BioNumerics Tutorial:

Follow-up analysis of Spa typing data

1 Introduction

In this tutorial we will perform some analyses on our imported Spa trace files: we will screen the trimmed consensus sequences for the repeat information and use this information to cluster and match our samples.

2 Preparing the database

1. Create a new database and install the *Spa typing plugin* as described in the tutorial: "Installation and setup of the Spa Typing plugin".

2. Import and assembly the demo Spa trace files as described in the tutorial: "Importing and assembling Spa trace files in batch".

3 Spa-Typing in BioNumerics

In the *Main* window, a **Spa-typing** experiment is present for each contig project (see colored dot in the **Spa-typing** column in the *Experiment presence* panel). Screening for spa repeats and types based on the sequences stored in the **Spa-typing** can be done for all entries present in the database, or for any selection of entries.

1. Make a selection in the *Main* window. To select a single entry in the *Database entries* panel hold the **Ctrl**-key and left-click on the entry. In order to select a group of entries, hold the **Shift**-key and click on another entry. With ***Edit*** > ***Select all*** (**Ctrl+A**) all entries are selected at once.

2. Select ***Spa-Typing*** > ***Assign Spa types*** in the *Main* window.

3. Press <***OK***>.

If entries are detected with sequence assembly problems or unknown repeats, the *Errors occurred* dialog box pops up, listing all these entries with a description of the detected problems. Entries can be selected and their assemblies can be opened in Assembler.

The *Spa typing plugin* uses a 2-step approach when the command ***Spa-Typing*** > ***Assign Spa types*** is selected:

Step 1: The assembly is screened for repeats

The repeat succession is displayed in the database information field that holds the repeat succession information (default name: **RepeatSuccession**). If the Character data module is present in the BioNumerics configuration, the repeat succession is also stored in the character type **Spa-repsuc**.

Step 2: Repeat type (if available) is assigned to each selected entry

The Spa type is displayed in the information field that holds the Spa Type information (default name: **SpaType**. The Spa type is denoted as "???" if the repeat succession is incomplete. When the repeat information is currently not linked to a Spa type in the database, "Unknown" is displayed in the spa type information field. If no repeats are found, "NA"(Not Available) is displayed.

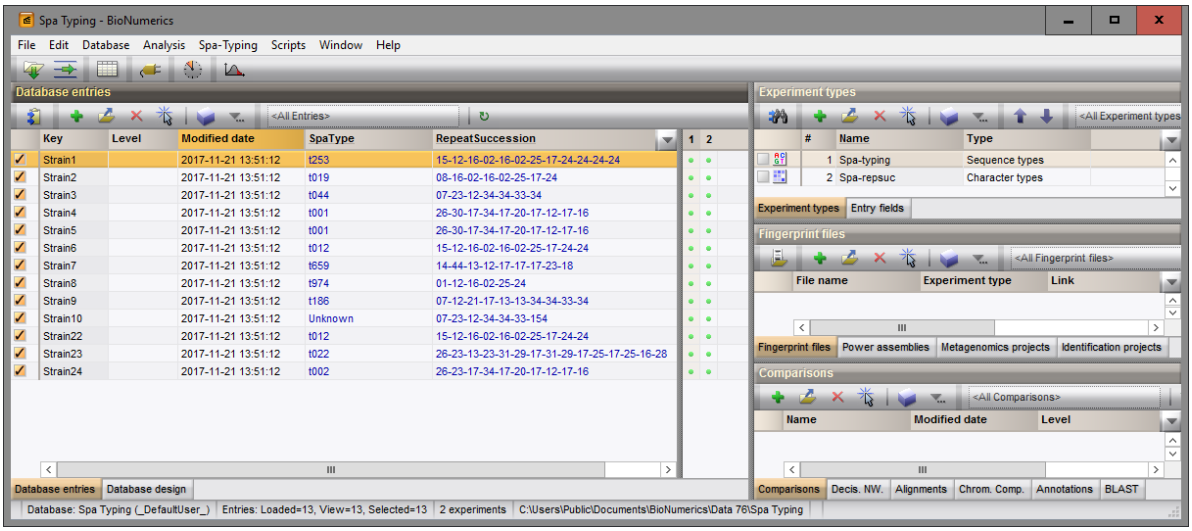


Figure 1: The *Main* window after repeat and type assignment.

4 Cluster analysis of Spa types

4.1 Introduction

In this section, we are going to take a look at the evolutionary relationship between the Spa sequences by means of the construction of a dendrogram and a minimum spanning tree.

