BioNumerics Tutorial:

Importing and assembling Spa trace files in batch

1 Introduction

With the BioNumerics batch assembly import routine, hundreds of sequence trace files can be imported in batch and assembled automatically into contigs. This batch tool is very flexible and highly automated and allows the direct import of sequencer trace files from Applied BioSystems, Amersham and Beckman automated sequencers. In this tutorial you will learn how to use this batch tool by importing and assembling some Spa trace files.

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".

Example .AB1 trace files that will be used in this tutorial can be downloaded from the Applied Maths website:

2. Go to the ***Sample data*** section on our website (http://www.applied-maths.com/download/sample-data), and click on "SPA typing data files". Download and unzip the files.

3 Import routine

1. In the BioNumerics *Main* window select ***File*** > ***Import...*** (, **Ctrl+I**) to call the *Import* dialog box.

2. Select ***Import and assemble trace files*** under ***Sequence type data*** and press <***Import***>.

When importing (assembled) FASTA files, choose ***Import and assemble traces from FASTA text files***.

3. Select the <***Browse***> button, navigate to the correct path, select all the sequence trace files and press <***Open***>.

4. Press <***Next***> to go the next step.

The way the information should be imported in the database can be specified with an import template. In the example data set, the ***Key*** is provided in the trace file name.

5. Make sure the ***Example import 1*** template is selected and press the <***Preview***> button.

The ***Example import 1*** template will parse the ***Key*** from the file names.

6. Close the preview.

7. Make sure the ***Example import 1*** template is selected, and select the ***Spa-typing*** from the ***Experiment type*** list (see Figure 1).

8. Press <***Next***>.

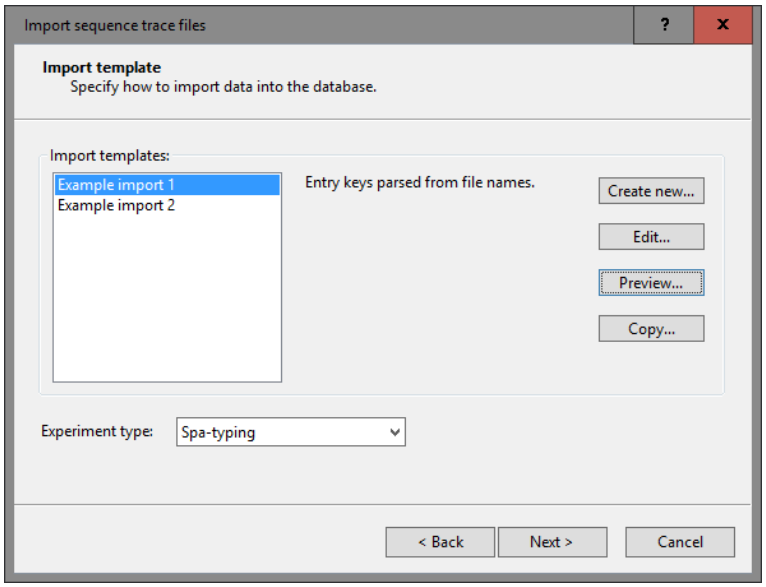

Figure 1: Import sequence trace files.

9. Press <***Next***> once more to confirm the creation of 13 new entries.

The *Processing* wizard page opens.

10. Press <***Trimming settings***> to pop up the *Assembly trimming settings* dialog box (see Figure 2).

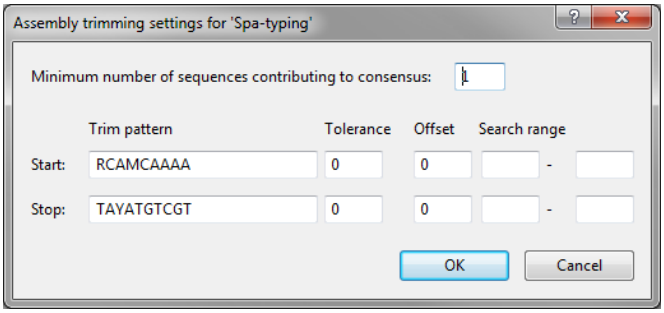

Figure 2: The *Assembly trimming settings* dialog box.

