BioNumerics®

MICROBIAL DATA ANALYSIS SOFTWARE

RDP plugin

PLUGINS

VERSION 7.6

www.applied-maths.com

Contents

NOTES

SUPPORT BY APPLIED MATHS

While the best efforts have been made in preparing this manuscript, no liability is assumed by the authors with respect to the use of the information provided.

Applied Maths will provide support to research laboratories in developing new and highly specialized applications, as well as to diagnostic laboratories where speed, efficiency and continuity are of primary importance. Our software thanks its current status for a part to the response of many customers worldwide. Please contact us if you have any problems or questions concerning the use of BioNumerics®, or suggestions for improvement, refinement or extension of the software to your specific applications:

Applied Maths NV
Keistraat 120
9830 Sint-Martens-Latem
Belgium
PHONE: +32 9 2222 100
FAX: +32 9 2222 102
E-MAIL: info@applied-maths.com
URL: http://www.applied-maths.com

Applied Maths, Inc.
11940 Jollyville Road, Suite 115N
Austin, Texas 78759
U.S.A.
PHONE: +1 512-482-9700
FAX: +1 512-482-9708
E-MAIL: info-US@applied-maths.com

LIMITATIONS ON USE

The BioNumerics® software, its plugin tools and their accompanying guides are subject to the terms and conditions outlined in the License Agreement. The support, entitlement to upgrades and the right to use the software automatically terminate if the user fails to comply with any of the statements of the License Agreement. No part of this guide may be reproduced by any means without prior written permission of the authors.

BioNumerics® uses following third-party software tools and libraries:

- The Python® 2.7.4 release from the Python Software Foundation (http://www.python.org/).
- A library for XML input and output from the Apache Software Foundation (http://www.apache.org).
- NCBI toolkit version 2.2.10 (http://www.ncbi.nlm.nih.gov/BLAST/).
- The Boost c++ libraries (http://www.boost.org/).
- Samtools for interacting with SAM / BAM files (http://www.htslib.org/download/)
- The 7-Zip command line version (7za.exe) from 7-Zip, copyright 1999-2010 Igor Pavlov. http://www.7-zip.org/
- Velvet for Windows, source code can be downloaded from http://www.applied-maths.com/download/open-source.
- Ray for Windows, source code can be downloaded from http://www.applied-maths.com/download/open-source.
- Mothur for Windows, source code can be downloaded from http://www.applied-maths.com/download/open-source.
- Cairo 2D graphics library version 1.12.14 (http://cairographics.org/).
- Crypto++ Library version 5.5.2 (http://www.cryptopp.com/).
- libSVM library for Support Vector Machines (http://www.csie.ntu.edu.tw/~cjlin/libsvm/).
- SQLite version 3.7.17 (http://www.sqlite.org/).
- Gecko engine version 21 (https://developer.mozilla.org/en-US/docs/Mozilla/Gecko).
- pymzML Python® module for high throughput bioinformatics on mass spectrometry data (https://github.com/pymzml/pymzML).
- Numpy Python® library version 1.8.1 (http://www.numpy.org/).
- BioPython Python® library version 1.64 (http://www.biopython.org/).
- PIL Python library® version 1.1.7 (http://www.pythonware.com/products/pil/).
- The SPAdes genome assembler version 3.7.1 (http://bioinf.spbau.ru/spades).

Chapter 1

Starting and setting up BioNumerics

1.1 Startup program

When BioNumerics is launched from the Windows start panel or when the BioNumerics shortcut ( ) on your computer's desktop is double-clicked, the **Startup program** is run. This program shows the *BioNumerics Startup* window (see Figure 1.1).

