Binning microbial long reads: learning bacterial family signatures with XGBoost

Siegfried Dubois[1] and Jacques Nicolas[1]

Univ Rennes1, INRIA, Rennes, France

Abstract. Next-Generation Sequencing technologies such as MinION sequencers enabled the possibility for on-field sequencing, but as of now still lacks software support. We have developped a quantitative approach, based upon small words (k-mers) and a straight-forward alphabet of nucleic acids (A,T,C,G), allowing to bin reads of a sample inside a user-defined classification. Relying on model of regression trees boosting, thanks to the XGBoost library, our software is seeking to serve a coherent tool suite which can be integrated inside a pipeline ; covering encoded databases creation, error processing, creation of reports, quality analysis of predictions and user-friendly parameters interface, the whole project is pushing towards being a fast alternative to bacteria families determination, in a sense of rejecting numerous upstream false guesses before a strain identification with more powerful but slower algorithms. Determination relies on three keystones : a divide and conquer approach which considers each level of classifications with regard to previous predictions, a data filtering shifting all low significance subreads from predictions, and a formation of splitting rules based on quantitative signature heuristics.
Source code, installation instructions and data are freely available here: https://github.com/Tharos-ux/wisp

Keywords: Machine learning · Binning · Metagenomics · Long reads · 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 [13], 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 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 [14], 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 [2,4,10]. 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).

We propose to mitigate these limitations by using a two-step identification process. The first step is to identify a genomic sequence down to the taxonomic rank of families using compositional analysis. The second step seeks to refine the identification based on the presence of kmers with gaps (spaced seeds), assuming the family determined. This paper concern the first step, the second one being the subject of a method called ORI (Oxford nanopore Reads Identification) [12] for identification down to the strain level from long reads.

2 Methods

2.1 Protocol

Our first step is to create a database we can use with our algorithm. We compute kmers counts in both ways ($5' \rightarrow 3'$ and $3' \rightarrow 5'$) for all of our reference genomes, sampled in long reads (10.000 bp), with a seed of 11111 for domain and phylum levels, 1111 for group, order and family, a shift of 1 between all kmers, and sampled 50 times for domain and family, 100 times for phylum, group and order for database type $v1$; whereas we sampled 100 times for domain and 500 times for phylum, group, order and family in database type $v2$.

- **Guillaume's dataset** is issued from *143 Prokaryote genomes* [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 sub-collection of the NCBI public database *refseq*, made with random sampling of 3 reference genomes for all families counting at least 3 representatives : it is made of 129 archaeal and bacterial genomes, divided under 2 domains, 8 phyla, 14 groups, 26 orders and 43 families.

We started investigating on database parameters impact on classification, but we settled on similar results as initial PhyloPhyta's experiments [9]. As of

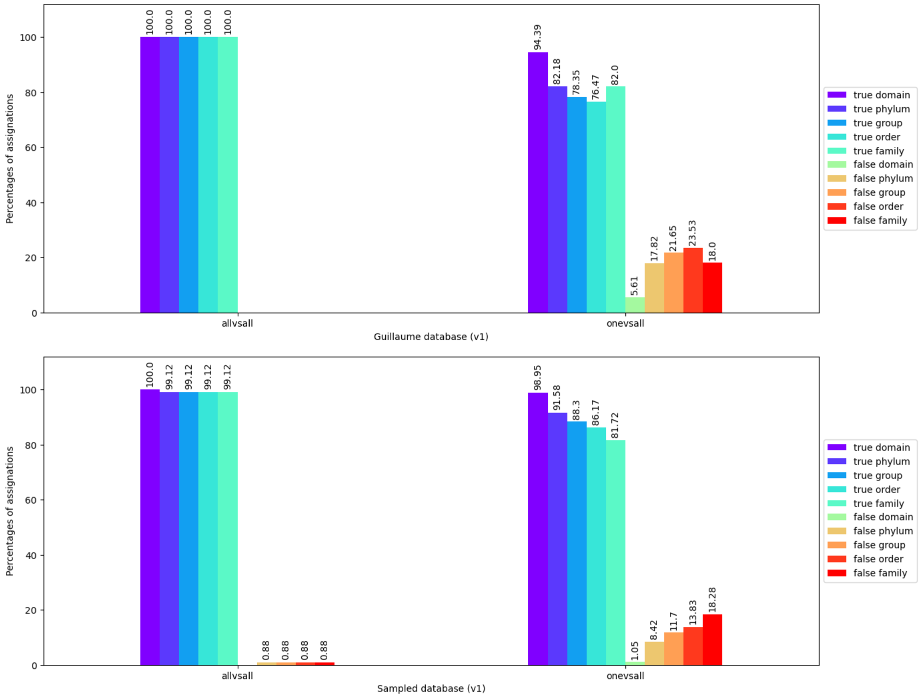

Fig. 1. Global classification results in baseline and leave-one-out scenarios, for Guillaume's and supported databases at resp. top and bottom positions. Those results are both issued form the v1 style of both databases.

our unknown sequence, we sampled randomly 400 fragments of 10000 bp, with the same seed and step, and tried to assign each of those fragments to a domain, then to a phyla knowing determined domain, down to the family level. For all experiments described here, exploration threshold (minimum number of reads to consider investigating lower level) was set to 10%, and filtering of non-significant fragments was set to reject all reads where $max(v_x) - \mu(v_x) \leq 0.1$, v_x being the vector of predictions for the read.

