BioNumerics®

MICROBIAL DATA ANALYSIS SOFTWARE

E. coli functional genotyping 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 Introduction

This guide is designed as a tutorial for the *E. coli functional genotyping plugin* of BioNumerics. This plugin allows you to screen *Escherichia coli* whole genome sequences to predict phenotypic traits. The genome sequences can be imported in BioNumerics using one of the import routines available in the software or can be the result of a de novo assembly performed in BioNumerics on a sequence reads set.

The plugin contains public databases for serotype, virulence and resistance prediction, as well as plasmid and prophage detection. The reference databases are based on the data sets available from the Center for Genomic Epidemiology, DTU (https://cge.cbs.dtu.dk/services/data.php), combined with reference sequences kindly provided by Dr. Rebecca Lindsey (Enteric Diseases Laboratory Branch, Centers for Disease Control and Prevention, Atlanta, GA, USA). An in silico PCR tool is also implemented, making it possible to detect e.g. Shiga toxin gene subtypes and virulence genes, mimicking the wet lab PCR.

The minimal configuration for the installation of the *E. coli functional genotyping plugin* includes the Sequence data module (SQ), Character data module (CH) and the Genome Analysis Tools module GA.

1.2 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).

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.

1.3 Installing the E. coli functional genotyping 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...*** ().

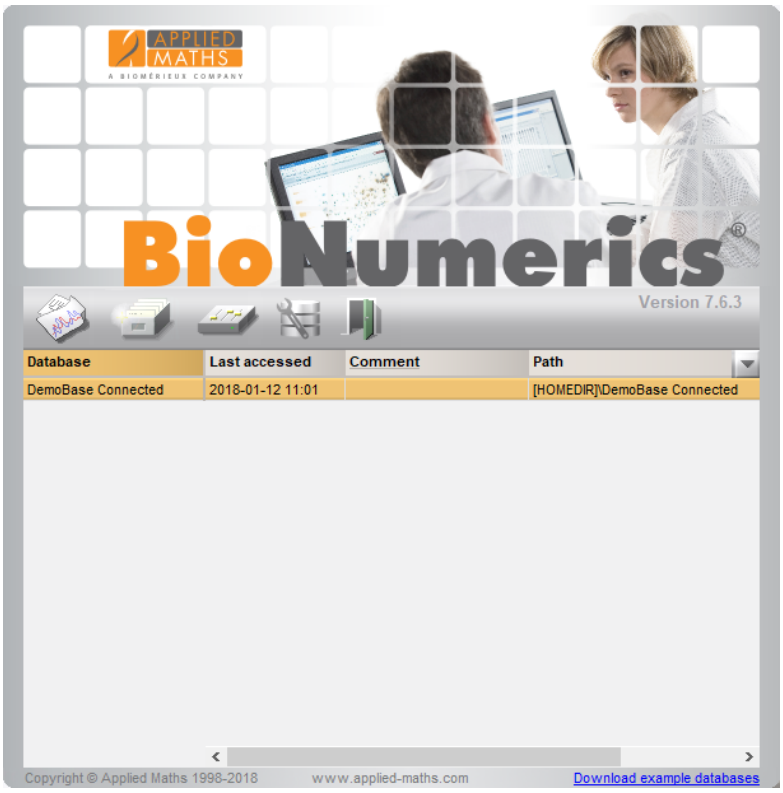

Figure 1.1: The *BioNumerics Startup* window.

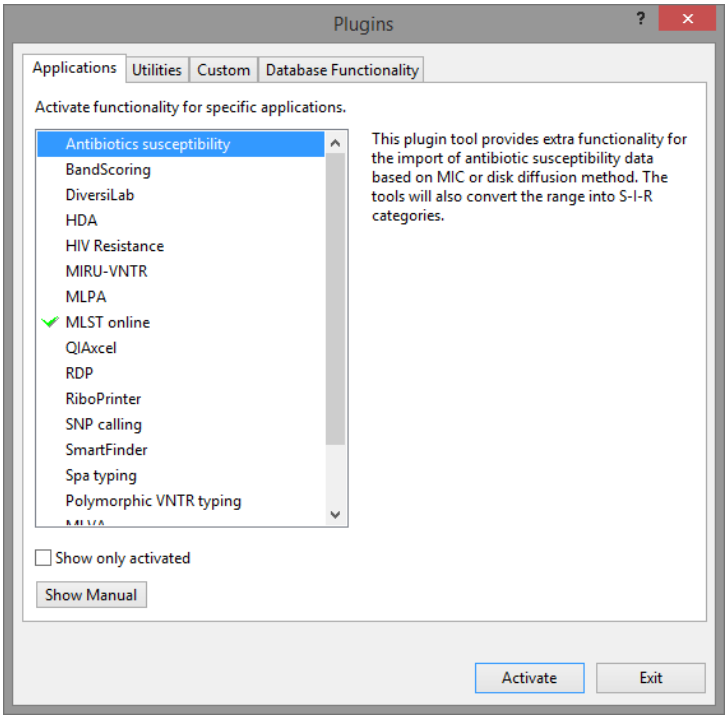

Figure 1.2: The *Plugins* dialog box.

