BIONUMERICS Tutorial:

Importing MLST data from an Excel file

1 Aims

This tutorial shows how to import MLST allele information from a MS Excel file as character data in a BIONUMERICS database. It illustrates the use of *import templates* in the software. Import templates specify from which external field – in a file or a database – information should be imported into which information field in BIONUMERICS (e.g. an entry field, character value or character experiment field). Furthermore, it specifies whether the information should simply be copied or that information should be ”extracted” (i.e. parsed) from the external field via simple or complex parsing instructions.

The idea behind import templates is to create them once and use them many times. Furthermore, existing templates can be copied and edited to match similarly, but not identically, structured external data files. Import templates can be shared among database users and even exported as XML files and imported again in a different database.

2 Example data

The example Excel file from which we will import data in this tutorial contains preprocessed MLST information for about 500 *Neisseria meningitidis* strains and can be downloaded from https://www.applied-maths.com/download/sample-data (click on ”Neisseria MLST data”).

1. Open the file `Neisseria MLST.xlsx` in Excel to examine the data that will be imported.

The first sheet contains for 500 isolates following information: a unique identifier (”Key”), a strain number, an MLST sequence type that was deduced from the analysis (”ST”), the clonal complex information (”CC”), the serogroup, the country where the strains originate from, the year of isolation, the species and the disease in which the strains were involved. This information should be imported as entry information fields in BIONUMERICS. Note that information is missing for some fields. Finally, the allele number is reported for each of the seven loci sequenced (”abcZ”, ”adk”, ”aroE”, ”fumC”, ”gdh”, ”pdhC” and ”pgm”) for all 500 strains.

2. Close the Excel file again.

Using the *MLST online plugin*, the complete MLST analysis can be performed in BIONUMERICS, starting from sequencer trace files.

3 Creating a new character type

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

Since we will be importing MLST allelic profiles as character data, we will first create a character type to hold this data. The steps below can be skipped if a suitable character type is already present in the database.

2. In the *Main* window, click on + in the toolbar of the *Experiment types* panel and select ***Character type*** from the list. Press <***OK***>.

The *New character type* wizard prompts you to enter a name for the new character type.

3. Enter a name, for example “MLST” and press <***Next***>.

In the next step of the wizard, the choice is offered between ***Numerical values*** and ***Binary data***.

4. Choose ***Numerical values***.

5. Since we only want to use integer values, leave the number of decimal digits unaltered (zero). Press <***Next***>.

The wizard asks if the character type has an open (***Yes***) or closed (***No***) character set.

6. Answer ***No*** and make sure the ***Number of rows*** and ***Number of columns*** is set to zero.

7. Press the <***Finish***> button to complete the setup of the new character type.

The *Experiment types* panel now lists the new character type **MLST**.

4 Import procedure

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

2. Choose the option ***Import fields and characters (Excel file)*** under the ***Character type data*** item in the tree (see Figure 1) and press <***Import***>.

3. Press <***Browse***> and browse for the downloaded `Neisseria MLST.xlsx` file. Next, press <***Open***>.

4. Specify the ***Data range*** that contains the MLST allelic profiles (i.e. “Sheet1”) (see Figure 2) and press <***Next***>.

As this is the first time we import character data from Excel into the database, we need to create a new import template by specifying ***Import rules***.

5. Select ”Key” in the list and click <***Edit destination***> or simply double-click on ”Key”. Select ”Key” as the BIONUMERICS destination field in the *Edit data destination* dialog box and press <***OK***>.

6. Double-click on ”Strain” in the list.

Since there is no field present yet in the database to hold this information, the database information field needs to be created first.

7. Select ”<Create new>” under ”Entry info field” and press <***OK***>.

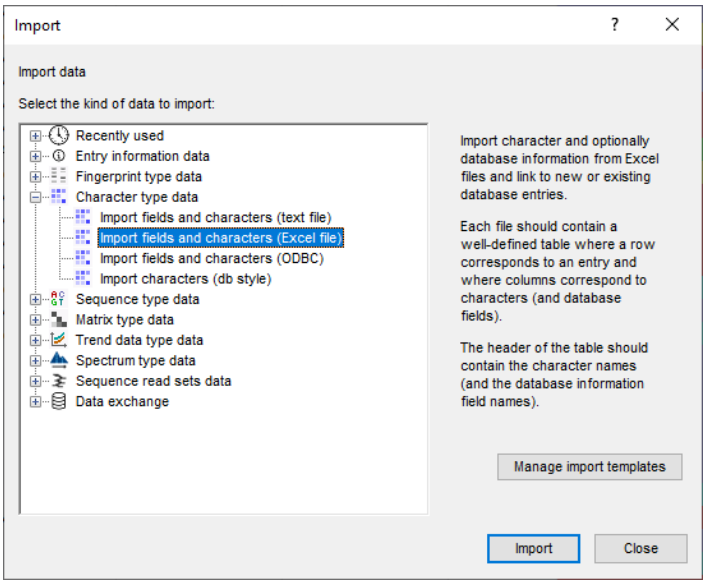

Figure 1: The *Import* dialog box.

