APPLIED MATHS
A BIOMÉRIEUX COMPANY

BioNumerics Tutorial:

Importing sequences from GenBank/EMBL files

1 Aim

With the BioNumerics GenBank/EMBL import routine, sequences in GenBank/Embl format can be imported into BioNumerics. In this tutorial you will learn how to use this import tool by importing sequences from an example file.

2 Example data

As an example we will import Influenza A sequences of the hemagglutinin (HA) gene into a new or existing BioNumerics database. The example GenBank file (`H1N1 HA gene.gb` file) can be found on the download page on our website (http://www.applied-maths.com/download/sample-data, ”GenBank file”.). Note that the steps for the import of an EMBL formatted file are the same.

3 The Import wizard

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

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

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

4. Press <***Browse***>, select the `H1N1 HA gene.gb` file and press <***Open***>.

5. With the option ***Preview sequences*** checked, press <***Next***>.

The import wizard now displays a preview of the sequence data in the file (see Figure 1).

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 GenBank 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.

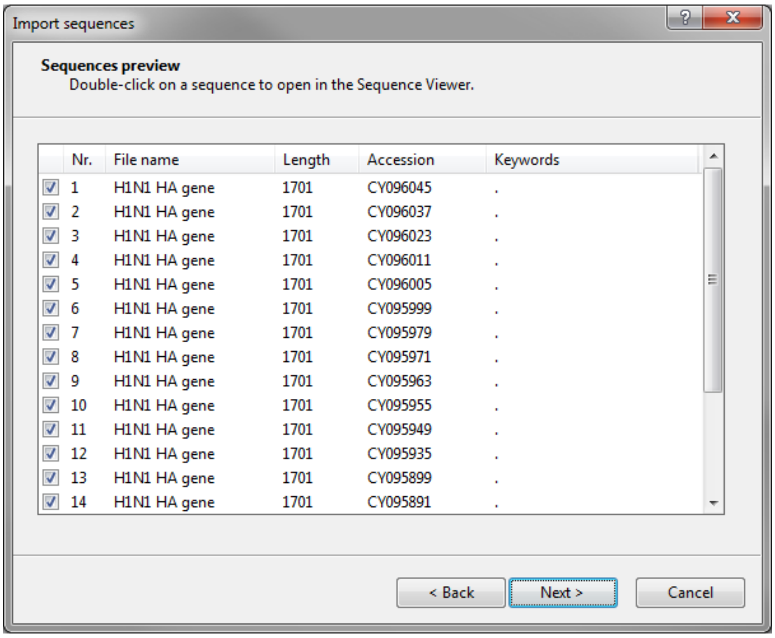
8. Select ”OS - SOURCE” in the list and click <***Edit destination***>.

9. Select ”[Create new]” under ***Entry info field*** and click <***OK***>.

10. Enter “Strain” as name for the new information field, press <***OK***> and press <***Yes***>.

The accession number is valuable information which is specific for the sequence (not for the strain), therefore we will store the accession number as a sequence information field.

11. Select ”AC - ACCESSION” in the list and click <***Edit destination***> or double-click on ”AC - ACCESSION”. Under ***Sequence info field***, select ”[Create new]” and press <***OK***>.



