BioNumerics Tutorial:

Identifying unknown samples based on 16s rDNA sequences

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 datatypes or combinations of datatypes can be used for identification. In this tutorial we will use the 16s sequences available in the **Demobase Connected**.

2 Preparing the database

The **DemoBase Connected** will be used in this tutorial and can be downloaded directly from the *BioNumerics Startup* window or restored from the back-up file available on our website:

1. To download the database directly from the *BioNumerics Startup* window, click the ***Download example databases*** link, located in the lower right corner of the *BioNumerics Startup* window. Select **DemoBase Connected** from the list and select ***Database*** > ***Download***. Confirm the download action.
2. To restore the database from the back-up file, first download the file `DemoBase_Connected.bnbk` from `http://www.applied-maths.com/download/sample-data`, under 'DemoBase Connected'.

   In the *BioNumerics Startup* window, press the  button, select ***Restore database***, browse for the downloaded file and select ***Create copy***. Specify a name and click <***OK***>.

In contrast to other browsers, some versions of Internet Explorer rename the `DemoBase_Connected.bnbk` database backup file into `DemoBase_Connected.zip`. If this happens, you should manually remove the `.zip` file extension and replace with `.bnbk`. A warning will appear ("If you change a file name extension, the file might become unusable."), but you can safely confirm this action. Keep in mind that Windows might not display the `.zip` file extension if the option "Hide extensions for known file types" is checked in your Windows folder options.

3 Creating the reference comparison

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

1. In the *BioNumerics Startup* window, double-click on the **DemoBase Connected** database to open it.

Select all samples with an identification down to species level. This can be done as follows:

2. Selecting ***Edit*** > ***Search...*** (**Ctrl+F**) to open the *Find objects in view* dialog box (see Figure 1).
3. Select the ***Genus*** row in the grid, and specify "STANDARD" in the text box. Press the <***NOT***> button in the *Group by* panel.

4. Select the ***Species*** row in the grid, and specify "sp." in the text box. Press the <***NOT***> button in the *Group by* panel.

Two rules are now defined in the query, one that the field **Genus** should not contain **STANDARD** and a second that the field **Species** should not contain **sp.** (see Figure 1).

