BioNumerics®

MICROBIAL DATA ANALYSIS SOFTWARE

Sequence extraction 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 *Sequence extraction plugin* of BioNumerics. This plugin allows you to extract a subsequence from an *origin* experiment type (typically a whole genome sequence) and store the sequence into a *destination* experiment type (see 2.1).

The sequence extraction functionality provided by the *Sequence extraction plugin* has several possible applications. For example, it could be used to extract the sequences of genes with a particular function (resistance genes, virulence genes, etc.) from complete genome or plasmid sequences and save them in separate sequence experiments. This action will make the gene sequences amenable to a more in-depth study via multiple alignments and/or mutation searches. Another common application is to perform Multi Locus Sequence Typing (MLST) based on draft genomes. In this scenario, the sequences corresponding to the seven housekeeping genes are extracted and stored in their own sequence types. Via the *MLST online plugin*, allelic profiles can then be determined based on these sequences.

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 Sequence extraction 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...*** ().

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

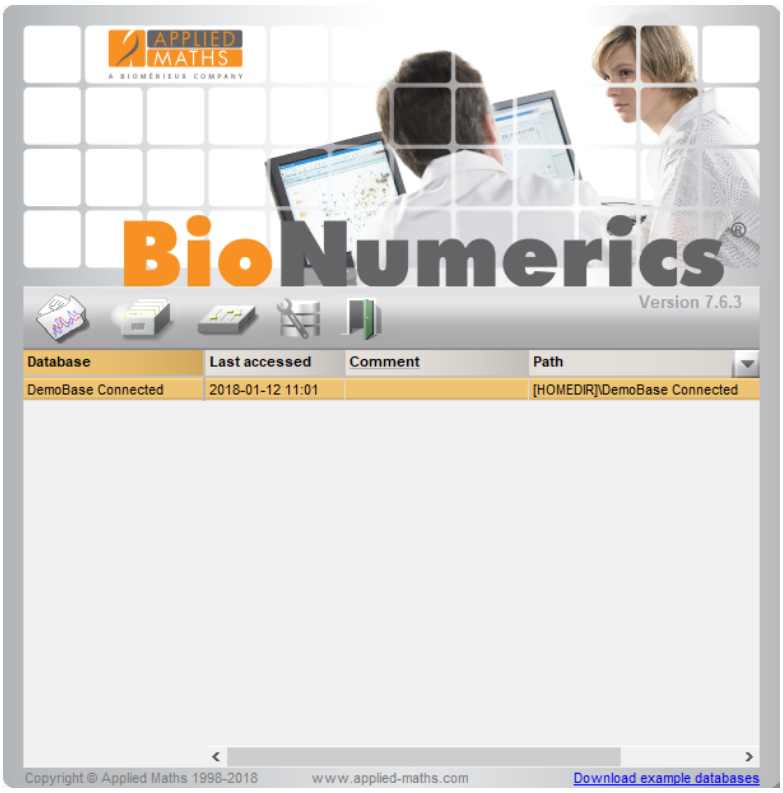

Figure 1.1: The *BioNumerics Startup* window.

