BIONUMERICS Tutorial:

Analyzing spectrum data in a database with levels

1 Aim

In this tutorial a few follow-up analysis tools applicable to spectra are illustated. The steps are illustrated using a BIONUMERICS database that is structured in levels.

2 Preparing the demo database

1. Create a new database and import the example raw spectra files as described in the tutorial: ”Importing spectrum data in a database with levels”. Make sure the *Database design* panel is docked above the *Database entries* panel.

Entries belonging to a selected level in the *Database design* panel are displayed in the *Database entries* panel.

2. Click on the ”Isolate” level in the *Database design* panel (see Figure 1).

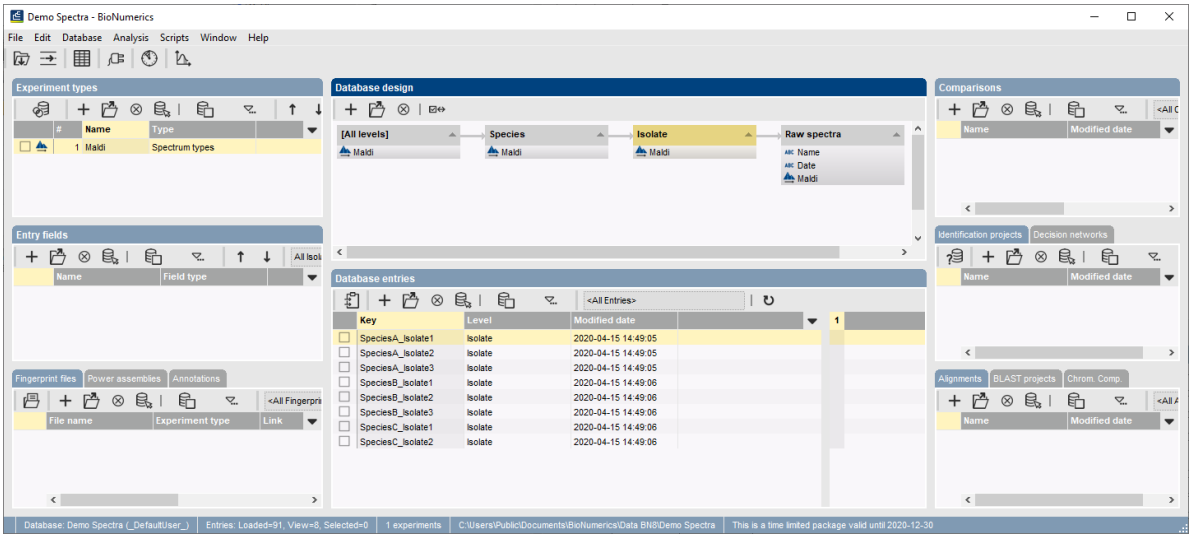

Figure 1: *Database design* panel shown above the *Database entries* panel: the entries belonging to the level ”Isolate” are shown in the *Database entries* panel.

3. Verify the creation of the correct dependencies in each level, by double-clicking on an entry to open the *Entry* window. Click the tab of the *Dependencies* panel to display the dependencies (see Figure 2 for an example).

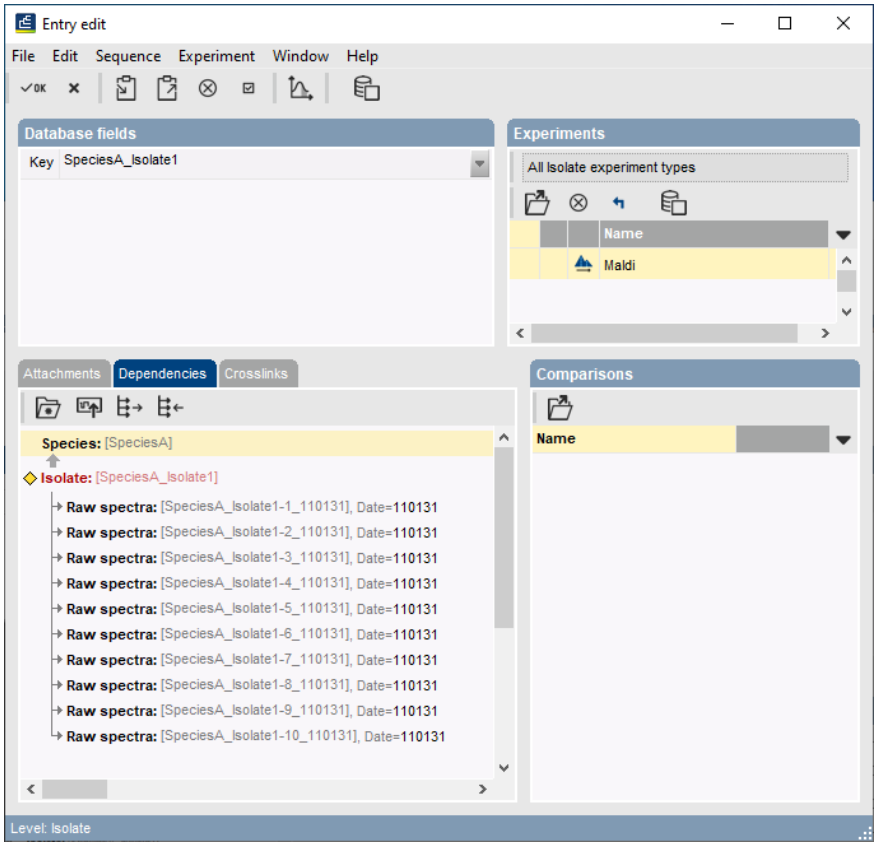

Figure 2: Example of dependencies at a database level.

3 Peak matching and follow up analysis of spectra

3.1 Introduction

This section aims to familiarize the user with the process of peak matching and also to give the user some examples of possible follow up analyses available with peak matching.

There are three important terms for the peak matching: peak, peak class and peak class view.

