Ribopeaks User Manual

Welcome to the Ribopeaks release 2 user's manual. This manual brings the novelties implemented in the Ribopeaks which improved its accuracy and brought clinical information about the identified bacteria. The software characteristics and concepts of each parameter/button are indicated from 1 to 20 in Figure 1. For more information, please contact us at labmom@uepg.br, or access http://sites.uepg.br/labmom/contato.php.

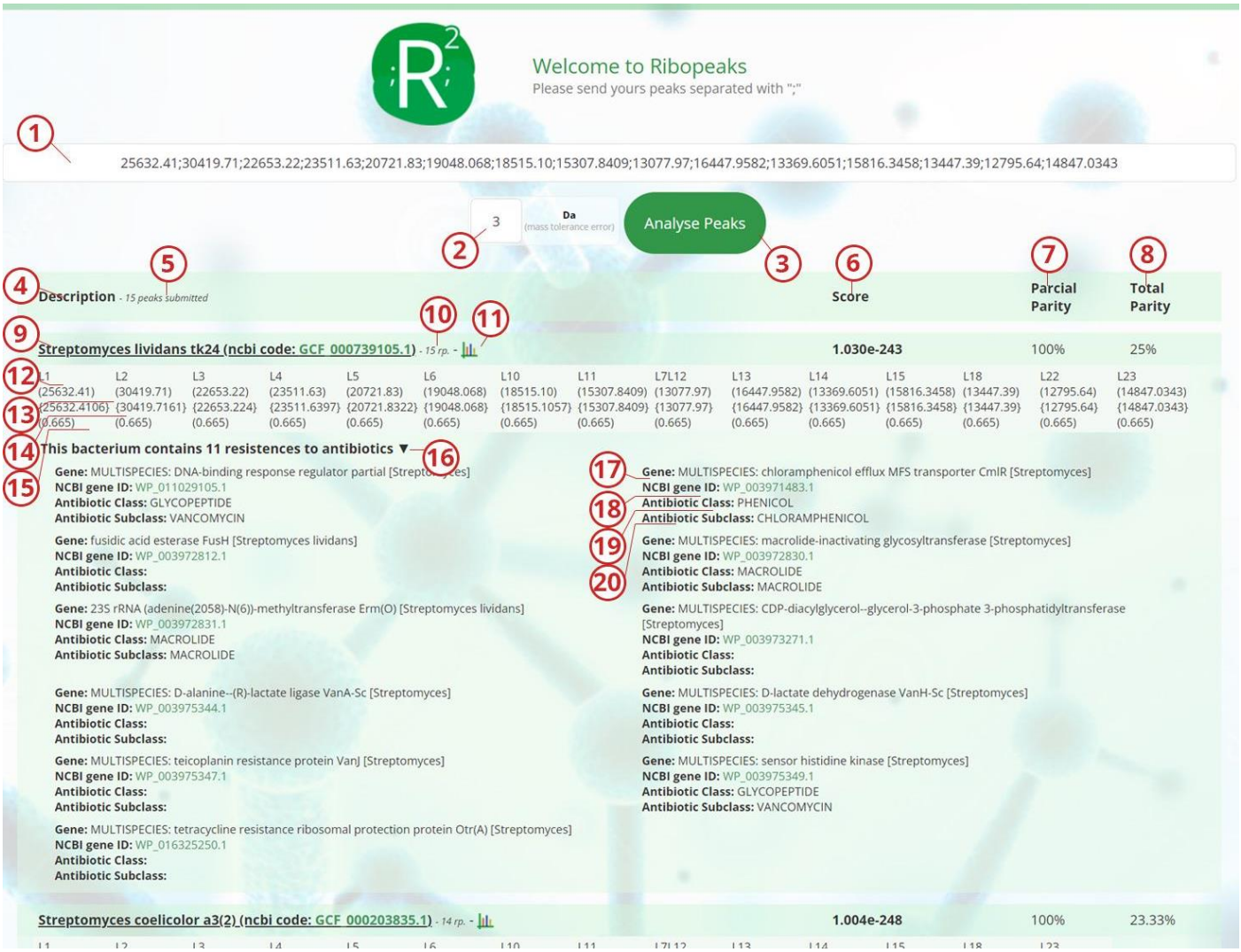

Figure 1 - A view of a Ribopeaks II example result. This example can be accessed by typing "@"+"space" at the mass value box (1) and clicking "Analyse Peaks" (6).

1. Type or paste in the indicated box the mass values of the bacterium r-protein in Da. The system will only allow numbers separated with the character "." (International standard) or "," (Brazilian standard). Between mass values, add the character ";" (semicolon). Letters, special characters and symbols, are not permitted. You can access an example analysis by typing "@"+"space" at the mass value box and clicking "Analyse Peaks". This will allow you to check the correct format of the inputted mass values (query peaks) and to see an example of the result (**Figure 1**). We recommend the addition of at least 10 (ten) r-protein peaks to reach a more confident classification.