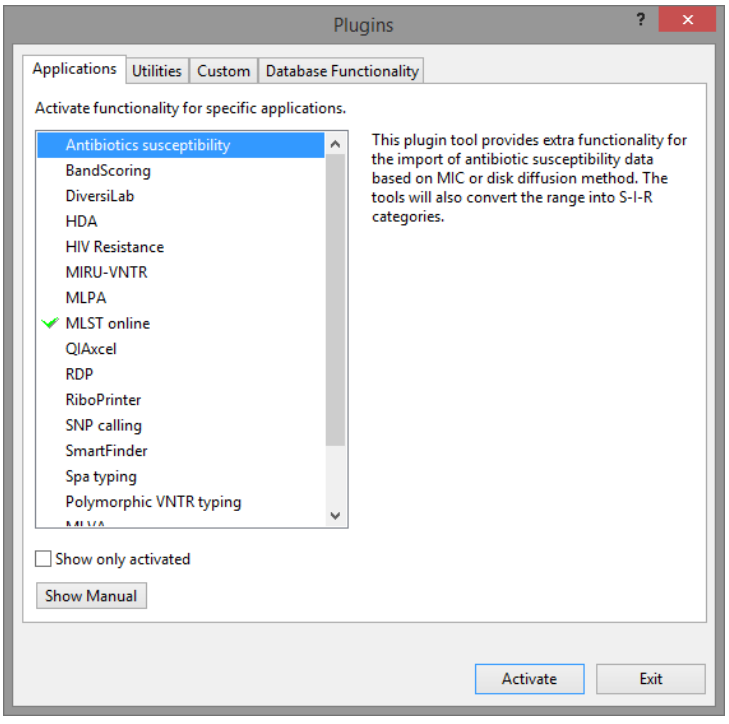

Figure 1.2: The *Plugins* dialog box.

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.

3.1 Select the *Sequence extraction plugin* from the list in the *Utilities tab* and press the <***Activate***> button.

The program will ask to confirm the installation of the plugin.

3.2 Press <***OK***> to continue with the installation of the plugin.

3.3 When the installation is complete, press <***Exit***> to close the *Plugins* dialog box.

The plugin provides three additional menu items in the *Main* window (see Figure 1.3).

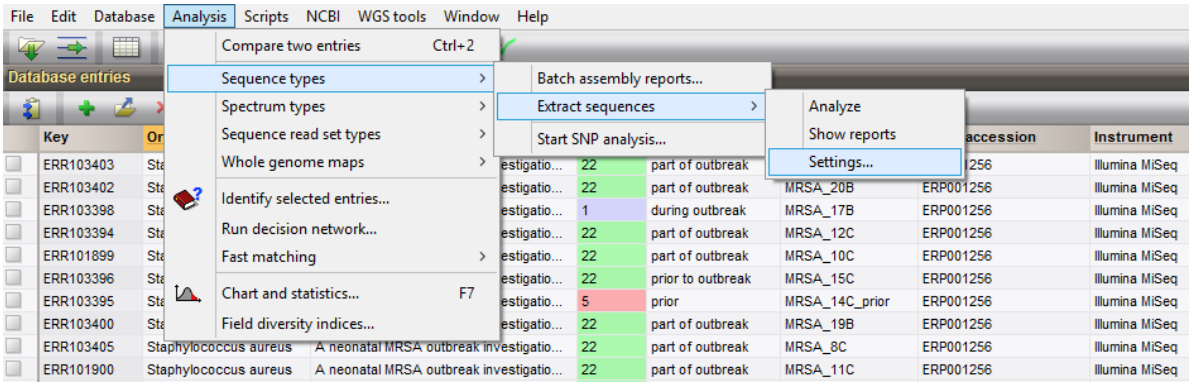

Figure 1.3: Additional menu items installed by the *Sequence extraction plugin*.

The commands ***Analysis*** > ***Sequence types*** > ***Extract sequences*** > ***Analyze*** and ***Analysis*** > ***Sequence types*** > ***Extract sequences*** > ***Settings...*** can also be executed from the *Process data* dialog box (see Figure 1.4). This dialog is called via ***File*** > ***Process...*** ( ).