- A *peak* is defined on the level of the spectrum during preprocessing, performing a peak matching does not make any changes to the defined peaks.
- A *peak class* is defined on a group of spectra and is similar to the band classes for fingerprint types. During peak matching, peaks will be assigned to a peak class.
- A set of peak classes can be stored as a *peak class view*. Several peaks class views containing different peak classes can be defined and stored in your database.

3.2 Information fields

In this section we will be working on the data in the lowest level, ”Raw spectra”, though the techniques may also be applied to summary spectra. As most analysis will be performed to distinguish between the species, it is easy to have this information in a field at the level ”Raw spectra”.

1. Click on the ”Isolate” level in the *Database design* panel to make it the active level.

2. Press ***Edit*** > ***Create new object...*** (+) in the *Entry fields* panel.

3. In the *Create new entry information field* dialog box, fill in the name **Species** and check the second option, ***Calculate field content from other fields***. Press <***Edit***> and press <***Add information field***>. Select ’<Parent Key>’ and then press <***OK***> three times to create the information field.

4. Right-click on this new information field in the *Database entries* panel and select ***Field properties*** to open the *Database field properties* dialog box and press <***Level assignment***>.

5. In the *Level assignment* dialog box, use the drop down arrow in the **Assignment** column at the level ”Raw spectra” to assign the field as ***Replicated*** (see Figure 3). Press <***OK***> twice.

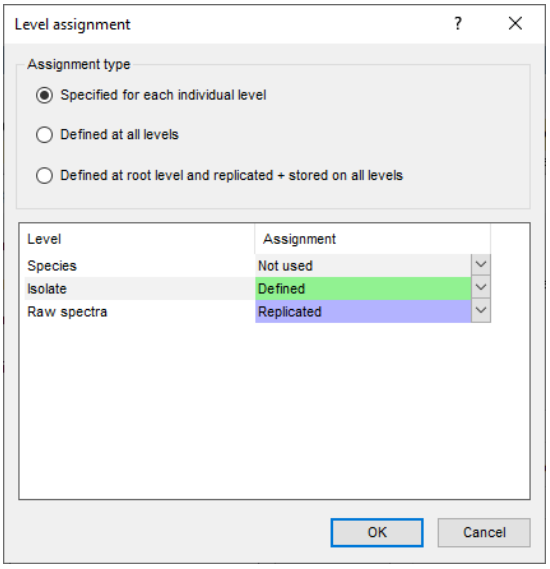

Figure 3: Level assignment of information field ***Species***.

3.3 Peak matching

6. Click on the ”Raw spectra” level in the *Database design* panel to make it the active level.

7. Select all 80 entries (***Edit*** > ***Select all*** (**Ctrl+A**)) in the *Database entries* panel.

8. Press ***Edit*** > ***Create new object...*** (+) in the *Comparisons* panel to create a comparison containing all the imported spectra.

9. 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.

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

This pops up the *Peak class matching* wizard. The only option currently available is ***Recreate peak classes***. This will create new peak classes and add these to the default peak class type.

11. Press <***Next***>.

12. Fill in a constant tolerance of “1.9”, a linear tolerance of “550” and a peak detection rate of “10%” (see Figure 4) and press <***Finish***>.

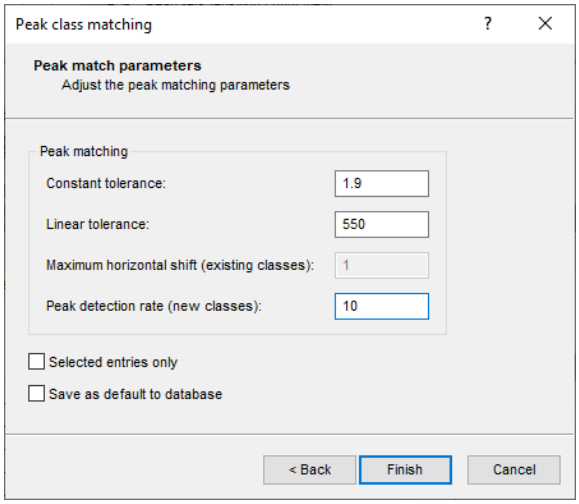

Figure 4: Second page of *Peak class matching* wizard

With a peak matching present, the user can proceed to a follow up analysis. In order to visualize the results better, groups can be created based on the ***Species*** information field:

13. Right-click on the header of the information field ***Species*** in the *Comparison* window and select ***Groups*** > ***Create groups from database field***. Leave all settings at default and press <***OK***> to create three groups based on the three species.

3.4 Exporting a peak matching table

Exporting a peak matching table cannot be performed directly on the spectral experiment. The peak matching table is character data derived from the spectral type and can be accessed using composite datasets.

14. Save the comparison and close the *Comparison* window with ***File*** > ***Exit***.

15. In the *Experiment types* panel, select ***Edit*** > ***Create new object...*** (+) to create a new experiment type, select ***Composite data set*** from the list and press <***OK***>. Name the new composite dataset “MALDI” and press <***OK***>.

16. Double-click on the ***MALDI*** experiment in the *Experiment types* panel.

17. In the *Composite data type* window, select the spectral type **Maldi** and select ***Experiment*** > ***Include experiments*** (👁) to base the composite dataset on our spectral type. Close the *Composite data type* window.

18. Double-click on the saved comparison in the *Comparisons* panel.

19. Click on the composite dataset **MALDI** in the *Experiments* panel and select ***Layout*** > ***Show image*** (👁) or press the eye button (👁) next to the experiment name in the *Experiments* panel.

The icons displayed at the top of the *Experiment data* panel will determine how the data is visualized and exported. The first icon will result in a binary representation, with only absence and presence of the peak classes shown (see Figure 5). The second and third icon result in a representation of the intensity values as color or as value respectively.

