BIONUMERICS Tutorial:

Importing and assembling sequences in batch

1 Aim

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 example trace files.

2 Example data

Example .SCF trace files that will be used in this tutorial can be downloaded from the Applied Maths website (https://www.applied-maths.com/download/sample-data, click on "Batch assembly and alignment data"). The trace files originate from influenza A virus strains and represent partial sequences of the haemagglutinin (HA) and neuraminidase (NA) genes. These publicly available trace files were downloaded from the NCBI Trace Archive (http://0-www.ncbi.nlm.nih.gov.catalog.llu.edu/Traces/trace.cgi?).

3 Import and assembly

1. Create a new database (see tutorial "Creating a new database") or open an existing database.
2. Select ***File*** > ***Import...*** (, **Ctrl+I**) to call the *Import* dialog box.
3. Select ***Import and assemble trace files*** under ***Sequence type data*** and press <***Import***> to start the batch import routine.
4. Browse for the folder, select all .SCF trace files, press <***Open***> and press <***Next***>.

As this is the first time we import and assemble trace files in the database, we need to create a new import template by specifying ***Import rules***.

5. Select <***Create new***>.

The only source of information available in the newly created import template is the file name. The text between the underscore (_) and hyphen (-) holds the strain information and will now be linked

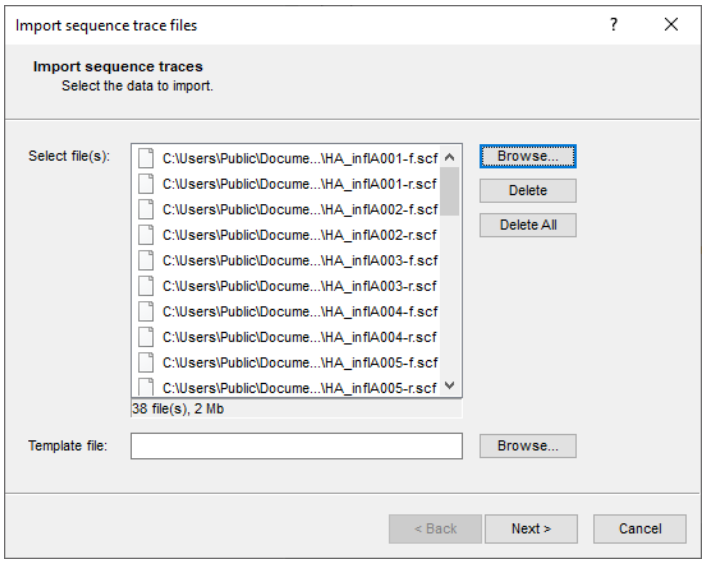

Figure 1: Select all trace files.

to the ***Key*** field in the database:

6. Double-click on the only line in the grid, or press <***Edit Destination***>. Select ***Key*** in the *Edit data destination* dialog box (see Figure 2) and press <***OK***>.

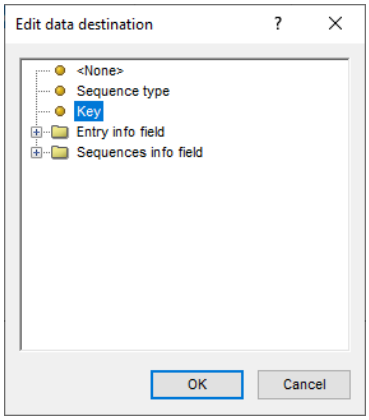

Figure 2: Select destination

7. Visualize the advanced options for the *Import template* dialog box by clicking on the check box next to **Show advanced options** and press <***Edit parsing***> to open the *Data parsing* dialog box.

8. In the *Data parsing* dialog box, fill in following data parsing string: **_[DATA]-**. The asterisk will serve as wildcard.