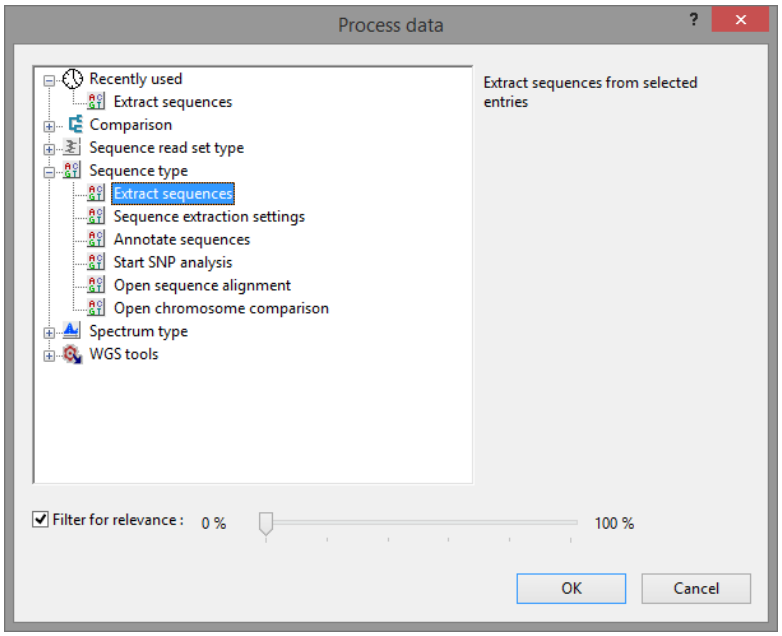

Figure 1.4: The *Process data* dialog box, displaying the two items (***Extract sequences*** and ***Sequence extraction settings***) that are injected by the *Sequence extraction plugin*.

Chapter 2

Extracting subsequences

2.1 Principles

The *Sequence extraction plugin* uses a BLAST approach to extract subsequences in batch from sequences stored in an origin experiment type and saves the retrieved subsequences in a destination experiment type.

Typically, the *origin* experiment will contain a sequence spanning a complete DNA replicon such as a chromosome or plasmid that consists of a single (closed genome) or multiple contigs (draft genome).

The BLAST search is based on a single *query sequence* per destination experiment type. The query sequence is specified by picking a database entry that has the query sequence stored in the destination experiment type. In some conditions, it might be necessary to correct the length of the retrieved subsequence. Length correction is possible by specifying PCR primers (trimming positions) or can be based on start and stop codons, retrieving a full protein coding sequence (CDS).

The subsequence that was retrieved by the BLAST search after length correction (if specified) will be stored in the *destination* experiment. Note that the destination experiment will always be overwritten, even in case no hit was found.

To use the *Sequence extraction plugin*, sequence extraction settings should first be specified and stored per experiment type (see 2.2). When an analysis is performed (see 2.3), sequences are extracted for the selected entries and for all sequence experiment types where sequence extraction settings are available for. Optionally, a report (see 2.4) can be displayed.

2.2 Sequence extraction settings

Since the plugin retrieves its query sequence from the sequence experiment stored with an entry, we need to make sure (1) the destination experiment type is created in the database and (2) that the query sequence is stored with an entry in the destination experiment type before specifying any sequence extraction settings.

To add or to change sequence extraction settings, select ***Analysis*** > ***Sequence types*** > ***Extract sequences*** > ***Settings...*** in the *Main* window. This action opens the *Sequence extraction settings* dialog box (see Figure 2.1).

This dialog box gives access to the actual ***Sequence extraction*** settings per sequence experiment type and the general ***Reports*** settings. Initially, the tree control on the left will be empty.

Sequence extraction settings can be added for one or more sequence experiment types by pressing the <***Add***> button. This action starts the *Extract gene from genome* (see Figure 2.2).

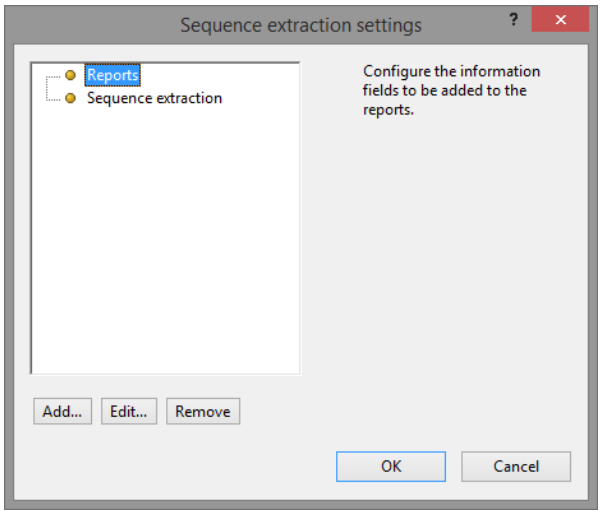

Figure 2.1: The *Sequence extraction settings* dialog box: initially, no experiment types are listed.

It is convenient to add sequence types in batch if most of their settings are the same. However, if for example the query sequences are stored with different entries, the sequence types should be added one by one.

