Binning microbial long reads: learning bacterial family signatures with XGBoost

Siegfried Dubois[1] and Jacques Nicolas[1]

Univ Rennes1, INRIA, Rennes, France

Abstract. Genomic and metagenomic data are flowing into microbiology thanks to advances in recent sequencers such as the ONT Minion that allow low-cost, in-field access to this technology. In this context, this work addresses a key problem, binning, which consists of grouping sequenced reads into taxonomically coherent sets. We have learned genomic signatures on two databases of microbial genomes and for various taxonomic levels, relying on a model of regression trees boosting, thanks to the XGBoost library. We used as attributes the frequencies of small k-mers (in range 4-6) on 10,000b fragments sampled along the genomes. Each level of the taxonomy (until the family level) is predicted assuming the previous level is known. The prediction was made at the scale of single reads and groups of 400 reads, with a preprocessing step discarding low signifiance fragments. Overall, the level of accuracy achieved is very satisfactory, with known isolated problems such as for the *Fusobacteria* phylum and the *Deltaproteobacteria* group. Apart from the domain level (bacteria or archae), the most coherent taxonomic level seems to be that of the order. We propose a software suite, Wisp, covering database creation, error processing, creation of reports, quality analysis of predictions and a user-friendly parameters interface. Source code and data are freely available at https://github.com/Tharos-ux/wisp

Keywords: Machine learning · XGBoost · Metagenomics · Binning · Long reads · Genomic signatures · k-mers.

1 Introduction

Progress in sequencing technologies and metagenomics increased drastically the amount of available microbial genomic sequences, and allows to train automatic classifiers for various prediction tasks, e.g. taxonomic classification of metagenomic short reads [5], prediction of phenotypic traits [15], or prediction of antimicrobial resistance profiles [8]. Oxford Nanopore Technologies (ONT) MinION sequencers offer interesting possibilities in microbial genomics since they are cost-efficient, allow on-field sequencing, and produce multi-kilobase reads. Their main current limitation is the error level of sequences, with up to 6% of errors represented as indels and substitutions [3].

This work elaborates on the relatively old topic of taxonomic classification from genomic composition in the light of these recent long read sequencers and the development of efficient learning tools. The specificity of codon usage in each genome has been noticed early in bioinformatics studies [6] but the interest

in this type of research dates back to the seminal article by Karlin et al [7], where some dinucleotide and tetranucleotide frequencies were found to have a small variation within fragments of a bacterial genome and a greater variation between different genomes. This was shown for a small number of genomes at the time (15, including incomplete genomes) and few short patterns on 50kb fragments. Since then, advances in indexing have made it possible to rely on longer words called kmers for classification from genomic data and to move from composition analysis to presence analysis. For instance Kraken [16], likely the most popular tool for species identification from short reads, relies on long kmers (k=31 by default). More generally, using kmers as identification signatures is one of the most used alignment-free approaches [1,4,10]. Its limitations are the size of the kmer indexes needed to represent a large set of species or strains, as well as the degradation of the results according to the error rate in the sequences (kmers are searched exactly). These limitations can be mitigated by using a two-step identification process, one for higher taxonomic ranks using compositional analysis and a second one for finer levels, based on the presence of kmers. This paper concerns the first step.

Various machine learning algorithms have been tried to train a genomic model from available reference genomes : naive Bayes classifier (NBC) [13], support vector machines (SVMs) [9,12,14], and convolutional neural networks (CNN) and deep belief networks (DBN) [5]. They have been designed for the classification from short reads with very few errors.

2 Learning taxonomic signatures with decision forests

2.1 Data sets and compositional attributes

We have worked on two data sets, considering 5 taxonomic levels:

- **Bernard's dataset** is issued from [1] and is made of 143 archaeal and bacterial genomes. We removed records which were discarded since by the NCBI, and we kept 107 genomes, divided under 2 domains, 16 phyla, 24 groups, 44 orders and 76 families.
- **Supported dataset** is a selection of the NCBI *refseq* database limited to families counting at least 3 representatives, and for each of which 3 genomes have been randomly drawn: it is made of 129 archaeal and bacterial genomes, divided under 2 domains, 8 phyla, 14 groups, 26 orders and 43 families.