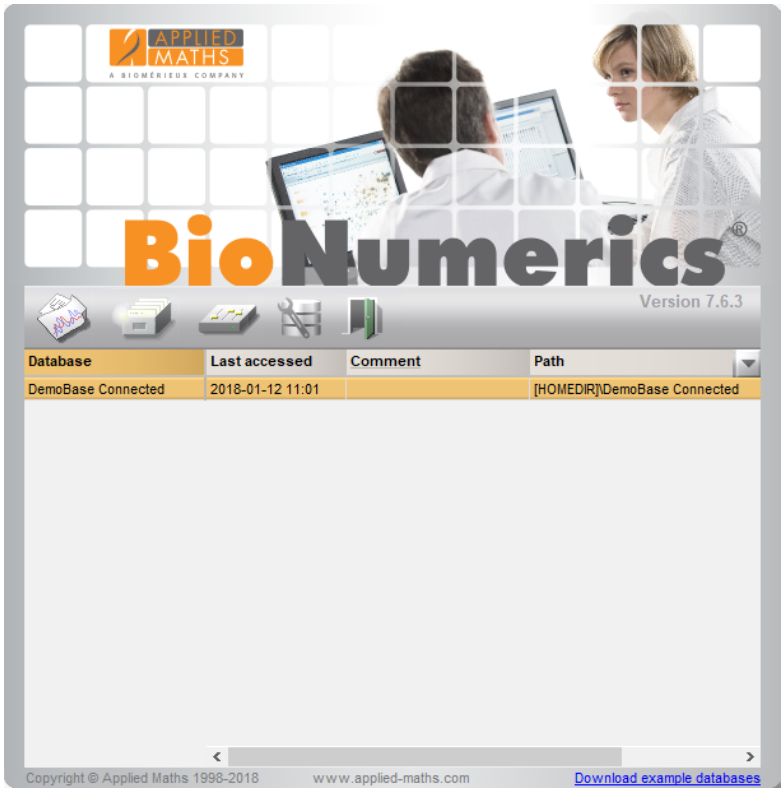

Figure 1.1: The *BioNumerics Startup* window.

A new BioNumerics database is created from the Startup program by pressing the button.

An existing database is opened in BioNumerics with or by simply double-clicking on a database name in the list.

1.2 Installing the RDP plugin

If a database is opened for the first time, the *Plugins* dialog box will appear by default (see Figure 1.2).

If the database has already been opened previously, the *Plugins* dialog box can be called from the *Main* window by selecting ***File*** > ***Install / remove plugins...*** ( ).

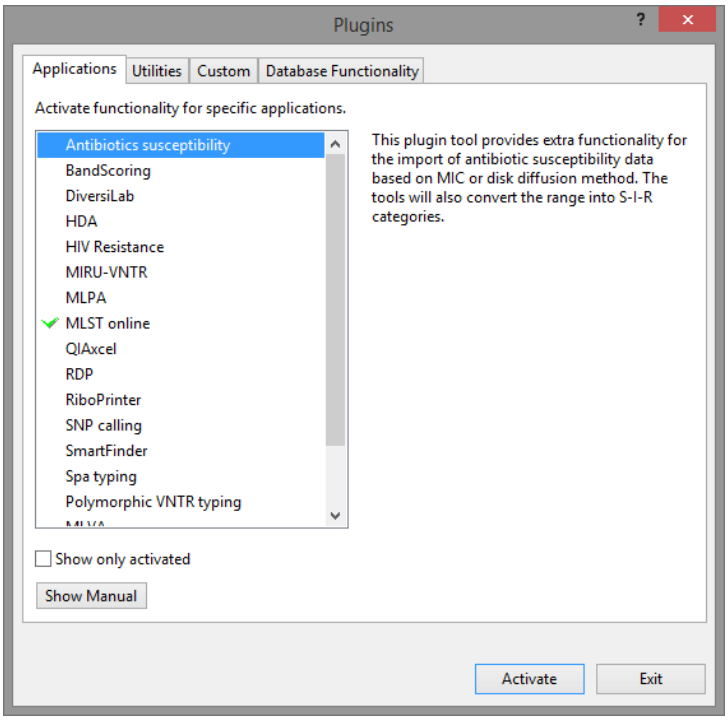

Figure 1.2: The *Plugins* dialog box.

When a particular plugin is selected from the list of plugins, a short description appears in the right panel.

A selected plugin can be installed with the <***Activate***> button. The software will ask for confirmation before installation. Some plugins depend on functionality offered by specific BioNumerics modules. If a required module is missing, the plugin cannot be installed and an error message will be generated.

Once a plugin is installed, it is marked with a green V-sign. It can be removed again with the <***Deactivate***> button.

If the selected plugin is documented, pressing <***Show Manual***> will open its manual in the *Help* window.

2.1 To install the *RDP plugin* in your database, select the *Applications tab* and select the *RDP plugin* from the list of plugins.

2.2 Press the <***Activate***> button, confirm the installation of the plugin and close the *Plugins* dialog box.

The *RDP plugin* installs menu items in the *Main* window under the menu item ***Analysis*** > ***Sequence types*** and in the *Sequence editor* window under ***RDP***.