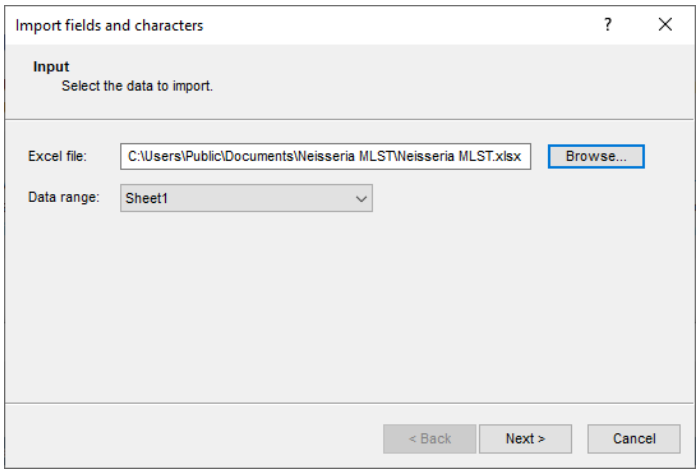

Figure 2: Select Excel file and data range.

8. In the dialog box that appears, press <***OK***> to accept the suggested name (by default the same of the corresponding column in the Excel file) and confirm this modification to the database with <***Yes***>.

Instead of assigning the data destinations one by one, the external fields that should be imported as entry information can be assigned in bulk:

9. Click on ”ST” and whilst holding the **Shift**-key, click on ”Disease” to make a multiple selection of the external fields ”ST”, ”CC”, ”Serogroup”, ”Country”, ”Year”, ”Species” and ”Disease”.

10. Press <***Edit destination***>, select ”Entry info field” as destination and click <***OK***>.

The software tries to map the external fields to a BIONUMERICS information field. When there is no field present with the same name (as is the case here, since we are importing in an empty database), you will be prompted to create the new information fields.

11. Press <***OK***> and then <***Yes***> to confirm the creation of the new entry information fields.

12. Make a multiple selection for all MLST housekeeping genes. Do this by selecting ”abcZ”, scroll down and while holding the **Shift**-key, click on ”pgm”. Press <***Edit destination***>, select ”MLST” under ***Character value*** as destination (see Figure 3) and click <***OK***>.

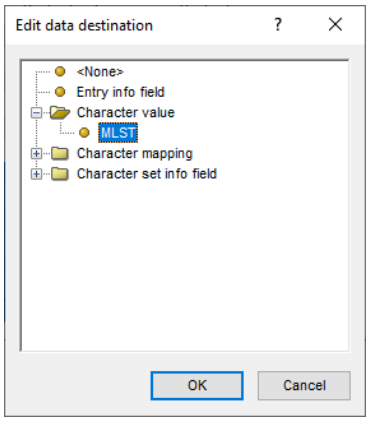

Figure 3: Link to the MLST character type experiment.

13. Press <***OK***> and then <***Yes***> to confirm the creation of new characters.

The grid panel is updated (see Figure 4).