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.

The *E. coli functional genotyping plugin* is provided as an *online plugin*. Online plugins are available from the Applied Maths website, from which they can be downloaded and installed in the database in just a few mouse clicks. Since administrator rights are not required for installation of a plugin in the database, online plugins can be easily updated to take advantage of the latest improvements in program code and search data.

Proceed as follows to install the *E. coli functional genotyping plugin*, starting from the *Plugins* dialog box:

3.1 Select the *Database Functionality tab* in the *Plugins* dialog box and press the <***Add / Update...***> button.

3.2 Check the box that corresponds to the *E. coli functional genotyping plugin* (see Figure 1.3) and press <***OK***>.

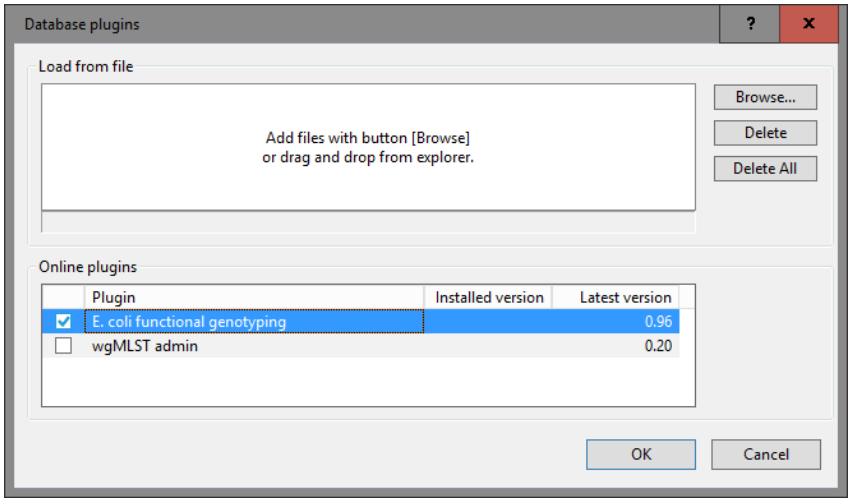

Figure 1.3: Adding an online plugin.

The actual download of the plugin file can take a couple of minutes, depending on the speed of your internet connection.

The first page of the *E. coli genotyping settings* wizard prompts for some general settings (see Figure 1.4):

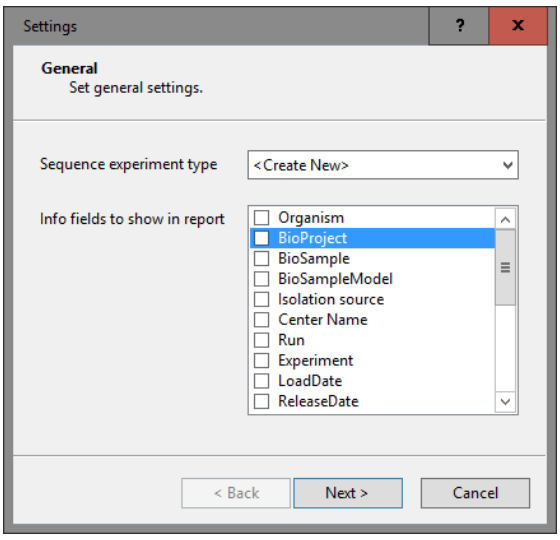

Figure 1.4: The *General settings* wizard page.

- The ***Sequence type experiment*** that holds the (whole) genome sequences that will be screened. A new (<***Create new***>) or existing sequence experiment can be selected from the list.

- The ***Information fields*** that will appear in the report (see 2.2).

The next steps group the settings for each possible search: serotype, pathotype, resistance, virulence, plasmid, prophage, complete plasmid, and in silico PCR search (see Figure 1.5 for an example).