3 Results

We validated our classifier with two complementary steps. Firstly, we simulated a *baseline* scenario, where each of the reference genomes was evaluated against the full database. This experiment gives us the best results we can aim for when we will do leave-one-out simulations. Secondly, we created a *one-against-all* scenario, where for each reference genome, we compute it as a unknown sample, removing all its references from the training database, re-calculate the model from the remaining data, and compute the prediction against the model.

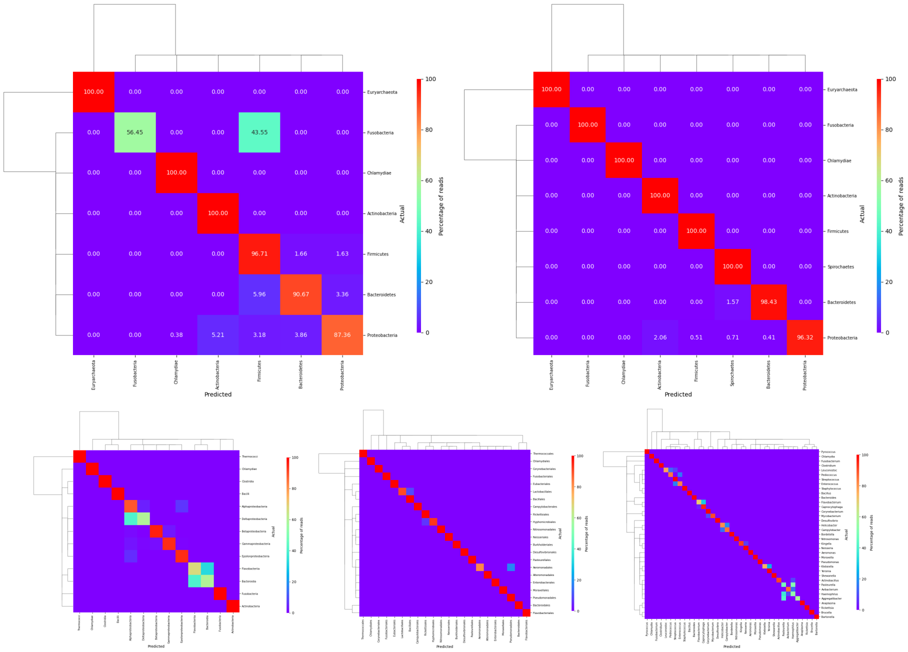

Fig. 2. Individual read assignations at phylum level for leave-one-out and baseline scenarios (resp. left and right) with the supported database. Dendrogram is the taxonomy used at higher levels. Bottom figures are results in leave-one-out, with the same database, on resp. from left to right group, order and family levels.

We aim by this method emulate the determination of *de-novo* taxa, species that are not in database yet and which we try to attribute to the closest relative.

We evaluated assignment accuracy across our different scenarios and databases. For simulated *de-novo* organisms, read attributions results 1 shows a +9.58% 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 2.94% read attribution performance decrease for Guillaume's database and supported one.

4 Discussion

Across all our experiments, phylum is the taxonomic unit that seems the hardest to capture, as displayed in 1. Moreover, phyla are quite subject to changes [11], but in this model we use them as splitting points, to reduce the number of classes at lower ranks. Having family-level relatives does not entirely solves this issue 2, at least with our parameters ; however, error rates are significantly lower.

Table 1. Accurate assignations percentages of individual reads at different levels, for splits that can be determined (splits which has relatives in train set). Global category depicts the loss we accumulate through our iterations.

	Guillaume's database				Supported database			
	baseline		leave-one-out		baseline		leave-one-out	
	v1	v2	v1	v2	v1	v2	v1	v2
Domain	99.63	99.89	94.15	94.06	99.70	99.92	99.24	99.31
Phylum	99.24	99.24	82.77	82.77	96.37	97.92	87.26	82.61
Group	99.71	99.71	90.74	90.74	99.29	99.54	92.25	90.51
Order	99.68	99.51	92.91	92.41	99.10	99.46	95.92	97.19
Family	99.02	98.69	90.05	89.42	99.15	99.65	89.69	91.18
Global	97.31	97.06	59.16	58.37	93.74	96.54	68.74	65.80

Errors are condensed inside specific clades ; from our experiments, we can learn additional rules to suppresss those errors.