Chapter 2

Searching the nearest 16S rRNA sequence from the Ribosomal Database Project

2.1 Introduction

The *RDP plugin* will search for each query sequence the nearest 16S rRNA sequences in the Ribosomal Database Project (http://rdp.cme.msu.edu/) by using a k-nearest-neighbor classifier. The algorithm uses a word-matching strategy not requiring alignment to determine the percentage of shared seven-character words between a query sequence and members of a database of sequences i.e. the RDP database. The RDP plugin functionality is similar to the SeqMatch functionality on the RDP website (http://rdp.cme.msu.edu/seqmatch).

2.2 Setting up the RDP analysis

The RDP matching analysis can be initiated from the *Sequence editor* window for one entry, or from the *Main* window for one or multiple database entries. When analyzing only one sequence from the *Sequence editor* window, ***RDP*** > ***Submit to RDP...*** should be used to start the analysis. When starting from the *Main* window, first the database entries should be selected. One entry can be selected by hitting the space bar or clicking the check box in front of the entry. Multiple entries can be selected by **Ctrl**-clicking or by holding the **Shift**-key to select a range of entries. The analysis is automatically launched when selecting ***Analysis*** > ***Sequence types*** > ***Submit to RDP...***.

If only one sequence type experiment is present for these entries, no further questions are asked. If multiple sequence experiments are defined for the selected entries, the *Select sequence experiment type* dialog appears where the sequence experiment type that should be used for the analysis can be selected from the drop-down list. After confirmation, the RDP analysis is automatically launched and once all data is retrieved from the RDP web service, the *RDP overview* window opens.

2.3 The RDP matching results

The *RDP overview* window (see Figure 2.1) displays the results of the RDP matching analysis.

The *RDP overview* window consist of three panels, from left to right:

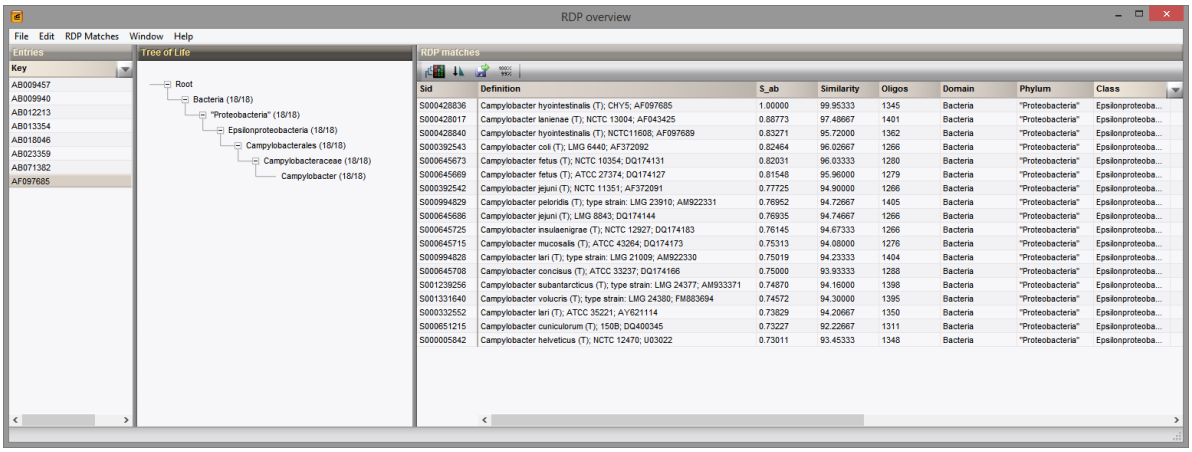

Figure 2.1: The *RDP overview* window.

- The *Entries* panel: This panel lists all the entries that were submitted for analysis. When selecting one of the entries, the panels at the right are automatically updated with the relevant information for that entry. When the analysis was launched from the *Sequence editor* window, the *Entries* panel is not present in the *RDP overview* window as only one entry was submitted for analysis.

- The *Tree of Life* panel: This panel shows the results in a hierarchical tree based on the nomenclatural or NCBI taxonomy, depending on which nomenclature was defined in the plugin settings (see 2.4). For each taxon, the number of taxon-specific RDP matches and the total number retrieved matched is indicated next to the taxon name. When selecting a taxon name, the sequences displayed at the right are updated. Only sequences belonging to the taxon at hand are displayed. To close or open all sequences below a taxon, click the <-> or <+> sign, respectively, in front of that taxon.

