BioNumerics Tutorial:

Identifying unknown samples based on peak data

1 Aim

BioNumerics contains powerful tools for the identification of unknown samples against a reference set. With the internal validation options, the user knows exactly how reliable the identification is and which type of errors can be expected. Different data types or combinations of data types can be used for identification. In this tutorial we will use peak data as dataset for the identification.

2 Preparing the sample database

1. Create a new database and import the example raw spectra files as described in the tutorial: ”Importing and preprocessing raw spectrum data”.

2. Do a peak matching on the raw spectra as described in the tutorial: ”Peak matching and follow up analysis of spectra”.

3. Select all peak classes that distinguish Species C from Species A and B and store the selected peak classes in a new peak class *view* as described in the tutorial: ”Peak matching and follow up analysis of spectra”.

4. Double-click the spectral experiment **Maldi** in the *Experiment types* panel of the *Main* window.

5. Click on the *Peak Classes* tab in the *Spectrum type* window.

The <***All peak classes***> view displays all peak classes saved after peak matching. A second view is available, called ***Distinguishing Species C***, containing peak classes that reliably distinguish Species C from Species A and B (see Figure 1).

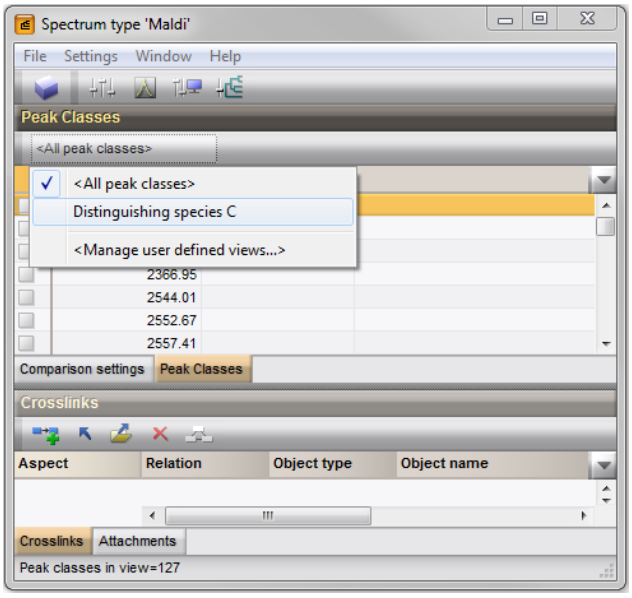

Figure 1: Two peak class views.

6. Select the view ***Distinguishing Species C*** from the drop-down list in the toolbar of the *Peak Classes* panel.

The peak class list is updated.

7. Close the *Spectrum type* window.

3 Creating the reference comparison

Before creating an identification project, we first need to create a comparison containing the *reference set* against which our *unknown samples* will be identified.

1. Make sure the ***Raw spectra*** level is selected in the *Database design* panel and click anywhere in the *Database entries* panel to make it the active panel.

2. Select all 80 entries at the lowest level 'Raw spectra' with ***Edit*** > ***Select all*** (**Ctrl+A**).

3. Unselect two entries belonging to Species A. Use the check boxes next to the entries to unselect an entry.

4. Unselect two entries belonging to Species B and do the same for two entries belonging to Species C.

74 entries are now selected. This is our *reference set*. The 6 entries - not included in the reference set - are our *unknown samples*.

5. Click on the <***All Entries***> view in the toolbar of the *Database entries* panel and choose <***Manage user defined views...***> from the drop-down list.

6. Click the <***Add***> button, specify a name (e.g. **Reference set**), make sure ***Subset based*** is checked, and press <***OK***> and <***Exit***> (see Figure 2).

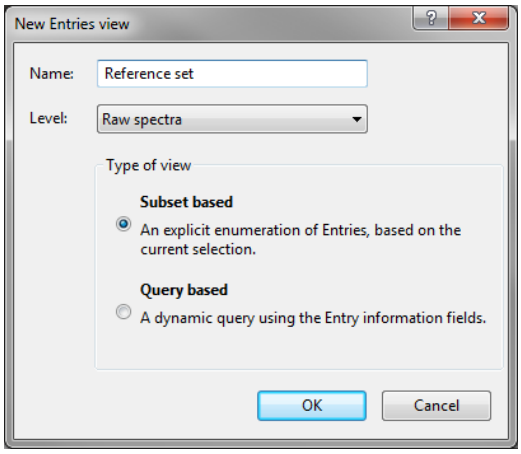

Figure 2: Create a new entry view.

The new view is added to the drop-down list in the *Database entries* panel and is automatically selected.

7. Select in the *Comparisons* panel.

The *Comparison* window opens containing the 74 selected spectra.