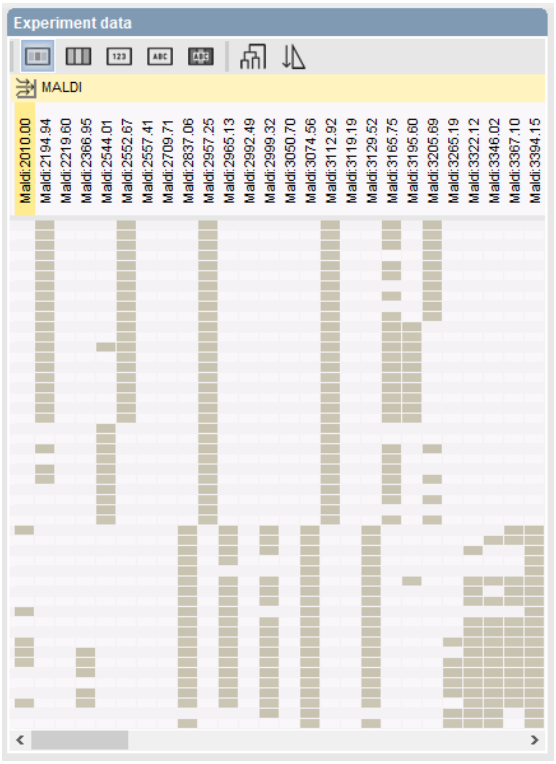

Figure 5: Binary peak table.

20. The information displayed can be exported to a tab delimited file by selecting ***Composite*** > ***Export character table...***.

Depending on the visualization, the exported file will contain either a binary peak matching table (presence/absence of peaks) or a peak matching table containing the intensity values.

3.5 Principal component analysis of peak classes

Principal component analysis (PCA) is a powerful technique that can be used in this context to reduce the complexity of the data and make it easier to identify groups and visualize the data in two or three dimensions. PCA works on character sets only, so to apply this technique to spectral types, it is necessary to perform the peak matching first.

21. To perform a PCA analysis on our spectra, make sure the spectral experiment ***Maldi*** is highlighted in the *Experiments* panel of the *Comparison* window.

22. Select ***Statistics*** > ***Principal Components Analysis...*** ( ). Leave all settings at default (see Figure 6) and press <***OK***>.

This will start the calculation of the PCA and result in Figure 7.

In the 2 dimensional images, the spectra from different species are clearly separated. For species A and B, there are peaks that are specifically linked to these species. This can be seen by looking at the coordinates of both the entries and the characters. All entries for species A can be found in the second quadrant. There is also a group of peaks located in this second quadrant, they are

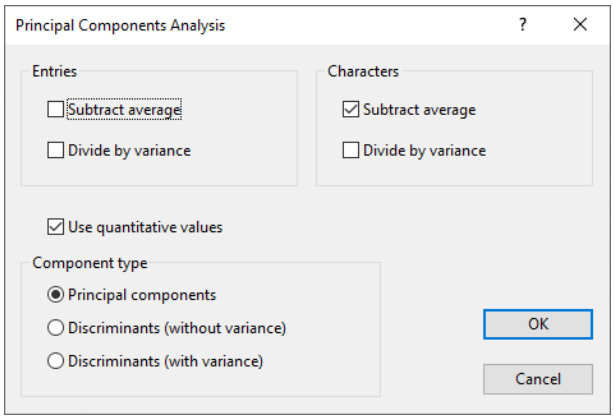

Figure 6: Settings to perform a PCA

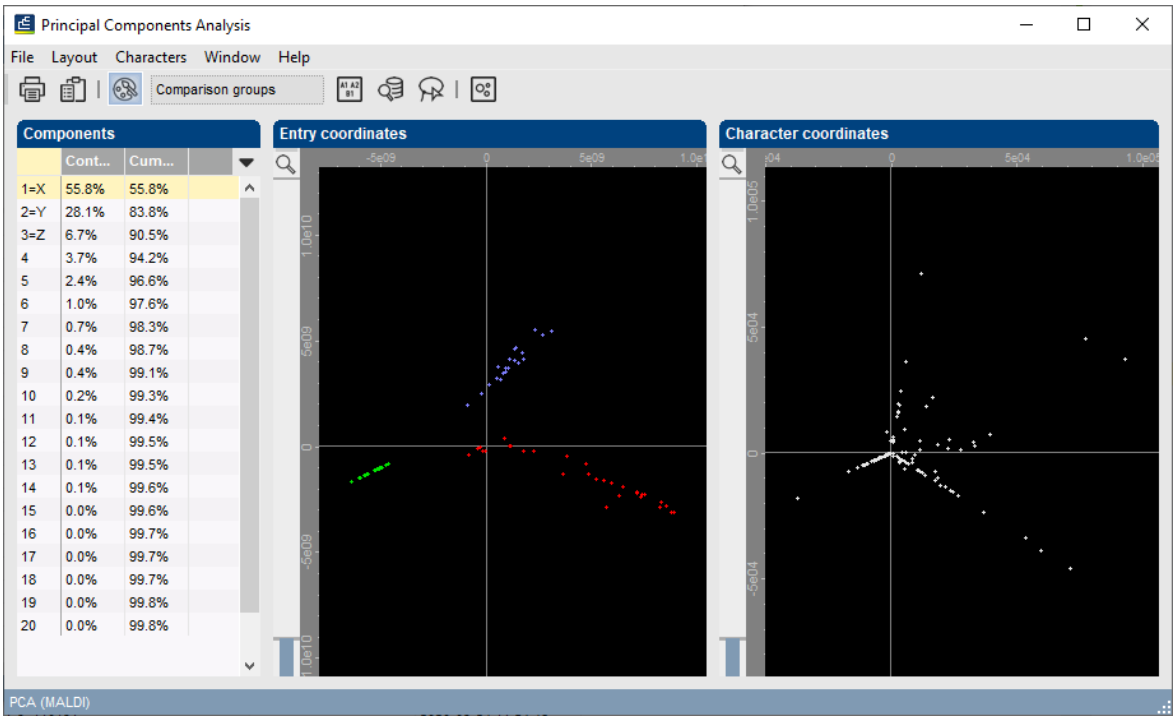

Figure 7: 2D results of PCA

likely linked specifically to species A. The same can be seen for species B, but not for species C. This species will likely be defined by the combination of several peaks, that can also be present individually in other species.

