BITS :: Call for Abstracts 2024 - Oral communication

Type	Oral communication
Session	Bioinformatics AI, Models and Tools
Title	Fragment length distribution fingerprints for microbial activity prediction via unbalanced machine-learning models
All Authors	Vincenzo Bonnici (1), Fiorenzo Tittaferrante (1), Alessia Levante (2), Valentina Bernini (2), Camilla Lazzi (2), Erasmo Neviani (2), Alessandro Dal Palù (1)

Affiliation

(1) Department of Mathematical, Physical and Computer Sciences, University of Parma, Italy
(2) Department of Food and Drug, University of Parma, Italy

Motivation

Microorganisms serve as reservoirs of biodiversity across diverse ecological habitats, with lactic acid bacteria (LAB) comprising a broad spectrum of genera capable of inhabiting various niches. LAB exhibit adaptability to a range of environmental conditions owing to the versatility of their metabolic processes [1]. The combination of genotypic and phenotypic data offers insights into the metabolic potential of industrially relevant LAB strains, indicating their suitability for biotransformation applications involving diverse substrates such as agro-industrial wastes and by-products.
Amplified Fragment Length Polymorphism (AFLP) [2] stands as a well-established PCR-based technique employed for the selective amplification of a subset of digested DNA fragments. This process generates genomic fingerprints utilized in genomic, transcriptomic, and epigenetic studies of plants [3], as well as in exploring microbial diversity [4]. Mathematically, AFLP signals correspond to recurrence distance distributions, reflecting the frequency of occurrence of specific k-mers at defined distances within a genomic sequence [5]. Such a distribution is used for detecting repetitive regions of the genomes or for extracting statistically significant words from them[6,7,8]. However, despite AFLP's efficacy in providing genomic fingerprints through wet lab analyses, there is a lack of computational models that can exploit such information in a machine-learning fashion to predict metabolic activities of microbes. The development of such models is also affected by an implicit unbalancing of the training sets for such phenomena.

Methods

In this study, our focus was on exploring the utilization of fragment length distribution profiles obtained from bacterial genomes through AFLP technology as features for computational predictive models. AFLP profiles of bacterial genomes were transformed into a 0/1 representation such that the value assigned to that specific length was set to 1 for each fragment length showing a minimum amount of amplification, 0 otherwise. Lengths from 50 to 500 were taken into account. Machine-learning models were trained with data regarding a specific metabolic activity observed for the analysed organisms.
Because of the possible unbalance of the training sets, our solution involves three different approaches depending on the level of unbalance of the training set. For balanced data sets, well-known supervised machine-learning models, whose implementation is currently publicly available via the scikit-learn python library (https://scikit-learn.org), were considered for the task. For semi-unbalanced sets, we apply downsampling to the training set in order to partially restore the balance, and then treat the data sets as a balanced one. Lastly, for highly unbalanced data sets we apply an outlier detection strategy based on feature selection. In particular, we extract the most common features among the positive strains that are uncommon for negative samples, and vice versa. For all the approaches, the power of the model in predicting the metabolic activity of organisms not included in the training set was assessed via a leave-one-out procedure.

Results

We implemented the proposed approach on two distinct benchmarks. The first benchmark utilized publicly available data sourced from the Bacterial Diversity Metadatabase (BacDIVE) (https://bacdive.dsmz.de), specifically focusing on API-50 tests used for bacterial species identification via the assessment of specific carbohydrate metabolism. In total, 509 genomes representing various species were selected, each with at least one of the 50 metabolic activities of API-50 recorded. For this benchmark, AFLP profiles were generated in silico by executing the AFLPInSilico software [9] on the corresponding genomic sequences obtained from NCBI and Patric databases.
The second benchmark comprised 141 bacterial isolates from the University of Parma Cultural Collection (UPCC) [10]. These bacteria underwent testing using the GEN III MicroPlate provided by BIOLOG, which includes 71 carbon source utilization assays and 23 chemical sensitivity assays.
For balanced and semi-balanced data sets, Random Forest was the best classification model with an average accuracy of 0.85 and an average f1-score of 0.77. Similar results were obtained for the majority of the API-50 activities in which outlier detection was performed as a classification method. Regarding the second benchmark, we obtained an average accuracy of 0.72, with 9 assays had F1-score greater than 0.7.
We further investigated the coverage of the AFLP fragments over the complete genomic sequence of the BacDIVE benchmark and found that such coverage is between 10 and 20 per cent of the total genomic sequence of each isolate. Such a result motivates future studies in

which different restriction sites of the AFLP can be exploited for capturing a wider genomic information.

Info

[1] Duar, R. M. et al (2017). Lifestyles in transition: evolution and natural history of the genus Lactobacillus. FEMS microbiology reviews, 41(Supp_1), S27–S48. https://doi.org/10.1093/femsre/fux030
[2] Vos, Pieter, et al. "AFLP: a new technique for DNA fingerprinting."NAR 23.21 (1995): 4407-4414. https://doi.org/10.1093/nar/23.21.4407.
[3] Paun, O. and Schönswetter, P., 2012. Amplified fragment length polymorphism: an invaluable fingerprinting technique for genomic, transcriptomic, and epigenetic studies. Plant DNA Fingerprinting and Barcoding: Methods Mol Biol. 2012; 862: 75–87. doi:10.1007/978-1-61779-609-8_7
[4] Bertani, G., Savo Sardaro, M.L., Neviani, E. and Lazzi, C., 2019. AFLP protocol comparison for microbial diversity fingerprinting. Journal of applied genetics, 60, pp.217-223. https://doi.org/10.1007/s13353-019-00492-0.
[5] Bonnici, V. and Manca, V., 2015. Recurrence distance distributions in computational genomics. American Journal of Bioinformatics and Computational Biology, 3(1), pp.5-23. doi:10.7726/ajbcb.2015.1002.
[6] Hackenberg, M., Previti, C., Luque-Escamilla, P.L. et al. CpGcluster: a distance-based algorithm for CpG-island detection. BMC Bioinformatics 7, 446 (2006). https://doi.org/10.1186/1471-2105-7-446.
[7] Ortuño, M., Carpena, P., Bernaola-Galván, P., Muñoz, E. and Somoza, A.M., 2002. Keyword detection in natural languages and DNA. Europhysics Letters, 57(5), p.759. DOI 10.1209/epl/i2002-00528-3
[8] Bonnici V, Franco G, Manca V. A word recurrence based algorithm to extract genomic dictionaries. BIOTECHNO 2021, The Thirteenth International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies. ISSN: 2308-4383. ISBN: 978-1-61208-859-4.
[9] Carretero-Campos, C., et al. "Improving statistical keyword detection in short texts: Entropic and clustering approaches." Physica A: Statistical Mechanics and its Applications 392.6 (2013): 1481-1492. https://doi.org/10.1016/j.physa.2012.11.052
[9] Koblihova, J., Srutova, K., Krutska, M., Klamova, H. and Machova Polakova, K., 2018. AFLP-AFLP in silico-NGS approach reveals polymorphisms in repetitive elements in the malignant genome. Plos one, 13(11), p.e0206620. doi: 10.1371/journal.pone.0206620.
[10]https://www.foodproject.unipr.it/en/research/special-projects-at-unipr/the-university-microbial-collection-university-of-parma-culture-collection/64/

filename	-

Figure

-

Availability	-

Dissemination Material

Social

-

Summary

-

Corresponding Author

Name, Surname	vincenzo, bonnici
Email	vincenzo.bonnici@unipr.it

Submitted on	26.04.2024