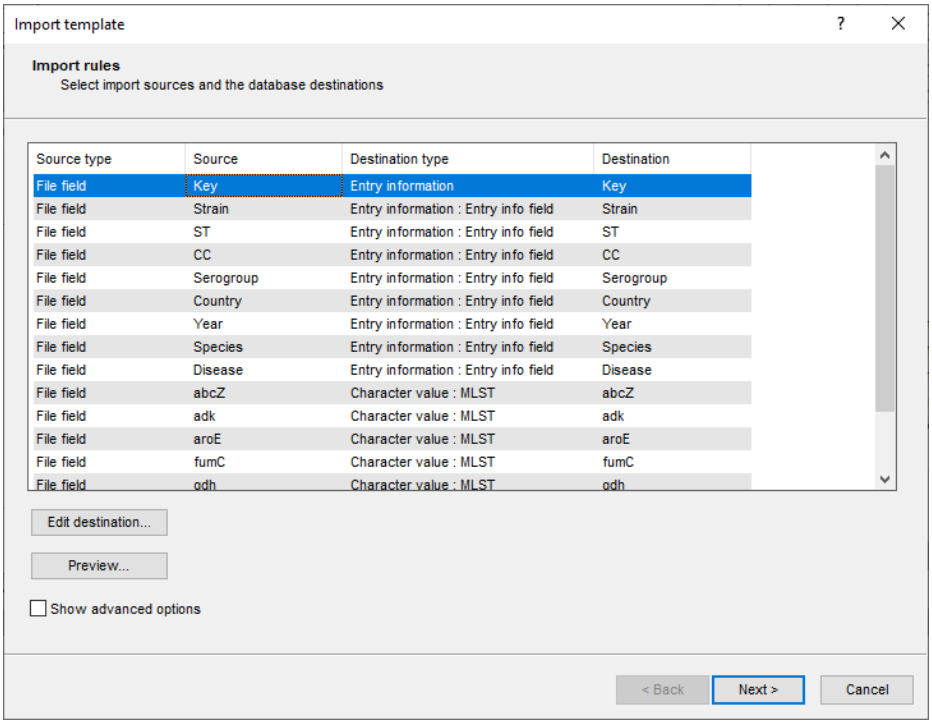

Figure 4: The import rules.

14. Press <***Preview***> to see what you are about to import.

From the preview, it can be seen that both the genus and species designation will be imported into the ’Species’ field. To split this information over a ’Genus’ and ’Species’ field, we will need to define *parsing rules*.

15. Press the <***Close***> button to close the preview and then check ***Show advanced options***.

The advanced options appear in the *Import rules* dialog box.

16. Press the <***Add rule***> button.

17. Under ***File field***, select the ”Species” data source and press <***Next***>.

18. Select ”<Create new>” under ”Entry info field” and press <***Next***>.

19. In the dialog box that appears, enter “Genus” and press <***OK***>. Confirm this modification to the database with <***Yes***>.

20. With ***Parse component*** checked, enter “[DATA] *” (i.e. ”DATA” between square brackets, followed by a space and an asterisk) as ***Data parsing string*** (see Figure 5). This parsing string instructs the software to import only that part of the information that occurs before the first space, which in this case corresponds to the genus designation.

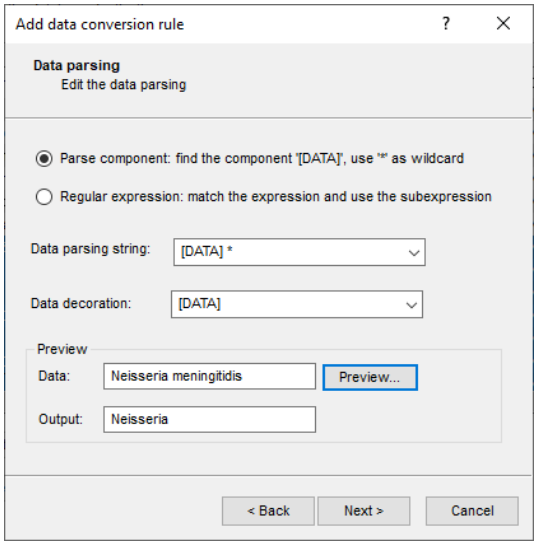

Figure 5: Data conversion rule.

21. Press <***Next***> and <***Finish***> to complete the parsing rule.

22. Highlight the ***Species*** row in the grid and press <***Edit parsing***>.

23. Enter “* [DATA]” as ***Data parsing string*** (see Figure 6) and press <***OK***>.

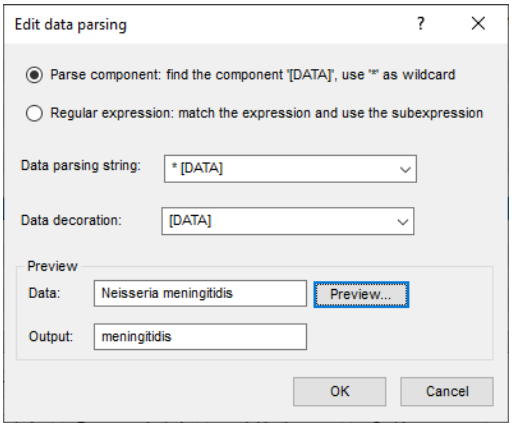

Figure 6: Data conversion rule.

Optionally, call the preview again to verify that the genus and species name now are correctly split up and will be imported in the corresponding information fields.

24. Press <***Next***> to proceed to the *Import links* dialog box.

25. Make sure ”Key” is checked as link and press <***Finish***>.

The import template needs to be saved to be able to use it again later on.