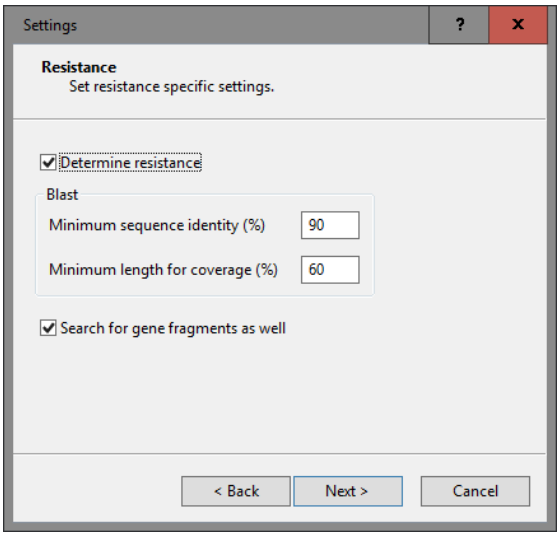

Figure 1.5: Blast search settings.

The ***BLAST*** search settings become available when checking the ***Determine*** (or ***Find***) check box. The ***BLAST settings*** include two thresholds that a BLAST hit should fulfill (see Figure 1.5):

- A ***Minimum sequence identity (%)*** between the subsequence found in the (whole genome) sequence and the sequence in the reference database, expressed as a percentage.
- A ***Minimum length for coverage (%)***, i.e. a minimum overlap between the subsequence found in the (whole genome) sequence and the sequence in the reference database, expressed as a percentage.

With the check box ***Search for gene fragments as well*** enabled in the ***Pathotype***, ***Resistance*** and ***Virulence*** dialogs, sequence fragments (with a maximum of 3 fragments) are also considered.

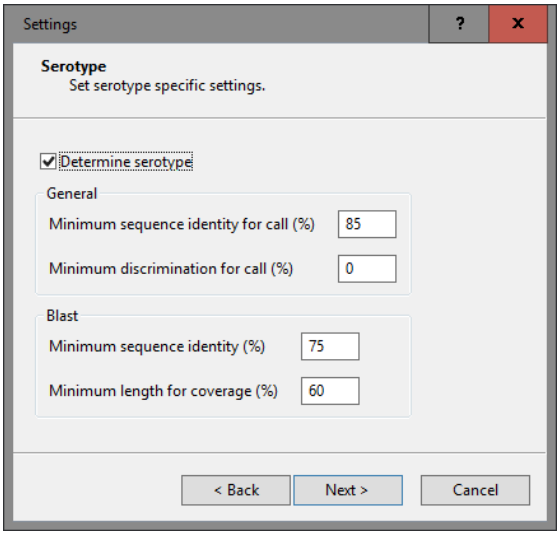

Figure 1.6: Serotype settings.

In the *Serotype settings* wizard page two additional settings are listed to check the discrimination between the two best hits that passed the BLAST criteria (see Figure 1.6). Only if the best hit passes these discrimination criteria, the best hit is used for serotype prediction.

- ***Minimum sequence identity for call (%)***: This is the minimum sequence identity the best BLAST hit should have to be considered for serotype prediction.
- ***Minimum discrimination for call (%)***: This is minimum required discrimination between the two best hits in order to predict the serotype based on the best BLAST hit.

The discrimination D is calculated as:

$$D = \frac{P_1 - P_2}{100 - P_{min}}$$

with P_{min} = minimum sequence identity percentage specified for BLAST, P_1 = sequence identity percentage of best BLAST hit, P_2 = sequence identity percentage of second best BLAST hit. When there is only one best hit, $P_2 = P_{min}$.

The last step prompts for the in silico PCR specific settings (see Figure 1.7).

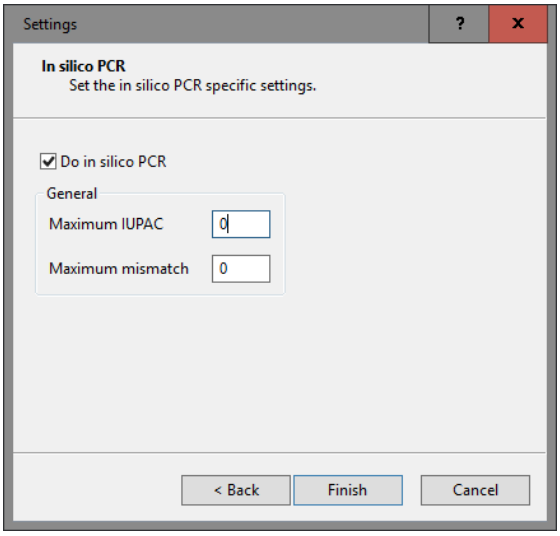

Figure 1.7: In silico PCR specific settings.