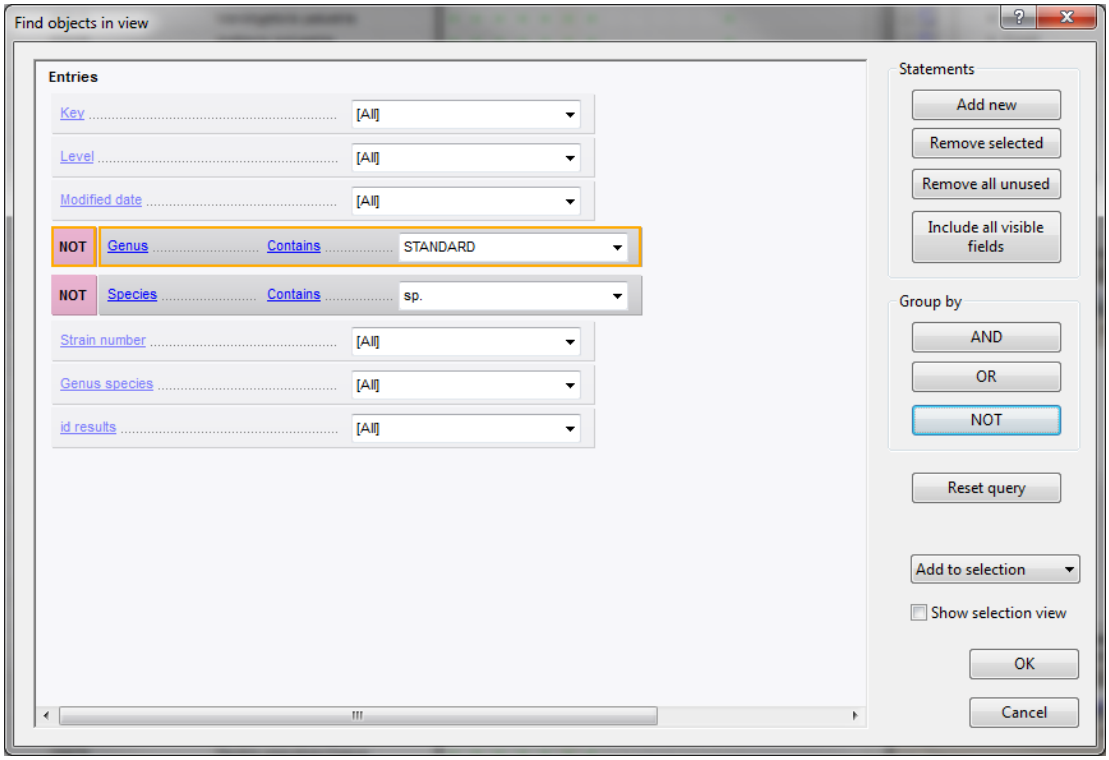

Figure 1: The *Find objects in view* dialog box with the query to select all samples identified down to species level.

5. Press <***OK***>.

If the query is run correctly, 42 entries are now selected. This is our reference set.

6. In the comparison panel, click on ***Edit*** > ***Create new object...*** ( ) to create a new comparison with our reference set.

A cluster analysis can be performed using the 16s sequences to evaluate how well separated the data is. This is not necessary to perform the identification though it can affect some of the options available in the identification project.

7. Select **16S rDNA** in the *Experiments* panel, select ***Clustering*** > ***Calculate*** > ***Cluster analysis (similarity matrix)...***, use the similarity coefficient based on the ***Standard*** pairwise alignment and leave the other settings at default.

8. Create groups that reflect the **Genus** and **Species** of each sample by first right-clicking on the column header of the field 'Genus' and selecting ***Groups*** > ***Create groups from database field***, leave the settings at default and pressing <***OK***>. Repeat this for the field **Species**, but this time select the option to subdivide the existing groups.

There are now six groups present in the comparison (see Figure 2).

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

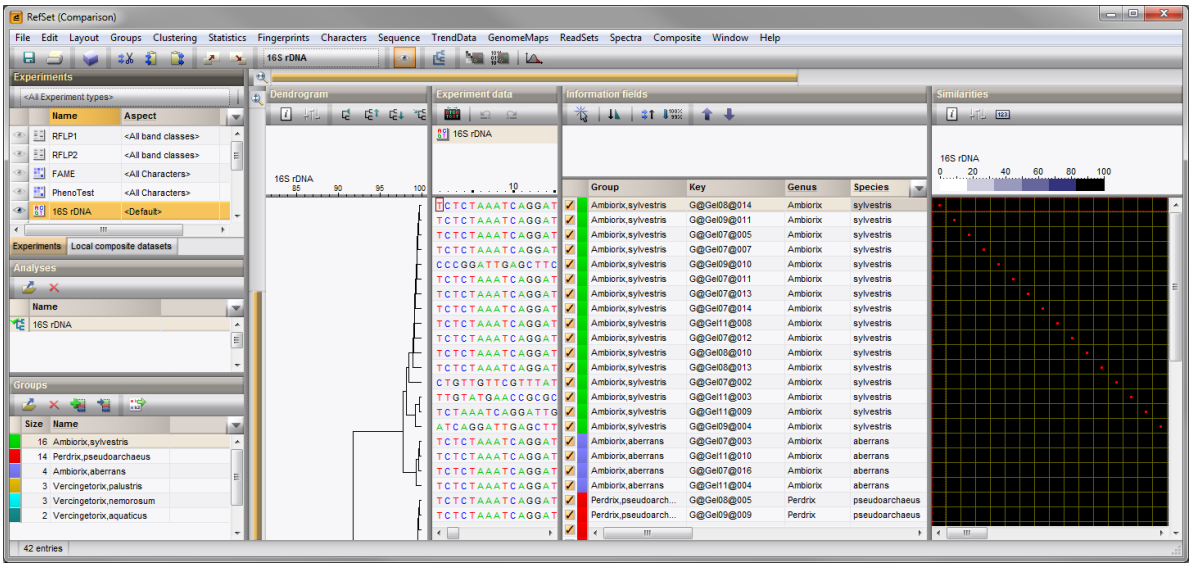

Figure 2: The *Comparison* window with groups defined.

