BACON: Bacterial Clone Recognition from Metagenomic Sequencing Data

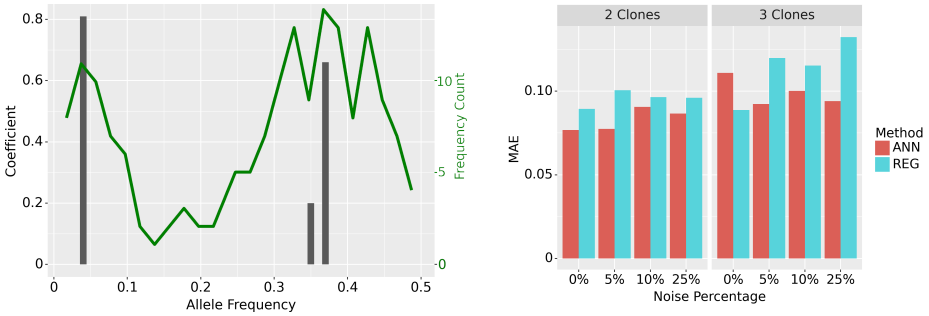
Fabian Müntefering[1 [0000−0002−4174−9662]],
Jörn Ostermann[1 [0000−0002−6743−3324]], and Jan Voges[1 [0000−0002−6080−660X]]

Institut für Informationsverarbeitung, Leibniz Universität Hannover, Appelstraße 9a, 30167 Hannover, Germany
{muenteferi,voges,ostermann}@tnt.uni-hannover.de
https://www.tnt.uni-hannover.de

The advent of metagenomic microbiome analysis has significantly advanced our understanding of microbial communities in various environments. It has been shown that the microbiome, for example in the human gut [1] or lung [2], can have a significant influence on health. Microorganisms with near-identical properties and DNA sequence are grouped under the concept of a *clone*. Important properties like pathogenicity are in some cases different between clones of the same species [3]. Hence, the analysis of the clonal composition of a microbial population is highly relevant for personalized medicine. With existing tools such as Wochenende [4] or Raspir [5], it is only possible to analyze the composition of a microbiome on the taxonomic level of species. We propose BACON (**BA**cterial **C**lone recogniti**ON**), a statistical approach to identify bacterial clones inside the population of a single species using the relative allele frequencies of single nucleotide polymorphisms (SNPs).

Bacterial clones can be distinguished by their set of alleles that are typically present in associated individuals. Given that the population of one microbial species consists of N clones C_n with a share p_n of the total population, $\sum_{i=1}^{n} p_i = 1$ must hold true. Furthermore, given that M different SNPs are present in the population, the share q_m of the total population carrying the allele A_m at the SNP locus m must be $q_m = \sum_{i=1}^{n} A_{m,n} p_n$ with $A_{m,n} = 1$ if C_n carries A_m and $A_{m,n} = 0$ otherwise. That means that for N clones, at most $|\mathcal{P}(C) \setminus \{C, \emptyset\}| = 2^N - 2$ discrete frequencies of alleles can be expected to occur. The direct analysis of q_m is not possible, as the true allele frequencies are unknown. What is available from the sequencing data, however, are L_m samples for each SNP m. Assuming the real allele frequency in the population is q_m, random sampling yields the allele $A_{m,0}$ with a probability of q_m and the allele $A_{m,1}$ with a probability of $1 - q_m$. The sampled allele frequencies $\tilde{q}_m$ can therefore be expected to form a binomial distribution around the real allele frequencies q_m. BACON decomposes the global distribution of allele frequencies into a weighted sum of binomial distributions, which can be used to determine the number and shares of all clones in the population. We developed multiple methodologies for this BACON decomposition concept, including a simple linear regression method as well as a multi-layer fully-connected end-to-end neural network.

To train the neural network, and to evaluate both methodologies, we generated synthetic data mimicking real-world SNP distributions. We evaluated both



(a) (b)

Fig. 1: **a:** Example of a distribution of allele frequencies (green) and predicted BACON coefficients (grey). **b:** Loss between the predicted and simulated clone distributions for different levels of noise, using the artificial neural network (ANN) and the linear regression model (REG).

methodologies by computing the mean absolute error (MAE) of the predicted w.r.t. the simulated distributions. Furthermore, we varied the number of clones as well as the noise that is incurred with the simulation.

In summary, we found that linear regression can already yield usable results, yielding an average MAE of 0.105 over all experiments. However, our neural network was able to reach a superior accuracy when predicting the clonal composition in a population, yielding an average MAE of 0.091. Additionally, the neural network is more resistant to noise and can also learn unexpected distributions that might occur in real-world data. Finally, with BACON we provide the first-of-its-kind methodology for computational sub-species analysis of metagenomics data.

Acknowledgements: Special thanks to Muneeb Mohammad for running extensive experiments with BACON in the context of his bachelor thesis.

References

1. Cresci, G.A. and Bawden, E. (2015), Gut Microbiome. Nutrition in Clinical Practice, 30: 734-746.
2. Li, W., Wang, B., Tan, M. et al. Analysis of sputum microbial metagenome in COPD based on exacerbation frequency and lung function: a case control study. Respir Res 23, 321 (2022).
3. Spratt, B.G. (2004). Exploring the Concept of Clonality in Bacteria. In: Woodford, N., Johnson, A.P. (eds) Genomics, Proteomics, and Clinical Bacteriology. Methods in Molecular Biology™, vol 266.
4. Rosenboom, I., Scheithauer, T. et al. Wochenende — modular and flexible alignment-based shotgun metagenome analysis. *BMC Genomics* **23**, 748 (2022).
5. Pust, MM., Tümmler, B. Identification of core and rare species in metagenome samples based on shotgun metagenomic sequencing, Fourier transforms and spectral comparisons. *ISME COMMUN.* **1**, 2 (2021).