9. Press the <***Preview***> button and press <***OK***> when the parsing is correct (see Figure 3).

The text before the underscore (_) holds the gene names (HA and NA) and will now be linked to sequence types experiments in the database:

10. Select <***Add rule***>, select ***Name*** under ***File*** (see Figure 4) and press <***Next***>.

11. Select ***Sequence type*** from the list (see Figure 5) and press <***Next***> once more.

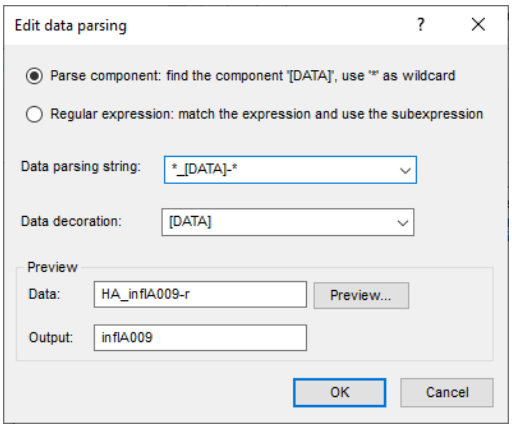

Figure 3: Parsing string.

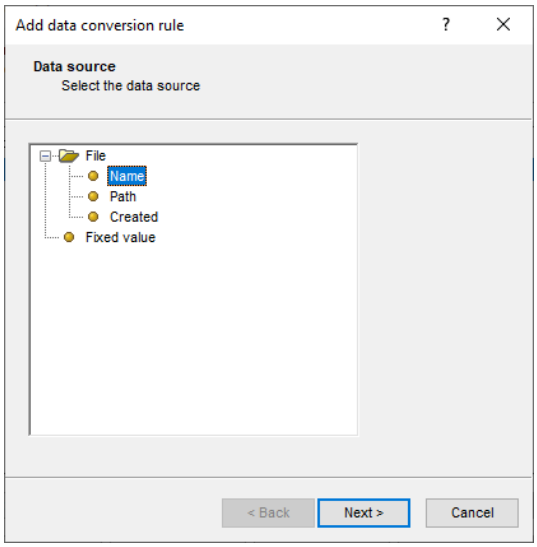

Figure 4: Add a new import rule.

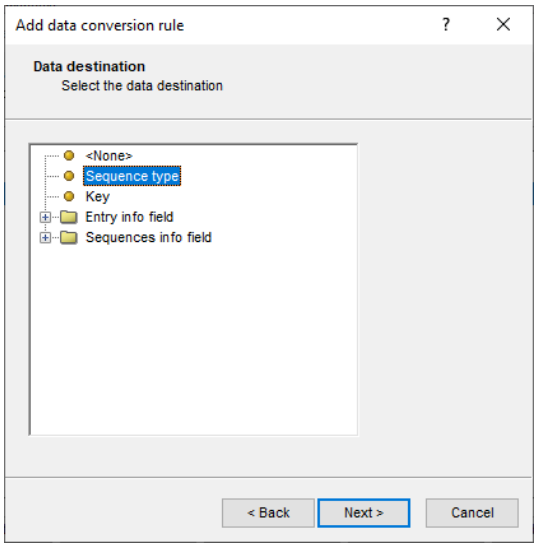

Figure 5: Link to a sequence type experiment.

12. In the *Data parsing* dialog box, fill in following data parsing string: *[DATA]_**. The asterisk will serve as wildcard.

13. Press the <***Preview***> button and press <***Next***> when the parsing is correct (see Figure 6) and <***Finish***>.