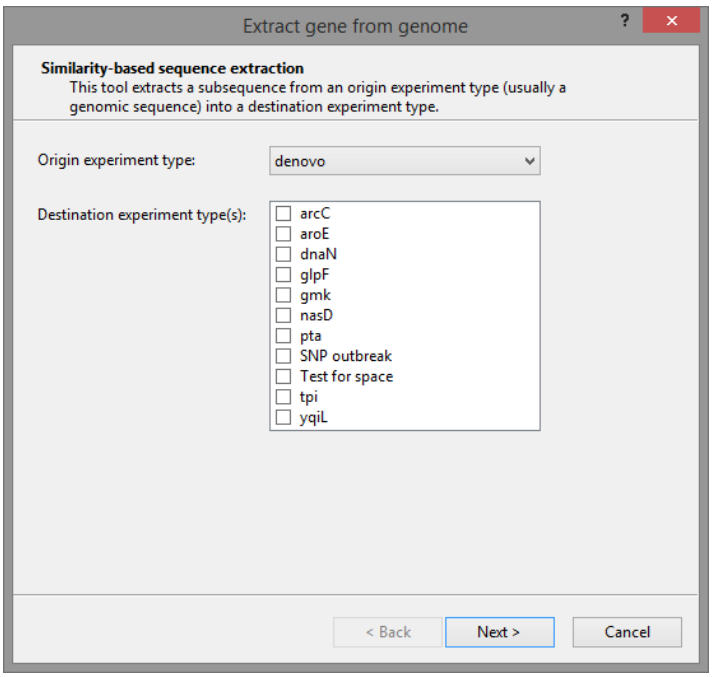

Figure 2.2: The *Similarity-based sequence extraction* wizard page from the *Extract gene from genome*.

The ***Origin experiment type*** is the sequence experiment type that will screened and from which a subsequence will be copied from. Usually, this will correspond to a genome sequence, i.e. either a draft genome consisting of several contigs or a fully closed genome. The ***Origin experiment type*** can be selected from the corresponding drop-down list.

The extracted subsequence will be copied to a destination experiment type. The ***Destination experiment type(s)*** list shows all remaining sequence experiment types in the database. Check the corresponding check boxes to add sequence extraction settings for one or more destination experiment types.

In case the selected destination experiment types already have sequence extraction settings, adding them again will overwrite the current sequence extraction settings.

Press <***Next***> to proceed to the *Settings* wizard page (see Figure 2.3).

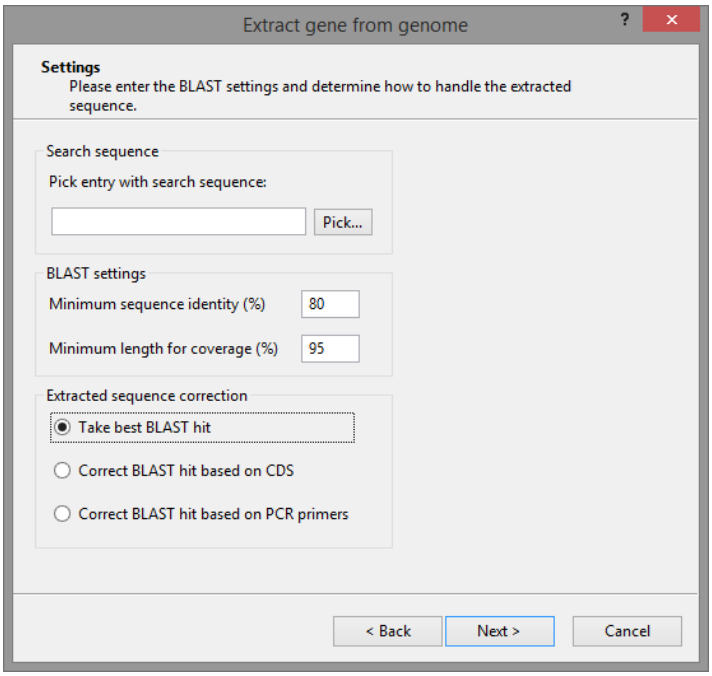

Figure 2.3: The *Settings* wizard page.

