BioNumerics Tutorial:

Streptococcus pyogenes emm typing

1 Introduction

This tutorial explains how to use the BioNumerics *emm typing* script, developed for BioNumerics users that want to type *Streptococcus pyogenes* based on the portion of the emm gene that dictates the M serotype. The script assigns the emm type and subtype to imported (and assembled) sequences, by querying the CDC M-type specific database (`ftp://ftp.cdc.gov/pub/infectious_diseases/biotech/tsemm/`).

2 Preparing the database

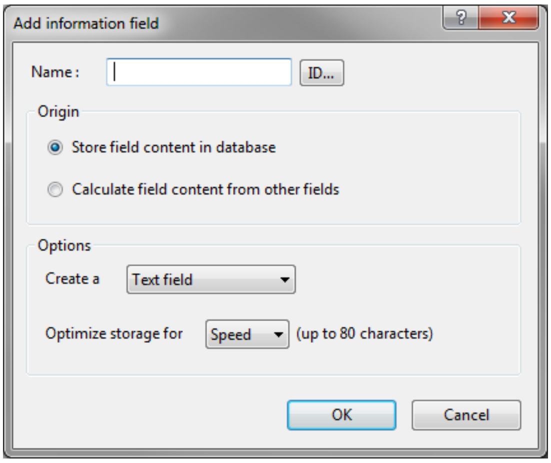
1. Create a new database (see tutorial ”Creating a new database”) or open an existing database.

Prior to running the emm typing script for the first time, folllowing items need to be created in the BioNumerics database: three information fields (**emmType**, **Number of hits**, **Comments** (see 2.1)), one sequence type experiment (**emm** (see 2.2)) and one BLAST database (**emm** (see 2.3)).

2.1 Create three entry information fields

2. In the *Main* window, select ***Edit*** > ***Information fields*** > ***Add information field...*** or highlight the *Entry fields* panel and select ***Edit*** > ***Create new object...*** (✚).

The *Create new entry information field* dialog box pops up (see Figure 1).



Name :

Origin

● Store field content in database

○ Calculate field content from other fields

Options

Create a Text field

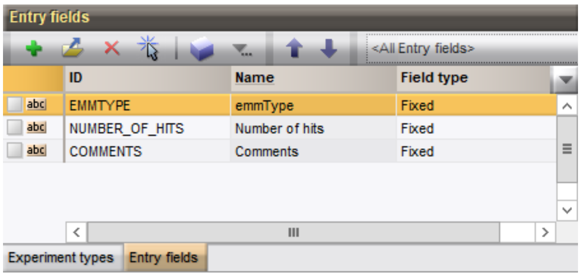
Optimize storage for Speed (up to 80 characters)

Figure 1: The *Create new entry information field* dialog box.

3. Specify the name “emmType” and press <***OK***>.

The new entry information field is created and is displayed in the *Database entries* panel (see Figure 2).

4. Repeat previous step to add two more information fields called “Number of hits” and “Comments” (see Figure 2).



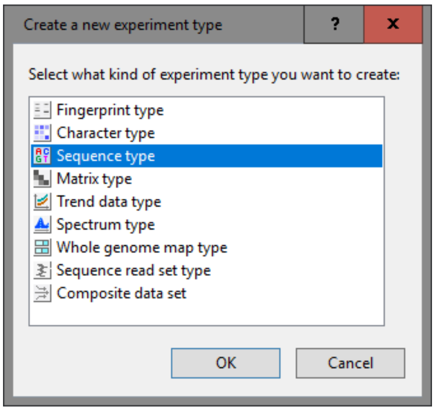
ID	Name	Field type
EMMTYPE	emmType	Fixed
NUMBER_OF_HITS	Number of hits	Fixed
COMMENTS	Comments	Fixed

Figure 2: Information fields.

The script requires the presence of the **emmType**, **Number of hits** and **Comments** information fields names. Make sure to use these exact names when creating the fields.

2.2 Create sequence type experiment

5. In the *Main* window, click on + in the toolbar of the *Experiment types* panel and select ***Sequence type*** from the list (see Figure 3).