23. To obtain a three dimensional view of the PCA analysis of the entries, select ***Layout*** > ***Show 3D plot*** ( ).

In the 3D view (see Figure 8) the same can be seen, the three species form distinct groups. Species B shows the lowest variance with the entries grouped closest together, species A shows the highest variance. For both species, distinct subgroups can be seen, possibly correlated to individual isolates.

A PCA allows for a good visualization of the data, resulting in a quick visual interpretation of the data.

24. Close the PCA analysis windows.

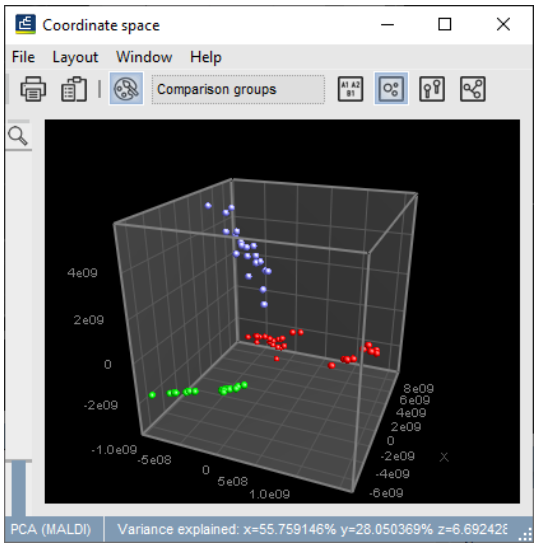

Figure 8: 3D results of PCA

3.6 Analysis of peak classes using the matrix mining

The functionalities available in the *Matrix Mining* window are very useful for spectra. It can be used to make sub-selections of peaks with certain characteristics and perform statistical analysis on the peak classes. Similar to the PCA, the *Matrix Mining* window works with character data, meaning a peak matching is required to use these functionalities. It is not the aim of this tutorial to offer an exhaustive description of all functionalities available in the *Matrix Mining* window, but to offer some examples that can provide useful conclusions for analysis of spectral data.

25. 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.

26. Select ***Statistics*** > ***Matrix mining...*** in the *Comparison* window. This will open the *Matrix Mining* window.

In the **Matrix panel**, the intensity of the peaks matched to the peak classes is represented by colors, green meaning low intensity, red high intensity. A two-way clustering can be performed to produce a heat map:

27. Select ***Analysis*** > ***Cluster analysis...*** ( ), make sure ***Characters*** is selected, leave the remaining settings at default and press <***Next***>.

28. In the *Calculate dendrogram* dialog box, select the similarity coefficient **Spearman Ranks** and press <***Next***>.

29. In the last page, leave the settings at default and press <***Finish***>.

30. Repeat these steps choosing ***Entries*** in the first step to obtain the two-way clustering (see Figure 9).

From the two-way clustering, it can very quickly be deduced that species B has very few peak classes in common with species A and C. Species A and C do have a set of peak classes in common, but each species also has a set of species specific peak classes. Some sets only belonging to a certain isolate can also be seen.

As the selection of peak classes is synchronized between the *Matrix Mining* window and the *Comparison* window, we can use the *Matrix Mining* window to select peaks sharing a certain

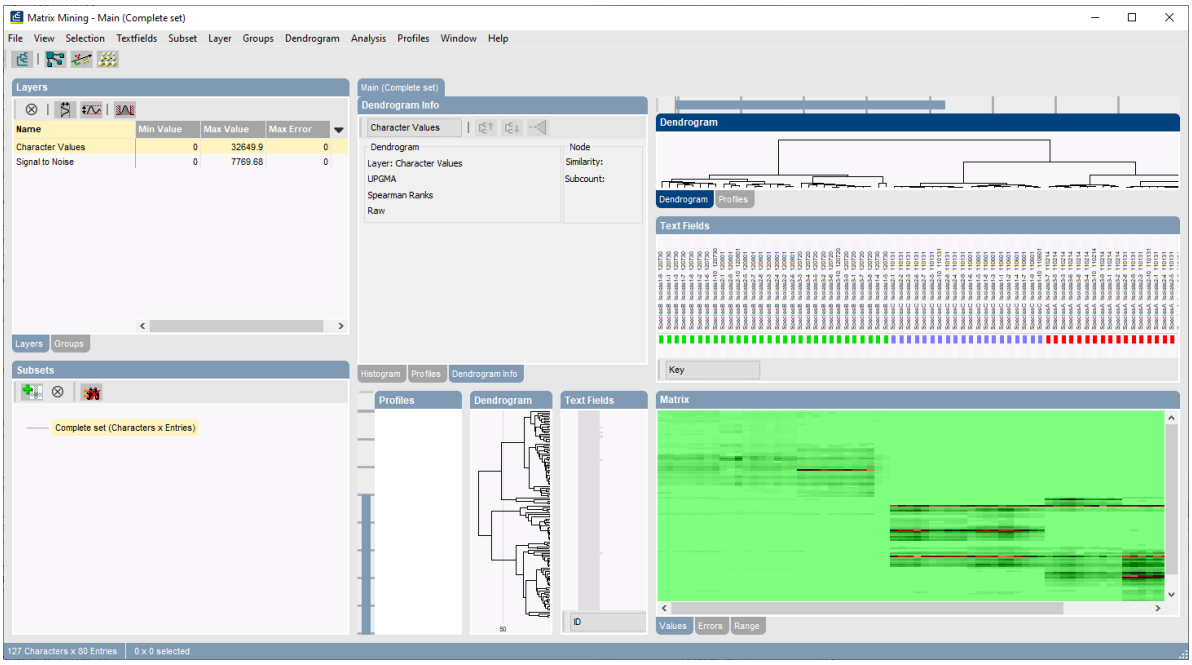

Figure 9: Two-way clustering.

characteristic. Selecting peak classes in the *Matrix Mining* window is very similar to selecting entries in the *Comparison* window.