4 Creating the identification project

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

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. In the first step of the *New identification project* wizard, select the comparison **RefSet** and leave the option to lock the reference comparison checked (see Figure 3). This will safeguard the comparison against any accidental changes that might affect the identification results. Press <***Next***>.

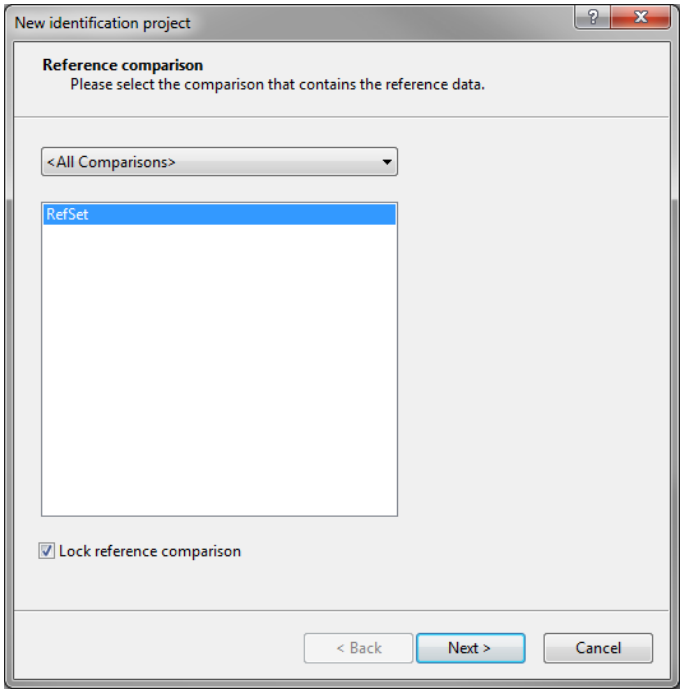

Figure 3: Select reference data.

If the reference comparison is locked, changes can only be applied if the comparison is unlocked first. To do this, open the comparison and select ***File*** > ***Object access status...*** (🔒) to open the *Object access* dialog box. Check the radio button next to Unlocked and click <***OK***>. It is strongly recommended that after applying the desired changes the comparison is locked again.

3. In the second window of *New identification project* wizard, make sure ***Comparison groups*** are checked as class labels and click <***Finish***> (see Figure 4).

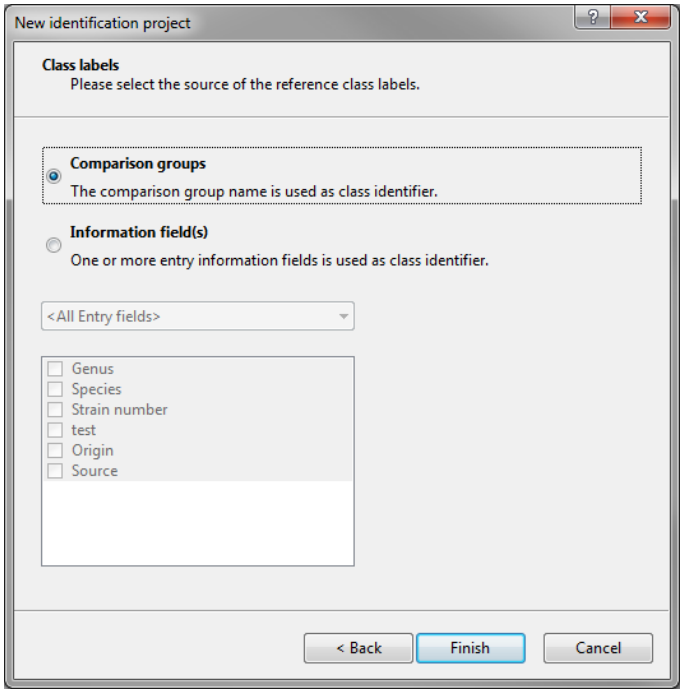

Figure 4: Select reference class labels.