2. Type the **mass tolerance error (Da)** allowed between the peak from the Ribopeaks Database or the Genus Model (subject peak) and the peak provided by the user (query peak). The software will consider *P* + *X* and *P* - *X*, in which *P* represents the mass of the peak from the Ribopeaks database or Genus Model and *X* the mass tolerance error provided by the user. The default was settled to 3 (three) Da, based on previous tests using the 116 strains described by Ziegler *et al.* (2015). At the mass tolerance error of ± 3 Da, high and equivalent levels of correct taxonomic classification were reached at species and genus levels. In addition, it seems that at ± 3 Da, there is less chance to accept false-positive peak matches and to reject positive matches.

3. Click on the "Analyse Peaks" button to start the analysis of the query data. The button will disappear while the mass values are being compared with the database and will reappear with the results. During analysis, the user can get information about the number of peaks added in box 1.

4. Beginning of the outcomes' description.

5. Indication of the number of peaks that was added in box 1.

6. The **Score (S)** indicates the confidence of the taxonomic classification. The Score is calculated by the following formula, in which *P(p|o)* is the *density probability* of each protein match; *P(o)* is the independent probability of each organism; and *P(p)* is the sum of the *density probability* of each protein match. This concept was based on John and Langley (1995), in which a Bayesian classifier was used. The higher the score you get, the better the taxonomic classification you reach. Note that the score is a very small number, represented by *e-N*, in which *N* is the exponential number of the *e* base.

$$S(o|p) = \frac{P(p|o)\ P(o)}{P(p)}$$

7. The **Partial Parity (Pp)** represents, in percentage, how close the query peaks are from the subject ones. Partial Parity is calculated by the following formula; in which *p* are the query peaks and μ, the matched peaks from the Ribopeaks Database or Genus/Specie Model (subject peaks).

$$Dif(p_1 \dots p_n, \mu_1 \dots \mu_n) = \frac{(\sum_{\mu_1}^{\mu_n} \mu - \sum_{p_1}^{p_n} p)}{\sum_{\mu_1}^{\mu_n} \mu}$$

$$Pp(p_1 \dots p_n, \mu_1 \dots \mu_n) = 100\ (1 - Dif)$$

8. The **Total Parity (Tp)** represents, in percentage, the coverage of the result based on the total information stored into the Ribopeaks Database or Genus/Specie Model. Total Parity value is calculated by the following formula, in which *i* represents the number of peaks found for the organism, *j* the number of total proteins stored in the model.

$$Tp(i,j) = \frac{i*100}{j}$$

9. It shows the bacterium taxonomy.

10. It indicates the number of query peaks that matched with the ones of the indicated bacterium.

11. Click on this icon to see the **interactive virtual spectral graph** of the indicated bacterium. In the spectral graph exemplified in **Figure 2**, query peaks from the Ribopeaks Database are distributed according to their *m/z* mass. In addition, both the *m/z* values and its corresponding r-protein are shown by positioning the mouse over the desired peak.