Create a new experiment type

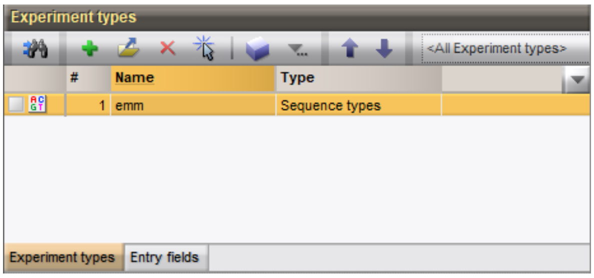
Select what kind of experiment type you want to create:

- Fingerprint type
- Character type
- Sequence type
- Matrix type
- Trend data type
- Spectrum type
- Whole genome map type
- Sequence read set type
- Composite data set

Figure 3: The *Create a new experiment type* dialog box.

6. Press <***OK***>, enter a name, for example **emm** and press <***Next***> and <***Finish***>.

The *Experiment types* panel now lists the sequence type **emm** (see Figure 4).



#	Name	Type
1	emm	Sequence types

Figure 4: Sequence type experiment.

The script requires the presence of the **emm** sequence type name. Make sure to use this exact name when creating the experiment.

2.3 Create BLAST database

In this section we will create a new BLAST database in BioNumerics, using a multi-fasta file containing the trimmed EMM variant sequences available in the CDC database:

7. Go to ftp://ftp.cdc.gov/pub/infectious_diseases/biotech/tsemm/trimmed.tfa and save the trimmed.tfa file to a location on your computer.

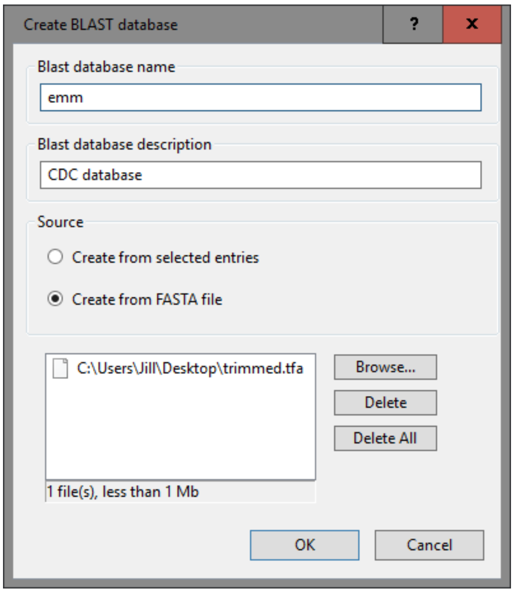
Alternatively, go to ftp://ftp.cdc.gov/pub/infectious_diseases/biotech/tsemm/, scroll down the list of variant sequences, and click the trimmed.tfa file at the bottom of the list and save the file to a location on your computer.

8. In the *Main* window, select ***Database*** > ***Sequence databases*** > ***BLAST databases...*** to call the *BLAST Database Tools* dialog box.

9. To create a new BLAST database, select <***New***> in the *BLAST Database Tools* dialog box. This pops up the *Create BLAST database* dialog box (see Figure 5).

A new BLAST database can be created from a selection of entries in the database or from a FASTA file.

10. Specify **emm** as ***Blast database name***, optionally update the ***Description*** and make sure ***Create from FASTA file*** is checked. Click <***Browse***> and browse for the trimmed.tfa file which contains the emm reference sequences.



Create BLAST database

Blast database name: emm

Blast database description: CDC database

Source

○ Create from selected entries

● Create from FASTA file

C:\Users\Jill\Desktop\trimmed.tfa

1 file(s), less than 1 Mb

Figure 5: Create emm BLAST database.