The trimming patterns specified in the *Spa typing settings* dialog box are shown in the ***Start pattern*** and ***Stop pattern*** text boxes.

11. Leave the predefined settings unaltered and press <***OK***> to close the trimming dialog box.

12. Press the <***Assembly settings***> button to call the *Assembly settings* dialog box.

The default Spa ***Quality*** assignment settings are required for submission of new types to the SpaServer.

13. For this exercise, do not change the settings and press <***OK***>.

14. Make sure the option ***Open assembly overview report*** is checked and press <***Finish***> to assemble the selected trace files from the example dataset into separate contig projects.

4 Checking the assemblies

When the assemblies are processed, an interactive report window appears. This window can also be displayed from the *Main* window with ***Analysis*** > ***Sequence types*** > ***Batch assembly reports...***.

The *Overview* panel displays the entries (keys) as rows and the experiments as columns (see Figure 3). Each cell, corresponding to a key/experiment pair, provides information about the current status of the contig project.

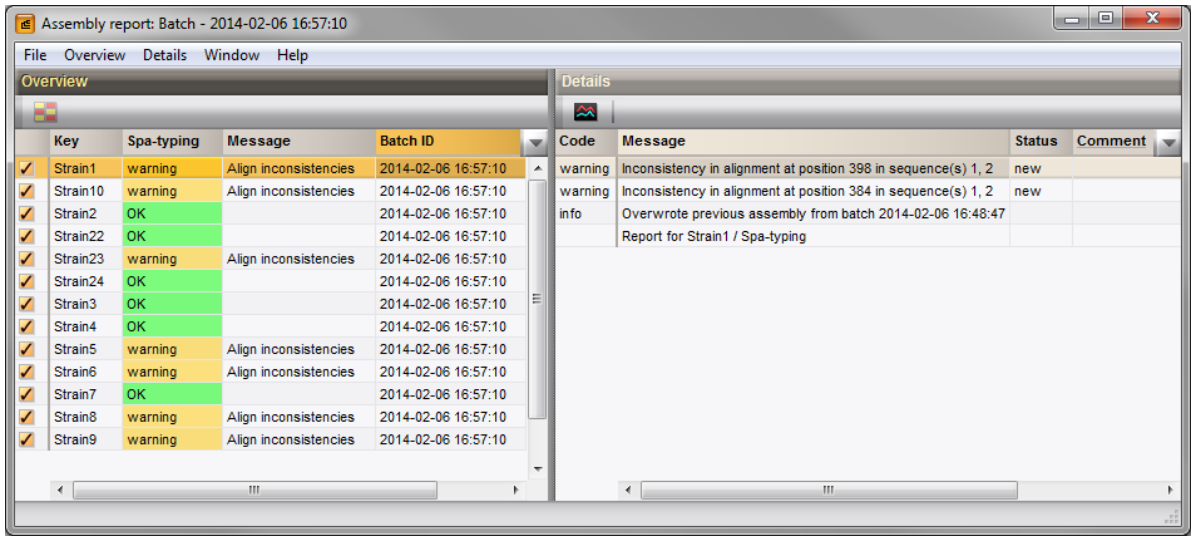

Figure 3: The *Batch sequence assembly report* window.

1. Click a cell, e.g. ***Strain23/Spa-typing*** to update the *Details* panel on the right-hand side.

The *Details* panel is organized in message rows with four columns.

- The first column displays a message **Code**, which can be either ”info”, ”warning” or ”error”.
- The second column shows the actual **Message**. Double-clicking on this cell opens the *Contig assembly* window (if not already open), with the corresponding position highlighted.
- The third column displays the **Status** of the message, which can be ”new”, ”read” or ”solved”. The status can be changed by the user.
- The fourth column is a **Comment** field. A comment can be entered by the user.

2. Open the *Contig assembly* window for the entry with key **Strain23** by double-clicking on the first message in the *Details* panel of the *Batch sequence assembly report* window.