31. For instance, select all peak classes belonging to species B by holding down the **Ctrl**-key and clicking on the node in the dendrogram of the peak classes that contains all peak classes specific for species B.

Go back to the *Comparison* window (leave the *Matrix Mining* window open).

32. Make sure the spectra are displayed (press the eye button ( ) next to the experiment name in the *Experiments* panel.

The peak classes selected in the *Matrix Mining* window are also selected in the *Comparison* window. We can modify the properties of these peak classes to define that these peak classes are specific for species B.

33. Select ***Spectra*** > ***Manage peak class types*** ( ), go to the tab *Custom Fields* and press <***Create New***>. Enter the name ***Species*** and press <***OK***> and <***Close***>.

34. Next, select ***Spectra*** > ***Display settings*** ( ), choose the field ***Species*** as display label and press <***OK***>.

35. Now we will edit this new field for the selected peak classes by pressing ***Spectra*** > ***Edit peak class properties*** ( ). Fill in ***Species B*** as property for Species and optionally change the color. Press <***OK***>.

All selected peak classes will have the label species B and will be colored, including the peaks matched to these classes. The same procedure can be followed for the peak classes common to both species A and C, specific for Species A and specific for species C.

Default all peak classes found after peak matching are displayed in the *Comparison* window: the *view* in the ***Aspect*** column in the *Experiments* panel is set to <***All peak classes***>. It is possible to store a subset of peak classes in a new peak class ***view***. Consider the case that these three species are hard to distinguish phenotypically, but species C is a pathogen and species A and B are harmless commensals. In this case, we will only be interested in the peak classes that reliably

distinguish species C from species A and B.

36. Go back to the *Matrix Mining* window and select ***Profiles*** > ***Statistics wizard...*** ().

This will open the first step of the *Statistics Wizard* dialog box.

37. Under **Orientation**, select the first option, to ***Calculate a statistic for each Character*** and press <***Next***>.

38. In the second step, select ***Mann-Witney test*** under **Independent tests (two groups)** and press <***Next***>.

39. In the third step, choose **Species C** as group 1 and **All other groups** as group 2 and press <***Next***> and <***Finish***>.

In the profiles panel, there is now a profile present with the p-value from our Mann-Witney test. All peak classes with a p-value lower than 0.05 are significantly different between species C and the other two species.

40. To select peak classes significantly different between species C and the other species, click on the profile with the p-values in the *Profiles panel*, right-click on the profile and choose ***To query*** (see Figure 10).

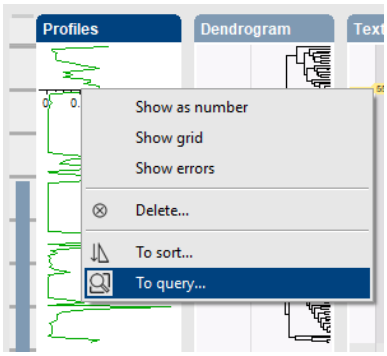

Figure 10: Profiles panel.

41. In the *Profile To Query* dialog box, select '<=' in the first box and fill in 0.05 in the second box and press <***OK***> (see Figure 11).

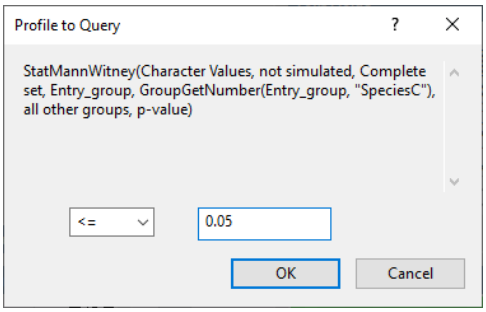

Figure 11: Query profile.

42. Go back to the *Comparison* window.

All peak classes now selected are significantly different and can be used to distinguish between species C and the other two species. This set can be stored as a new peak class *view*.

43. Select ***Spectra*** > ***Manage views...*** () and press <***Add***>. Name the new peak class view 'Distinguishing Species C' (see Figure 12) and press <***OK***> and <***Exit***>.

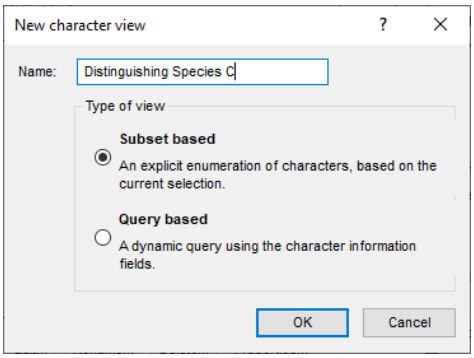

Figure 12: New subset based view.

The new *view* is automatically selected in the ***Aspect*** column in the *Experiments* panel. Defining peak class views containing a subset of peaks can be important for certain analyses, such as clustering with a composite dataset based on the peak matching. In identification projects based on spectra, peak class views can be used to base the classification on.

44. Save and close the *Comparison* window.

4 Identifying unknown samples based on peak data

4.1 Introduction

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 section we will use peak data as dataset for the identification.

4.2 Peak class views

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

2. 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 13).

3. 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.

4. Close the *Spectrum type* window.

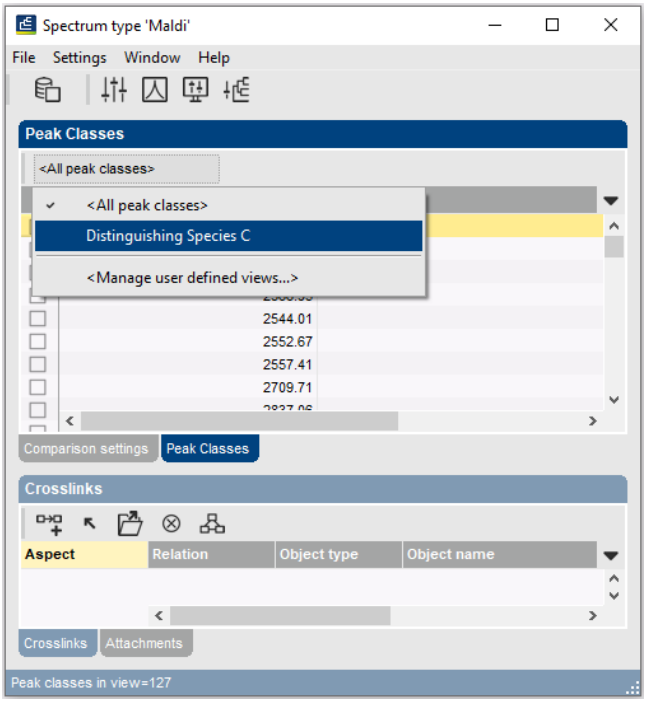

Figure 13: Two peak class views.