- The *RDP matches* panel: This panel presents the (taxon-specific) RDP matches for the sequence that was matched to the RDP database. Each match result contains, from left to right,

  - The *Sid*: This sequences identifier is used to uniquely identify the RDP sequence.
  - The *Definition*: This text field often contains organism information such as organism name, whether or not it is a type strain (T), the strain number and the sequence accession number.
  - The *S_ab* score: This score is calculated as the number of (unique) 7-base oligo mers shared between the query sequence and a given RDP sequence divided by the lowest number of unique oligos in either of the two sequences.
  - The *Similarity* score: This similarity score is the percent sequence identity over all pairwise comparable positions. Comparable positions are aligned positions containing a base in both sequences. By default, the similarity values are not calculated. To update this information, select the desired RDP matches and select ***RDP Matches*** > ***Calculate similarities*** (). This will fetch the sequences from the RDP database, next, BioNumerics will calculate the similarity scores and they are updated in the *RDP overview* window.
  - *Oligos*: This number indicates the uniquely occurring oligo mers within a given sequence. If the same oligo mer occurs more than once, they are counted only once. As such, this number only approximately reflects the sequence length.
  - Full name information fields: These fields contain the different hierarchical nomenclature ranks. Possible ranks include: superkingdom, kingdom, subkingdom, domain, superphylum, phylum, subphylum, superclass, class, subclass, infraclass, cohort, subcohort, superorder, order, suborder, infraorder, parvorder, superfamily, family, subfamily, tribe, subtribe, genus, subgenus, species group, species subgroup, species, subspecies, varietas, forma, unclassified.

Note that the rank order may differ between S_ab and the pairwise identity scores, but the top 20 S_ab scores will contain the closest sequence by pairwise identity about 95% of the time [2].

The RDP matches results can be sorted according to specific column information by highlighting the column and selecting ***RDP Matches*** > ***Sort*** (). Subsequently selecting ***RDP Matches*** > ***Sort*** () on the same column will change the sort direction between ascending and descending order.

The query sequence can easily be compared to (a subset of) the RDP matches in the *Sequence alignment* window. Thereto, first select the RDP matches that will be part of the comparison, second select ***RDP Matches*** > ***Open in alignment editor*** (). This will open the *Sequence alignment* window, where all sequences are imported, aligned and a clustering based on the sequence alignment similarity values is calculated. Apart from the pre-calculated analysis, all functionality in the *Sequence alignment* window is available for detailed analysis of the query sequence and the RDP match sequences.

This RDP plugin is very well suited for identification projects, and as such, there may be an interest in transferring the RDP match information for a specific entry to database fields linked to this entry. The export needs to be defined for each of the entries individually. Once the RDP match sequence is selected, press ***RDP Matches*** > ***Export to database...*** () to transfer the information to the database. Note that no new entry will be created but the information will be transferred to the query entry in the database. Selecting ***RDP Matches*** > ***Export to database...*** () opens the *Export* dialog (see Figure 2.2).

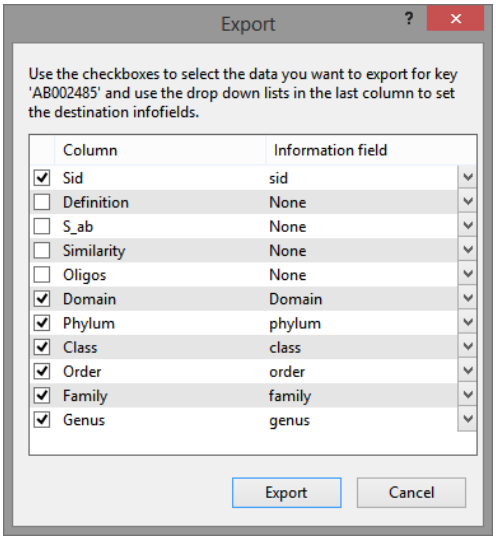

Figure 2.2: The *Export* dialog.

Within the *Export* dialog one can define which columns need to be exported, and if so, to which information field they should be linked. Checking the boxes in front of the column names determines whether or not this column information will be copied to the database. For the selected columns, the information field where the information should be copied to, needs to be selected from the drop-down list that appears when clicking the information field names. Selecting <***Export***> will copy the information to the database information fields of the query entry.

In addition, the information from the RDP matches results can easily be exported to the clipboard or to e.g. Excel by selecting the column properties and ***Copy content to clipboard*** or ***Copy content to file***, respectively.