The *Alignment* panel in the *Contig assembly* window shows the consensus sequence (upper line) and the individual trace sequences that contribute to the displayed consensus. The upper panel (*Alignment overview* panel) displays the aligned trace sequences. If the arrow points to the left, the program has invert-complemented the sequence to obtain the correct alignment. The upper left panel displays the selected consensus with its length and the number of sequences that are part of it.

3. Select the *Aligned traces* panel.

The bottom panel now displays the chromatogram files for both trace sequences (see Figure 4).

4. To obtain an optimal view of the curves, use the zoom sliders in the *Traces* panel or use the zoom buttons.

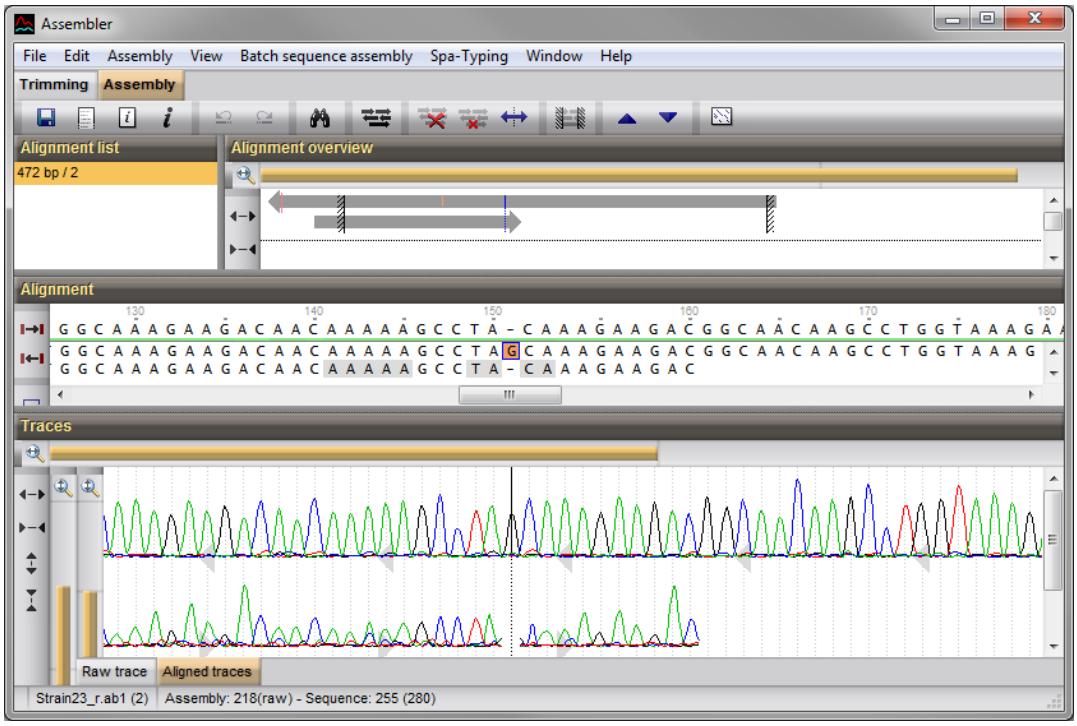

Figure 4: The *Aligned traces* panel.

The parameter ***Req. bases to include*** in the *Assembly settings* dialog box is by default set to 51%. This means that a gap in one sequence and a nucleotide in the other will insert a gap in the consensus sequence. If you take a closer look at the alignment inconsistencies of this assembly, two gaps are present in the forward sequence (at positions 160 and 218), resulting in two gaps in the consensus sequence. These positions will be further investigated in the next steps.

5. In the *Contig assembly* window, select ***Spa-Typing*** > ***Show repeats*** or use the shortcut **Shift+F5**.

Assembler screens the consensus sequence for repeats.

- Known repeats are shown in *green* (see Figure 5) and the name of the repeat is shown on top of the known repeat sequence.
- Bases in the repeat succession string that are not assigned to a known repeat are shown in *red*.
- The 5' and 3' signatures are displayed in *yellow*.

