

## Databases and ontologies

# Ribopeaks: a web tool for bacterial classification through $m/z$ data from ribosomal proteins

Douglas Tomachewski<sup>1,2</sup>, Carolina Weigert Galvão<sup>2</sup>,  
Arion de Campos Júnior<sup>1</sup>, Alaine Margarete Guimarães<sup>1</sup>,  
José Carlos Ferreira da Rocha<sup>1</sup> and Rafael Mazer Etto<sup>1,2,\*</sup>

<sup>1</sup>Postgraduate Program in Applied Computing, Department of Computer Science and <sup>2</sup>Microbial Molecular Biology Laboratory, Sector of Biological and Health Sciences, State University of Ponta Grossa, 84030-900 PR, Brazil

\*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on January 19, 2018; revised on March 17, 2018; editorial decision on April 1, 2018; accepted on April 4, 2018

## Abstract

**Summary:** MALDI-TOF MS is a rapid, sensitive and economic tool for bacterial identification. Highly abundant bacterial proteins are detected by this technique, including ribosomal proteins (r-protein), and the generated mass spectra are compared with a MALDI-TOF MS spectra database. Currently, it allows mainly the classification of clinical bacteria due to the limited number of environmental bacteria included in the spectra database. We present a wide-ranging bacterium classifier tool, called Ribopeaks, which was created based on r-protein data from the Genbank. The Ribopeaks database has more than 28 500 bacterial taxonomic records. It compares the incoming  $m/z$  data from MALDI-TOF MS analysis with models stored in the Ribopeaks database created by machine learning and then taxonomically classifies the bacteria.

**Availability and implementation:** The software is available at <http://www.ribopeaks.com>.

**Contact:** mazeretto@uepg.br

**Supplementary information:** [Supplementary data](#) are available at *Bioinformatics* online.

## 1 Introduction

Matrix-Assisted Laser Desorption Ionization–Time-Of-Flight Mass Spectrometry (MALDI-TOF MS) is a promising, rapid and quite inexpensive tool for the bacterial identification, based on the generation of mass spectra from whole cells (Hsieh *et al.*, 2008). In this technique, proteins with mass range of 2–20 kDa are used to identify a particular microorganism by matching its peptide mass fingerprint (PMF) pattern with the PMFs contained in a database (Šedo *et al.*, 2011; Singhal *et al.*, 2015; Welker and Moore, 2011). Several databases have been created and demonstrated to be suitable for high-throughput routine analysis in medical laboratories, replacing the traditional biochemical or molecular techniques (Sauget *et al.*, 2017). The limitation of this technology is that identification of new isolates is possible only if the spectra database contains the PMF of specific genera/species/subspecies/strains. Currently, the MALDI Biotyper (Bruker Daltonics), the largest and most elaborated spectral database, includes more than 1800 bacterial species in approximately 4381 registers (Böhme *et al.*, 2012).

Although microbial identification is carried out with a high percentage of correct identification, limited spectral information is provided for non-clinical samples. In addition, there are only few studies applying whole-cell MALDI-TOF MS to analyze microbial diversity of environmental samples (Dieckmann *et al.*, 2005; Ferreira *et al.*, 2011; Ghyselinck *et al.*, 2011; Munoz *et al.*, 2011; Stets *et al.*, 2013). The environmental whole-cell MALDI-TOF MS is a powerful technique for biotechnological applications due to its capability to rapidly characterize microorganisms in a number of areas such as biodefense, environmental monitoring, agricultural stewardship, food quality control, occupational safety and culture typing (Demirev and Fenselau, 2008).

Based on the fact that about 60–70% of the dry weight of a microbial whole-cell is represented by ribosomal proteins (r-proteins) and that the r-proteins create a characteristic pattern in MALDI-TOF spectra (Singhal *et al.*, 2015), we created a spectral database using r-protein data from the GenBank and used it for bacterial taxonomic classification. As the GenBank sequence database offers

an open access to all annotated public available nucleotide and protein sequences (Benson *et al.*, 2017), the developed tool, called Ribopeaks, provides a wide-range bacterium classification, including clinical and environmental bacteria samples.