8. Press **F4** to clear the selection.

9. Select all spectra belonging to Species A and B. Use the check boxes to select individual spectra, or use the **Ctrl-** and **Shift-** keys to select a range of spectra in the *Information fields* panel.

10. Select ***Groups*** > ***Create new group from selection*** (, **Ctrl+G**), enter a name (e.g. **Species A and B**) and press <***OK***>.

The 56 selected spectra are assigned to a new group and the group is added to the *Groups* panel (see Figure 3).

11. Press **F4** to clear the selection and select all spectra belonging to Species C.

12. Select ***Groups*** > ***Create new group from selection*** (, **Ctrl+G**), enter a name (e.g. **Species C**) and press <***OK***>.

The 18 selected spectra are assigned to a new group and the group is added to the *Groups* panel (see Figure 3).

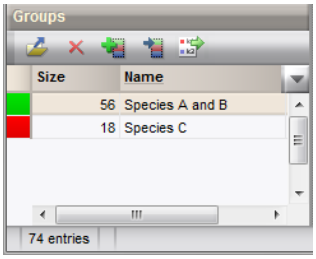

Figure 3: Two groups.

13. Press **F4** to clear the selection.

14. Click on the spectrum type **Maldi** in the *Experiments* panel and select ***Layout*** > ***Show image*** or press the eye button () next to the experiment name in the *Experiments* panel.

15. Select ***Spectra*** > ***Do peak matching*** ().

16. Select ***Existing peak classes only*** and press <***Next***>.

17. Fill in a constant tolerance of "1.9", a linear tolerance of "550" and press <***Finish***>.

18. Save the comparison with ***File*** > ***Save*** (, **Ctrl+S**), name it "RefSet" and close it with ***File*** > ***Exit***.

The reference set is now ready to base our identification project on.

4 Creating the identification project

In the *Main* window, the *Identification projects* panel is displayed in default configuration as a tab.

1. To create a new identification project, select the *Identification projects* tab in the *Main* window and select ***Edit*** > ***Create new object...*** ().

2. Select the comparison **RefSet** and leave the option to lock the reference comparison checked (see Figure 4). This will safeguard the comparison against any accidental changes that might affect the identification results. Press <***Next***>.

3. In the second window of *New identification project* wizard, make sure ***Comparison groups*** is checked as class labels (i.e. ***Species A and B*** and ***Species C*** in our **RefSet** comparison) and click <***Finish***>.

4. Optionally, change the name of the project and press <***OK***>.

We have now defined where our reference set is and what we wish to use as label for the identification. Next, we need to define the classifier(s).

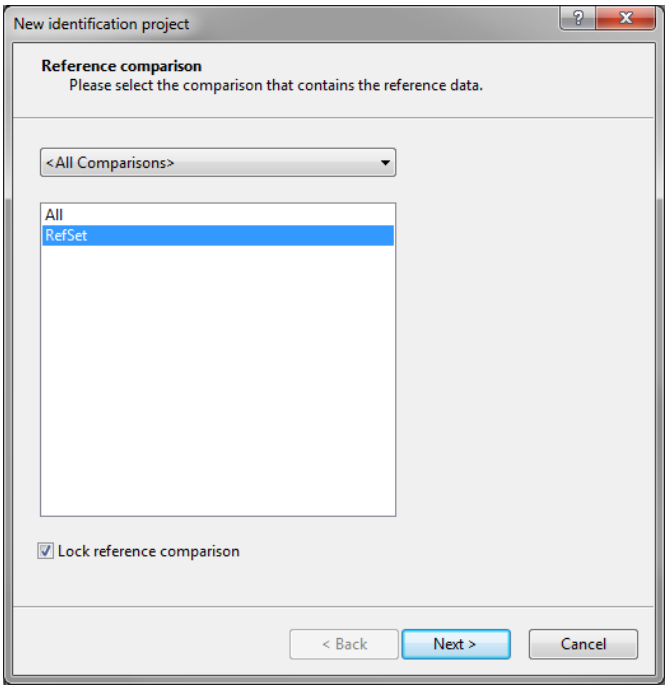

Figure 4: New identification project: step 1.

5 Selecting a classifier

Per identification project, several classifiers can be defined in order to compare identification results from different experiments and /or algorithms. In this tutorial, we will only define one classifier.

1. Create a new classifier by selecting ***Edit*** > ***Create new classifier...*** ( ) in the *Identification project* window.

This opens the *New classifier* wizard.

2. In the first step, select the spectral experiment **Maldi** and press <***Next***>.

In the second step, all algorithms compatible with the selected experiment are listed. This means that this list is different for different experiment types.

3. Select the method **(Distinguishing species C) Character values** and click <***Next***>.

4. In the third step of the *New classifier* wizard, choose **Support Vector Machine (Linear)** as scoring method and press <***Next***> (see Figure 5).