For the attributes, we used kmers counts computed on a sliding window of 10.000 bp on the direct and reverse strand of each reference genome. Exploring kmers of size 4,5 and 6 across all classification levels on Bernard's database. We got the best predictive accuracy with k=5 for high levels domain and phylum, and 4 for other levels ; 6 was the worst all along the line (PhyloPhyta achieves the best results with k=6 and 5). For each database, 2 training sets were built, $v1$ by sampling 50 times for domain and family, 100 for other levels, and $v2$ to test overfitting by sampling 100 times for domain and 500 for other levels.

2.2 Learning with XGBoost

We wanted to test the XGBoost learning method on this taxonomic identification problem from sequence composition. XGBoost is a popular gradient boosting technique working on regression tree ensembles, which offers an interesting tradeoff in terms of predictive power, explainability and efficiency [2]. Moreover, we were interested on testing the learned models on data typical from ONT Minion sequencers, that is, long and erroneous reads.

In order to have a certain stability of the classifiers according to the evolution of the databases, we built them in a hierarchical way, by assuming at each level that the taxon at the higher level is known (e.g. domain is known at the phylum level). This has consequences for the challenge of the learning task as the size of the learning samples decreases with the depth in the taxonomy. For instance, for the Supported database and $v1$ learning sets, and for the 5 levels domain, phylum, group, order and family, the mean size of the learning set is respectively based on 6450, 6450, 1613, 921 and 248 sequences.

The test sets were built from 400 fragments of 10000 bp randomly and independently drawn in the target genome for each taxonomic level, represented with the same attributes than the learning sets. For the $v1$ training set, we created two test sets, $v1$, a version without sequence errors, and $v1E$, a version simulating sequencing errors introducing an indel or substitution error at each position with probability 6%.

We used XGBoost version 1.6.1 ; models were made in approximate mode, 10 boostings with gbtree method, evaluation metric was mlogloss, a tree depth of 10, multi:softprob as classification objective, a step size shrinkage of 0.3, a minimal child weight of 1 and all other parameters set as default. We investigated tree depths between 6 and 12 and boostings between 8 and 12 on Bernard's database, using the baseline validation scenario. Best results were reached for value 10 for both and all results are presented using these values.

3 Validation results

The classifiers were cross-validated with a one-leave-out procedure. Tested sequences were either in the learning set (*baseline* scenario) sequences or removed from it (*one-against-all* scenario). The first scenario tests the most favourable case where the taxonomy of known genomes must be recovered, and the second the more interesting *de-novo* case where the genome is unknown.

For all experiments, prediction were made by considering for each read x and each level l, the vector of predictions v_x^l on possible taxons at this level. Spurious or ubiquitous reads were detected using a minimum threshold θ=10% for the value of $(max(v_x^l) - \mu(v_x^l))/\mu(v_x^l)$ on the set of classes at level l. Prediction is done on a read-by-read basis and at any level, all prediction results that relate to at least 10% of the reads in the test set are retained.

We evaluated assignment accuracy across our different scenarios and databases.

The results of the baseline scenario show a very high level of recall of the learning sequences, where 100% of the sequences are recognised at all levels,

whether or not they have errors for the Bernard database, and 100% also for the supported database for the domain and some misclassifications for the other levels (0.9% without errors, 2.6% with errors). Figure 1 shows global results with the leave-one-out validation procedure.

For simulated *de-novo* organisms, read attributions results 1 shows a +11.5% gain for having each time two other family representatives in database. In leave-one-out scenarios, sampling more our reference genomes during the database creation harmed the ability for our classifier to identify with success reads 1 ; our results on both databases displays respectively a overall 0.79% and 3.56% read attribution performance decrease for Bernard's and supported databases.

Database construction took respectively 15 and 17 minutes for Bernard's and supported databases with 4 cores and 50 Gb of RAM. Forty minutes were needed to learn all models, so a total of 1 hour to prepare the prediction. Prediction took an average of 12 seconds per sample (single core,6 GB of RAM). Across all our experiments, phylum is the taxonomic unit that seems the hardest to capture, as displayed in table 1. Moreover, phyla are quite subject to changes [11], but in this model we use them to reduce the number of classes at lower ranks.