4.3 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.

5. 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.

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

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

8. 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*.

9. 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.

10. Click the <***Add***> button, specify a name (e.g. **Reference set**), make sure ***Subset based*** is checked, and press <***OK***> and <***Exit***> (see Figure 14).

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

11. Select + in the *Comparisons* panel.

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

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

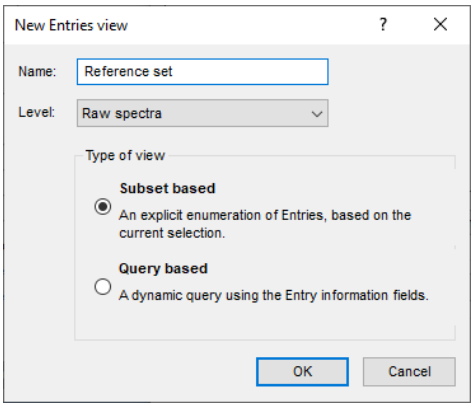

Figure 14: Create a new entry view.

13. 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.

14. 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 15).

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

16. 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 15).

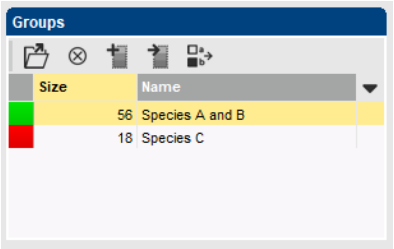

Figure 15: Two groups.

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

18. 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.

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

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

21. Fill in a constant tolerance of “1.9”, a linear tolerance of “550” and press <***Finish***>.

22. 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.

4.4 Creating the identification project

23. To create a new identification project, select + in the *Identification projects* panel of the *Main* window.

24. Select the comparison **RefSet** and leave the option to lock the reference comparison checked (see Figure 16). This will safeguard the comparison against any accidental changes that might affect the identification results. Press <***Next***>.

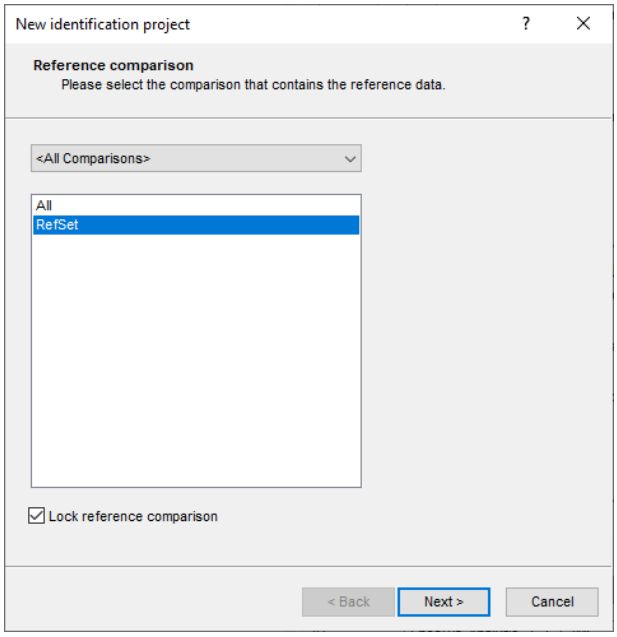

Figure 16: New identification project: step 1.

25. 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***>.

26. 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).

4.5 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.

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

This opens the *New classifier* wizard.

28. 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.

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

30. In the third step of the *New classifier* wizard, choose **Support Vector Machine (Linear)** as scoring method and press <***Next***> (see Figure 17).

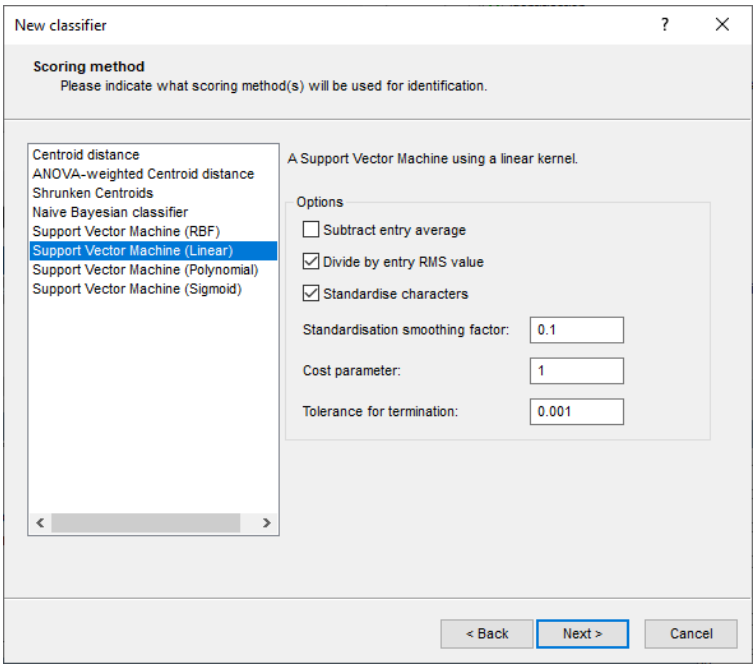

Figure 17: New classifier: step 3.