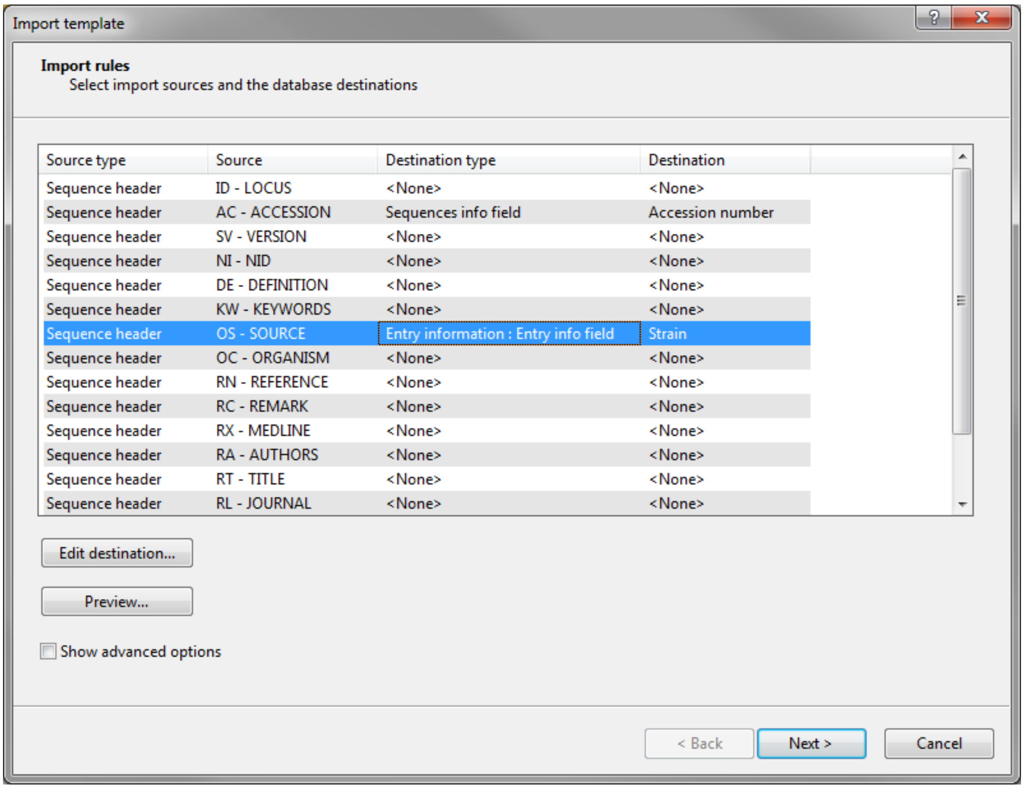
Import sequences

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

	Nr.	File name	Length	Accession	Keywords
☑	1	H1N1 HA gene	1701	CY096045	.
☑	2	H1N1 HA gene	1701	CY096037	.
☑	3	H1N1 HA gene	1701	CY096023	.
☑	4	H1N1 HA gene	1701	CY096011	.
☑	5	H1N1 HA gene	1701	CY096005	.
☑	6	H1N1 HA gene	1701	CY095999	.
☑	7	H1N1 HA gene	1701	CY095979	.
☑	8	H1N1 HA gene	1701	CY095971	.
☑	9	H1N1 HA gene	1701	CY095963	.
☑	10	H1N1 HA gene	1701	CY095955	.
☑	11	H1N1 HA gene	1701	CY095949	.
☑	12	H1N1 HA gene	1701	CY095935	.
☑	13	H1N1 HA gene	1701	CY095899	.
☑	14	H1N1 HA gene	1701	CY095891	.

Figure 1: Preview.

12. Change the suggested name (”AC - ACCESSION”) for the new information field into ”Accession number”, press <***OK***> and confirm with <***Yes***>.



Import template

Import rules
Select import sources and the database destinations

Source type	Source	Destination type	Destination
Sequence header	ID - LOCUS	<None>	<None>
Sequence header	AC - ACCESSION	Sequences info field	Accession number
Sequence header	SV - VERSION	<None>	<None>
Sequence header	NI - NID	<None>	<None>
Sequence header	DE - DEFINITION	<None>	<None>
Sequence header	KW - KEYWORDS	<None>	<None>
Sequence header	OS - SOURCE	Entry information : Entry info field	Strain
Sequence header	OC - ORGANISM	<None>	<None>
Sequence header	RN - REFERENCE	<None>	<None>
Sequence header	RC - REMARK	<None>	<None>
Sequence header	RX - MEDLINE	<None>	<None>
Sequence header	RA - AUTHORS	<None>	<None>
Sequence header	RT - TITLE	<None>	<None>
Sequence header	RL - JOURNAL	<None>	<None>