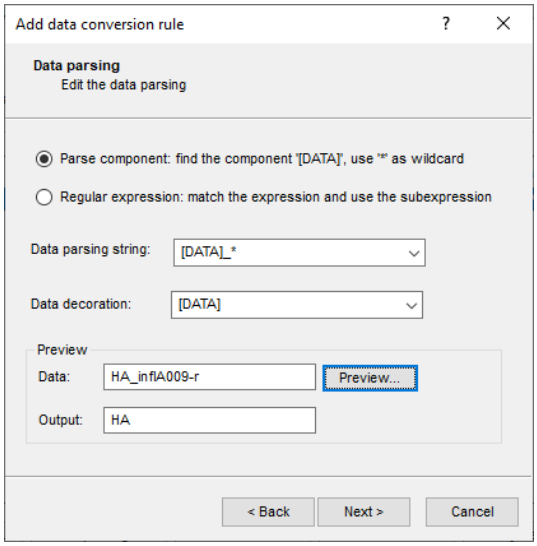

Figure 6: Parsing string.

The grid panel should now look like Figure 7.

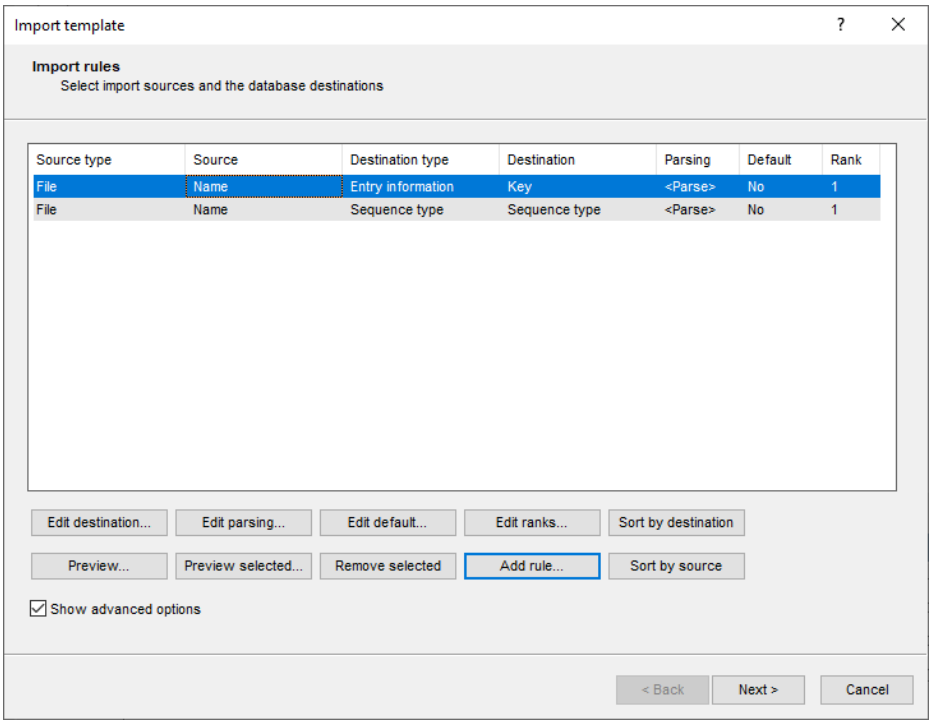

Figure 7: Import rules.

14. In the *Import template* dialog box, press <***Preview***> and verify the preview of the import (see Figure 8). If no errors occurred, press <***Next***> and <***Finish***>, else verify that the source, destination and parsing string of each rule has been entered correctly.