## 2 Approach

In order to generate a model based on r-proteins information, all of the data referring to r-proteins (30S and 50S) were downloaded from the GenBank on 06/13/2016 through the API Entrez Programming Utilities (E-utilities). A total of 2 807 341 amino acid sequences of 57 different r-protein families were downloaded utilizing the 'fasta' file format.

The amino acid sequences were converted into molecular masses using the atomic weights of the elements and considering the post-translational changes present in prokaryotic proteins as acetylation, methylation and glutamate addition (Yutin *et al.*, 2012). Subsequently, a training database was created with 28 505 taxonomic records belonging to 6936 species and 1949 genera. Paralogous r-proteins were analyzed at the specie level for determination of pattern data *m/z*.

The Weka program (Frank *et al.*, 2016), using the Naïve Bayes' algorithm, was applied to build the model, as it assumes complete independence among the different r-proteins (Supplementary Fig. S1A and Supplementary Table S1; Langley *et al.*, 1992). As the *m/z* data showed non-normal distribution, kernel density estimator was used (John and Langley, 1995). The algorithm generated a model for classification, providing a standard deviation and a group of means (kernels) for every r-protein of each specie or genus provided. Then, the outputted model, called Ribopeaks Genus or Specie model, was used to perform the classification in the web tool.

## 3 Description of software

Ribopeaks software searches for matches between the inputted r-proteins *m/z* data (query peaks) and the subjected peaks from Ribopeaks Genus or Specie model. In these models, all taxonomic records of the Ribopeaks Database are used to generate an r-protein mass map at genus or specie level through machine learning. In addition, the software can also perform the taxonomic classification at strain level. In this option, the software performs a Direct Match (DM) with the Ribopeaks Database. Results show the Deepest Taxonomic Classification (DTC) from the GenBank.

To find a match, Ribopeaks calls the function  $f(p)$  (see Equation 1).  $f(p)$  analyzes the value of each query peak ( $p$ ), its corresponding subjected peak ( $\mu$ ) in the Ribopeaks Genus or Specie model or in the Ribopeaks Database, and the mass tolerance error ( $\sigma$ ) informed by the user.

$$f(p) = \begin{cases} \text{true,} & \text{if } (p \in [(\mu - \sigma), (\mu + \sigma)]) \\ \text{false,} & \text{otherwise.} \end{cases} \quad (1)$$

Once there is a match, the software calculates the query peaks' probability of being the subjected ones from Ribopeaks Genus or Specie model or from Ribopeaks Database. Posteriorly, the ten bacterial taxa that presented more peaks in common (and less deviation) with the query data return to the user in descending order of probability.

Each result comes with four types of indicators: (i) score (indicates the confidence of the taxonomic classification); (ii) partial parity (indicates how close the query peaks are from the subject ones); (iii) total parity (indicates the coverage of matched masses with

the Ribopeaks Genus/Specie model or Ribopeaks Database); and (iv) density probability (indicates the contribution of each match to the final score) (Supplementary Fig. S1B). The Ribopeaks also generates a relative spectral graph for each taxonomic classification (Supplementary Fig. S1C).

The Ribopeaks interface is showed in Figure 1 and more details about the inputs, metrics and characteristics of the software are available in a User Manual at the Ribopeaks website (<http://www.ribopeaks.com>).

**Fig. 1.** Ribopeaks interface. Illustration of the software interface adapted to be used even in small screens such as mobile cellphones. There is a box to type or paste the MALDI-TOF *m/z* values result, three options of search (at the Genus, Specie or Strain level), and a box at which to add the mass tolerance error allowed in the current analysis. It is available at <http://www.ribopeaks.com>

The *m/z* values of 13 r-proteins from 116 environmental bacterial strains (Ziegler *et al.*, 2015) were analyzed by Ribopeaks. These data were not previously used to train the database learning model. As a result, the software correctly ranked 111 strains (95.68%) at the specie and genus levels. At the specie level, 102 strains (87.93%) were correctly classified in the first position, and nine strains (7.75%) were classified in the second to tenth positions. For the genus level, 105 strains (90.51%) were correctly classified in the first position, and six strains (5.17%) were classified in the second to tenth positions.