5. Check ***P Value*** and choose ***P Value*** as ***Rank score*** in the last step and press <***Next***>.

6. Optionally change the default suggested classifier name and click <***OK***>.

7. Press <***Yes***> to train the classifier.

The classifier is now present in our identification project and ready for use.

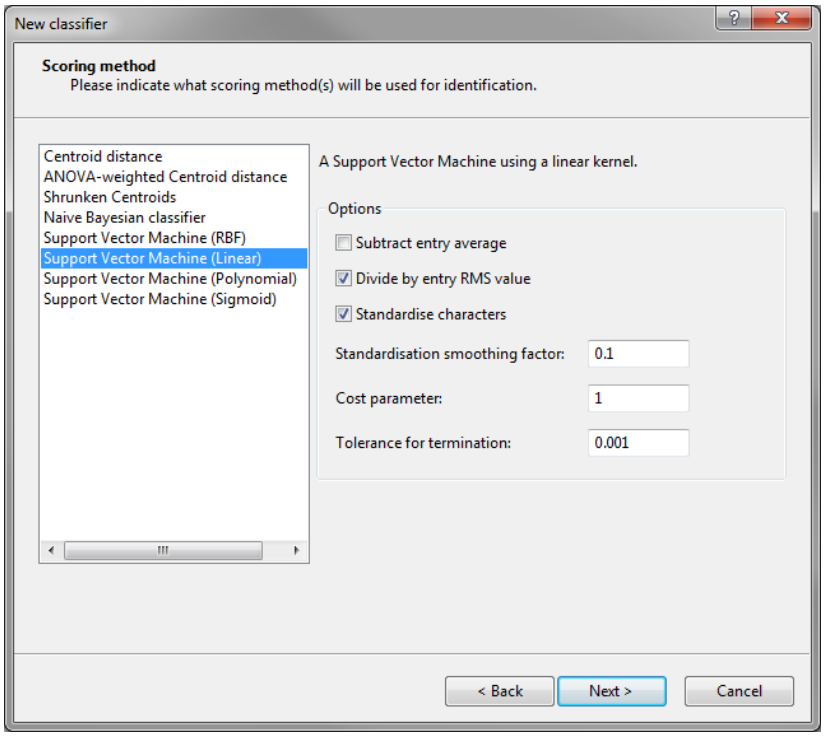

Figure 5: New classifier: step 3.

6 Validating a classifier

It is advised to run a validation on the classifier to check its performance before using it for identification purposes.

1. A tool for internal validation has been included in the software and can be run by selecting ***Edit*** > ***Cross-validation analysis...*** ( ).

2. Leave the settings at default and click <***OK***>.

The validation analysis can take quite some time, especially on large reference sets. In these cases it is advised to increase the test group size and decrease the coverage.

After the cross validation has finished, a detailed overview of the results are shown (see Figure 6). Clicking on a cell in the confusion matrix will give a detailed overview on the entries in this cell in the lower right panel.

3. Close the *Identification cross validation* window, save the identification project (***File*** > ***Save*** (, **Ctrl+S**)) and close it.

We are now ready to identify our unknown samples.

7 Identifying unknown samples

1. Make sure no entries are selected in the *Database entries* panel using ***Database*** > ***Entries*** > ***Unselect all entries (all levels)*** (, **F4**).

2. Select the ***Raw spectra*** level in the *Database design* panel and click anywhere in the *Database entries* panel to make it the active panel.

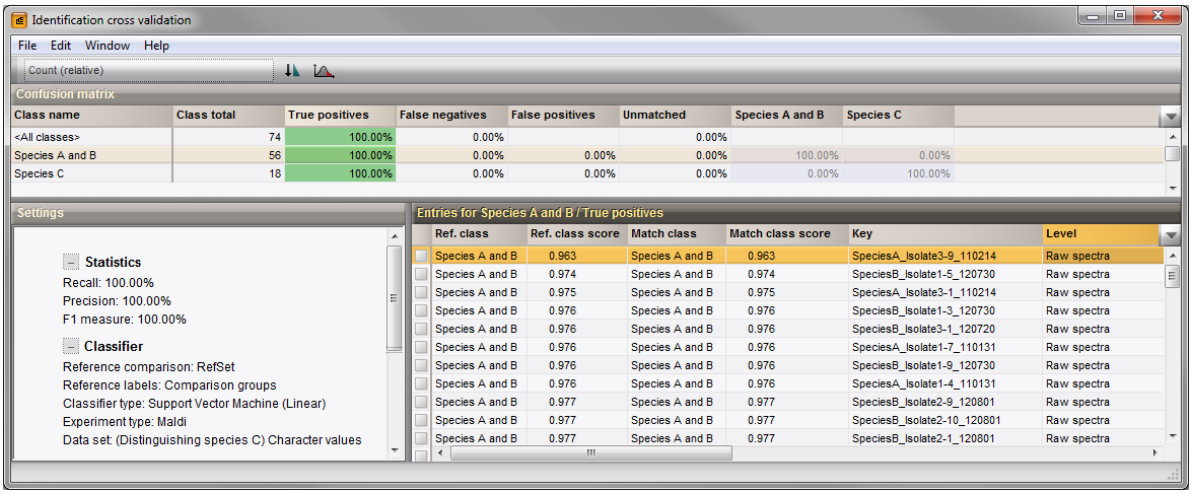

Figure 6: Validation analysis.