15. Name the import template (e.g. ”Import my SCF trace files”) and optionally give it a description. Press <***OK***>.

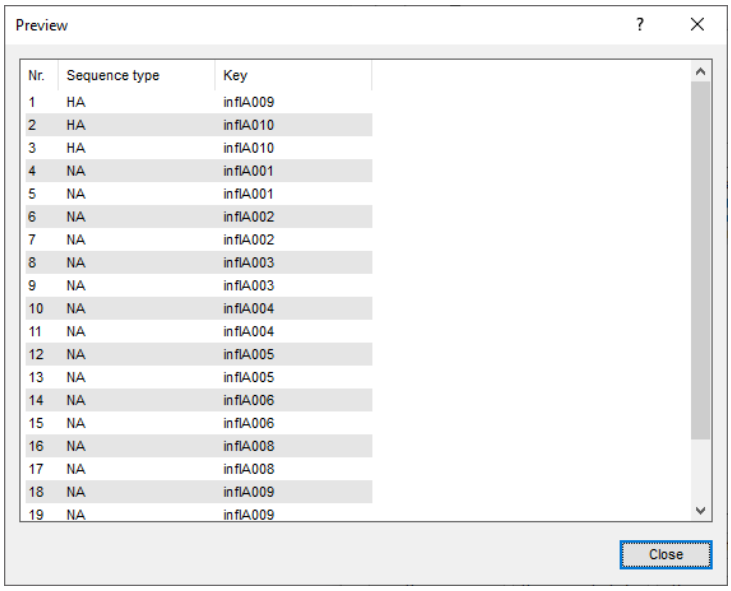

Figure 8: Preview of import.

The new import template is added to the template list and is automatically selected (see Figure 9).

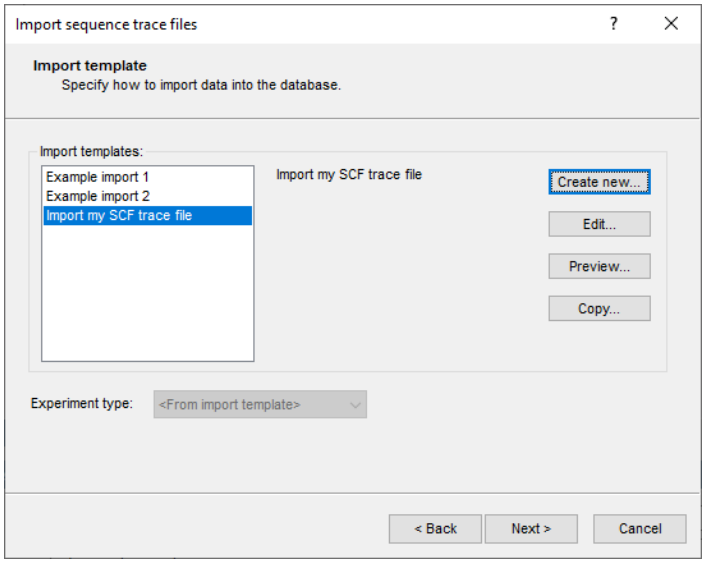

Figure 9: My new import template.

16. With the new import template highlighted, press <***Next***>.

BIONUMERICS will warn that the two sequence types are still missing in the database (see Figure 10).

17. Press <***Yes***> twice to have the two sequence type experiments created by the software.

In case there are no entries present with the same key as in the trace file names, the *Database links* wizard page will indicate that 10 new entries will be created during import.

18. Press <***Next***>.

The *Processing* wizard page opens (see Figure 12).

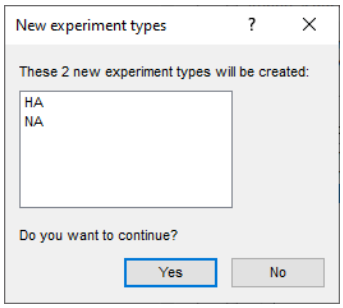

Figure 10: Missing experiments in the database.

