2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

PrimerEvalPy: A Tool for *In-Silico* Evaluation of Primers for Targeting the Microbiome

Lara Vázquez-González, Alba Regueira-Iglesias, Carlos Balsa-Castro, Nicolás Vila-Blanco, Inmaculada Tomás, and María J. Carreira, *Member, IEEE*

***Abstract*— The selection of primer pairs in a sequencing-based investigation of the microbiome can greatly influence the results, manifesting the need for a tool to analyze *in silico* its performance before the sequencing process.**

***Clinical Relevance*— PrimerEvalPy is relevant because its configurability enables the analysis of any target loci and any microbial ecosystem, detecting the best primer pairs to describe the microbiome diversity in any clinical condition.**

I. INTRODUCTION

The 16S rRNA gene is widely used in the identification of bacterial microorganisms and is considered a phylogenetic marker (1). Nevertheless, the selection of the primer pairs can greatly influence the results of any sequencing-based investigation. Testing the performance of these primer pairs against specific databases, prior to its use, can be of great interest to ensure the quality of the results, as they may vary depending on the ecosystem under study.

In this work we present PrimerEvalPy, a tool to evaluate *in silico* primers or primer pairs against specific sequences databases, testing the most used primers in the literature and pairing the best. These promising primer pairs were tested to assess their suitability in the oral cavity ecosystem, considering the non-covered taxa as well as the performance.

II. METHODS

A first iteration of the proposed Python-based package was developed to test the performance of any primer or primer pair against any sequence database. The package provides the calculation of the metric *species primer coverage*, which can be defined as the percentage of species covered, i.e., species containing at least one variant sequence captured by the primer pair evaluated.

PrimerEvalPy allows grouping sequences into different taxonomic ranks, calculating for each rank its corresponding primer coverage value. The main part of the analysis process integrated into this tool is the use of regular expressions to find the primers or primer pairs in the sequences under study. As well as the *primer coverage*, other relevant values are generated, such as the captured sequences and their mean initial/final positions.

As a practical case, PrimerEvalPy was used to test the most used primers in the literature against two 16S rRNA gene oral databases containing bacteria and archaea (2). To perform this experimentation, an existing bacterial database was improved to correct some annotation mistakes, and an archaeal database was created from complete genomes of the human oral archaeal species in the NCBI nucleotide database. PrimerEvalPy was also used to find the best primer pairs to detect oral bacteria and archaea.

III. RESULTS

A total of 456 bacterial and archaeal individual primers were analyzed with PrimerEvalPy at variant and species level (2) with some covering at the latest level a different domain than expected, as shown in Table 1. Those with coverage at species level ≥75% (148 bacterial and 65 archaeal primers) were selected to form valid primer pairs.

TABLE I. NUMBER OF PRIMERS COVERING BACTERIA, ARCHAEA, BOTH OR NONE, COMPARING THOSE DESCRIBED IN THE LITERATURE WITH THOSE CLASSIFIED BY PRIMEREVALPY

	Bacteria	***Archaea***	***Both***	***None***
Literature	356	79	21	0
PrimerEvalPy	200	64	166	26

IV. DISCUSSION & CONCLUSION

The results showed that some of the primer pairs with the highest coverage proposed by the literature did not cover many oral species that were covered by other primer pairs evaluated here. Therefore, using PrimerEvalPy we observed that not only global coverage at the variant level must be taken into account, but also coverage at the species level, and mainly an evaluation with ecosystem-specific sequences.

In conclusion, PrimerEvalPy is a useful tool that allows the analysis of *in silico* primers prior to any sequencing process, thus contributing to improve the quality and reliability of the microbial diversity results of any ecosystem.

* Research supported by the Instituto de Salud Carlos III and co-financed by FEDER Grant ISCIII/PI17/01722, and Consellería de Cultura, Educación e Ordenación Universitaria Grants ED431B 2017/029 and ED431G-2019/04.

LV-G, NV-B and MJC are with Centro de Investigación en Tecnoloxías Intelixentes (CiTIUS) and Dpt. Electrónica e Computación. AR-I, CB-C and IT are with Oral Sciences Research Group, Dept. Surgery and Medical-Surgical Specialties. Universidade de Santiago. SPAIN.

All authors are with Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), SPAIN.

Corresponding author: laram.vazquez@usc.es

REFERENCES

[1] Rajendhran J et al. Microbial phylogeny and diversity: small subunit ribosomal RNA sequence analysis and beyond. Microbiol. Res. 2011;166(2):99-110.

[2] Regueira-Iglesias A, Vázquez-González L et al. In-Silico Evaluation and Selection of the Best 16S rRNA Gene Primers for Use in Next-Generation Sequencing to Detect Oral Bacteria and Archaea. Microbiome 2021 (in review). DOI 10.21203/rs.3.rs-516961/v1