- ***Maximum IUPAC***: Maximum number of allowed IUPAC codes in the subsequence of the (whole genome) sequence. The different possibilities for the ambiguous positions are considered when performing the matching against the sequences in the reference database.
- ***Maximum mismatch***: Maximum number of allowed mismatches between the subsequence of the (whole genome) sequence and the sequences in the reference database.

3.3 In the last step press <***Finish***> to complete the installation of the plugin.

3.4 Press <***Exit***> to close the *Plugins* dialog box.

The *E. coli functional genotyping plugin* installs menu items in the main menu of the software under ***E. coli*** (see Figure 1.8).

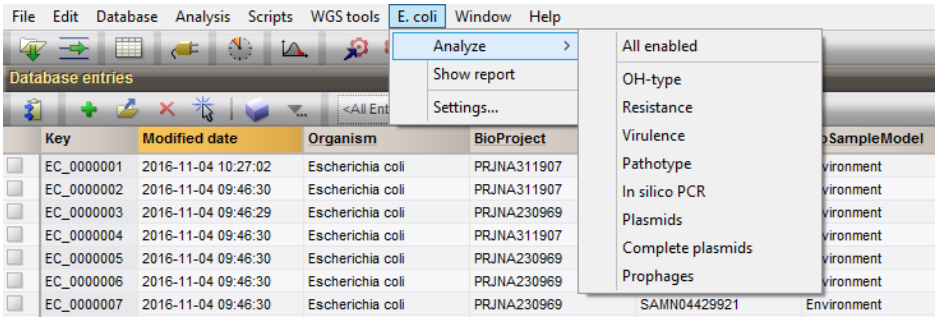

Figure 1.8: New menu items after installation of the plugin.

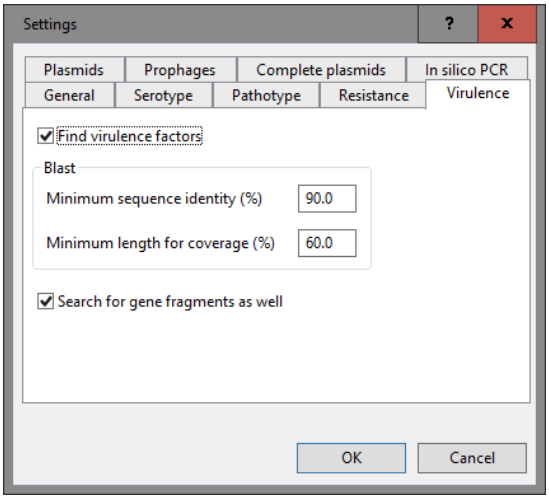

Figure 1.9: The *Settings* dialog box.

The settings specified during installation of the plugin can be called again at any time with ***E. coli*** > ***Settings...*** (see Figure 1.9).

When the ***In silico PCR*** search option is enabled (see Figure 1.7), an **In silico PCR overview** character type is added to the *Experiment types* panel. This experiment will summarize the in silico PCR search results. The predicted **In silico PCR** sequences will be stored in the corresponding sequence type experiments (see Figure 1.10).

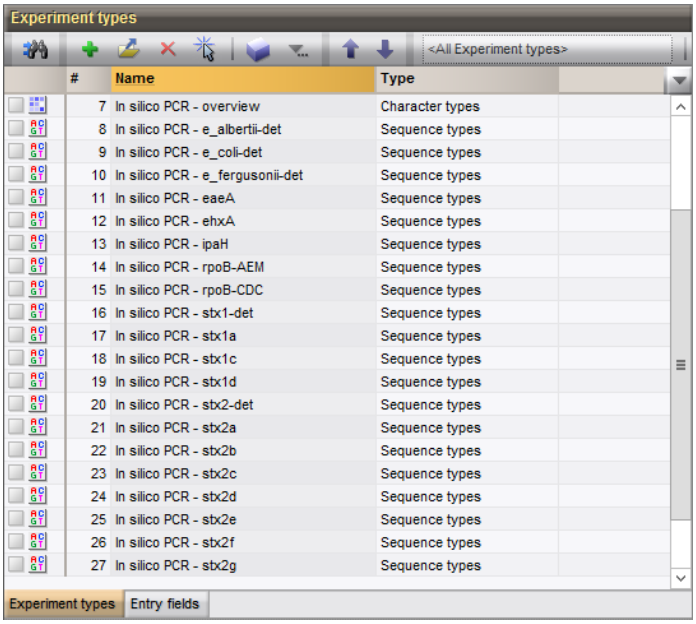

Figure 1.10: In silico PCR experiments.