☐ Show advanced options

Figure 2: Import template

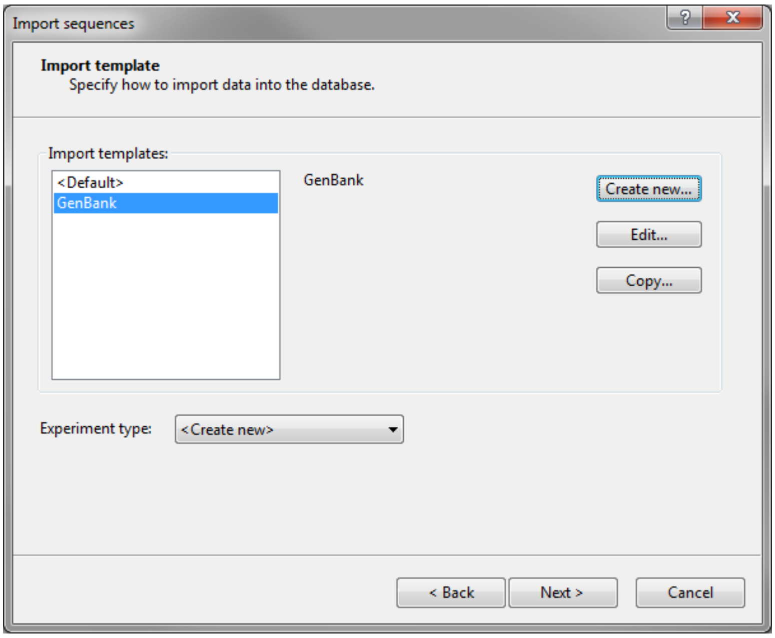
13. Optionally, you can press <***Preview***> to obtain a preview of the data you are about to import.

14. Click <***Next***> to go to the next step.

15. Do not select an ***Entry link field*** to have the database keys automatically generated. Press <***Finish***>.

16. Specify a template name (e.g. “GenBank”) and optionally enter a description. Press <***OK***>.

17. Highlight the newly created template and select ”Create new” as ***Experiment type*** (see Figure 3).



Import sequences

Import template

Specify how to import data into the database.

Import templates:

<Default>
GenBank

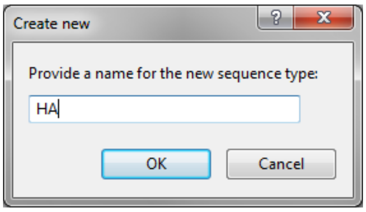
GenBank

Experiment type: <Create new>

Figure 3: Import template.

18. Press <***Next***>.

19. Specify a sequence type name (e.g. **HA** or **haemagglutinin**) and press <***OK***> and confirm the action.



Create new

Provide a name for the new sequence type:

HA

Figure 4: Create a new sequence type.

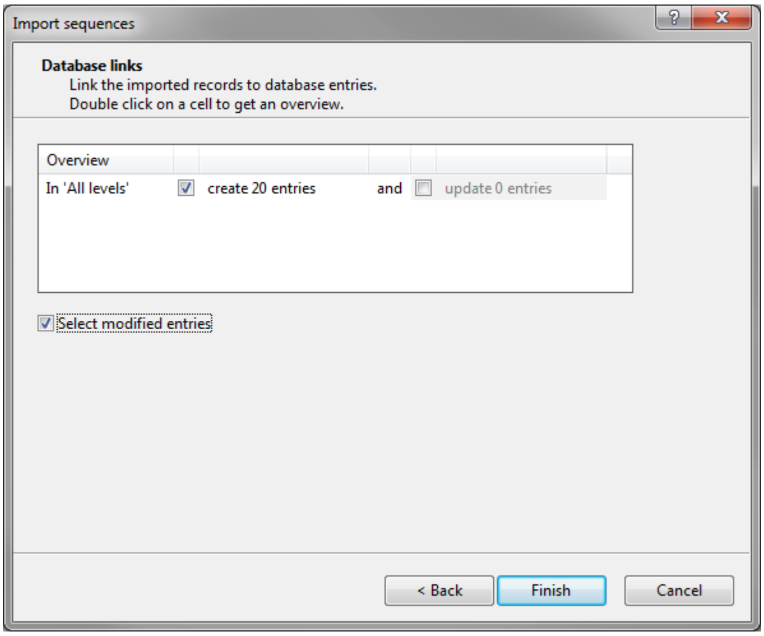
The *Database links* wizard page will indicate that 20 new entries will be created during import (see Figure 5).

