hyperparameters. Our group plans to create a random forest classifier based on different attributes of the sequence - including percentages of nucleotides and nucleotide pairs, but most importantly, percentage of k-mer appearances in sequences. The latter approach is done by taking the total counts of different k-mers in the DNA sequences, and dividing them by sequence length; normalized counts are preferred instead of raw counts since the total length of sequences varies dramatically in the MIBiG dataset and could thus overfit by raw counts alone. Similarly to the BLAST classifier, different hyperparameters - most notably the value of k in the k-mers - will be fine-tuned and reported in the final deliverable. To evaluate the random forest's results, we will use the macro F1 score as our primary metric, since that metric is what's used to evaluate the results in the DGEB paper.

For the foundation models, we focus on the results as reported from the DGEB paper. Because we are focusing on a DNA classification task, we are only to evaluate using the nucleic acid-based models. These include the models of the Nucleotide Transformer (NT) and Evo; these models have 5 and 2 different variations, respectively. The results reported for the MIBiG DNA Task in the DGEB paper are the macro F1 scores.

3. Results

After evaluating our two non-language classification models, we found that the BLAST k-NN classifier outperformed all benchmark models and the random forest algorithm performed similarly to the DGEB.

The BLAST k-NN classifier achieved the highest F1 score out of all of the classification models, with a macro F1 score of 0.629. Importantly, this model outperformed the top foundation model (NT v2-250m) by 24.3%. This demonstrates the efficacy of traditional alignment methods for sequence classification tasks! As seen in table 2, the BLAST k-NN performed well in five out of the six classes. Interestingly for the RiPP class, the random forest classifier out performed BLAST k-NN in both recall and F1 score.

For the number of nearest neighbors, k, there is an inverse relationship between k and macro F1 score. The classifier performs the best at $k = 3$, and the performance continually decreases until stabilizing at $k = 7$.

When comparing the random forest approach to the foundation models, it can be observed that some of the foundation models outperform even the best RF. The best performing random forest model evaluated on k-mers of $k = 4$, with additional hyperparameters of 200 estimators, a minimum sample split size of 5, a minimum number of 1 sample per leaf, no max features, a maximum tree depth of 30, and bootstrapping being enabled. However, even this combination of parameters yielded a macro F1 score of 0.485; while this significantly outperforms both Evo models and the simplest Nucleotide Transformer model (50m parameter model), it falls short of the remaining NT models by at least 0.014. The closeness to the results suggest that a more robust RF with more variables - particularly nucleotide indeces - and stronger hyperparameter tuning could be competitive with foundation models.

Model Type	Model/Key Hyperparameter	Macro F1 Score
Random Forest	$k = 3$	0.472
	$k = 4$	0.485
	$k = 5$	0.451
	$k = 6$	0.435
	$k = 7$	0.415
BLAST-kNN	$k = 2$	0.626
	$k = 3$	**0.629**
	$k = 4$	0.608
	$k = 5$	0.602
	$k = 6$	0.599
Foundation Model	NT v2-50m Multispecies	0.447
	NT v2-100m Multispecies	0.503
	NT v2-250m Multispecies	0.506
	NT v2-500m Multispecies	0.500
	NT 2.5b Multispecies	0.499
	evo-1 8k-base	0.426
	evo-1 131k-base	0.446

Table 1. Macro F1 Scores for the three different types of models, as well as specific models and key hyperparameters. The random forest's k's are for creating k-mers of length k, while the BLAST method's k designates the k in k-nearest neighbors. The best performing metric in each model type is designated in underline, while the highest overall metric is designated in **bold**.

MIBiG Class	# Training Instances	# Testing Instances	Method	Precision	Recall	Macro F1 Score
Alkaloid	48	12	Random Forest	0.00	0.00	0.00
			BLAST kNN	*0.50*	*0.17*	**0.25**
NRP	652	163	Random Forest	*0.68*	0.72	0.70
			BLAST kNN	0.60	*0.89*	**0.72**
Polyketide	551	138	Random Forest	0.63	0.72	0.67
			BLAST kNN	*0.82*	*0.73*	**0.77**
RiPP	266	67	Random Forest	0.64	*0.85*	**0.73**
			BLAST kNN	*1.00*	0.48	0.65
Saccharide	113	28	Random Forest	0.88	0.25	0.39
			BLAST kNN	*1.00*	*0.79*	**0.88**
Terpene	133	33	Random Forest	0.67	0.30	0.42
			BLAST kNN	*0.72*	*0.39*	**0.51**

Table 2. The six different MIBiG classes and metrics for the best performing Random Forest and BLAST kNN models. The exact hyperparameters can be found in the text. The model with a better precision and/or recall is designated with *italics & underscore*, while the best performing macro F1 score is designated in **bold**.