The ***Search sequence*** is what the BLAST algorithm will use to screen the origin experiment type for. Search sequences should be stored in the destination experiment type of an entry. An entry that contains the search sequence can be chosen. Pressing <***Pick***> will open the *Select entry* dialog box, from which an entry can be picked. See the Reference manual, Chapter Database objects for more information about the *Select entry* dialog box.

The ***BLAST settings*** include two thresholds that a BLAST hit should fulfill in order to be considered:

- A ***Minimum sequence identity (%)*** between the search sequence and the matched subsequence in the origin sequence experiment, expressed as a percentage.
- A ***Minimum length for coverage (%)***, i.e. a minimum overlap between the search sequence and the matched subsequence.

In case more than one BLAST result is found that fulfills both criteria, the best match will be copied to the destination experiment.

Optionally, the length of the extracted sequence can be corrected. Following ***Extracted sequence correction*** options are available:

- ***Take best BLAST hit***: With a ***Search sequence*** of the correct length, the BLAST result can in most cases be used as-is, i.e. without applying any correction.
- ***Correct BLAST hit based on CDS***: The BLAST result is extracted from the origin experiment, including a short leading and trailing subsequence. From this extended BLAST result, a coding DNA sequence (CDS) is extracted to the destination experiment, i.e. starting from the first encountered start codon and ending at the first encountered stop codon.

- ***Correct BLAST hit based on PCR primers***: The BLAST result is extracted from the origin experiment, including a short leading and trailing subsequence. This extended BLAST result is then trimmed based on PCR primers. The PCR primer settings are defined in the *Define primers* wizard page.

If no start/stop codon or PCR primers are found, the uncorrected BLAST result is returned.

With the option ***Take best BLAST hit*** or ***Correct BLAST hit based on CDS*** checked, pressing <***Next***> will complete the *Extract gene from genome*.

When ***Correct BLAST hit based on PCR primers*** was checked, pressing <***Next***> will show the *Define primers* wizard page (see Figure 2.4).

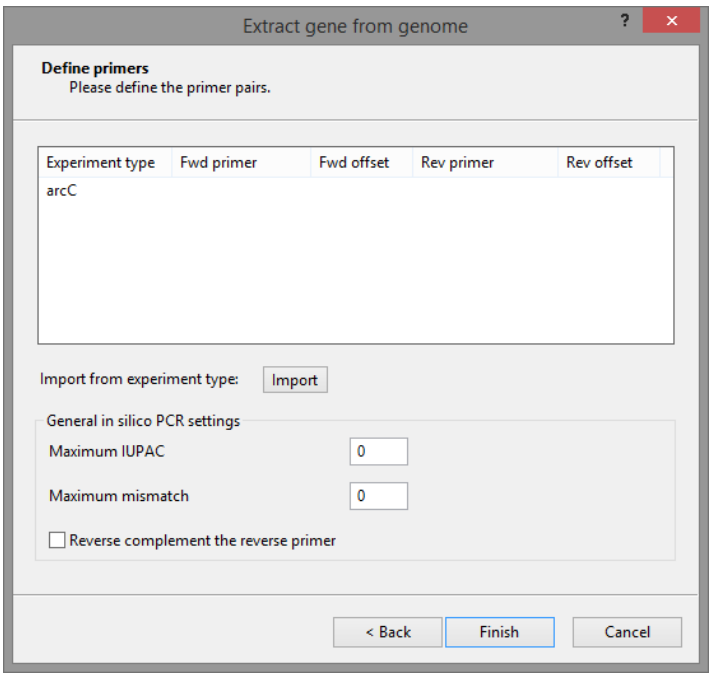

Figure 2.4: The *Define primers* wizard page.

The principle of length correction of extended BLAST results based on PCR primers is similar to the trimming step after assembly of Sanger sequences (see the Reference manual, Chapter Setting up sequence type experiments for more information).

The destination experiment types that were checked in the *Similarity-based sequence extraction* wizard page will now be listed under 'Experiment type'. For each experiment type, a forward primer ('Fwd primer'), forward offset ('Fwd offset'), reverse primer ('Rev primer') and a reverse offset ('Rev offset') can be entered in the list.