20. Press <***Finish***>.

Twenty sequences are imported in the database (see Figure 6). All entries for which information was imported are automatically selected.

21. Click on a green colored dot in the *Experiment presence* panel to open the *Sequence editor* window.

The sequence is displayed in the upper panel and a graphical representation of the sequence is displayed in the panel below. The *Annotation* panel (see Figure 7) holds the GenBank features, the header information is stored in the *Header* panel and the accession number is stored in the *Custom Fields* panel.

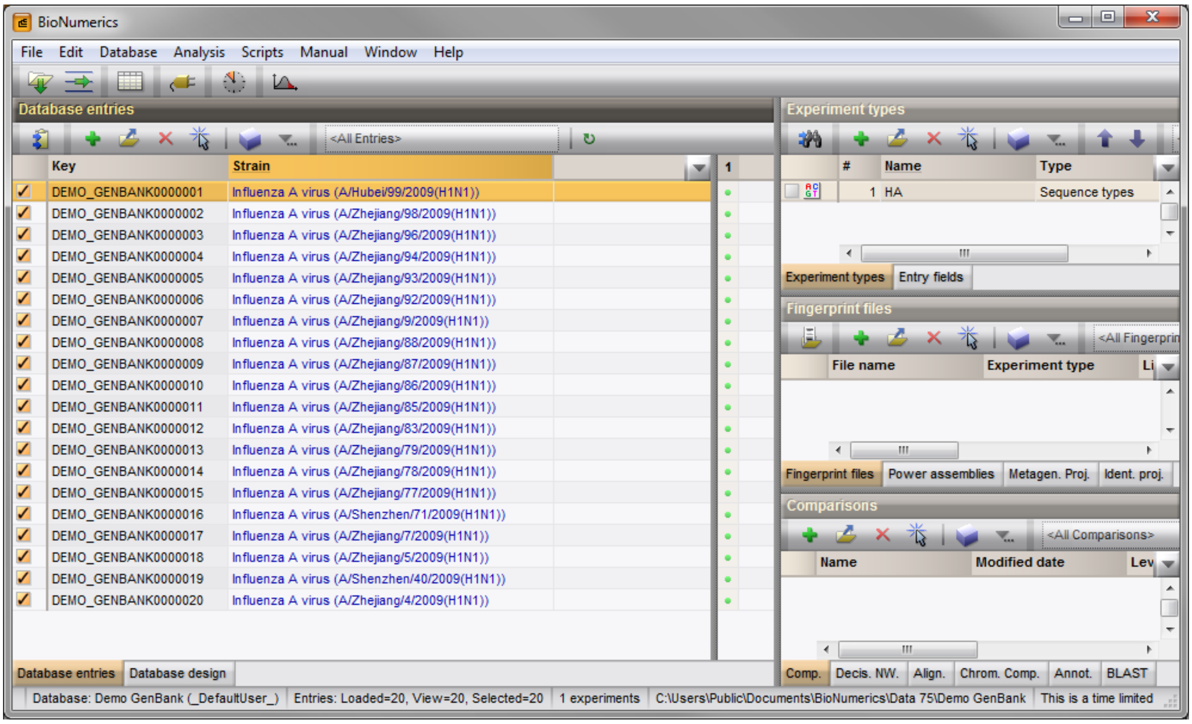


Database links
Link the imported records to database entries.
Double click on a cell to get an overview.