## Acknowledgements

We would like to thank the Brazilian Program of National Institutes of Science and Technology (INCT-FBN), National Council for Scientific and Technological Development (CNPq), Coordination for the Improvement of Higher Education Personnel (CAPES) and Fundação Araucária of the Paraná State for the financial support. We thank Microbial Molecular Biology Laboratory (LABMOM) team for technical support.

*Conflict of Interest:* none declared.

## References

- Benson, D.A. *et al.* (2017) GenBank. *Nucleic Acids Res.*, **45**, D37.
- Böhme, K. *et al.* (2012) SpectraBank: an open access tool for rapid microbial identification by MALDI-TOF MS fingerprinting. *Electrophoresis*, **33**, 2138–2142.
- Demirev, P.A. and Fenselau, C. (2008) Mass spectrometry for rapid characterization of microorganisms. *Annu. Rev. Anal. Chem.*, **1**, 71–93.
- Dieckmann, R. *et al.* (2005) Rapid screening and dereplication of bacterial isolates from marine sponges of the Sula Ridge by Intact-Cell-MALDI-TOF mass spectrometry (ICM-MS). *Appl. Microbiol. Biotechnol.*, **67**, 539–548.

- Ferreira, L. et al. (2011) MALDI-TOF mass spectrometry is a fast and reliable platform for identification and ecological studies of species from family Rhizobiaceae. *PLoS One*, **6**, e20223.
- Ghyselinck, J. et al. (2011) Evaluation of MALDI-TOF MS as a tool for high-throughput dereplication. *J. Microbiol. Methods*, **86**, 327–336.
- Hsieh, S.Y. et al. (2008) Highly efficient classification and identification of human pathogenic bacteria by MALDI-TOF MS. *Mol. Cell. Proteomics*, **7**, 448–456.
- John, G.H. and Langley, P. (1995) Estimating continuous distributions in Bayesian classifiers. In: *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., pp. 338–345.
- Langley, P. et al. (1992) An analysis of Bayesian classifiers. In: *Proceedings of the 10th National Conference on Artificial Intelligence AAAI Press/MIT Press*, pp. 223–228.
- Frank, E. et al. (2016) The WEKA Workbench. *Data Mining: Practical Machine Learning Tools and Techniques*, 4th edn. Morgan Kaufmann, MA, US.
- Munoz, R. et al. (2011) Evaluation of matrix-assisted laser desorption ionization-time of flight whole cell profiles for assessing the cultivable diversity of aerobic and moderately halophilic prokaryotes thriving in solar saltern sediments. *Syst. Appl. Microbiol.*, **34**, 69–75.
- Sauget, M. et al. (2017) Can MALDI-TOF mass spectrometry reasonably type bacteria? *Trends Microbiol.*, **25**, 447–455.
- Šedo, O. et al. (2011) Sample preparation methods for MALDI-MS profiling of bacteria. *Mass Spectrometry Rev.*, **30**, 417–434.
- Singhal, N. et al. (2015) MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis. *Front. Microbiol.*, **6**, 791.
- Stets, M.I. et al. (2013) Rapid identification of bacterial isolates from wheat roots by high resolution whole cell MALDI-TOF MS analysis. *J. Biotechnol.*, **165**, 167–174.
- Welker, M. and Moore, E.R. (2011) Applications of whole-cell matrix-assisted laser-desorption/ionization time-of-flight mass spectrometry in systematic microbiology. *Syst. Appl. Microbiol.*, **34**, 2–11.
- Yutin, N. et al. (2012) Phylogenomics of prokaryotic ribosomal proteins. *PLoS One*, **7**, e36972.
- Ziegler, D. et al. (2015) Ribosomal protein biomarkers provide root nodule bacterial identification by MALDI-TOF MS. *Appl. Microbiol. Biotechnol.*, **99**, 5547–5562.