4. Conclusion

In regards to the performance of the BLAST k-NN search and random forests, it demonstrates that our hypothesis was partially correct: a simple ML algorithm using traditional biology methods can outperform an NLP-based model, while the random forest model demonstrates that without these traditional algorithms it comes short in outperforming the embedding-based models. This suggests that while FMs demonstrate versatility and ease of use, it is still not quite as ready compared to dedicated traditional ML approaches without a plethora of training data.

A limitation of the study can be observed in regard to the random forests, particularly the lack of any variables tracking positions of nucleotides. Tracking percentages of k-mers caused information on order or index of nucleotides to get lost, potentially losing critical information. The small difference between the RF and FM performances leaves potential to explore this, and determine if positional information could lead to RFs outperforming the NLP models.

References

[1] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. Blast+: architecture and applications. *BMC bioinformatics*, 10:1–9, 2009. 2

[2] Fei Guo, Renchu Guan, Yaohang Li, Qi Liu, Xiaowo Wang, Can Yang, and Jianxin Wang. Foundation models in bioinformatics. *National Science Review*, page nwaf028, 2025. 1

[3] Qing Li, Zhihang Hu, Yixuan Wang, Lei Li, Yimin Fan, Irwin King, Gengjie Jia, Sheng Wang, Le Song, and Yu Li. Progress and opportunities of foundation models in bioinformatics. *Briefings in Bioinformatics*, 25(6):bbae548, 2024. 1

[4] Barbara R Terlouw, Kai Blin, Jorge C Navarro-Muñoz, Nicole E Avalon, Marc G Chevrette, Susan Egbert, Sanghoon Lee, David Meijer, Michael J J Recchia, Zachary L Reitz, Jeffrey A van Santen, Nelly Selem-Mojica, Thomas Tørring, Liana Zaroubi, Mohammad Alanjary, Gajender Aleti, César Aguilar, Suhad A A Al-Salihi, Hannah E Augustijn, J Abraham Avelar-Rivas, Luis A Avitia-Domínguez, Francisco Barona-Gómez, Jordan Bernaldo-Agüero, Vincent A Bielinski, Friederike Biermann, Thomas J Booth, Victor J Carrion Bravo, Raquel Castelo-Branco, Fernanda O Chagas, Pablo Cruz-Morales, Chao Du, Katherine R Duncan, Athina Gavriilidou, Damien Gayrard, Karina Gutiérrez-García, Kristina Haslinger, Eric J N Helfrich, Justin J J van der Hooft, Afif P Jati, Edward Kalkreuter, Nikolaos Kalyvas, Kyo Bin Kang, Satria Kautsar, Wonyong Kim, Aditya M Kunjapur, Yong-Xin Li, Geng-Min Lin, Catarina Loureiro, Joris J R Louwen, Nico L L Louwen, George Lund, Jonathan Parra, Benjamin Philmus, Bita Pourmohsenin, Lotte J U Pronk, Adriana Rego, Devasahayam Arokia Balaya Rex, Serina Robinson, L Rodrigo Rosas-Becerra, Eve T Roxborough, Michelle A Schorn, Darren J Scobie, Kumar Saurabh Singh, Nika Sokolova, Xiaoyu Tang, Daniel Udwary, Aruna Vigneshwari, Kristiina Vind, Sophie P J M Vromans, Valentin Waschulin, Sam E Williams, Jaclyn M Winter, Thomas E Witte, Huali Xie, Dong Yang, Jingwei Yu, Mitja Zdouc, Zheng Zhong, Jérôme Collemare, Roger G Linington, Tilmann Weber, and Marnix H Medema. Mibig 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic Acids Research*, 51(D1):D603–D610, 2022. 1

[5] Jacob West-Roberts, Joshua Kravitz, Nishant Jha, Andre Cornman, and Yunha Hwang. Diverse genomic embedding benchmark for functional evaluation across the tree of life. *bioRxiv*, pages 2024–07, 2024. 1