BioNumerics Tutorial:

Importing sequences from FASTA files

1 Aim

With the BioNumerics FASTA import routine, sequences in FASTA 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 sequences from the FASTA file `H1N1_HA.txt` into a new or existing BioNumerics database. This FASTA file contains the sequence of the HA (haemagglutinin) viral segment for a set of 200 Influenza A H1N1 strains. The example file can be found on the download page on our website (`http://www.applied-maths.com/download/sample-data`, "FASTA files").

3 The Import wizard

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

2. Select ***File*** > ***Import...*** (, **Ctrl+I**) to open the *Import* dialog box.

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

4. The import wizard allows you to browse for one or more text files as data source. Press <***Browse***>, navigate to the folder, select the `H1N1_HA.txt` file and press <***Open***> (see Figure 1).

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

The import wizard now displays a preview of the sequence data in the FASTA file (see Figure 2). From this preview, it is clear that the first FASTA field contains the accession number, the second field the strain number, the third field the date of isolation and the fourth field the country of origin.

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". Under ***Entry info field***, select "[Create new]" and press <***OK***>. Change the suggested name for the new field into "Accession number" and confirm the action twice.

9. Double-click on "Field 2". Select "[Create new]" under ***Entry info field*** and click <***OK***>. Enter "Strain" as name for the new information field, press <***OK***> and confirm the creation of the new field with <***Yes***>.

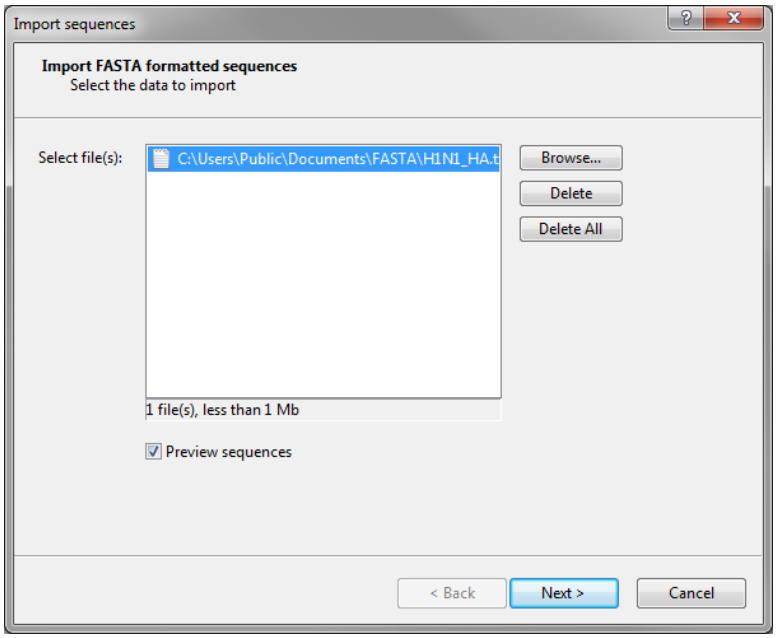

Figure 1: Select the FASTA file(s).

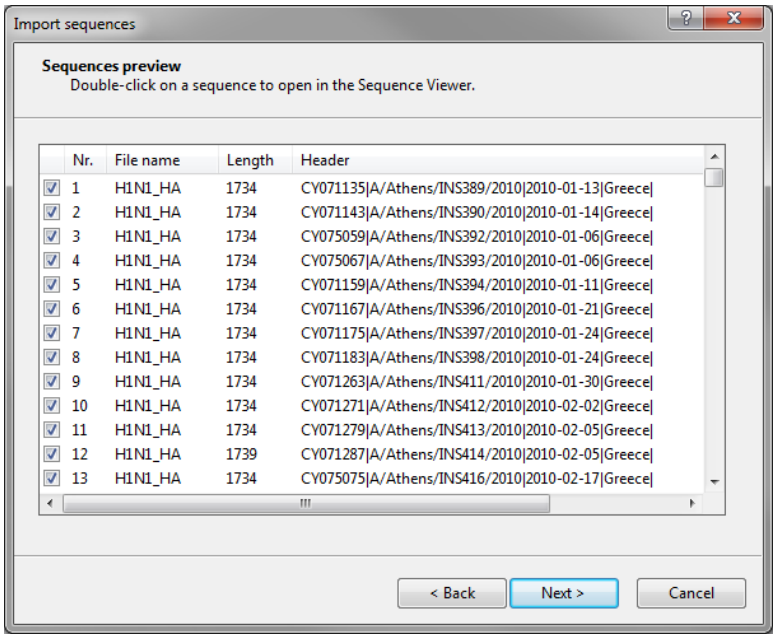

Figure 2: Sequence preview.

10. Using the **Ctrl**-key, highlight ”Field 3” and ”Field 4” in the list. Click <***Edit destination***> or double-click on ”Field 4”.

11. Highlight ***Entry info field*** and press <***OK***>.

12. Change the suggested names into ”Date” and ”Country” for ”Field 3” and ”Field 4”, respectively.

13. Press <***OK***> and confirm the creation of the new information fields with <***Yes***>.

The grid is updated (see Figure 4).