As our method relies on binary trees down to family level, we do not have as of now a early stopping approach for samples that cannot be assign due to lacking of relatives in database. User may define a rank he want to stop at, but all samples that could not be determined will be too specific anyways.

Groups *Clostridia* and *Fusobacteria* display very similar results as of kmer relative composition on genome scale, but are from different phyla (resp. *Firmicutes* and *Fusobacteria*) and as our cross-validation confusion matrices points out, we happen to have only 0.27% of *Firmicutes* reads mistaken to *Fusobacteria*. Even if this is not a problem in this particular case, having such similar signatures across different clades mean we can't rely solely on single read attribution by composition for binning metagenomics samples. ORI relies on a read/strain matrix, and thanks to ASP, aims to select the minimum of strains explaining the maximum of reads. A future improvement would be to implement our software at its interface, to apply this strategy at a family level and thus feed the data into ORI.

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** (2016)
2. Bernard, G., Ragan, M.A., Chan, C.X.: Recapitulating phylogenies using *k*-mers: from trees to networks. Tech. Rep. 5:2789, F1000Research (Dec 2016), https://f1000research.com/articles/5-2789, type: article
3. Delahaye, C., Nicolas, J.: Sequencing DNA with nanopores: Troubles and biases. PLOS ONE **16**(10), e0257521 (Oct 2021). https://doi.org/10.1371/journal.pone.0257521, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0257521

4. Déraspe, M., Raymond, F., Boisvert, S., Culley, A., Roy, P.H., Laviolette, F., Corbeil, J.: Phenetic Comparison of Prokaryotic Genomes Using k-mers. Molecular Biology and Evolution **34**(10), 2716–2729 (Oct 2017). https://doi.org/10.1093/molbev/msx200
5. Fiannaca, A., La Paglia, L., La Rosa, M., Lo Bosco, G., Renda, G., Rizzo, R., Gaglio, S., Urso, A.: Deep learning models for bacteria taxonomic classification of metagenomic data. BMC Bioinformatics **19**(7), 198 (Jul 2018). https://doi.org/10.1186/s12859-018-2182-6, https://doi.org/10.1186/s12859-018-2182-6
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), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC327256/
7. Karlin, S., Mrázek, J., Campbell, A.M.: Compositional biases of bacterial genomes and evolutionary implications. Journal of Bacteriology **179**(12), 3899–3913 (Jun 1997). https://doi.org/10.1128/jb.179.12.3899-3913.1997
8. Khaledi, A., Weimann, A., Schniederjans, M., Asgari, E., Kuo, T.H., Oliver, A., Cabot, G., Kola, A., Gastmeier, P., Hogardt, M., Jonas, D., Mofrad, M.R., Bremges, A., McHardy, A.C., Häussler, S.: Predicting antimicrobial resistance in Pseudomonas aeruginosa with machine learning-enabled molecular diagnostics. EMBO molecular medicine **12**(3), e10264 (Mar 2020). https://doi.org/10.15252/emmm.201910264
9. McHardy, A.C., Martín, H.G., Tsirigos, A., Hugenholtz, P., Rigoutsos, I.: Accurate phylogenetic classification of variable-length DNA fragments. Nature Methods **4**(1), 63–72 (Jan 2007). https://doi.org/10.1038/nmeth976, https://www.nature.com/articles/nmeth976
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). https://doi.org/10.3390/ijms21030944, https://www.mdpi.com/1422-0067/21/3/944
11. Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.A., Hugenholtz, P.: A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nature Biotechnology **36**(10), 996–1004 (Nov 2018). https://doi.org/10.1038/nbt.4229, https://www.nature.com/articles/nbt.4229
12. Siekaniec, G.: Identification of strains of a bacterial species from long reads. phdthesis, Université Rennes 1 (Dec 2021), https://tel.archives-ouvertes.fr/tel-03510672
13. Weimann, A., Mooren, K., Frank, J., Pope, P.B., Bremges, A., McHardy, A.C.: From Genomes to Phenotypes: Traitar, the Microbial Trait Analyzer. mSystems **1**(6), e00101–16 (Dec 2016). https://doi.org/10.1128/mSystems.00101-16, https://journals.asm.org/doi/10.1128/mSystems.00101-16
14. Wood, D.E., Salzberg, S.L.: Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology **15**(3), R46 (Mar 2014). https://doi.org/10.1186/gb-2014-15-3-r46, https://doi.org/10.1186/gb-2014-15-3-r46