Each of the experiment match results is saved in the database. If one prompts an analysis that has already been run, the cached results will be displayed. To reinforce a new RDP match analysis, the cached results need to be cleared from the database first. This can be done by selecting ***Edit*** > ***Clear stored results*** in any *RDP overview* window. After confirmation, all stored results will be cleared.

Note that this action clears all cached RDP match results in the database, for all entries that have been analyzed, so not only the entry selection present in the opened *RDP overview* window.

2.4 RDP SeqMatch settings

The setting for the SeqMatch web service can be entered by ***Edit*** > ***SeqMatch settings***. This opens the *SeqMatch settings* dialog box (see Figure 2.3).

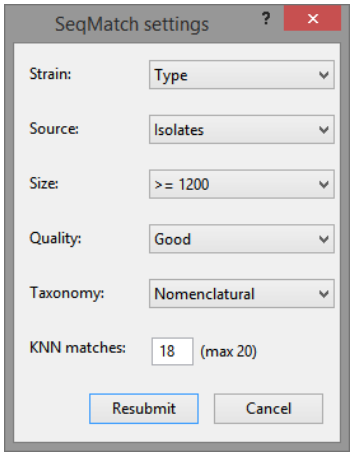

Figure 2.3: The *SeqMatch settings* dialog box.

In the *SeqMatch settings* dialog box, one can defined to which sequences should be matched, what the quality requirements for the sequences should be and which taxonomy is used to report the results.

The following parameters can be set:

- ***Strain***: Selecting ***Type*** restricts the results to only sequences of known type strains. The option ***Both*** will take type as well as non-type strains into account for matching analysis.
- ***Source***: Selecting ***Uncultured*** restricts the display to only sequences of environmental samples. Selecting ***Isolates*** restricts the display to only sequences from individual isolates. Selecting ***Both*** will combine data of uncultured and cultured isolates.
- ***Size***: Selecting ***>1200*** bases restricts the display to only near-full-length sequences. Smaller sequences can be targeted with the option ***<1200***, or both can be used in the analysis.
- ***Quality***: In this context, quality is related to the detection of chimeras by UCHIME [3] and the detection of systematic errors identified by RDP SeqMatch [2] and Pintail [1]. The options are to use only good quality sequences, only suspect quality sequences, or both. Sequences of suspect quality were flagged ().
- ***Taxonomy***: This option defines in which taxonomy sequences will be displayed. The ***Nomenclatural taxonomy*** displays sequences in a hierarchy based on a schema closely matching that proposed in the new phylogenetically consistent higher-order bacterial taxonomy, using a naive Bayesian classifier trained on sequences from known type strains to assign sequences. ***NCBI*** displays sequences as classified in the NCBI taxonomy. This information is directly obtained from the sequence record.
- ***kNN matches***: This parameter controls the number of matches displayed per sequence. The maximum value for k is 20.

If one or more parameters were altered in this dialog, <***Cancel***> will close the dialog without saving the parameter changes and <***Resubmit***> will update the parameter settings and recalculate the analyses displayed in the *RDP overview* window.

Bibliography

[1] K.E. Ashelford, N.A. Chuzhanova, J.C. Fry, A.J. Jones, and A.J. Weightman. At least 1 in 20 16s rrna sequence records currently held in public repositories is estimated to contain substantial anomalies. *Applied and Environmental Microbiology*, 71(12):7724–7736, 2005.

[2] James R Cole, B Chai, Ryan J Farris, Q Wang, SA Kulam, DM McGarrell, George M Garrity, and James M Tiedje. The ribosomal database project (rdp-ii): sequences and tools for high-throughput rrna analysis. *Nucleic acids research*, 33(suppl 1):D294–D296, 2005.

[3] Robert C Edgar, Brian J Haas, Jose C Clemente, Christopher Quince, and Rob Knight. Uchime improves sensitivity and speed of chimera detection. *Bioinformatics*, 27(16):2194–2200, 2011.

APPLIED MATHS

A BIOMÉRIEUX COMPANY

Please contact us for any additional information you might require, we will gladly help you!

Headquarters
Keistraat 120 • 9830 Sint-Martens-Latem • Belgium
+32 922 22 100 info@applied-maths.com

USA and Canada
11940 Jollyville Rd., Suite 115N • Austin, TX 78750 USA
+1 512 482 9700 info-us@applied-maths.com