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

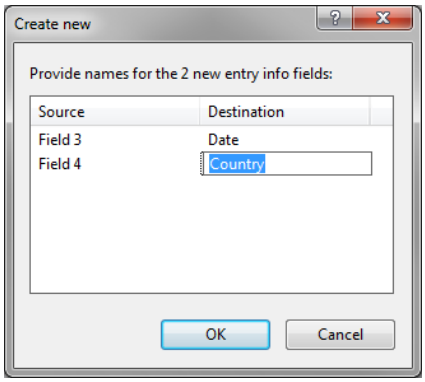

Figure 3: Create new entry information fields.

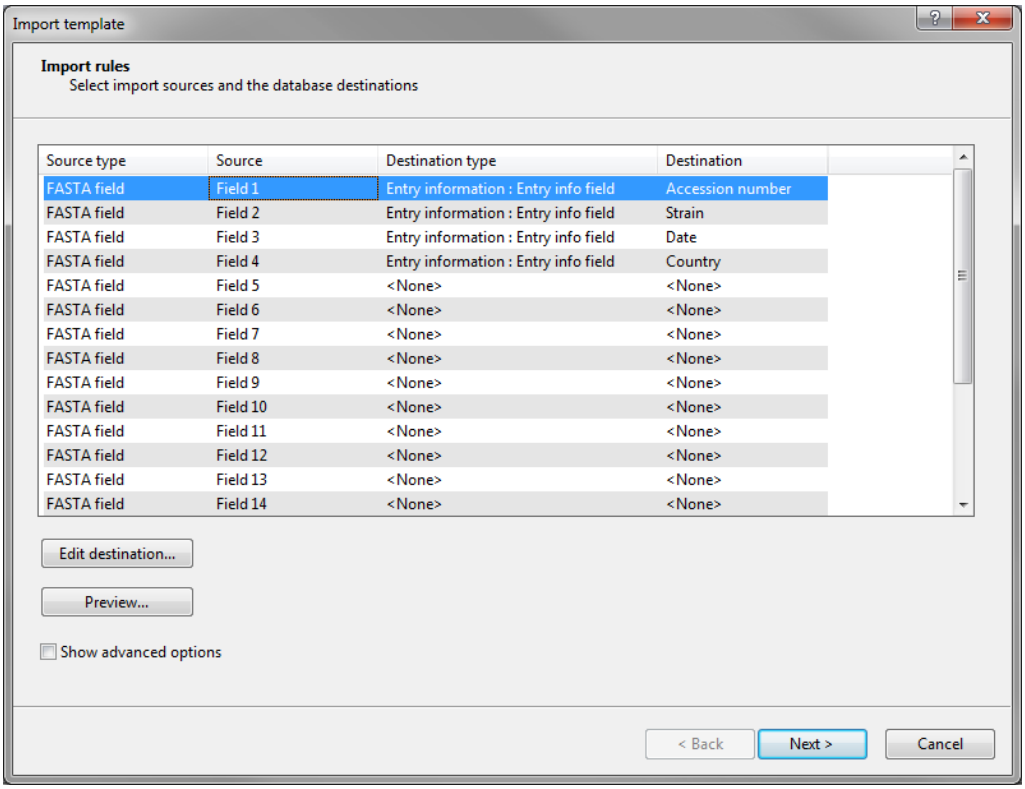

Figure 4: Import rules.

15. Click <***Next***>.

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

17. Specify a template name, e.g. ”FASTA”, and optionally enter a description. Press <***OK***>.

18. Highlight the newly created template and select ”Create new” as ***Experiment type*** (see Figure 5).

19. Press <***Next***>.

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

The *Database links* wizard page will indicate that 210 new entries will be created during import (see Figure 6).

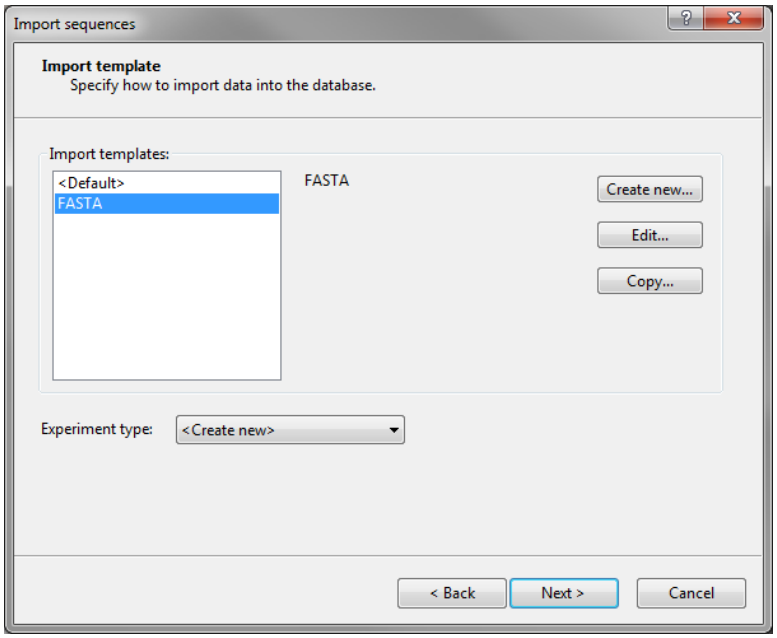

Figure 5: Import template.