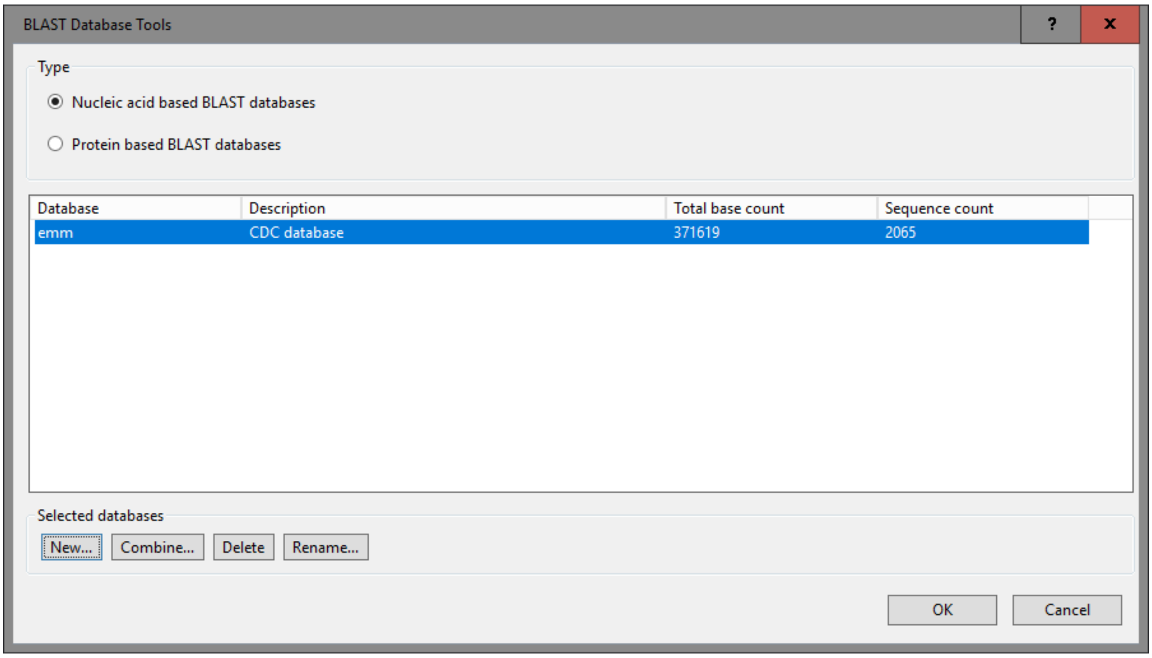
11. Press <***OK***> to create the BLAST database.

The script requires the presence of the **emm** BLAST database name. Make sure to use this exact name when creating the BLAST database.

The new database is now added to the list in the *BLAST Database Tools* dialog box (see Figure 6).

To keep up with the updates in the CDC database, the trimmed.tfa file should be downloaded at regular intervals and used to create an updated BLAST database in BioNumerics.

12. Close the *Create BLAST database* dialog box.



● Nucleic acid based BLAST databases

○ Protein based BLAST databases

Database	Description	Total base count	Sequence count
emm	CDC database	371619	2065

Figure 6: Emm BLAST database created.

3 Importing sequence data

1. In the *Main* window, select ***File*** > ***Import...*** (, **Ctrl+I**) to open the *Import* dialog box.

All sequence import routines are grouped under the import topic ***Sequence data*** in the *Import* dialog box. Sequence data can be imported in BioNumerics in several ways:

1. Importing sequences in GenBank and EMBL format from a text file.
2. Importing sequences in FASTA format from a text file.
3. Importing sequences from online repositories.
4. Importing and assembling ABI and SCF sequence traces in batch.
5. Importing and assembling FASTA sequence traces in batch.

In this tutorial we will import some sample sequences in FASTA format from a text file. The sample file `samples_emm.fa` can be found on the download page on our website: http://www.applied-maths.com/download/sample-data, ”emm typing”. More information about the other import routines can be found in the BioNumerics reference manual and tutorials on our website.

