decOM: Similarity-based microbial source tracking of ancient oral samples using k-mer-based methods

Camila Duitama
Sequence Bioinformatics
Department of Computational Biology
Institut Pasteur, Université Paris Cité
Paris, France
cduitama@pasteur.fr

Riccardo Vicedomini
Sequence Bioinformatics
Department of Computational Biology
Institut Pasteur
Paris, France
rvicedom@pasteur.fr

Nicolás Rascovan
Microbial Paleogenomics
Department of Genomes and Genetics
Institut Pasteur
Paris, France
nicolas.rascovan@pasteur.fr

Rayan Chikhi
Sequence Bioinformatics
Department of Computational Biology
Institut Pasteur
Paris, France
rayan.chikhi@pasteur.fr

Téo Lemane
Université de Rennes
INRIA, CNRS, IRISA
Rennes, France
teo.lemane@inria.fr

Hugues Richard
Bioinformatics unit (MF1)
Robert Koch Institute
Berlin, Germany
RichardH@rki.de

Abstract

The analysis of ancient oral metagenomes from archaeological samples is largely confounded by contaminant DNA.We developed a novel method called decOM for Microbial Source Tracking and classification of ancient and modern metagenomic samples using k-mer matrices. decOM outperforms two state-of-the-art machine learning methods for source tracking, FEAST and mSourceTracker. We anticipate that decOM will be a valuable tool for ancient metagenomic studies.

Background Analysing ancient DNA (aDNA) is particularly challenging due to contamination with environmental and modern DNA sequences. The task of Microbial Source Tracking (MST) is to quantify the proportion of different microbial environments (sources) in a target microbial community (sink) [3]. MST allows to quantify contamination [1] in metagenomics sequencing data and to predict the metadata class of a given microbial sample. Two of the most widely-used methods today for MST in metagenomic data are metagenomic-SourceTracker (mSourceTracker)[2] and FEAST [3]. In this study we developed a novel reference-free and k-mer-based method called decOM to perform MST and environmental type prediction of a given microbial sample.

Implementation Dental calculus is one of the richest sources of aDNA in the archaeological record [6]. The microbial composition of a given aOral sample isolated from dental calculus has been modelled as a mixture of DNA originating from dental plaque, skin bacteria, soil and other sources [4, 5]. For this reason, we gathered 360 metagenomic data sets of diverse environment types to build a k-mer matrix of sources: ancient oral (aOral), sediment/soil, skin, or modern oral (mOral).

Workshop Learning Meaningful Representations of Life (LMRL) at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

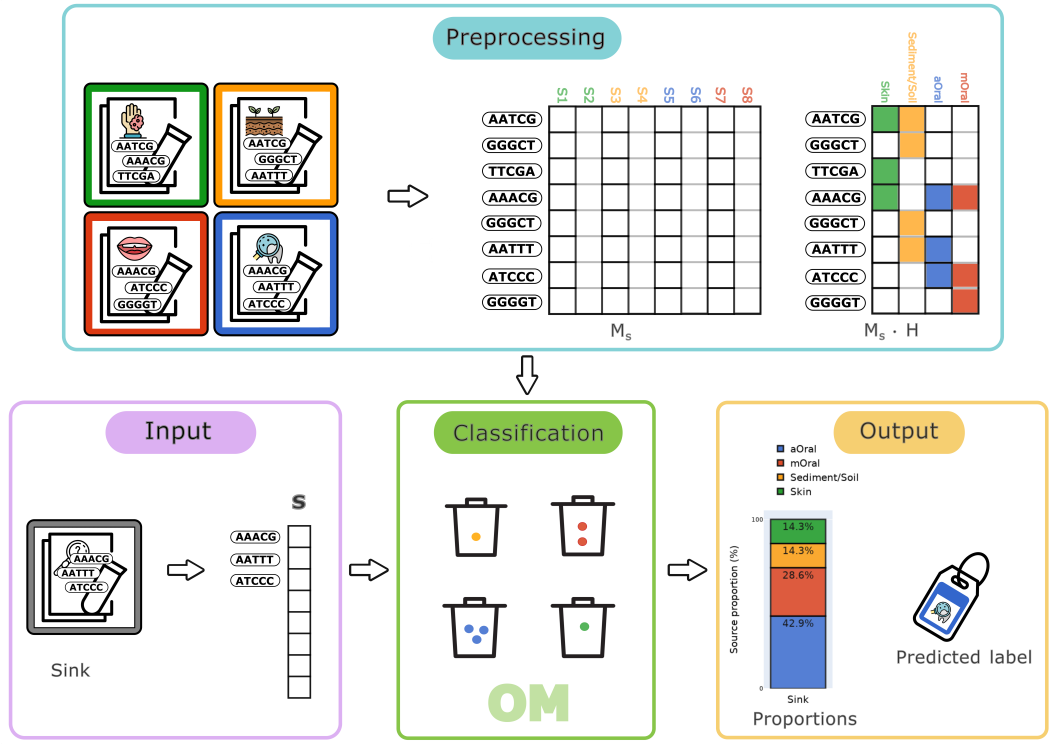

Figure 1: decOM preprocesses an input k-mer matrix of aOral metagenomic samples and its possible contaminants, divides it into sinks and sources and then estimates and outputs the proportions of each source environment in the sink. The core idea in the classification step is that if a k-mer is present in the sink s represented by the vector $\mathbf{m}^{(s)}$, and in the source vector $\mathbf{m}^{(j)}$ with environment label l_j, then a ball is added to the bin with label l_j (Ex: K-mer AAACG is present in the input sink S and in source S1 labelled as skin, S5 labelled as aOral and S7 labelled as mOral, hence one ball is added to the bin of skin, aOral and mOral respectively). After every entry in the the sink vector is compared against every entry of every vector in the sources, decOM outputs the estimated environment proportions and the hard label assigned to the sink s is that of the environment with the highest contribution.