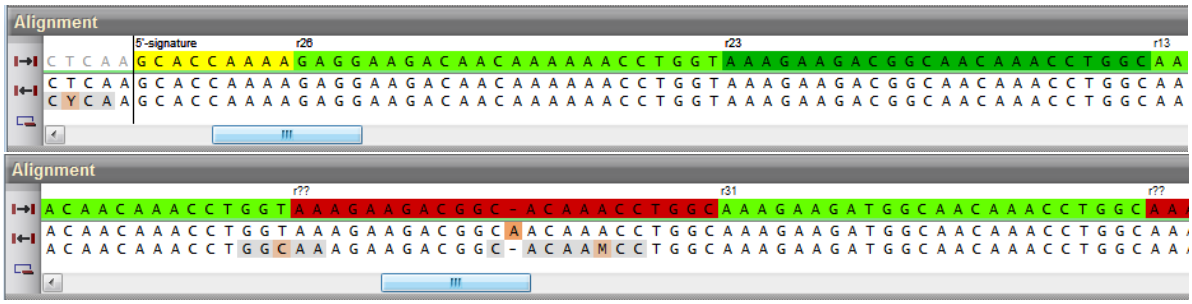

Figure 5: Showing the repeats on the consensus sequence.

The repeat succession string and the corresponding Spa type (if known) are displayed in the caption of the *Contig assembly* window (see Figure 6).

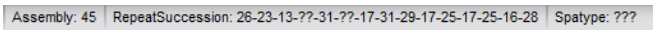

Figure 6: Repeat succession string.

6. Select ***Spa-Typing*** > ***Show repeats plot*** or use the shortcut **Shift+F6**.

The repeats are displayed in the *Spa repeat plot window* (see Figure 7).

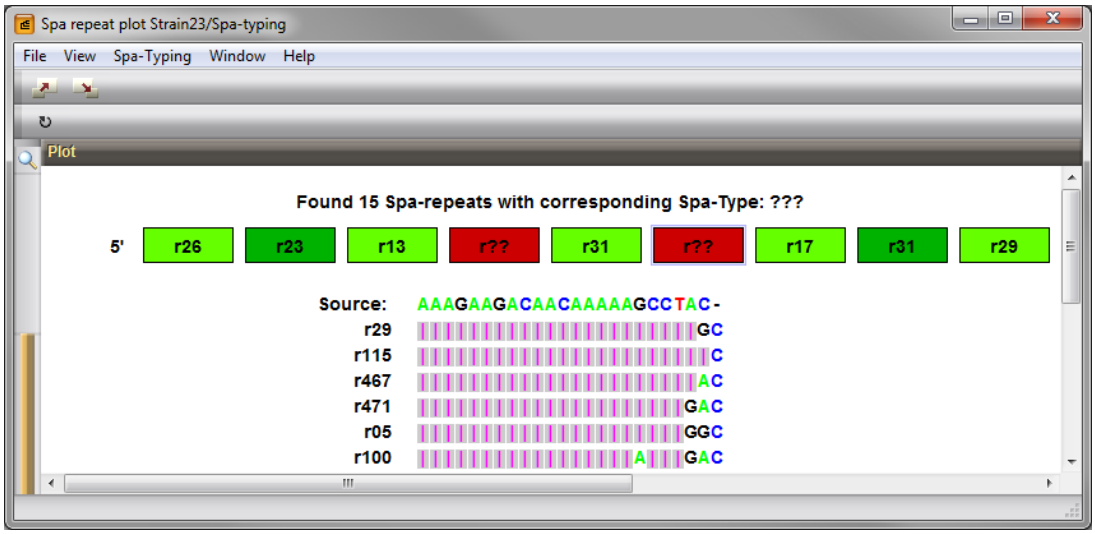

Figure 7: The repeat plot with editing suggestions for the first unknown repeat.

7. Click on the second unknown red "r??" repeat. A table is displayed with suggestions to edit the sequence. In the left column, the repeat is shown.

8. Use the zoom functions and (***View*** > ***Zoom in*** and ***View*** > ***Zoom out***) to obtain the best view of the plot.