The *Spa typing plugin* uses a multi-step approach for this cluster analysis:

- The plugin uses an algorithm based on a DSI model [1] for the pairwise alignment of the Spa repeats. This model considers three mutational events: Duplication of tandem repeats, Substitutions and Indels.
- Next, the cost matrix is used to correct for the evolutionary distances between the repeats.

Taking these costs into account, the output of the DSI model is a similarity matrix. From this similarity matrix a dendrogram and/or a minimum spanning tree can be constructed.

4.2 Comparison window

1. For this exercise, make sure all entries are selected in the *Main* window (**Ctrl+A**).
2. Highlight the *Comparisons* panel in the *Main* window and select ***Edit*** > ***Create new object...*** ( ) to create a new comparison for the selected entries.

If the Character data module is present in the BioNumerics configuration, the repeat information stored in the **Spa-repsuc** character type will be used when using the clustering tools. The repeat succession stored in the associated repeat information field is only used when the Character data module is not present in the BioNumerics configuration.

3. Click on the eye button ( ) of the character type **Spa-repsuc** in the *Experiments* panel.

The pattern images are displayed in the *Experiment data* panel. Initially, the character values are displayed as colors.

4. Select ***Characters*** > ***Show mappings*** ( ) or ***Characters*** > ***Show mappings+colors*** ( ) to display the mapped name for each character value (see Figure 2).

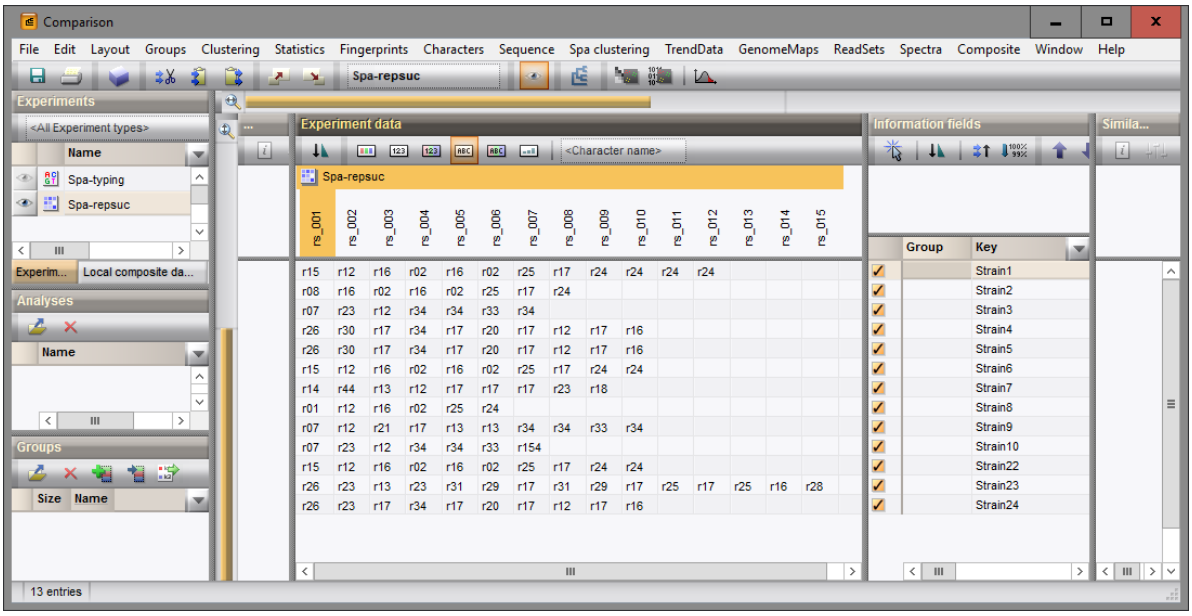


Figure 2: The *Comparison* window.

4.3 Similarity based clustering

5. Select ***Spa clustering*** > ***Cluster Spa types*** in the *Comparison* window to call the *Spa Clustering* dialog box.

In the *Matrix panel*, the default cost matrix or a custom cost matrix can be selected from the drop-down menu.

Cluster analysis *sensu stricto* is based upon the similarity matrix and a subsequent algorithm for calculating bifurcating dendrograms to cluster the entries. In the *Spa typing plugin* you can choose between the following four methods: Unweighted Pair Group Method using Arithmetic averages (***UPGMA***), the ***Neighbor Joining*** method and two variants of UPGMA: ***Single linkage*** and ***Complete linkage***.

6. Select ***UPGMA***, use the default alignment settings and default cost matrix and press <***OK***>.

The dendrogram is shown in the *Comparison* window (see Figure 3).