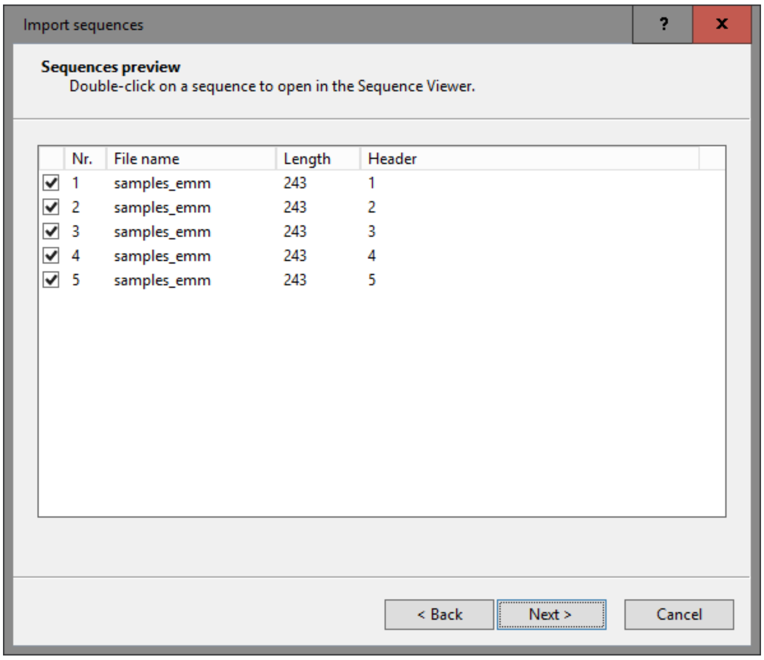
2. Download the sample file from our website and unzip the file.

3. Choose the option ***Import FASTA sequences from text files*** under the ***Sequence type data*** item in the tree and click <***Import***>.

4. Press <***Browse***>, select the `samples_emm.fa` file and press <***Open***>.

5. Press <***Next***>.

The import wizard now displays a preview of the sequence data in the FASTA file. From this preview, it is clear that the first and only FASTA field contains the sample number.



Sequences preview
Double-click on a sequence to open in the Sequence Viewer.

	Nr.	File name	Length	Header
☑	1	samples_emm	243	1
☑	2	samples_emm	243	2
☑	3	samples_emm	243	3
☑	4	samples_emm	243	4
☑	5	samples_emm	243	5

Figure 7: Sample preview.

6. Press <***Next***>.

The next step of the import wizard lists the templates that are present to import sequence information in the database. As this is the first time we import FASTA formatted sequences in the database, we need to create a new import template by specifying ***Import rules***.

7. Click <***Create new***> to create a new import template.

8. Select ”Field 1” in the list and click <***Edit destination***> or simply double-click on ”Field 1”. Select ***Key*** and press <***OK***>.

9. Optionally, you can press <***Preview***> to obtain a preview of the data you are about to import.

10. Click <***Next***> and <***Finish***>.

11. Specify a template name, e.g. ”Import fasta”, and optionally enter a description. Press <***OK***>.

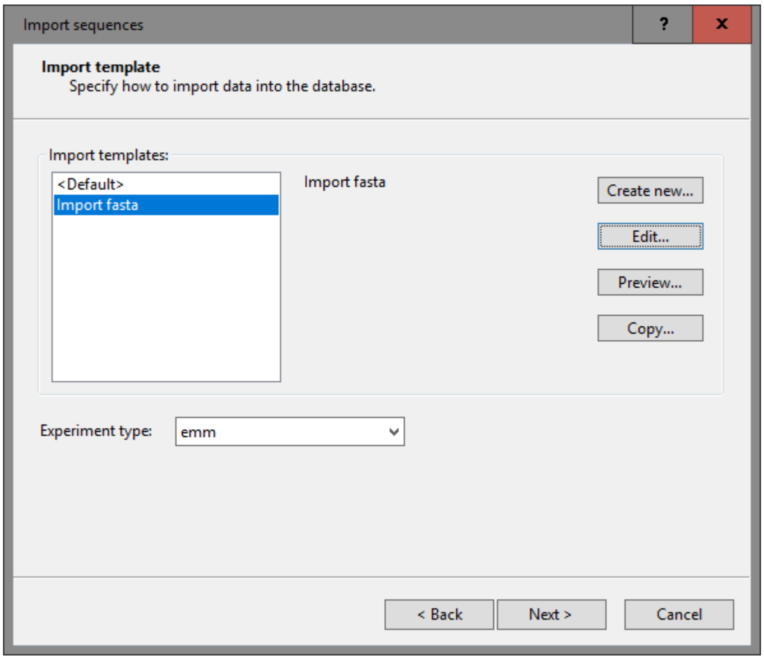
12. Highlight the newly created template and select ***emm*** as ***Experiment type*** (see Figure 8).