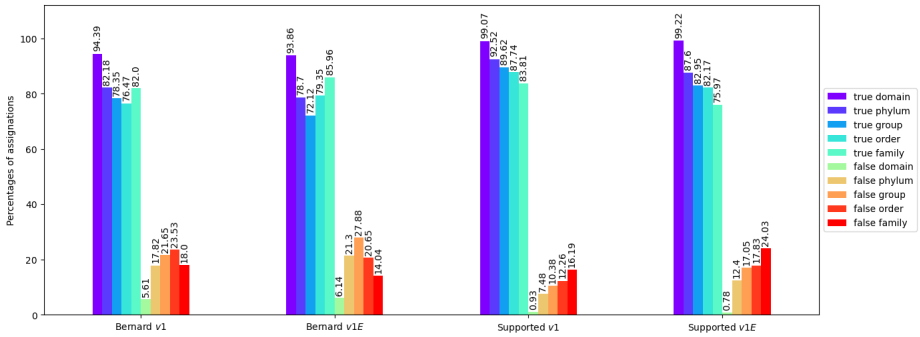

Fig. 1. Global prediction leave-one-out results, for Bernard's and supported databases and $v1$ sampling, with ($v1E$) and without errors.

Having family-level relatives does not entirely solves this issue (cf Figure 2), at least with our parameters ; however, error rates are significantly lower. Errors are concentrated within specific clades. From our experiments, we can learn additional rules to suppress those errors.

The groups *Clostridia* and *Fusobacteria* show very similar kmer composition at the genome scale, but belong to different phyla (*Firmicutes* and *Fusobacteria* respectively). The cross-validation confusion matrices show that only 0.27% of the *Firmicutes* reads are confused with *Fusobacteria*. However, for new genomes, 43.55% of the Fusobacteria reads confounded with Firmicutes (see Figure 2). Having such similar signatures across different clades means we can't rely solely on single read attribution for binning metagenomics samples.

Table 1. Accuracy of assignations of individual reads at different levels, for splits that can be determined (splits which have relatives in train set). The Global lines gives the cumulated loss through all levels.

	Bernard's database						Supported database					
	baseline			leave-one-out			baseline			leave-one-out		
	v1	v2	v1E	v1	v1E	v2	v1	v1E	v2	v1	v1E	v2
Domain	99.6	99.9	99.3	94.2	93.8	94.1	99.7	99.9	99.9	99.3	99.2	99.4
Phylum	99.2	99.2	96.6	82.8	82.4	82.8	96.4	93.4	97.9	88.2	83.6	84.5
Group	99.7	99.7	98.5	90.7	91.9	90.7	99.3	98.7	99.5	93.0	91.4	89.0
Order	99.7	99.5	98.4	92.9	91.1	92.4	99.1	98.5	99.5	95.4	95.1	97.4
Family	99.0	98.7	97.0	90.1	89.6	89.4	99.2	98.3	99.7	90.9	88.8	92.1
Global	97.3	97.1	90.1	59.2	57.9	58.4	93.7	89.1	96.5	70.7	64.0	67.1

To put our results in perspective, we compared the accuracy of PhyloPhytia and our method on fragments of size 10,000. For known genome samples (baseline), we obtained at all levels for sensitivity the range 99.4-100% (PhyloPhytia 97.7-98.5%). and for specificity 100% (PhyloPhytia 83.7-98.0%). For unknown genomes (leave-one-out), we obtained at all levels 90.0-99.9% sensitivity (PhyloPhytia 80.0-87.9%) and 98.5-99.8% specificity (PhyloPhytia 81.7-92.1%).

Acknowledgements We acknowledge the GenOuest bioinformatics core facility (https://www.genouest.org) for providing the computing infrastructure.

References

1. Bernard, G., Ragan, M.A., Chan, C.X.: Recapitulating phylogenies using k-mers: from trees to networks. F1000Research **5** (Dec 2016)
2. Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. In: Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining. pp. 785–794. KDD '16 (2016)
3. Delahaye, C., Nicolas, J.: Sequencing DNA with nanopores: Troubles and biases. PLOS ONE **16**(10), e0257521 (Oct 2021)
4. Déraspe, M., Raymond, F., et al.: Phenetic Comparison of Prokaryotic Genomes Using k-mers. Molecular Biology and Evolution **34**(10), 2716–2729 (2017)
5. Fiannaca, A., La Paglia, L., et al.: Deep learning models for bacteria taxonomic classification of metagenomic data. BMC Bioinformatics **19**(7), 198 (2018)
6. Grantham, R., Gautier, C., Gouy, M., Mercier, R., Pavé, A.: Codon catalog usage and the genome hypothesis. Nucleic Acids Research **8**(1), r49–r62 (Jan 1980)
7. Karlin, S., Mrázek, J., Campbell, A.M.: Compositional biases of bacterial genomes and evolutionary implications. Journal of Bacteriology **179**(12), 3899–3913 (1997)
8. Khaledi, A., Weimann, A., et al.: Predicting antimicrobial resistance in Pseudomonas aeruginosa with machine learning-enabled molecular diagnostics. EMBO molecular medicine **12**(3), e10264 (2020)
9. McHardy, A.C., Martín, H.G., et al.: Accurate phylogenetic classification of variable-length DNA fragments. Nature Methods **4**(1), 63–72 (2007)
10. Panyukov, V.V., Kiselev, S.S., Ozoline, O.N.: Unique k-mers as Strain-Specific Barcodes for Phylogenetic Analysis and Natural Microbiome Profiling. International Journal of Molecular Sciences **21**(3), 944 (Jan 2020)