4. Optionally, you can change the name of the project. 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 experiment and the algorithm, this is stored in the classifier. Per identification project, several classifiers can be defined in order to compare identification results from different experiments and /or algorithms, but in this tutorial, we will only create one classifier.

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

This opens the *New classifier* wizard.

6. In the first step, select **16S rDNA** and press <***Next***>.

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

7. Select the first method **Similarity values (Standard)** and click <***Next***>.

8. In the third step of the *New classifier* wizard, choose **Balanced Similarity** as scoring method and press <***Next***>.

9. In the fourth step, define a threshold of 98 % similarity with a minimum difference of 1 and click <***Next***>.

10. You will be prompted for a name of the classifier, just leave the name at the default suggestion and click <***OK***>.

5 Validating a classifier

The classifier is now present in our identification project and ready for use. However, it is advised to optimise the classifier parameters and to run a validation on the classifier to check its performance before using it for identification purposes.

1. For determining the optimal parameters, select ***Edit*** > ***Optimise classifier parameter...*** () to open the *Classifier parameter optimization* dialog box.

2. Select the **Max. similarity weight** and click <***OK***> to start optimizing this parameter.

The calculation of the optimization may take several minutes, depending on the capacity of your PC. The result of the optimization is a maximum similarity weight of 0.85. This value will be automatically filled in, in the classifier.

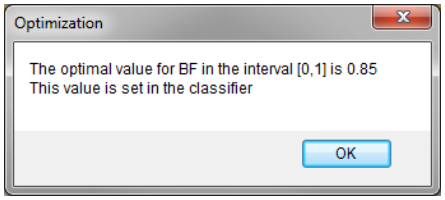

Figure 5: Optimized parameter.

3. A tool for internal validation has been included in the software and can be run by selecting ***Edit*** > ***Cross-validation analysis...*** (). This will open the *Cross-validation* dialog box, 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.

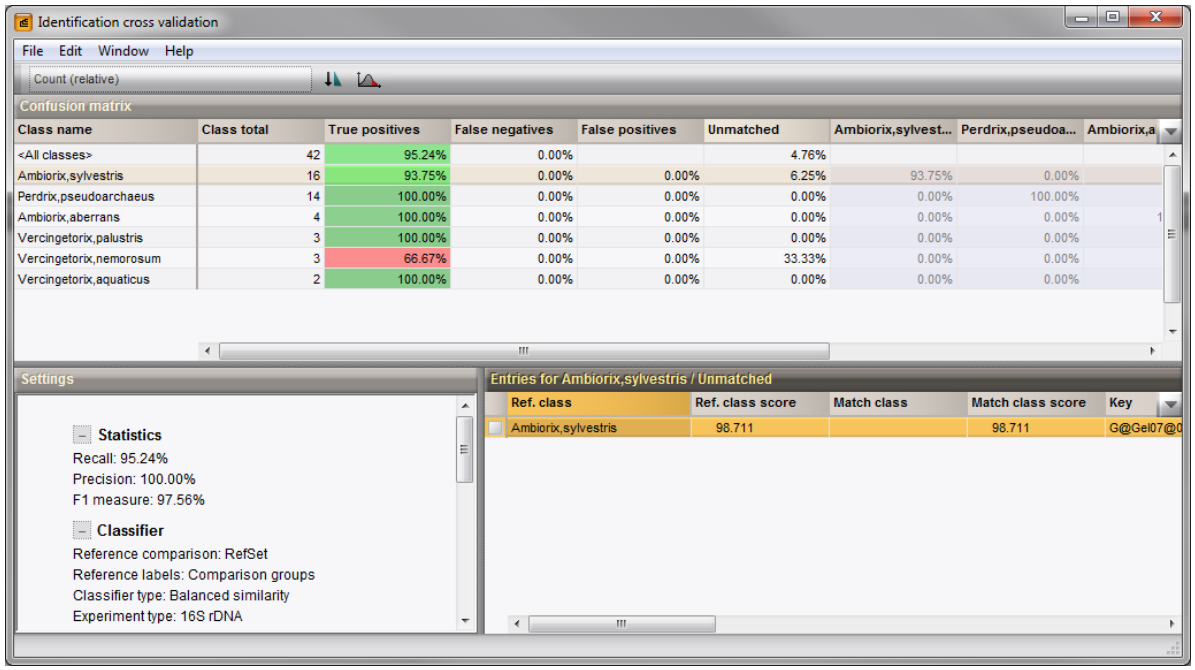

Figure 6: The *Identification cross validation* window with the results of the internal validation.