26. Enter a ***Name*** for the import template (e.g. “MLST Excel”) and optionally a ***Description***. Next, press <***OK***>.

27. In the *Import template* wizard page, highlight the newly created template and click <***Next***>.

In case there are no entries present with the same key as in the external file, the *Database links* wizard page will indicate that 500 new entries will be created during import.

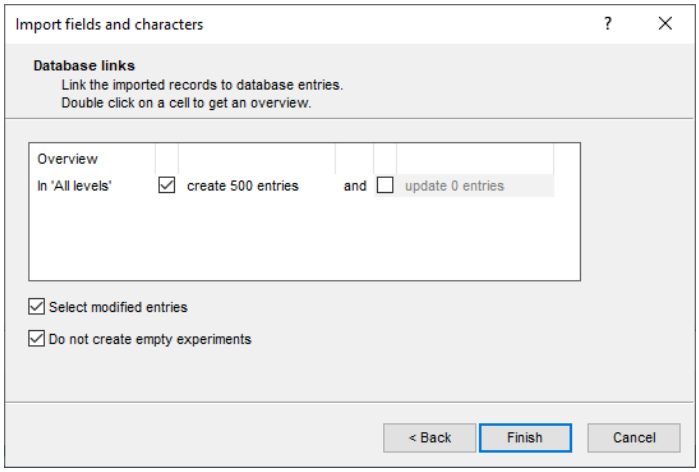

Figure 7: Create 500 new entries.

28. Press <***Finish***> to start the actual import. The progress of the import is shown while database information is added to the BIONUMERICS database.

The entry information data is displayed in the *Database entries* panel and all entries are automatically selected (see Figure 8).

The character data is stored in the character type **MLST**.

29. To view the values in a list, double-click on the experiment **MLST** in the *Experiment types* panel, select ***Settings*** > ***General settings...*** (), select the *Experiment card* tab and change the representation to ***List***. Close the two windows.

30. Click on a green colored dot in the *Experiment presence* panel to open the experiment card for an entry.

The imported allele numbers are displayed in the experiment card next to the corresponding housekeeping gene names (see Figure 9).

31. Close the experiment card by clicking in the left upper corner of the card.

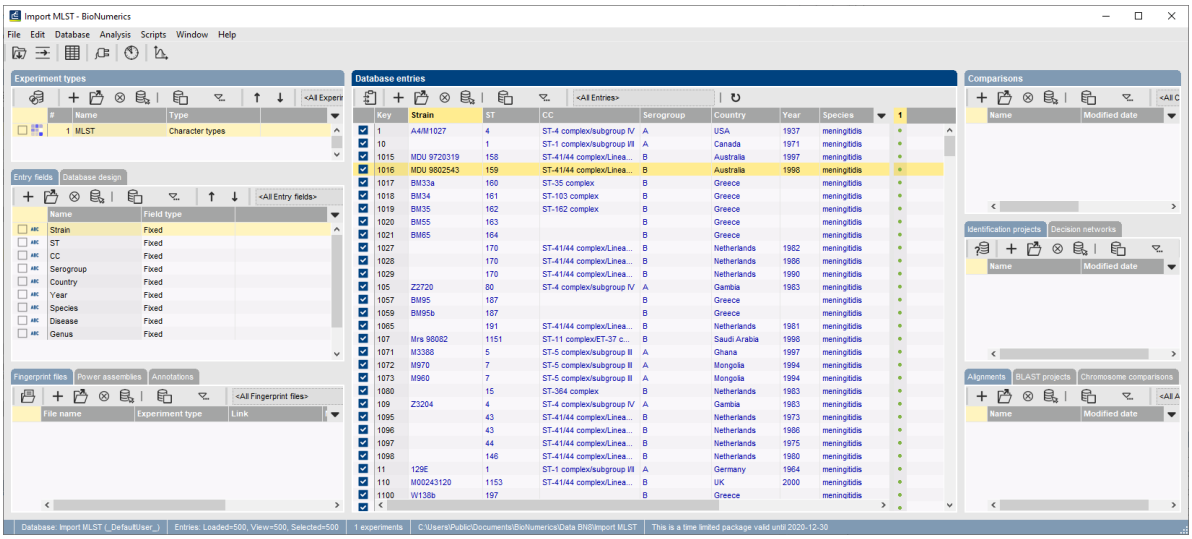

Figure 8: The *Main* window after import of the data.

1

Character	Value	Mapping
abcZ	1	<+>
adk	3	<+>
aroE	3	<+>
fumC	1	<+>
gdh	4	<+>
pdhC	2	<+>
pgm	3	<+>

Figure 9: The experiment card.