Chapter 2

E. coli genotyping in BioNumerics

2.1 E. coli genotyping analysis

Once the *E. coli functional genotyping plugin* is installed and the settings have been specified, the actual screening of the sequences of the selected entries is an easy process.

Selecting a single entry in the *Database entries* panel can be done by holding the **Ctrl**-key and clicking on the entry. Alternatively, use the **space bar** to select a highlighted entry or click the ballot box next to the entry. In order to select a group of entries, hold the **Shift**-key and click on another entry.

Screening for the phenotypic traits can be done for all tools checked in the *Settings* dialog box (using ***E. coli*** $>$ ***Analyze*** $>$ ***All enabled***) or for each tool separately with the corresponding command.

A progress bar appears when executing the screening. The analysis time depends on the number of selected entries and may take up to several minutes or even hours. When the analysis is finished, the progress bar disappears. The screening results are stored in the database.

The **Resistance**, (**Complete**) **Plasmids**, **Virulence**, **Prophages** character types are created and contain the predicted traits that passed the BLAST settings. Clicking on a green colored dot for one of these character types opens the character card. The identity percentages are displayed in the ***Value*** column (see Figure 2.1 for a some character card examples).

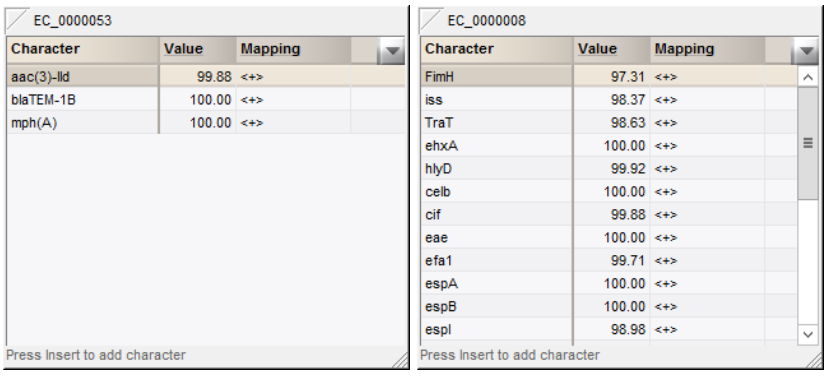

Figure 2.1: Resistance and virulence experiments.

The **In silico PCR - overview** character type summarizes the in silico PCR search results (see Figure 2.2).

The predicted **In silico PCR** sequences are stored in the corresponding sequence type experiments. Clicking on a green colored dot for an in silico experiment opens the *Sequence editor* window displaying the sequence (see Figure 2.3).

The **H and O serotypes** that passed the call criteria (see Figure 1.6) and the **pathotypes** that passed the

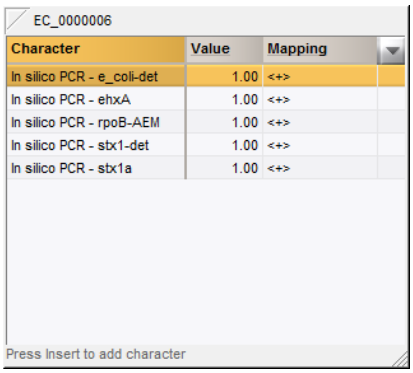

Figure 2.2: Overview of the in silico PCR search.