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

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

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

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

4.6 Validating a classifier

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

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

35. 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 18).

36. Clicking on a cell in the confusion matrix will give a detailed overview on the entries in this cell in the lower right panel.

37. 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.

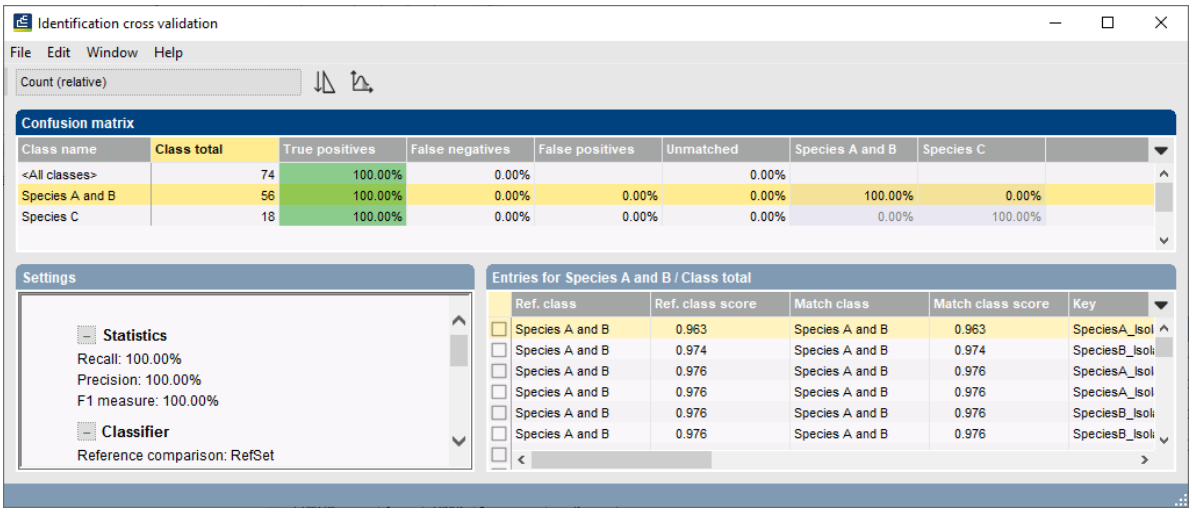

Figure 18: Validation analysis.

4.7 Identifying unknown samples

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

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

40. 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.

41. 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.

42. 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.

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

44. 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 19).

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.

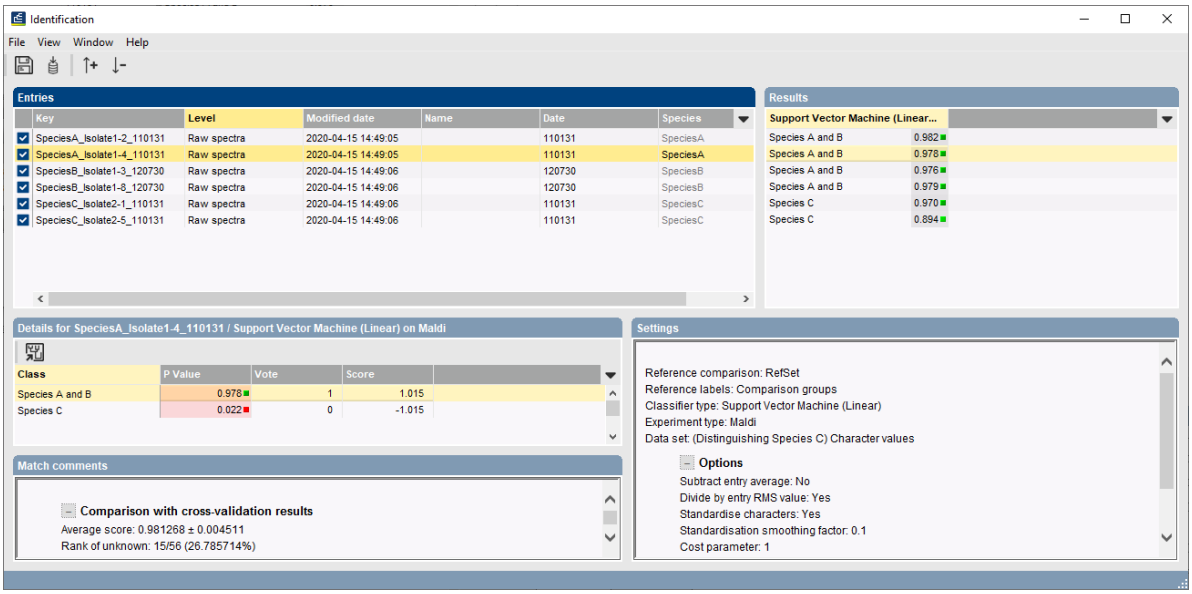

Figure 19: Identification results.

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*** ( ).

45. Close the *Identification* window.

5 Comparing spectra

1. Double-click on the saved comparison - created in 3 - in the *Comparisons* panel.

2. If groups are not present based on the ***Species*** information, create the three groups as described in 3.3.

3. Click on the spectrum type **Maldi** in the *Experiments* panel.

4. Optionally, display the spectra by pressing the eye button ( ).

5. Select ***Clustering*** > ***Calculate*** > ***Cluster analysis (similarity matrix)...*** to call the *Similarity coefficient* wizard page displaying all similary coefficients applicable to spectrum data.

All coefficients from the ***Curve based*** category provide similarities based upon densitometric curves.

All coefficients from the ***Peak based*** category measure the similarity based upon common and different peaks.

6. Select a coefficient, e.g. the ***Dice*** coefficient, click <***Next***>, make sure ***UPGMA*** is selected and press <***Finish***>.

The spectra are clustered based on the selected coefficient and the dendrogram is displayed in the *Dendrogram* panel.

7. Select ***Clustering*** > ***Dendrogram display settings...*** ( ) to call the *Dendrogram display settings* dialog box.