Table 1: Performance in validation set.

Method	Recall
decOM	**0.8654**
FEAST	0.6692
metaSourceTracker	0.6346

At a high level, decOM represents every sink the user inputs as a presence/absence vector and then estimates and outputs the proportions of each source environment in the sink as well as a hard label. Such estimation is achieved by comparing every k-mer in the sink vector against every other entry of every vector in the matrix of sources. Our method was implemented in Python 3.6 as a conda package and installation instructions are available in a GitHub repository `https://github.com/CamilaDuitama/decOM`. Figure 1 provides a graphical representation of our method.

Results We prove that decOM outperforms FEAST and mSourceTracker, two of the most commonly used ML methods for MST in two cross-validation settings and with an external validation set. In a leave-one-out cross-validation, decOM outperforms both mSourceTracker (+3% Accuracy, +8% Precision, +3% Recall, +5% F1 score) and FEAST (+19% Accuracy, +37% Precision, +12% Recall, +33% F1 score). In a stratified 5-fold cross-validation, decOM outperforms both methods in each of the five sink/sources folds for performance metrics such as Accuracy, Precision, Recall and F1 Score and when metrics are averaged across groups (see Table 2 in Supplementary File). Finally, when tested using an independent validation set of 254 aOral samples, decOM outperforms mSourceTracker and FEAST by classifying most of the samples as aOral (see Table 1)

Discussion We proposed and evaluated decOM as a tool to predict the metadata class of a given metagenomic sample by using a MST framework, in order to help paleogeneticists better assess the source content of their ancient samples. In principle, decOM can also help determine the composition of any other microbial community (not necessarily ancient or of oral origin). It would be interesting to study the impact on the classification performance of varying the hyperparameters for the construction of the k-mers matrix, such as the size of the k-mers, minimum recurrence or minimum abundance. We anticipate that the incorporation of decOM into paleogenomic analyses will prevent erroneous results and help identify contaminated metagenomic samples and ensure their validity.

Broader Impact

Any paleogeneticist can freely benefit from this research by using decOM and contributing to its improvement, as it is an open source software. The field of paleogenomics has been heavily criticised for reproducing colonial and extractivist practices. Despite not being able to influence in the monetary barrier local communities face to extract and sequence their own samples, we hope that with an open source code that does not require many computational resources, people from all over the world will be able to evaluate their local samples of ancient dental calculus, regardless of their origin and without any significant monetary or technological impediments.

If decOM is used as the only method to assess the contamination of ancient samples, there could be an inefficient use of resources if scientist discard samples that are useful (false negatives) or use samples that are actually highly contaminated (false positives).

Abbreviations

aDNA ancient DNA.

aOral ancient oral.

mOral modern oral.

mSourceTracker metagenomic-SourceTracker.

MST Microbial Source Tracking.

References

[1] KNIGHTS, Dan ; KUCZYNSKI, Justin ; CHARLSON, Emily S. ; ZANEVELD, Jesse ; MOZER, Michael C. ; COLLMAN, Ronald G. ; BUSHMAN, Frederic D. ; KNIGHT, Rob ; KELLEY, Scott T.: Bayesian community-wide culture-independent microbial source tracking. In: *Nature methods* 8 (2011), Nr. 9, S. 761–763

[2] MCGHEE, Jordan J. ; RAWSON, Nick ; BAILEY, Barbara A. ; FERNANDEZ-GUERRA, Antonio ; SISK-HACKWORTH, Laura ; KELLEY, Scott T.: Meta-SourceTracker: application of Bayesian source tracking to shotgun metagenomics. In: *PeerJ* 8 (2020), S. e8783

[3] SHENHAV, Liat ; THOMPSON, Mike ; JOSEPH, Tyler A. ; BRISCOE, Leah ; FURMAN, Ori ; BOGUMIL, David ; MIZRAHI, Itzhak ; PE'ER, Itsik ; HALPERIN, Eran: FEAST: fast expectation-maximization for microbial source tracking. In: *Nature Methods* 16 (2019), Nr. 7, S. 627–632

[4] WARINNER, Christina ; RODRIGUES, Joã. M. ; VYAS, Rounak ; TRACHSEL, Christian ; SHVED, Natallia ; GROSSMANN, Jonas ; RADINI, Anita ; HANCOCK, Y ; TITO, Raul Y. ; FIDDYMENT, Sarah u. a.: Pathogens and host immunity in the ancient human oral cavity. In: *Nature genetics* 46 (2014), Nr. 4, S. 336–344

[5] ZIESEMER, Kirsten A. ; MANN, Allison E. ; SANKARANARAYANAN, Krithivasan ; SCHROEDER, Hannes ; OZGA, Andrew T. ; BRANDT, Bernd W. ; ZAURA, Egija ; WATERS-RIST, Andrea ; HOOGLAND, Menno ; SALAZAR-GARCIA, Domingo C. u. a.: Intrinsic challenges in ancient microbiome reconstruction using 16S rRNA gene amplification. In: *Scientific Reports* 5 (2015), Nr. 1, S. 1–20

[6] ZIESEMER, Kirsten A. ; RAMOS-MADRIGAL, Jazmín ; MANN, Allison E. ; BRANDT, Bernd W. ; SANKARANARAYANAN, Krithivasan ; OZGA, Andrew T. ; HOOGLAND, Menno ; HOFMAN, Courtney A. ; SALAZAR-GARCÍA, Domingo C. ; FROHLICH, Bruno u. a.: The efficacy of whole

human genome capture on ancient dental calculus and dentin. In: *American journal of physical anthropology* 168 (2019), Nr. 3, S. 496–509