Alternatively, by pressing the <***Import***> button next to ***Import from experiment type***, the relevant parameters of the assembly trimming settings (as stored with the sequence experiment type; see the Reference manual, Chapter Setting up sequence type experiments) can be copied to the grid.

The ***General in silico PCR settings*** include:

- ***Maximum IUPAC***: The maximum number of IUPAC ambiguous bases in the matching sequence, i.e. the overlap between the primer and the extended BLAST result. Setting this to a low number avoids that very degenerated sequences (e.g. `NNNNNNNNNNNNN`) will match with the primer sequences.

- ***Maximum mismatch***: The maximum number of mismatches allowed between a primer and the target sequence.

- The option ***Reverse complement the reverse primer*** needs to be checked if the reverse primer is specified in the reverse direction (by default, BioNumerics expects both primer sequences in the forward direction).

Pressing <***Finish***> will complete the *Extract gene from genome*.

When sequence extraction settings are added for one or more destination experiment types, the latter will be listed in the *Sequence extraction settings* dialog box (see Figure 2.5):

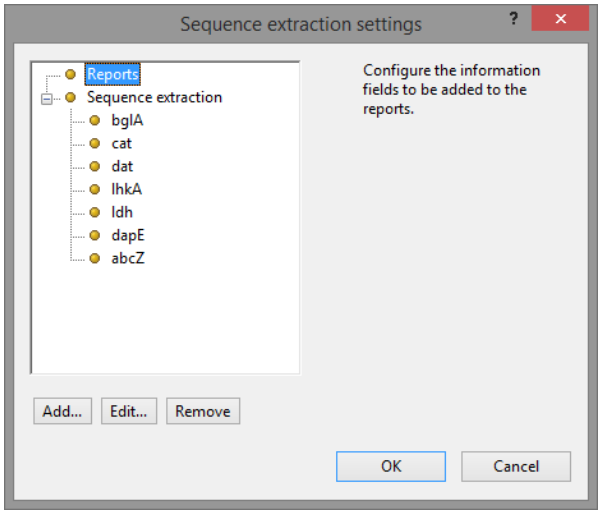

Figure 2.5: The *Sequence extraction settings* dialog box, listing experiments types for which sequence extraction settings are added.

The sequence extraction settings for a highlighted experiment type can be removed by pressing <***Remove***> button. The settings will be deleted after confirmation.

Existing sequence extraction settings can be edited by highlighting the corresponding item in the *Sequence extraction settings* dialog box and pressing <***Edit***>. This action opens the *Edit sequence extraction* dialog box (see Figure 2.6).

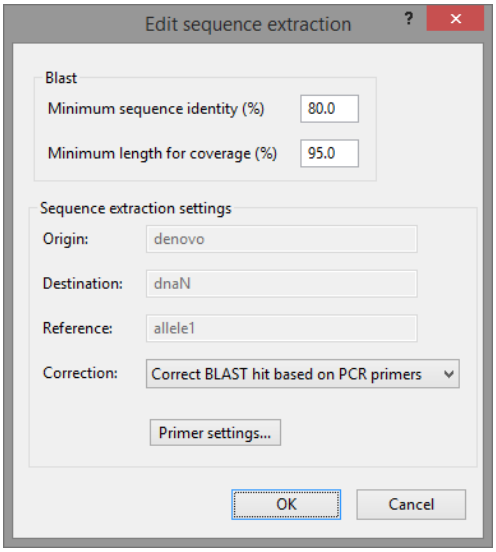

Figure 2.6: The *Edit sequence extraction* dialog box.

Only the ***BLAST*** thresholds and the length ***Correction*** can be edited, other parameters are read-only in this dialog.

With ”Correct BLAST hit based on PCR primers” selected for ***Correction***, an additional button <***Primer settings...***> becomes available. Pressing this button opens the *PCR primer settings* dialog box (see Figure 2.7).