Overview				
In 'All levels'	☑ create 20 entries	and	□ update 0 entries	

☑ Select modified entries

Figure 5: The *Database links* wizard page.



	Key	Strain	1
✓	DEMO_GENBANK0000001	Influenza A virus (A/Hubei/99/2009(H1N1))	●
✓	DEMO_GENBANK0000002	Influenza A virus (A/Zhejiang/98/2009(H1N1))	●
✓	DEMO_GENBANK0000003	Influenza A virus (A/Zhejiang/96/2009(H1N1))	●
✓	DEMO_GENBANK0000004	Influenza A virus (A/Zhejiang/94/2009(H1N1))	●
✓	DEMO_GENBANK0000005	Influenza A virus (A/Zhejiang/93/2009(H1N1))	●
✓	DEMO_GENBANK0000006	Influenza A virus (A/Zhejiang/92/2009(H1N1))	●
✓	DEMO_GENBANK0000007	Influenza A virus (A/Zhejiang/9/2009(H1N1))	●
✓	DEMO_GENBANK0000008	Influenza A virus (A/Zhejiang/88/2009(H1N1))	●
✓	DEMO_GENBANK0000009	Influenza A virus (A/Zhejiang/87/2009(H1N1))	●
✓	DEMO_GENBANK0000010	Influenza A virus (A/Zhejiang/86/2009(H1N1))	●
✓	DEMO_GENBANK0000011	Influenza A virus (A/Zhejiang/85/2009(H1N1))	●
✓	DEMO_GENBANK0000012	Influenza A virus (A/Zhejiang/83/2009(H1N1))	●
✓	DEMO_GENBANK0000013	Influenza A virus (A/Zhejiang/79/2009(H1N1))	●
✓	DEMO_GENBANK0000014	Influenza A virus (A/Zhejiang/78/2009(H1N1))	●
✓	DEMO_GENBANK0000015	Influenza A virus (A/Zhejiang/77/2009(H1N1))	●
✓	DEMO_GENBANK0000016	Influenza A virus (A/Shenzhen/71/2009(H1N1))	●
✓	DEMO_GENBANK0000017	Influenza A virus (A/Zhejiang/7/2009(H1N1))	●
✓	DEMO_GENBANK0000018	Influenza A virus (A/Zhejiang/5/2009(H1N1))	●
✓	DEMO_GENBANK0000019	Influenza A virus (A/Shenzhen/40/2009(H1N1))	●
✓	DEMO_GENBANK0000020	Influenza A virus (A/Zhejiang/4/2009(H1N1))	●