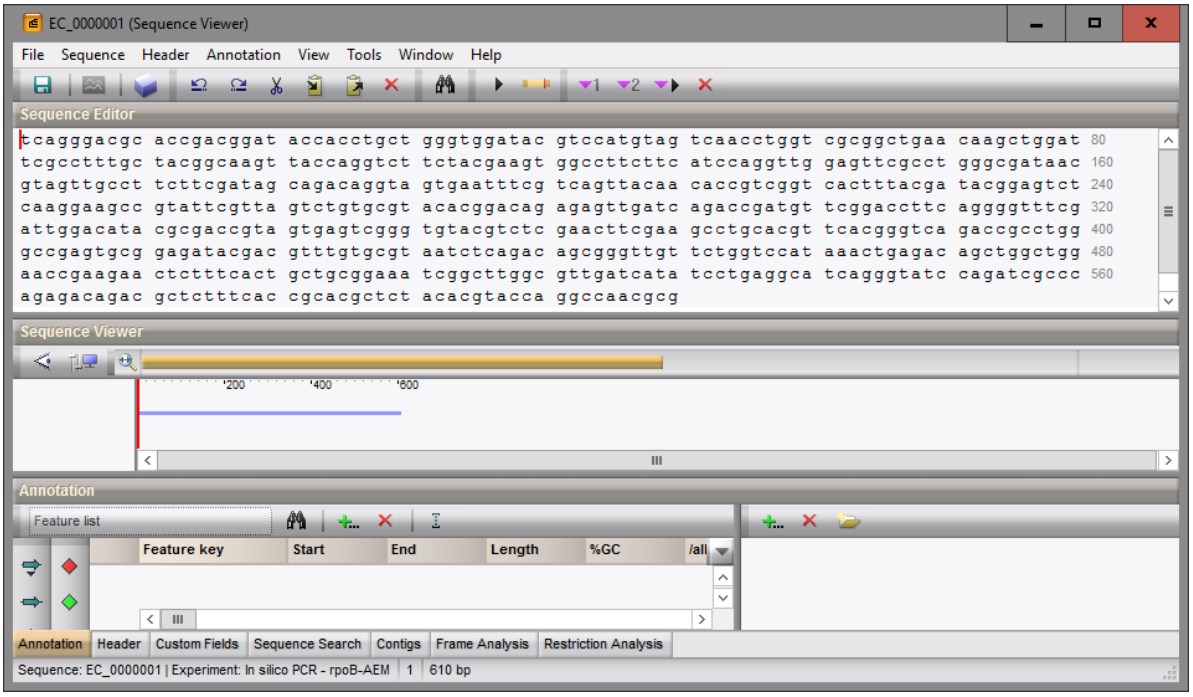

Figure 2.3: In silico PCR sequence.

BLAST criteria are displayed in the corresponding information fields in the *Database entries* panel (see Figure 2.4).

Predicted pathotype	Predicted H serotype	Predicted O serotype
STEC	H2	
STEC	H2	
STEC	H2	
STEC	H2	
STEC	H2	
STEC	H19	
STEC	H19	O121
STEC	H19	O121
STEC	H19	
STEC	H19	O121
STEC	H19	O121
STEC	H19	O121
EPEC	H19	O121
EPEC	H19	O121
STEC	H7	O157
STEC	H7	O157

Figure 2.4: Information fields displaying the predicted pathotype and serotypes.

2.2 E. coli genotyping reports

A genotype report can be opened for the selected entries with ***E. coli*** > ***Show report***.

The *Report* window contains a genotype report for each of the selected entries (see Figure 2.5). Selecting another entry in the *Entries* panel updates the results in the *Report* panel.