13. Press <***Next***> and <***Finish***>.

Five new entries are created in the *Database entries* panel and the sequence data is linked to the ***emm*** sequence experiment (see Figure 9).

14. Click on a green colored dot for one of the entries to display the imported sequence in the *Sequence editor* window.

15. Close the *Sequence editor* window.



Import sequences

Import template

Specify how to import data into the database.

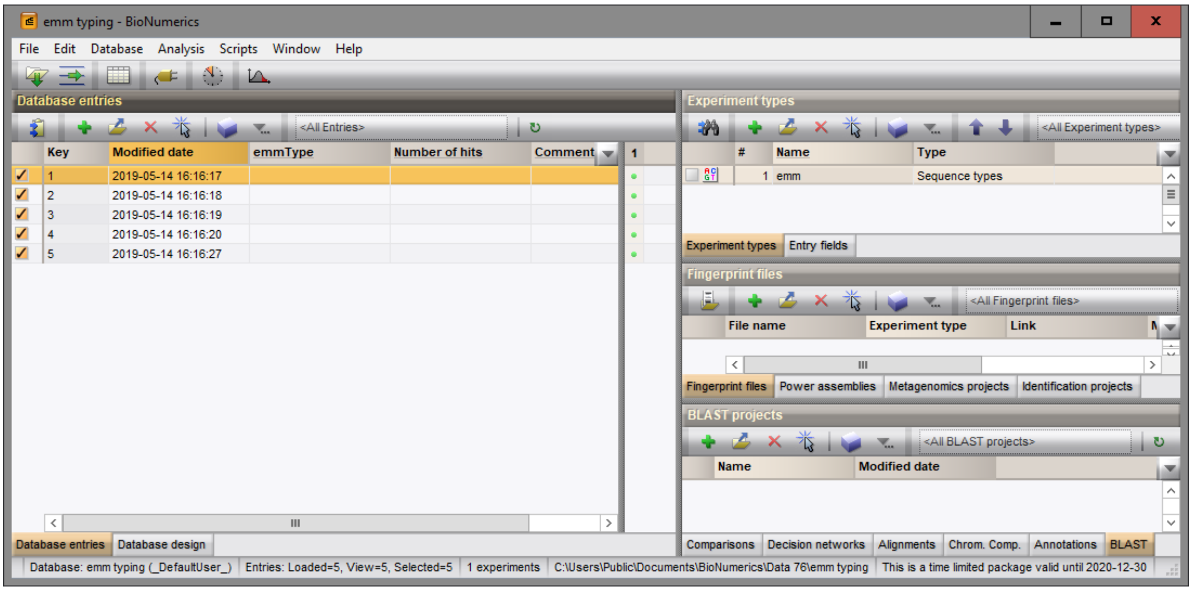
Import templates:

<Default>
Import fasta

Import fasta

Experiment type: emm

Figure 8: Import template.



emm typing - BioNumerics

Database entries

	Key	Modified date	emmType	Number of hits	Comment	1
✓	1	2019-05-14 16:16:17				●
✓	2	2019-05-14 16:16:18				●
✓	3	2019-05-14 16:16:19				●
✓	4	2019-05-14 16:16:20				●
✓	5	2019-05-14 16:16:27				●

Experiment types

	#	Name	Type
□	1	emm	Sequence types

Fingerprint files

File name	Experiment type	Link

BLAST projects

Name	Modified date

Database: emm typing (_DefaultUser_) | Entries: Loaded=5, View=5, Selected=5 | 1 experiments | C:\Users\Public\Documents\BioNumerics\Data 76\emm typing | This is a time limited package valid until 2020-12-30

Figure 9: The *Main* window.

4 Perform emm typing

The BioNumerics *emm typing* script can be found on the download page on our website: http://www.applied-maths.com/download/sample-data, ”emm typing”.

1. Download the sample file from our website and unzip the file.

Emm typing can be done on any selection of entries in database that have linked sequence data in the **emm** sequence type experiment.