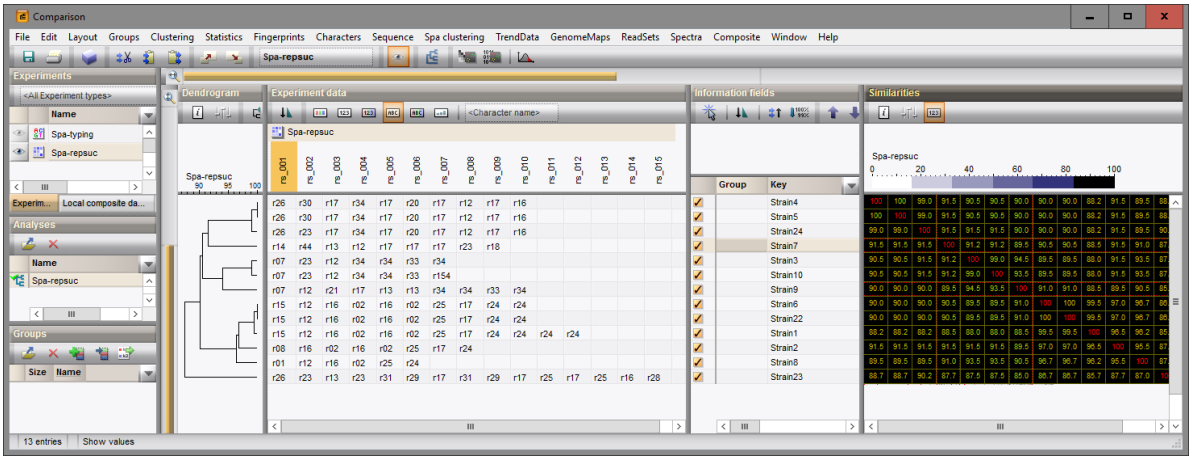


Figure 3: UPGMA tree.

7. Click on the dendrogram to place a cursor on any node or tip (where a branch ends in an individual entry). The average similarity at the cursor's place is shown in the upper part of the *Experiment data* panel. You can move the cursor with the arrow keys.

More detailed information about the *Comparison* window can be found in the manual.

4.4 Minimum spanning tree

Minimum spanning trees possess the property of having a total branch length that is as small as possible. A MST chooses the sample with the highest number of related samples as the root node, and derives the other samples from this node. This results in trees with star-like branches and allows for a correct classification of population systems that have a strong mutational or recombinational rate.

8. Select ***Spa Clustering*** > ***Cluster Spa types*** in the *Comparison* window and select ***Minimum Spanning Tree*** in the *Cluster Method panel*.

An additional setting called ***Distance bin size*** is displayed in the ***MST panel***. Based on this setting, the software creates bins of certain distance intervals, that are converted into distance units. When for example the distance bin size is set to 1%, two entries having a similarity of 99.6% will have a distance of 0 (interval 100%-99% = distance 0). Two entries that have a similarity of 98.7% will have a distance of 1 (interval 99%-98% = distance 1). The default setting is 1%.

9. Leave the settings unaltered and press <***OK***>.

The *Advanced cluster analysis* window pops up. The *Network panel* displays the minimum spanning tree, the upper right panel (*Entry list*) displays the entries that are present in the tree. The *Selection entry list* lists the entries that are present in the selected node(s).

10. Select a node or branch by clicking on them, or several nodes/branches by holding the **Shift**-key while clicking.

As an exercise we will change some display settings. More detailed information about the *Advanced cluster analysis* window can be found in the manual.

11. Press  or choose ***Display*** > ***Display settings*** to open the *Display settings* dialog box.

12. In the *Node labels and sizes tab*, select ***Show node labels*** and select ***SpaType*** from the list.

13. In the *Node colors tab*, select ***Number of entries*** from the drop-down list.

14. In the *Branch styles tab*, select ***branch length*** from the drop-down list.

15. In the *Branch labels and sizes tab*, select ***Show branch labels*** and ***branch length***.

16. Press <***OK***> to apply the new settings.

17. In the *Advanced cluster analysis* window, select ***Display*** > ***Zoom to fit*** or press  to optimize the view of the tree in the current window.

The *Advanced cluster analysis* window should now look like Figure 4.