Editing the sequence as suggested by the first row will give repeat "r29" (see Figure 7). Looking at this position in the *Contig assembly* window gives additional information about the missing base: in the chromatogram of the forward sequence, there is a missing "G".

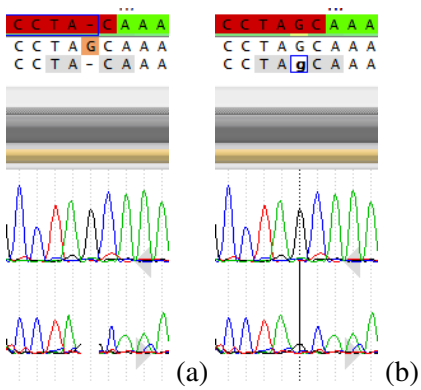

Figure 8: Missing peak in the chromatogram (a), Editing the trace sequence (b).

9. Place the cursor on the gap in the trace sequence and type "G". The consensus sequence is automatically updated (see Figure 8 (b)).

10. Select ***Spa-Typing*** > ***Show repeats***.

The repeat assignment in the consensus sequence is updated. The ”r29” repeat is displayed in green in the *Assembly view*.

11. Select ***Spa-Typing*** > ***Refresh*** in the repeat plot to update the information.

The corrected repeat is displayed in green.

12. Click on the remaining unknown repeat in the repeat plot.

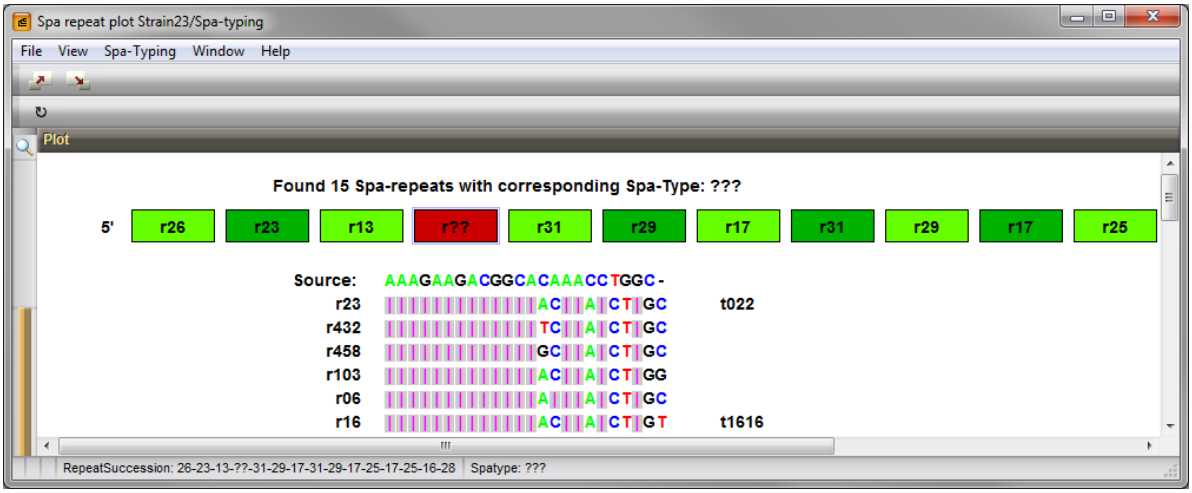

Figure 9: The repeat plot with editing suggestions for the second ”unknown” repeat.

The table with suggestions is displayed. In the left column, the repeat is shown. In the right column, the associated spa type is displayed (Figure 9). Editing the sequence as suggested by the first row will give repeat ”r23” and type ”t022”. Looking at this position in the *Contig assembly* window gives additional information: in the chromatogram of the forward sequence, there is a missing ”A” and based on the default ***Consensus determination*** parameters, this leads to a gap in the consensus sequence (see Figure 10 (a)).