2. Select a single entry in the *Database entries* panel by holding the **Ctrl**-key and left-clicking on the entry. Alternatively, use the **space bar** to select a highlighted entry or click the ballot box next to the entry.

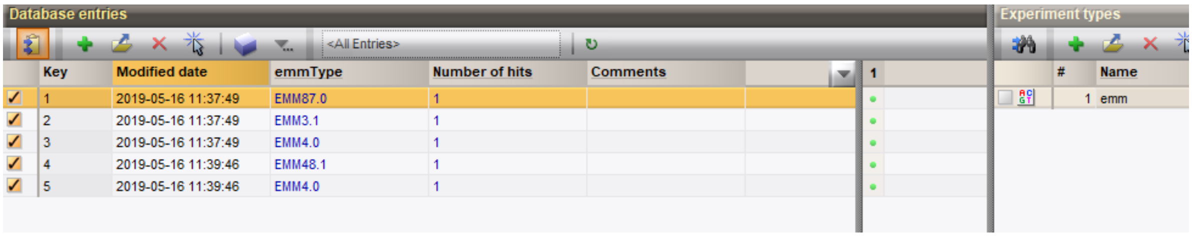
Selected entries are marked by a checked ballot box (☑) and can be unselected in the same way.

3. In order to select a group of entries, hold the **Shift**-key and click on another entry.

4. Make sure a few entries are selected in the *Database entries* panel that have a linked **emm** sequence type experiment.

5. Select ***Scripts*** > ***Run script from file*** and browse for the `emm_typing.py` script.

The emm sequences of all selected entries are screened against the ***emm*** BLAST database. The result, i.e. the ***emm type***, the ***Number of hits*** and optionally a ***Comment*** (hit too long, hit too short, no blast hits), is written in the corresponding entry information fields in the *Database entries* panel (see Figure 10).



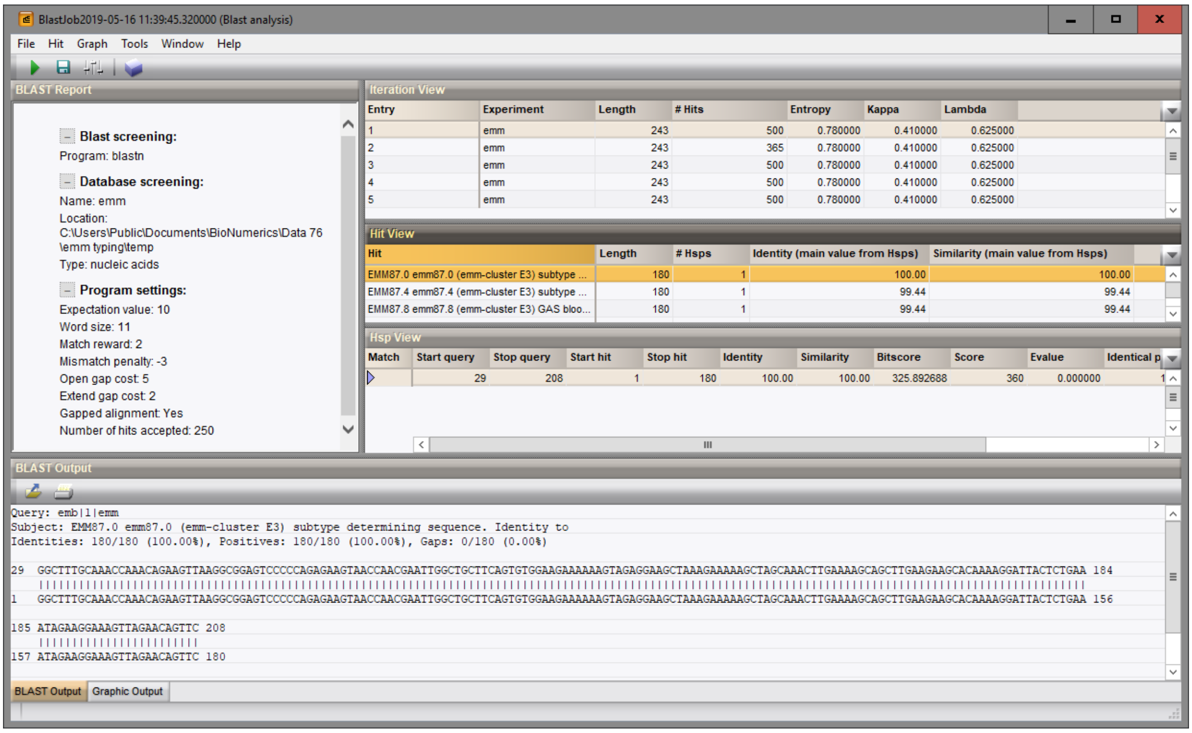
	Key	Modified date	emmType	Number of hits	Comments	1
✓	1	2019-05-16 11:37:49	EMM87.0	1		●
✓	2	2019-05-16 11:37:49	EMM3.1	1		●
✓	3	2019-05-16 11:37:49	EMM4.0	1		●
✓	4	2019-05-16 11:39:46	EMM48.1	1		●
✓	5	2019-05-16 11:39:46	EMM4.0	1		●