3. Make sure the ***Reference set*** view is selected in the toolbar of the *Database entries* panel and select ***Edit*** > ***Select all*** (**Ctrl+A**).

The 74 entries included in the reference set are now selected.

4. Select the <***All Entries***> view in the toolbar of the *Database entries* panel.

80 entries are now listed in the *Database entries* panel, of which 74 entries are selected. To select the 6 spectra that are not included in the reference set, we simply need to invert the selection.

5. Make sure the *Database entries* panel is the active panel and choose ***Edit*** > ***Invert selection***.

Our 6 *unknown* samples are now selected. There is only one identification project present in our database and this project is automatically selected in the *Identification projects* panel.

6. Select ***Analysis*** > ***Identify selected entries...*** () to start the identification wizard.

7. Make sure the option ***Stored classifier*** is checked in the first step and press <***Next***> twice.

The *Identification* window will open with the results of the identification (see Figure 7).

The *Entries* panel lists the unknown entries that were selected for identification. The *Results* panel contains the name of the best matching classes and their identification score. The identification scores of the classifier are obtained using the settings specified in the *Settings* panel. Colored squares appear next to the identification scores. They range from red (improbable identification) over orange, yellow (doubtful identification) to green (faithful identification).

The *Result details* panel lists the best matching classes for the selected unknown entry / classifier combination, ranked by their identification score. The normalized distances and p-values are displayed here as a number. Clicking in the *Entries* panel or *Results* panel updates the *Result details* panel with the information of the newly selected unknown entry / classifier combination.

It can be useful to store the identification results for each unknown entry. It is recommended to first create a dedicated field for this purpose in the database. Results can be transferred to an entry field with ***File*** > ***Transfer results to database*** ().

8. Close the *Identification* window.

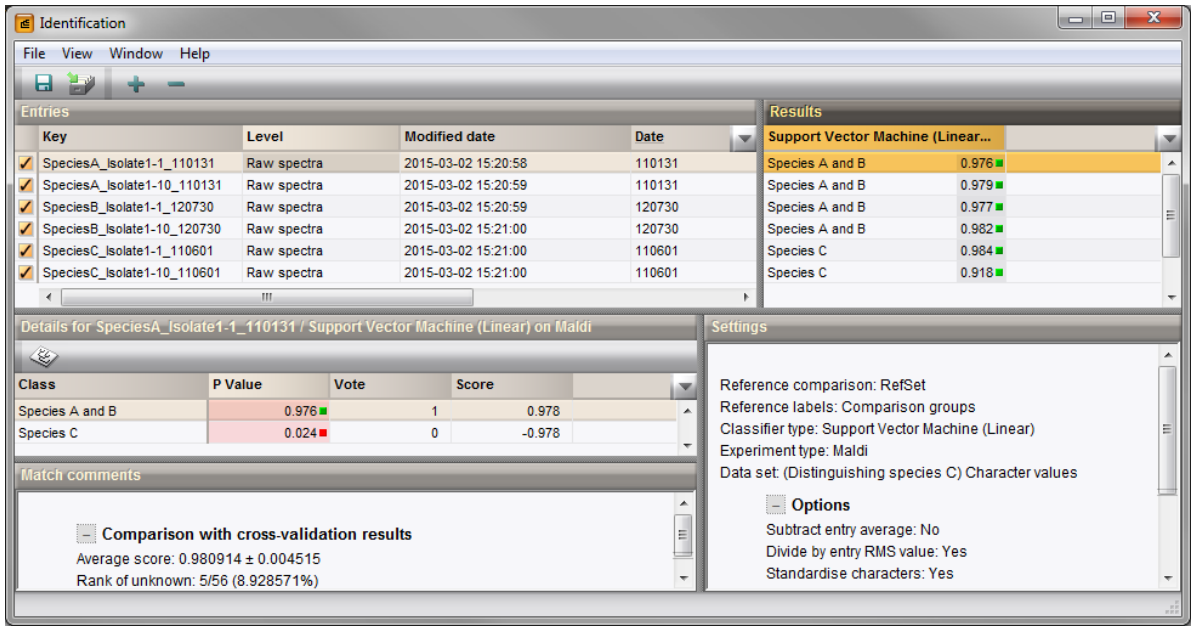

Figure 7: Identification results.