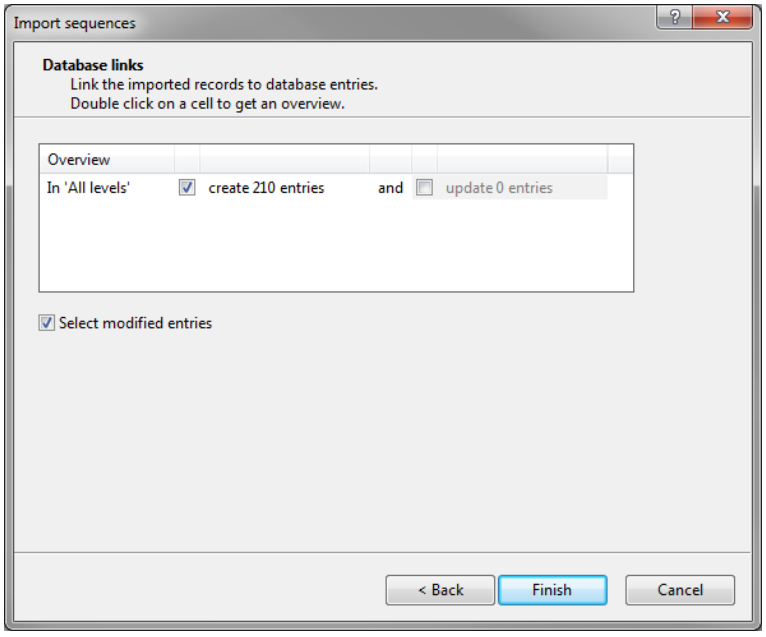

Figure 6: The *Database links* wizard page.

21. Press <***Finish***> to start the import into the database.

For 210 strains, strain information and sequences for the HA genome segment are imported in the database (see Figure 7).

4 Conclusion

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

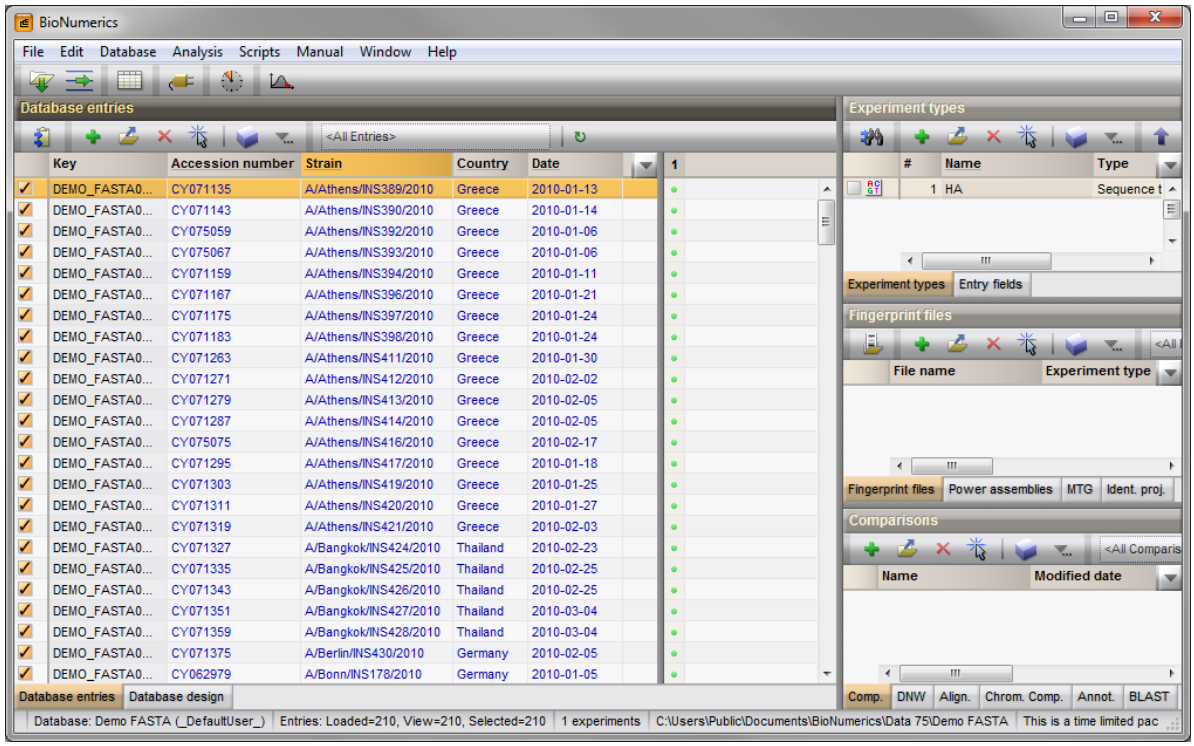

Figure 7: The *Main* window after import of the sequences.

our website.