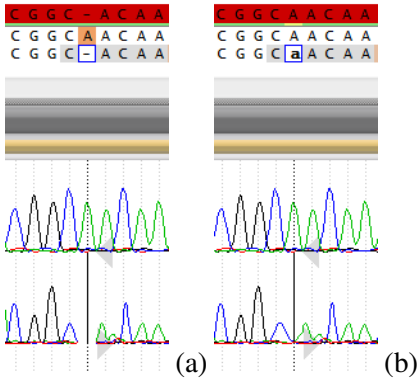

Figure 10: Missing peak in the chromatogram (a), Editing the trace sequence (b).

13. To insert the base in the trace sequence, place the cursor on the gap in the trace sequence and type ”A”.

The consensus sequence is automatically updated (see Figure 10 (b)).

14. Select ***Spa-Typing*** > ***Show repeats***.

The repeat assignment in the consensus sequence is updated. All repeats are now displayed in green in the *Assembly view*.

15. Select ***Spa-Typing*** > ***Refresh*** in the repeat plot to update the information.

16. Close the *Repeat plot window* with ***File*** > ***Exit***.

The two warning messages (***Inconsistency in alignment at position 218*** and ***Inconsistency in alignment at position 160***) are checked and corrected for **Strain23**).

17. Select ***Batch sequence assembly*** > ***Set report to solved, save and close*** (**Ctrl+Shift+S**) in the *Contig assembly* window.

The corresponding key/experiment cell in the overview *Batch sequence assembly report* window is updated and displayed in green. The status ”Solved” is displayed in the key/experiment field.

18. As an exercise, check some more assemblies with a warning status.

19. Close the *Overview* panel.

In the *Main* window, a **Spa-typing** experiment is present for each entry that was added to the database during import (see colored dot in the **Spa-typing** column in the *Experiment presence* panel).

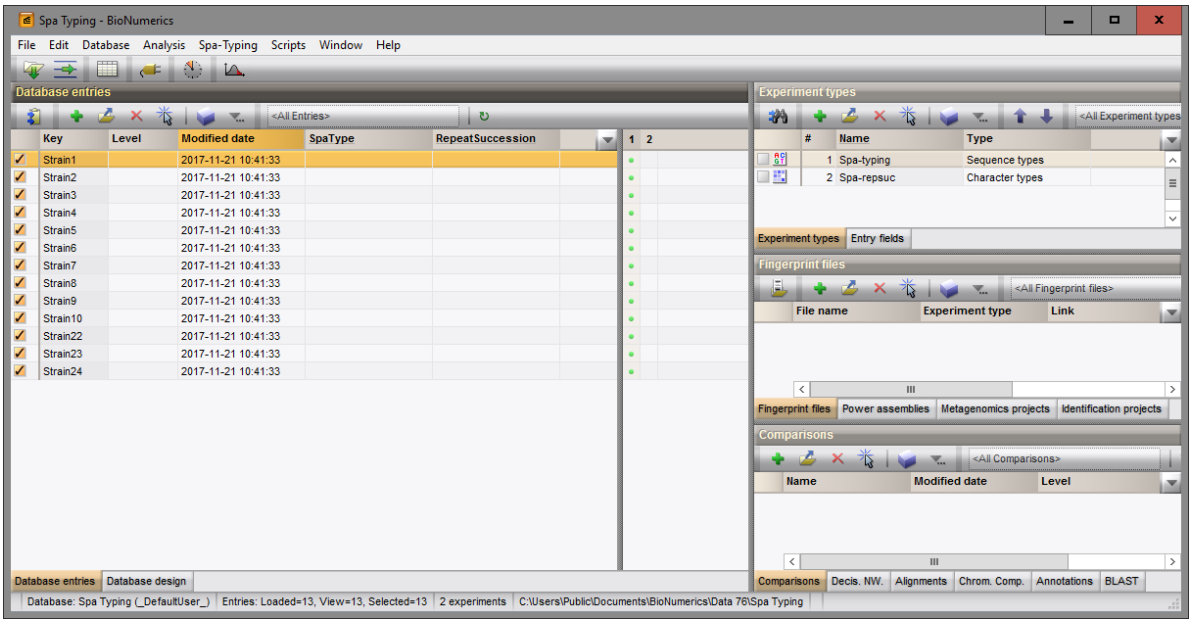

Figure 11: The *Main* window after import.