18. Close the *Advanced cluster analysis* window and the *Comparison* window.

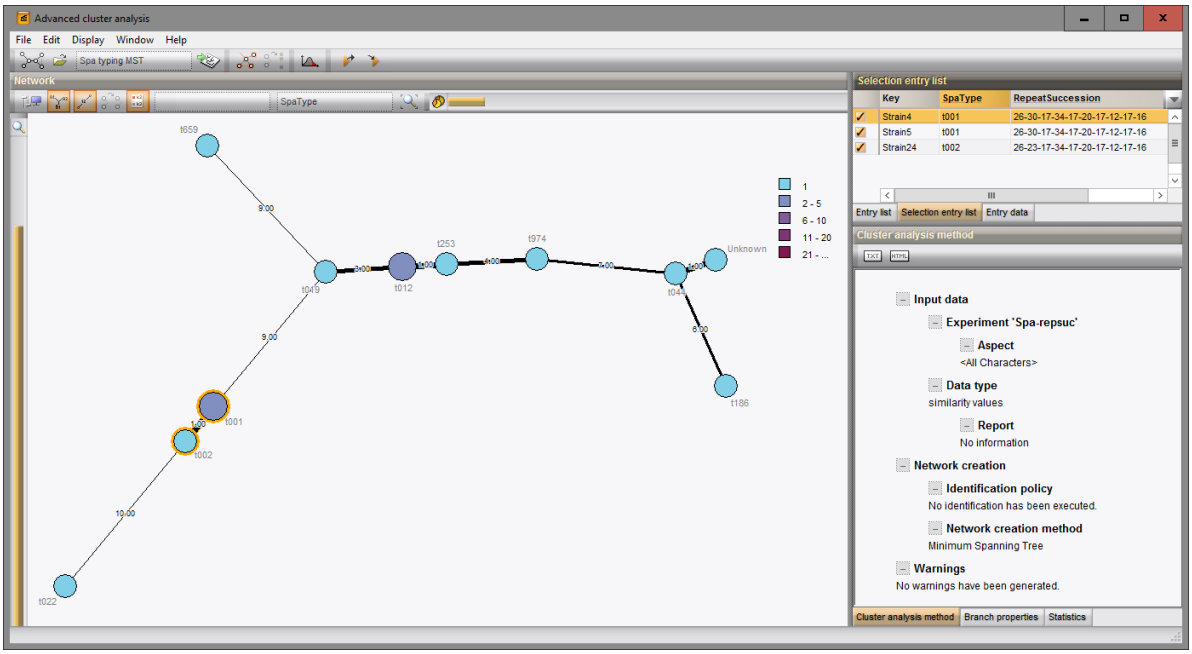


Figure 4: The *Advanced cluster analysis* window.

5 Matching Spa types

One or more selected Spa types can be matched (identified) against the complete database, all Spa types, or a selection in the database.

1. As an exercise, select a few entries in the *Main* window (e.g. **Strain22**, **Strain23**, and **Strain 24**). To unselect all entries press <***F4***>. Selecting entries one-by-one is done with the **Ctrl**-key or with the checkboxes.

2. Call the *Spa matching* dialog box with ***Spa-Typing*** > ***Match Spa types***.

3. For this exercise, choose <***All Entries***> from the ***Match against*** menu, leave all other settings at their defaults and press <***OK***>.

BioNumerics tries to find the best matches for the selected entries based on their repeats. The *Spa matching window* appears (see Figure 5).