	#	Name
□	1	emm

Figure 10: Emm typing results.

Detailed information about the BLAST results per selected entry is displayed in the *BLAST* window (see Figure 11 for an example). More information about this window can be found in the BioNumerics reference manual.

The results in the *BLAST* window can be saved with ***File*** > ***Save...*** (💾) and will become available in the *BLAST projects* panel in the *Main* window.



Blast screening:
Program: blastn

Database screening:
Name: emm
Location:
C:\Users\Public\Documents\BioNumerics\Data 76
\emm typing\temp
Type: nucleic acids

Program settings:
Expectation value: 10
Word size: 11
Match reward: 2
Mismatch penalty: -3
Open gap cost: 5
Extend gap cost: 2
Gapped alignment: Yes
Number of hits accepted: 250

Entry	Experiment	Length	# Hits	Entropy	Kappa	Lambda
1	emm	243	500	0.780000	0.410000	0.625000
2	emm	243	365	0.780000	0.410000	0.625000
3	emm	243	500	0.780000	0.410000	0.625000
4	emm	243	500	0.780000	0.410000	0.625000
5	emm	243	500	0.780000	0.410000	0.625000

Hit	Length	# Hsps	Identity (main value from Hsps)	Similarity (main value from Hsps)
EMM87.0 emm87.0 (emm-cluster E3) subtype ...	180	1	100.00	100.00
EMM87.4 emm87.4 (emm-cluster E3) subtype ...	180	1	99.44	99.44
EMM87.8 emm87.8 (emm-cluster E3) GAS bloo...	180	1	99.44	99.44

Match	Start query	Stop query	Start hit	Stop hit	Identity	Similarity	Bitscore	Score	Evalue	Identical p
▷	29	208	1	180	100.00	100.00	325.892688	360	0.000000	1

```
Query: emb|1|emm
Subject: EMM87.0 emm87.0 (emm-cluster E3) subtype determining sequence. Identity to
Identities: 180/180 (100.00%), Positives: 180/180 (100.00%), Gaps: 0/180 (0.00%)

29   GGCTTTGCAAACCAAACAGAAGTTAAGGCGGAGTCCCCCAGAGAAGTAACCAACGAATTGGCTGCTTCAGTGTGGAAGAAAAAAGTAGAGGAAGCTAAAGAAAAAGCTAGCAAACTTGAAAAGCAGCTTGAAGAAGCACAAAAGGATTACTCTGAA 184
     ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
1    GGCTTTGCAAACCAAACAGAAGTTAAGGCGGAGTCCCCCAGAGAAGTAACCAACGAATTGGCTGCTTCAGTGTGGAAGAAAAAAGTAGAGGAAGCTAAAGAAAAAGCTAGCAAACTTGAAAAGCAGCTTGAAGAAGCACAAAAGGATTACTCTGAA 156

185 ATAGAAGGAAAGTTAGAACAGTTC 208
    ||||||||||||||||||||||||
157 ATAGAAGGAAAGTTAGAACAGTTC 180
```

Figure 11: BLAST project.