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. Some predefined graphs are also available to get a better understanding on the performance of the classifier.

4. To open a predefined graph, select ***Edit*** > ***Chart and statistics...*** ().

5. Choose the first graph **Cross validation class score contrasts** and click <***OK***>.

This graph (see Figure 7) compares the average reference score with the average best non-reference score, showing also the standard deviations. A good classifier shows a clear separation of these two scores for all groups. In this case, the contrast between these scores is highest for Perdrix pseudoarchaeus. This is to be expected as this is the only species belonging to genus Perdrix.

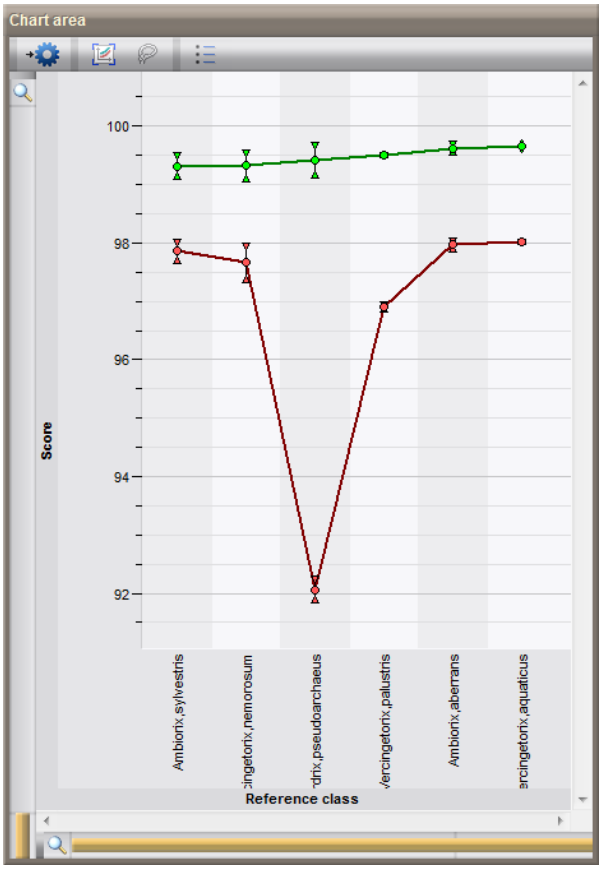

Figure 7: A graph comparing the reference class score (green) with the best non-reference class score (red).

The reliability of the cross validation depends greatly on the reference set. For smaller sets or sets that do not contain all representative samples, the results of the cross validation should be interpreted with care. For example, in this set we only have one species of the genus Perdrix, the cross validation shows us how well it can be identified compared to the other five species in the dataset, but it does not provide information on how well Perdrix pseudoarchaeus is separated from other Perdrix species.

6. Close both the graph and the *Identification cross validation* window, save the identification project (***File*** > ***Save*** ( , **Ctrl+S**)) and close it. We are now ready to identify unknown samples.

6 Identifying unknown samples

In the database, there are several entries that have not been identified down to species level, for these entries the field **Species** contains **sp.**.

Before running the identification, we will create a field to contain the results.

1. Select ***Edit*** > ***Information fields*** > ***Add information field...*** and type in **ID results** as name. Leave all other settings at default and click <***OK***>.

2. Make sure no entries are selected using ***Database*** > ***Entries*** > ***Unselect all entries (all levels)*** (, **F4**).

3. Select all entries with **sp.** for the field **Species** by selecting ***Edit*** > ***Find object in list...*** (, **Ctrl+Shift+F**) and typing in **sp.**. Click <***Select all***>.

There should now be five entries selected in the *Database entries* panel.