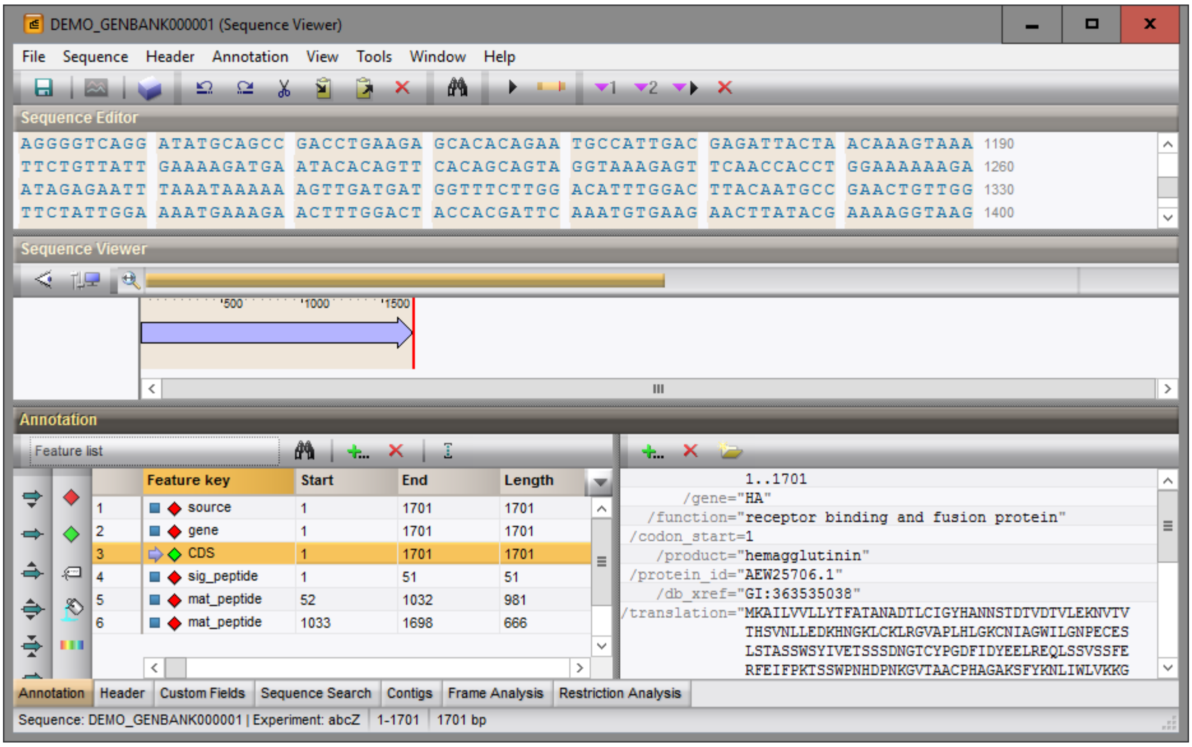
#	Name	Type
1	HA	Sequence types

Database: Demo GenBank (_DefaultUser_) | Entries: Loaded=20, View=20, Selected=20 | 1 experiments | C:\Users\Public\Documents\BioNumerics\Data 75\Demo GenBank | This is a time limited

Figure 6: The *Main* window.

4 Conclusion

In this tutorial you have seen how easy it is to import GenBank/Embl formattted sequences in BioNumerics. The sequences can now be analyzed in BioNumerics. More information can be found in the analysis tutorials on our website.



```
AGGGGTCAGG ATATGCAGCC GACCTGAAGA GCACACAGAA TGCCATTGAC GAGATTACTA ACAAAGTAAA 1190
TTCTGTTATT GAAAAGATGA ATACACAGTT CACAGCAGTA GGTAAAGAGT TCAACCACCT GGAAAAAAGA 1260
ATAGAGAATT TAAATAAAAA AGTTGATGAT GGTTTCTTGG ACATTTGGAC TTACAATGCC GAACTGTTGG 1330
TTCTATTGGA AAATGAAAGA ACTTTGGACT ACCACGATTC AAATGTGAAG AACTTATACG AAAAGGTAAG 1400
```

	Feature key	Start	End	Length
1	source	1	1701	1701
2	gene	1	1701	1701
3	CDS	1	1701	1701
4	sig_peptide	1	51	51
5	mat_peptide	52	1032	981
6	mat_peptide	1033	1698	666

```
                1..1701
           /gene="HA"
       /function="receptor binding and fusion protein"
 /codon_start=1
     /product="hemagglutinin"
 /protein_id="AEW25706.1"
     /db_xref="GI:363535038"
/translation="MKAILVVLLYTFATANADTLCIGYHANNSTDTVDTVLEKNVTV
              THSVNLLEDKHNGKLCKLRGVAPLHLGKCNIAGWILGNPECES
              LSTASSWSYIVETSSSDNGTCYPGDFIDYEELREQLSSVSSFE
              RFEIFPKTSSWPNHDPNKGVTAACPHAGAKSFYKNLIWLVKKG
```

Sequence: DEMO_GENBANK000001 | Experiment: abcZ | 1-1701 | 1701 bp

Figure 7: The *Sequence editor* window.