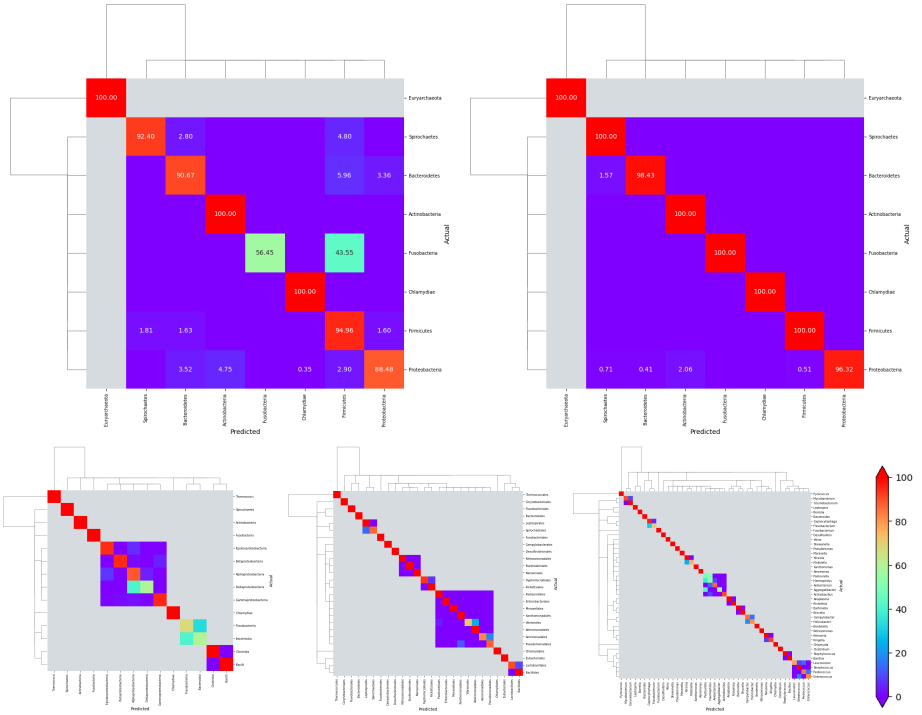

Fig. 2. Top: individual read assignations at phylum level for leave-one-out and baseline scenarios (resp. left and right) with the supported database. The dendrograms show taxonomy at higher levels. Bottom: from left to right, the leave-one-out results on the supported database for the group, order and family levels. Grey areas are the prediction cases our algorithm will not encounter due to its hierarchical approach.

11. Parks, D.H., Chuvochina, M., et al.: A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nature Biotechnology **36**(10), 996–1004 (2018)
12. Patil, K.R., Roune, L., McHardy, A.C.: The PhyloPythiaS Web Server for Taxonomic Assignment of Metagenome Sequences. PLOS ONE **7**(6), e38581 (Jun 2012)
13. Rosen, G.L., Reichenberger, E.R., Rosenfeld, A.M.: NBC: the Naïve Bayes Classification tool webserver for taxonomic classification of metagenomic reads. Bioinformatics **27**(1), 127–129 (11 2010)
14. Vervier, K., Mahé, P., Tournoud, M., Veyrieras, J.B., Vert, J.P.: Large-scale machine learning for metagenomics sequence classification. Bioinformatics **32**(7), 1023–1032 (Apr 2016)
15. Weimann, A., Mooren, K., et al.: From Genomes to Phenotypes: Traitar, the Microbial Trait Analyzer. mSystems **1**(6), e00101–16 (2016)
16. Wood, D.E., Salzberg, S.L.: Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology **15**(3), R46 (Mar 2014)