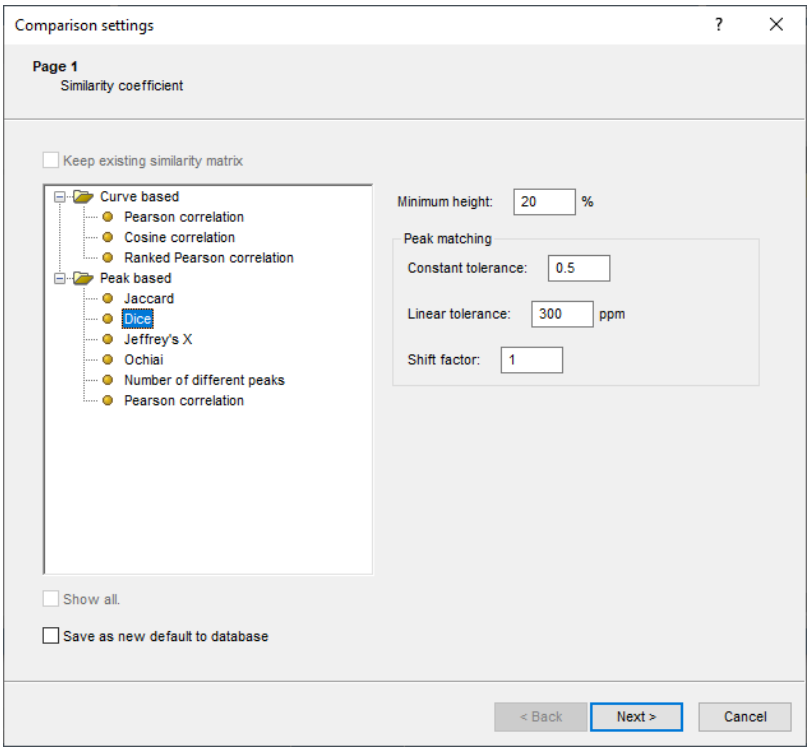

Figure 20: The *Similarity coefficient* wizard page, which deals with the choice of the similarity coefficient.

8. Enable ***Show group colors*** and press <***OK***>.

The dendrogram branches are now colored according to the group colors (see Figure 21).

9. Save and close the comparison.

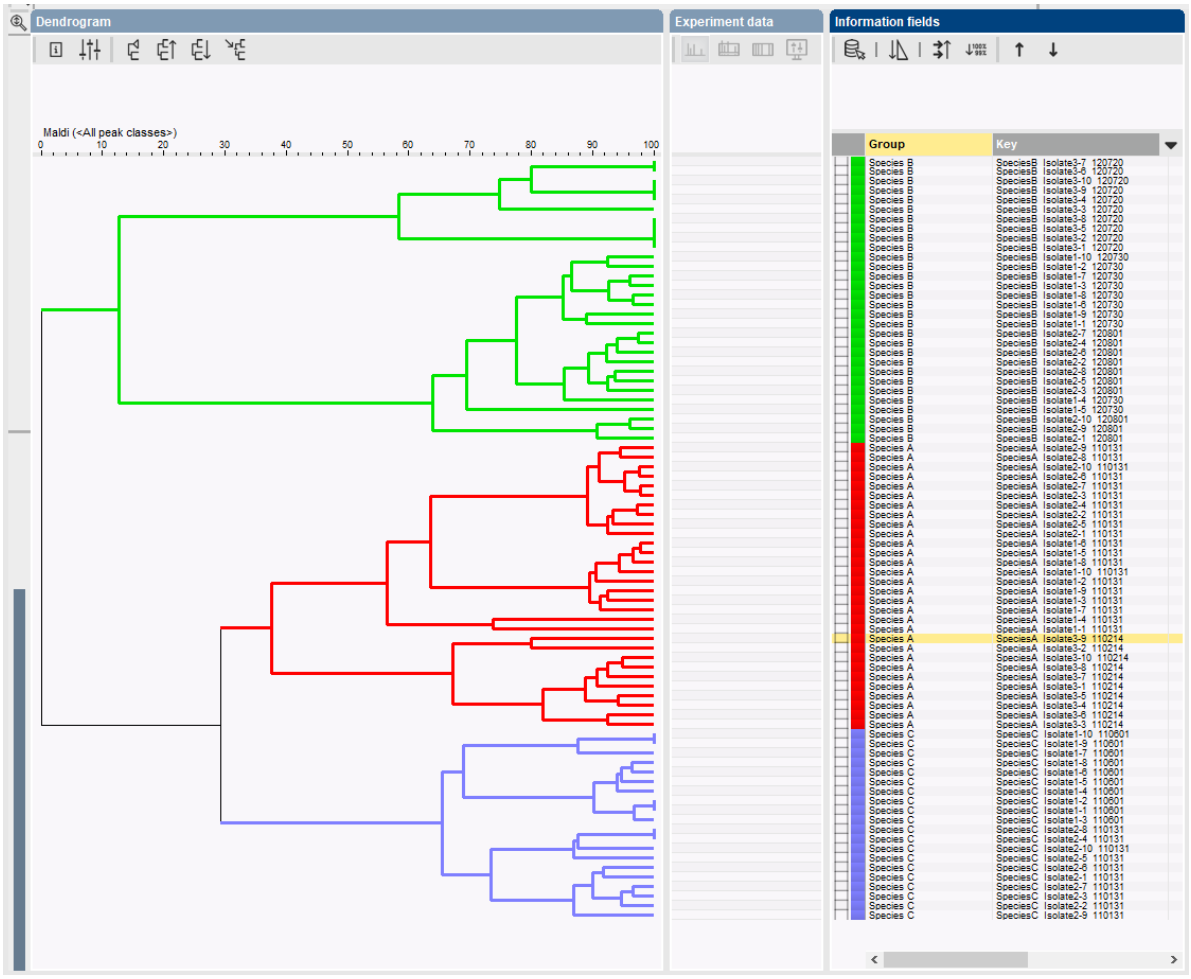

Figure 21: Group colors shown on the dendrogram.