4. Start the identification wizard using ***Analysis*** > ***Identify selected entries...*** ( ).

5. 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 8).

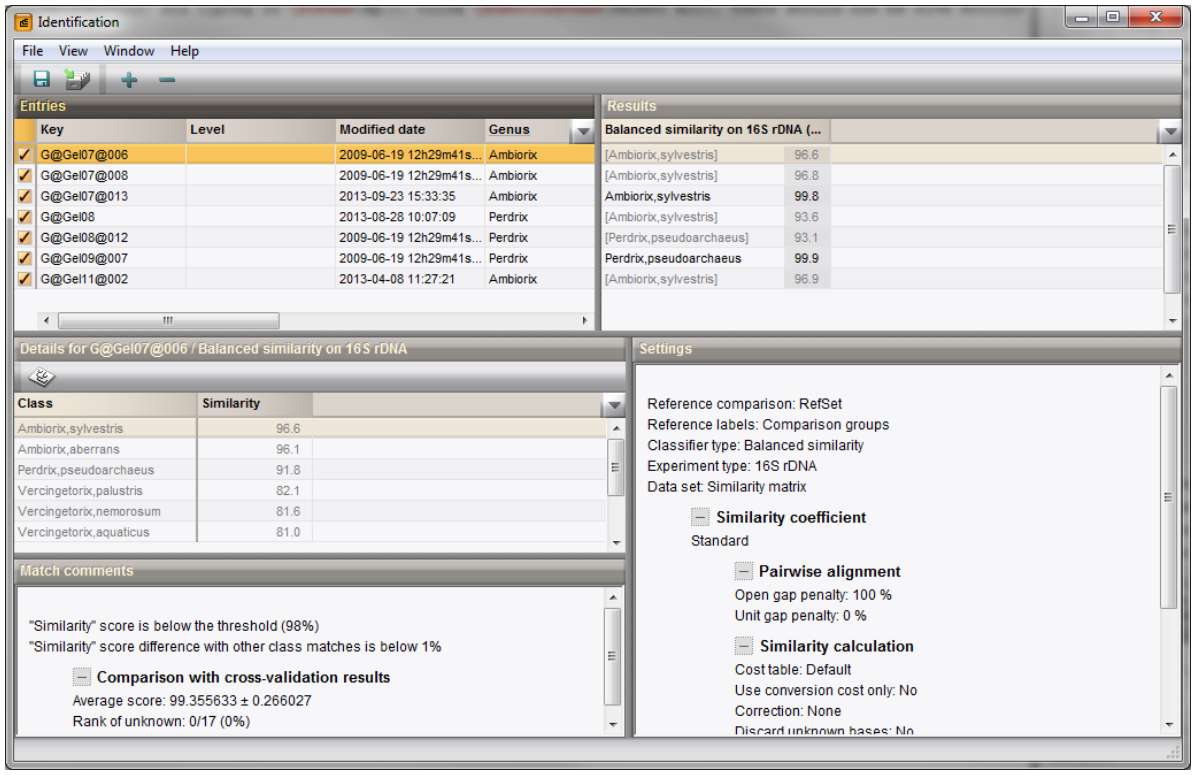

Figure 8: The identification results.

In this case, it is clear that based on the 16s rDNA sequences, the unknown samples cannot be assigned to any species as they do not meet the similarity threshold. The highest hit is shown in the right panel, but is in grey between brackets. For Figure 8 two samples that had already been identified were added to illustrate the difference between an assigned hit and no assignment. When transferring identification results to the database, results between brackets will be transferred as <**Undetermined**>. More detailed results on the highlighted entry are shown in the center panel. A comparison with this entry and all members of a certain group in the reference set can also be opened from this panel, this allows the user to evaluate the results in

more detail.

6. To open a comparison with a certain entry and the members of a certain group, highlight this entry, select a group in the panel with the detailed results (center panel) and select ***View*** > ***Open in comparison window*** ().

A comparison window opens with all members of the highlighted group and the highlighted entry.

7. The results of the identification can be transferred to the database using ***File*** > ***Transfer results to database*** (), select the field **ID results** and click <***OK***>.