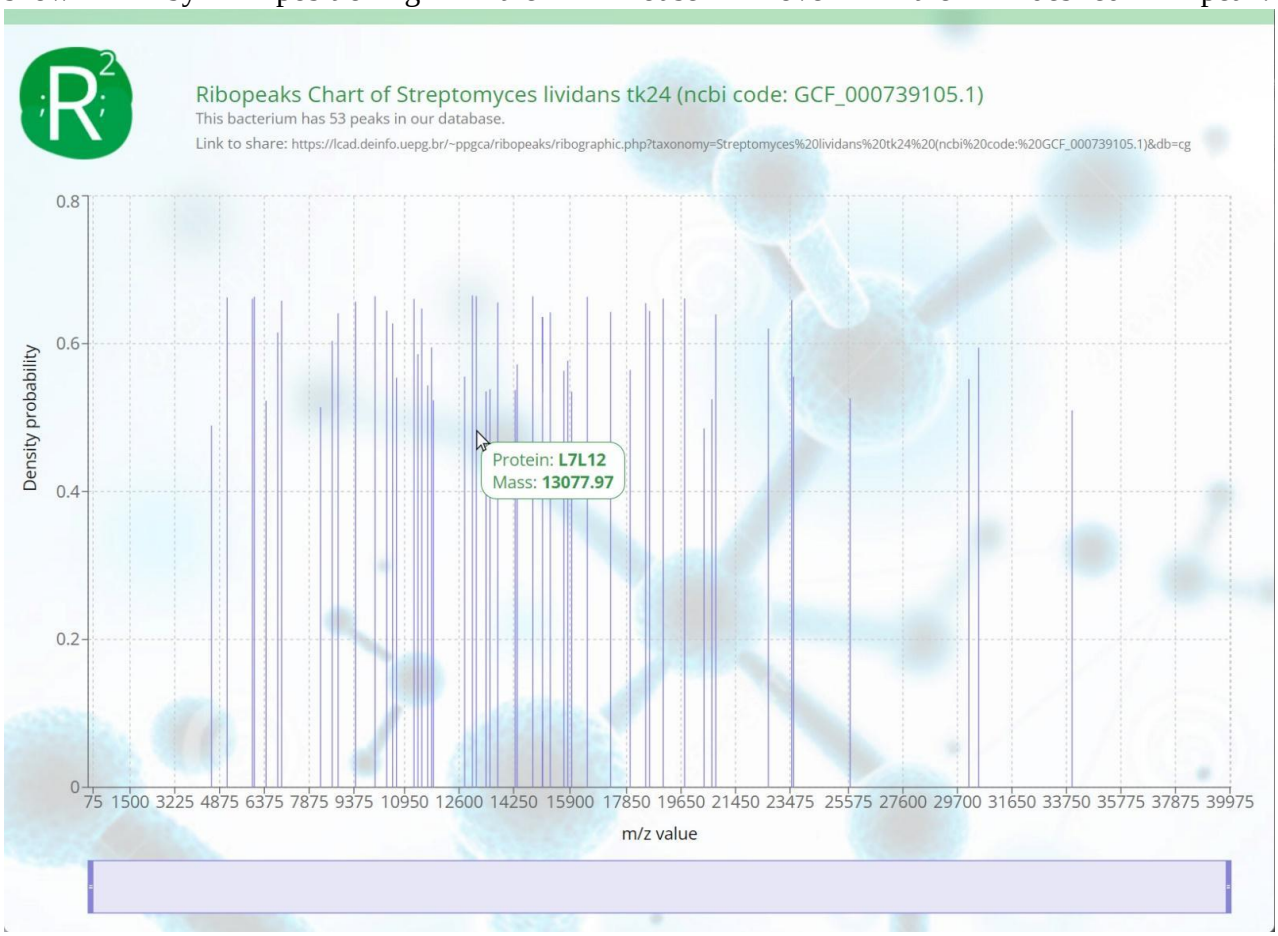

Figure 2 - Interactive virtual spectral graph of *Streptomyces lividans tk24*. The spectrum contains the peaks from the Ribopeaks II Database. The X-axis represents the *m/z* data from each peak, and the Y-axis represents the *density probability* of each peak. The bar below the X-axis allows the user to zoom in on peaks of greatest interest, to check the precise value increasing the accuracy of m/z values.

12. R-protein, from the Ribopeaks Database or Genus/Specie Model, currently used in the taxonomy. It presents an m/z value coincident with one of the query peaks. The acronyms “L” and “S” refer to the Large or Small unit of the ribosome, respectively.

13. It is one of the mass values of the query peaks informed by the user.

14. It is one of the r-proteins mass values from the Ribopeaks II Database when the analysis is settled for the Strain (DTC) taxonomy level. However, it is one of the r-proteins mass values from the Ribopeaks AI models when the analysis is generated for Specie or Genus taxonomic level.

15. **Density probability (Dp)** indicates the contribution of each match to the final score. Density probability is calculated by the following formula, in which *x* is the value to be tested; μ, the reference from the match; σ^2, the value of the standard deviation from the match; and *e*, the Euler constant.

$$Dp(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

16. Click on the arrow close to the statement "This bacterium contains X resistances to antibiotics" to access more details related to the antibiotic resistance presented by the identified bacterium. Below the arrow, there are clinical details about the bacterium resistance. If the statement and arrow do not appear in the outcome description, there are no antibiotic resistance genes related to the identified bacterium on our current Ribopeaks database. Here the user can expand and close the clinical details about the bacterium.

17. The **Gene** indicated the gene responsible for antibiotic resistance.

18. The **NCBI gene ID** informs the gene ID on NCBI and by clicking on the ID, it is possible to access the report on the website.

19. The **Antibiotic Class** informs the drug class.

20. The **Antibiotic Subclass** informs the antibiotic subclass.

References

John, G.H. and Langley, P., 1995, August. Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence* (pp. 338-345). Morgan Kaufmann Publishers Inc.

Ziegler, D., Pothier, J.F., Ardley, J., Fossou, R.K., Pflüger, V., De Meyer, S., Vogel, G., Tonolla, M., Howieson, J., Reeve, W. and Perret, X., 2015. Ribosomal protein biomarkers provide root nodule bacterial identification by MALDI-TOF MS. *Applied microbiology and biotechnology*, 99(13), pp.5547-5562.

National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2022 Apr 06]. Available from: https://www.ncbi.nlm.nih.gov/