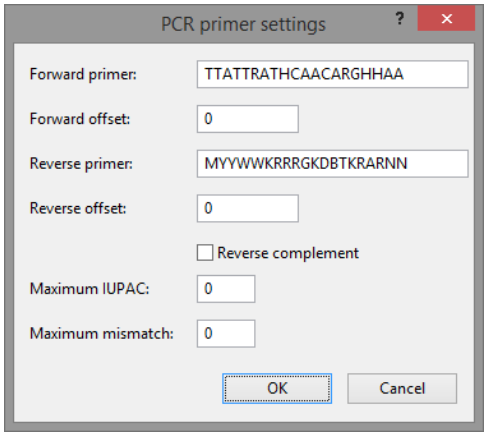

Figure 2.7: The *PCR primer settings* dialog box.

This dialog box allows you to enter or edit primer sequences, offsets and general in silico PCR primer settings as described for the *Define primers* wizard page.

To modify the report settings for the *Sequence extraction plugin*, highlight ***Reports*** in the tree and press <***Edit***>. The *Reports* dialog box pops up (see Figure 2.8).

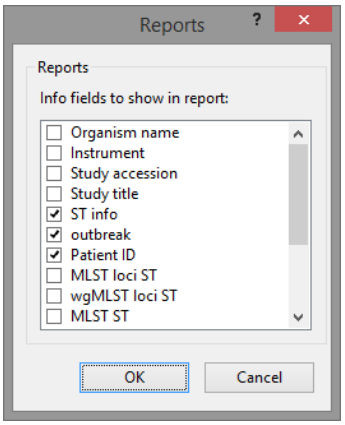

Figure 2.8: The *Reports* dialog box.

This dialog lists all entry information fields and all entry field views that are available in the database. Check the corresponding check box to include the individual field or the set of fields included in the view in the genotyping report.

When done editing report and/or sequence extraction settings, press <***OK***> in the *Sequence extraction settings* dialog box.

If an entry selection is present, the question ”Do you want to analyze the selected entries?” pops up. When you confirm by pressing <***Yes***>, a sequence extraction will start for the entry selection (see 2.3).

2.3 Sequence extraction analysis

Once the *Sequence extraction plugin* is set up by specifying sequence extraction settings for one or more destination experiment types (see 2.2), the actual sequence extraction is an easy process:

3.1 Select the entries for which to extract sequences.

3.2 Select ***Analysis*** > ***Sequence types*** > ***Extract sequences*** > ***Analyze*** or use the *Process data* dialog box: select ***File*** > ***Process...*** (), highlight ***Extract sequences*** under ***Sequence type*** and press <***OK***>.

When attempting to run an analysis on the selected entries when no sequence extraction settings are specified, the message "Please define sequence extraction settings first." pops up. Pressing <***OK***> will open the *Sequence extraction settings* dialog box (see 2.2).

A progress bar appears. Depending on the number of sequences to be extracted, the complete analysis may take up to several minutes. When the analysis is finished, the question "Do you want to open the reports?" pops up. Pressing <***Yes***> opens a *Report* window with a summary of the results (see 2.4).

The extracted sequences can easily be visualized in the *Comparison* window:

3.3 Leaving the entry selection unaltered, highlight the *Comparisons* panel in the *Main* window and select ***Edit*** > ***Create new object...*** ().

A *Comparison* window opens with the selected entries and the extracted sequences are displayed by clicking the corresponding eye icon () in the *Experiments* panel. A sequence multiple alignment can be calculated as described in the Reference manual, Chapter Sequence alignment and mutation analysis.

2.4 Sequence extraction reports

After sequence extraction (see 2.3), BioNumerics will prompt to open a report. Alternatively, a sequence extraction report can be opened for the selected entries with ***Analysis*** > ***Sequence types*** > ***Extract sequences*** > ***Show reports*** (see Figure 2.9). If entries are selected for which no sequences are extracted, the analysis will be ran before the report data are displayed.

The *Report* window contains a gene extraction report for each of the selected entries. For each destination experiment type ('Locus') that has sequence extraction settings, a result line is reported. See Table 2.1 for a description.