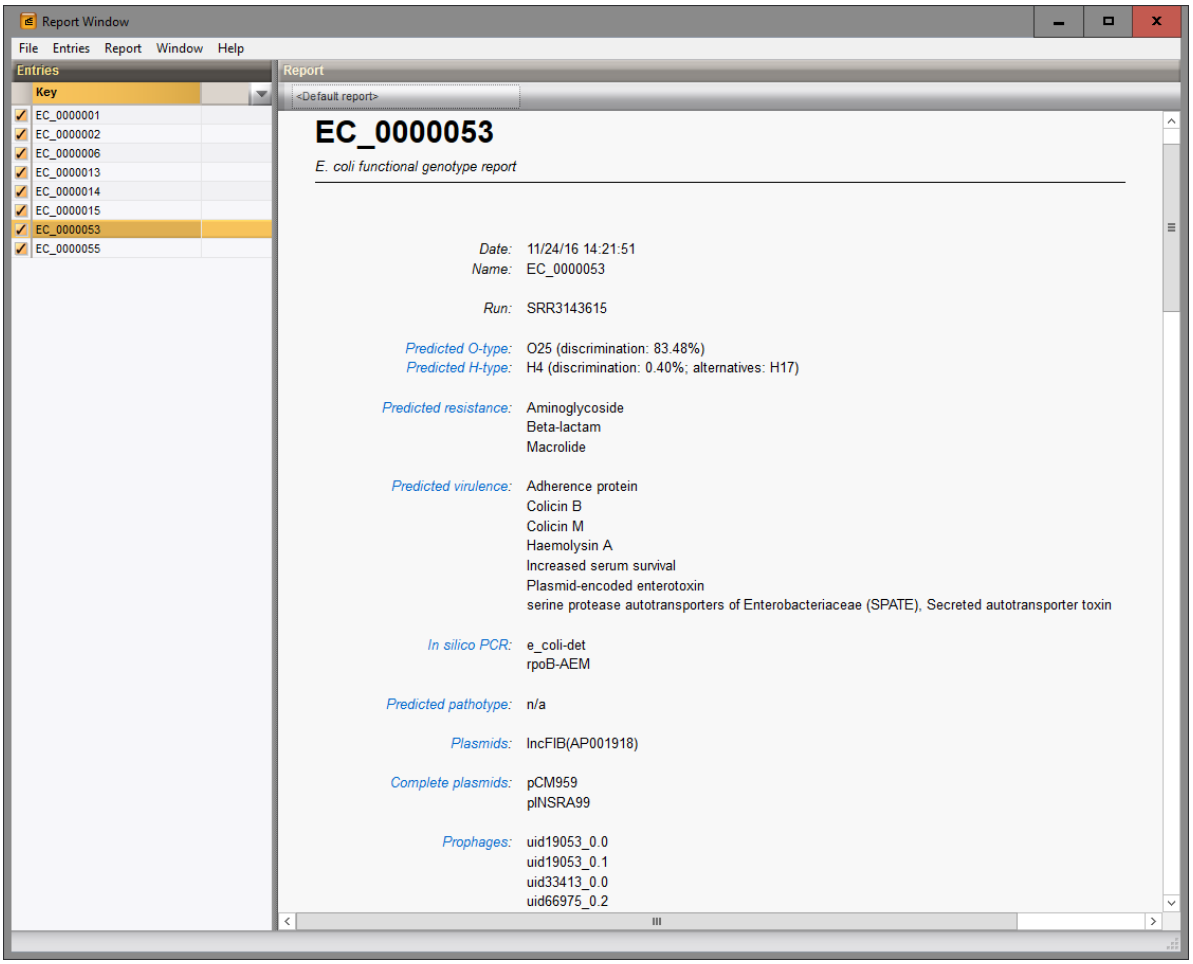

Figure 2.5: Genotype report: overview of the predicted traits.

The data the report was run (***Date***), the Key (***Name***), and information fields checked in the *Settings* dialog box (see Figure 1.9) are displayed in the *Report* panel, followed by a summary of the results of all analyzed traits.

The discrimination calculated for serotype prediction is displayed next to the predicted serotypes, optionally followed by the second best BLAST hit that passed the BLAST criteria (***alternatives***).

All hits that passed the settings for **Resistance**, **Virulence**, **In silico PCR**, **Pathotype**, (**Complete**) **Plasmids** and **Prophages** screening are listed.

Hyperlinks are available linking to the detailed results of the predicted traits at the bottom of the report. Detailed BLAST results include locus identifiers, BLAST similarity scores and descriptive information on the detected genes (see Figure 2.6). The date the analysis was launched and the version number of the *E. coli functional genotyping plugin* that was used to perform the analysis are indicated.

The genotype information displayed in the *Report* panel is exported with ***Report*** > ***Export current report***. A Comma Separated Values (CSV) file is created for each trait and stored in the `Export` subdirectory of the database folder. The location of the files can be opened after confirmation of the export.