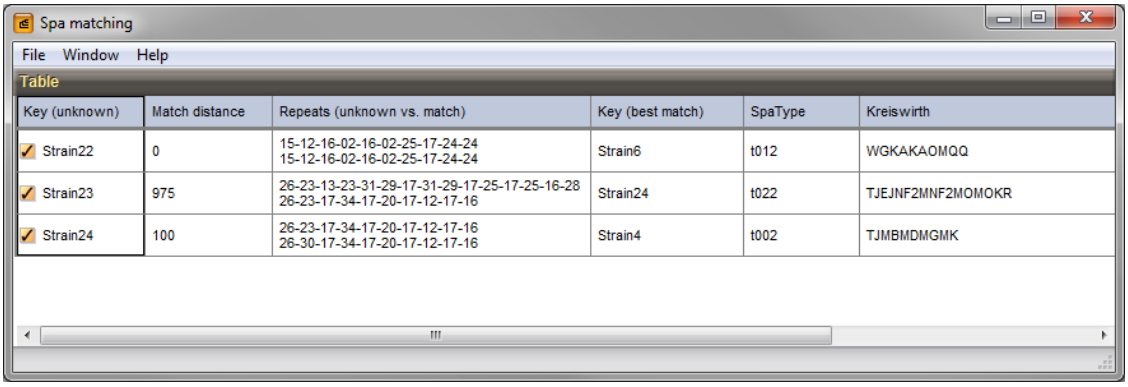


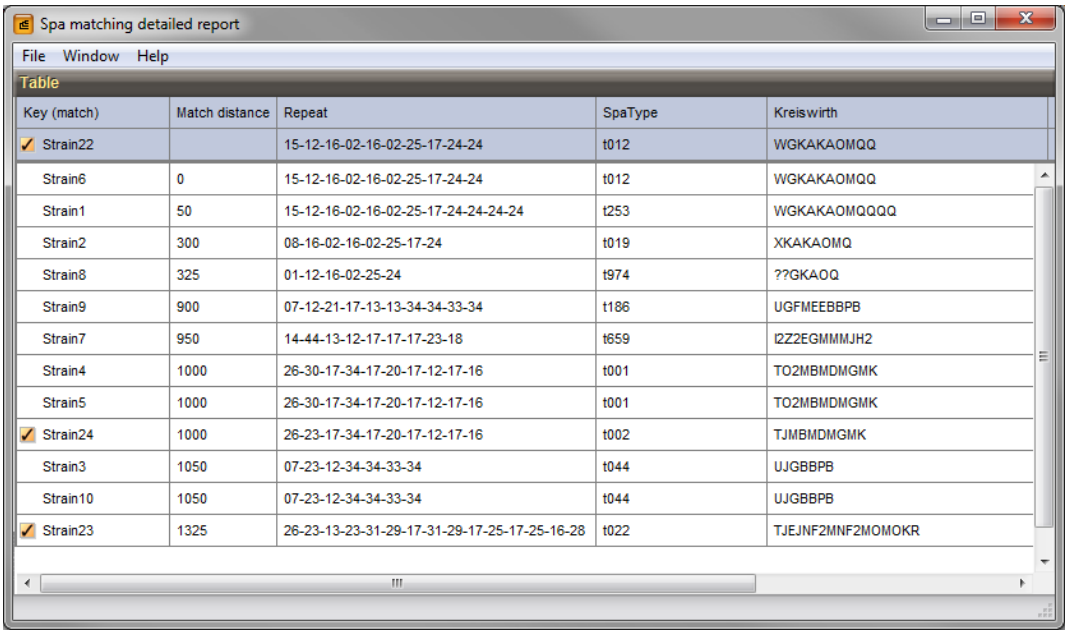
Figure 5: The *Spa matching window*.

- In the first column, the keys of the selected ”unknown” entries are shown.

- The fourth column displays the best matching entry.
- In the last column(s), the repeat type and Kreiswirth information of the unknown entry is listed.
- The repeats of the selected entries and their matches are shown in the third column.
- The distance between the selected entry and its match is displayed in the second column. The smaller the value, the better the match with ”0” being an exact match.

4. Double-click on an entry in the *Spa matching window* (e.g. entry with key **Strain22**).

A detailed report pops up (see Figure 6). The best matching entries are shown in descending order.



Key (match)	Match distance	Repeat	SpaType	Kreiswirth
Strain22		15-12-16-02-16-02-25-17-24-24	t012	WGKAKAOMQQ
Strain6	0	15-12-16-02-16-02-25-17-24-24	t012	WGKAKAOMQQ
Strain1	50	15-12-16-02-16-02-25-17-24-24-24-24	t253	WGKAKAOMQQQQ
Strain2	300	08-16-02-16-02-25-17-24	t019	XKAKAOMQ
Strain8	325	01-12-16-02-25-24	t974	??GKAOQ
Strain9	900	07-12-21-17-13-13-34-34-33-34	t186	UGFMEEBBPB
Strain7	950	14-44-13-12-17-17-17-23-18	t659	I2Z2EGMMMJH2
Strain4	1000	26-30-17-34-17-20-17-12-17-16	t001	TO2MBMDMGMK
Strain5	1000	26-30-17-34-17-20-17-12-17-16	t001	TO2MBMDMGMK
Strain24	1000	26-23-17-34-17-20-17-12-17-16	t002	TJMBMDMGMK
Strain3	1050	07-23-12-34-34-33-34	t044	UJGBBPB
Strain10	1050	07-23-12-34-34-33-34	t044	UJGBBPB
Strain23	1325	26-23-13-23-31-29-17-31-29-17-25-17-25-16-28	t022	TJEJNF2MNF2MOMOKR

Figure 6: Detailed report of the *Spa matching window*.

In both report windows, you can select or unselect entries by pressing the **Ctrl-** or **Shift-** key while holding the left mouse button.

5. Close both windows containing the Spa match results.

Bibliography

[1] G. Benson. Sequence alignment with tandem duplication. *Journal of Computational Biology*, 4(3):351–367, 1997.