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

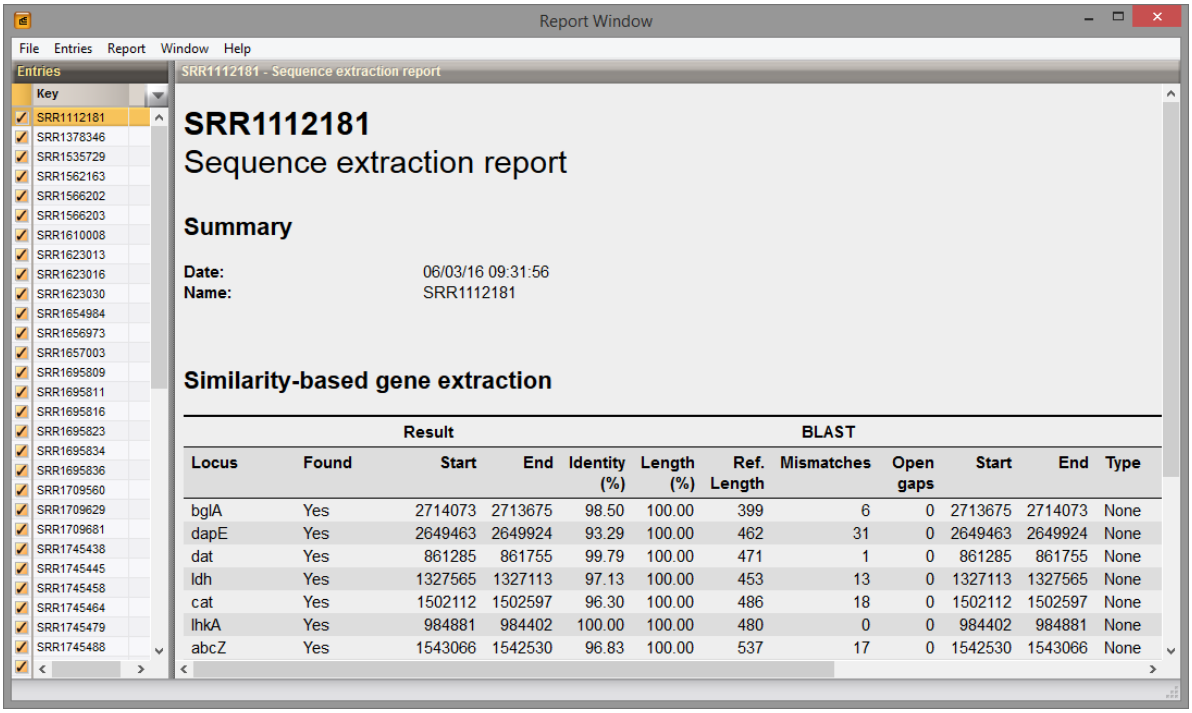

Figure 2.9: A gene extraction report displayed in the *Report* window.

Name	Result group	Explanation
Locus		The name of the extracted locus (i.e. the destination experiment type name)
Found	Result	Whether or not the sequence was found (Yes/No)
Start	Result	The start position of the actual retrieved sequence (i.e. after length correction, if applied) on the origin sequence
End	Result	The end position of the actual retrieved sequence (i.e. after length correction, if applied) on the origin sequence
Identity (%)	BLAST	Similarity (in %) of the query sequence with the best BLAST hit
Length (%)	BLAST	Length of the best BLAST hit, relative to the query sequence (in %)
Ref. length	BLAST	The length of the query sequence in bases
Mismatches	BLAST	The number of non-matching bases between the query sequence and the best BLAST hit
Open gaps	BLAST	The number of gaps between the query sequence and the best BLAST hit
Start	BLAST	The start position of the best BLAST hit on the origin sequence
End	BLAST	The end position of the best BLAST hit on the origin sequence
Type	Correction	The type of length correction applied (None/CDS/PCR) on the best BLAST hit
Start	Correction	Whether or not the start codon or forward primer was found (Yes/No) or a hyphen (-) if no length correction was applied
End	Correction	Whether or not the stop codon or reverse primer was found (Yes/No) or a hyphen (-) if no length correction was applied

Table 2.1: Explanation of the sequence extraction results reported in the *Report* window.