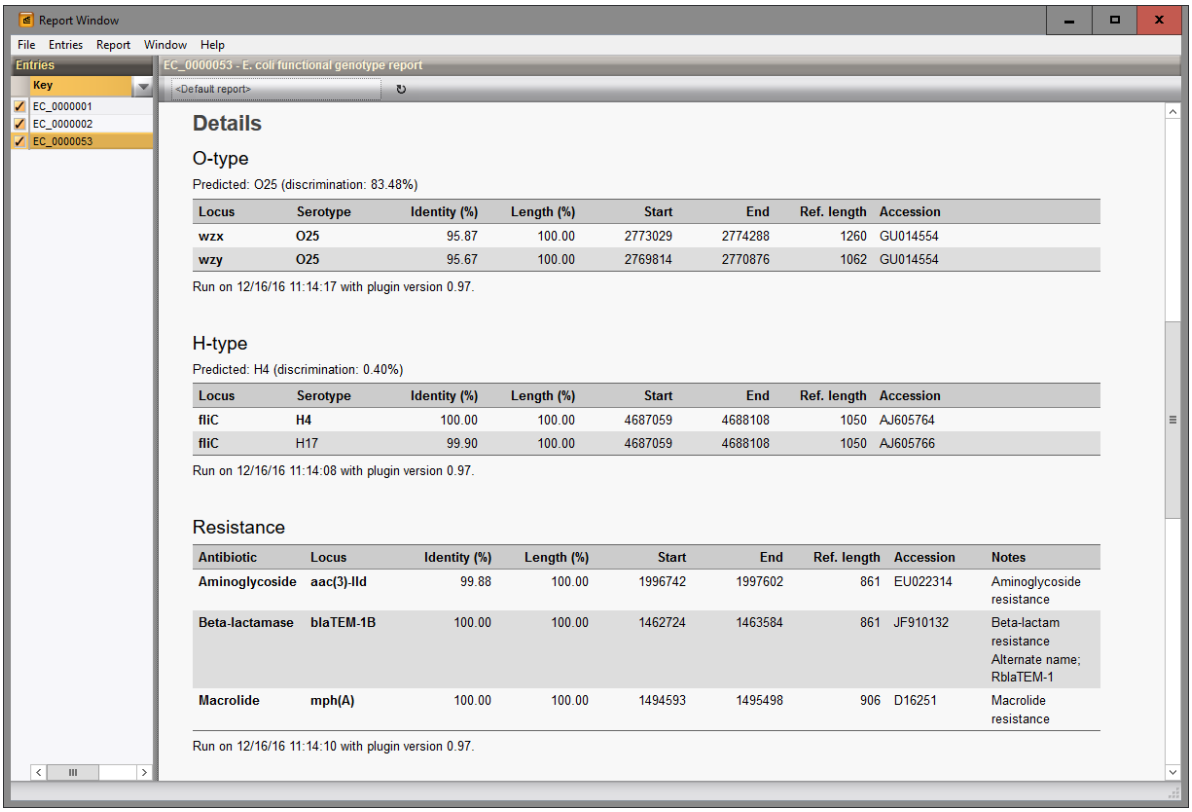

Figure 2.6: Genotype report: details.

The genotype information displayed in the *Report* panel is reanalyzed with ***Report*** > ***Reanalyze current report***. The settings specified in the *Settings* dialog box are used.

The information in the *Report* panel can be printed directly to a printer using ***Report*** > ***Print...***. The dialog box that appears is the standard Windows Print dialog box, allowing you to choose a printer and change the properties.

The genotype information for all entries present in the *Report* window is exported to CSV files with ***Entries*** > ***Export all***. The files are stored in the `Export` subdirectory of the database folder.

The genotype information for all entries present in the *Report* window is reanalyzed with ***Entries*** > ***Reanalyze all***. The settings specified in the *Settings* dialog box are used.

Selecting ***File*** > ***Exit*** closes the *Report* window.

Chapter 3

Appendix: Customizing genotyping reports

It is possible to customize the layout of any report shown in the *Report* window. Customizations can comprise minor modifications such as adding company logos, changing font types and sizes, background colors, etc., but could also include more fundamental changes as to how the information is being organized in the report and even report interactivity. Users with a basic knowledge of Hyper Text Markup Language (HTML) and Cascading Style Sheets (CSS) will probably be able to make any necessary customizations themselves. Alternatively, Applied Maths staff can generate report templates according to your specifications as a custom service.

On a technical level, reports are generated as eXtensible Markup Language (XML) files, which are transformed into HTML via eXtensible Stylesheet Language Transformations (XSLT). The generated HTML is styled using CSS and rendered in the *Report* panel, which is in fact a build-in browser that is included in the BioNumerics software setup.

Proceed as specified below to create one or more custom report templates:

Locate the source files directory via ***Database*** > ***Database settings...***. The *Database settings* dialog box will show the path to the source files directory in the *Files* tab. [DBPATH] is a token that indicates the database directory. The latter can be opened in Windows Explorer by clicking the <***Open database directory***> button.

The source files directory will contain a sub-directory that corresponds to the plugin file name and in which the relevant BLAST search data are stored. For example, if the *E. coli functional genotyping plugin* is installed in the database, a sub-directory called `GenotypingEcoli` will exist in the source files directory.

In this plugin-specific directory, create a folder called `ReportTemplates`.

Each folder within `ReportTemplates` that has at least a `report.css` or `report.xsl` file will be picked up as a report template and available in the *Report* window via ***Report*** > ***Report templates*** and from the drop-down list in the toolbar of the *Report* panel.

To obtain the default `report.css` and `report.xsl` files as a starting point for your customizations, they can be copied from the Windows temp folder: each time a report is displayed, a folder is created in the Windows temp directory with the necessary files.

The `report.css` and / or `report.xsl` files from a report template can then be edited to reflect the required customizations. Any image file included via the `report.css` style sheet or any additional style sheet and JavaScript file included via the `report.xsl` file should reside in the corresponding report template folder or any sub-folder thereof to ensure that it is copied along and included in the final report.

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