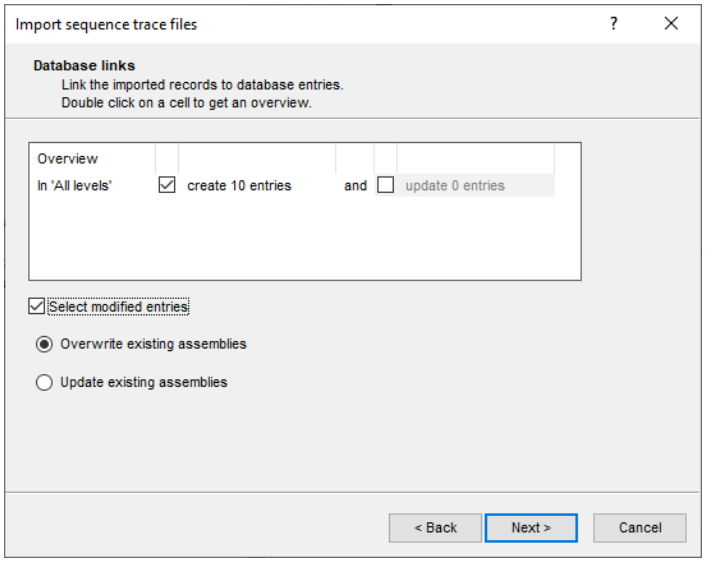

Figure 11: Create 10 new entries.

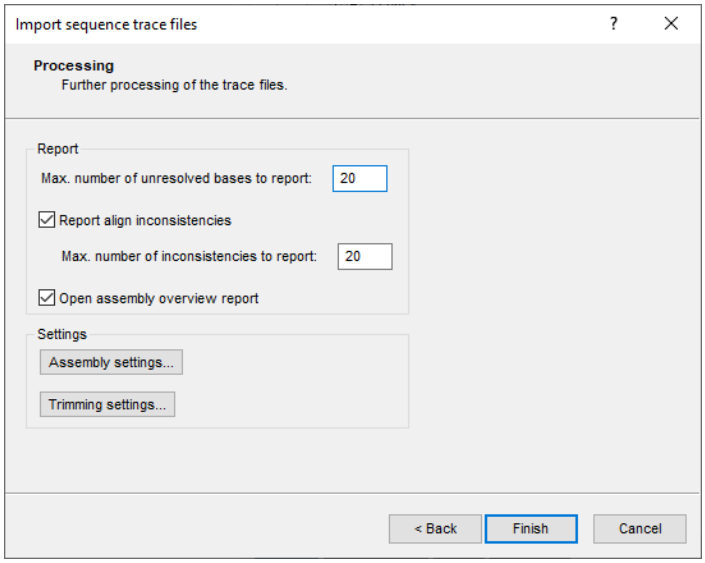

Figure 12: The *Processing* wizard page.

19. Press <***Trimming settings***> to pop up the *Assembly trimming settings* dialog box.

20. Double-click on the <***Edit***> button for experiment **HA** and enter the trimming settings as specified in Figure 13 and press <***OK***>.

When an ***Offset*** is specified, the consensus is trimmed at that offset from the trimming target positions.

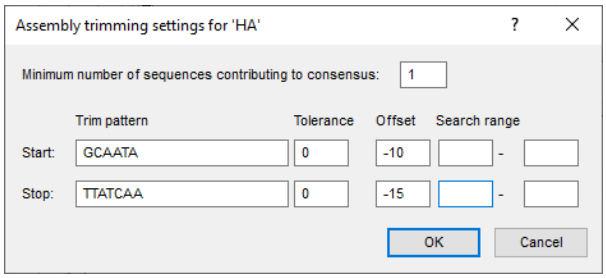

Figure 13: The *Assembly trimming settings* dialog box displaying the trimming settings for the HA sequence example data.

21. Double-click on the <***Edit***> button for experiment **NA** and enter the trimming settings as specified in Figure 14. When completed, press <***OK***>.

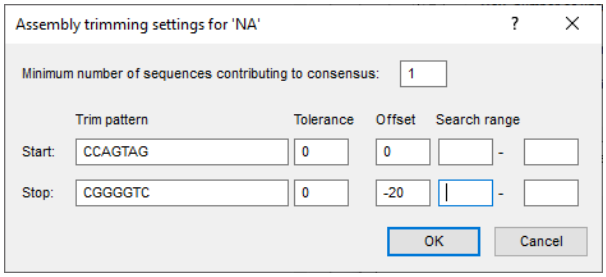

Figure 14: The *Assembly trimming settings* dialog box displaying the trimming settings for the NA sequence example data.

22. Press <***Close***> to close the *Assembly trimming settings* dialog box (see Figure 15).

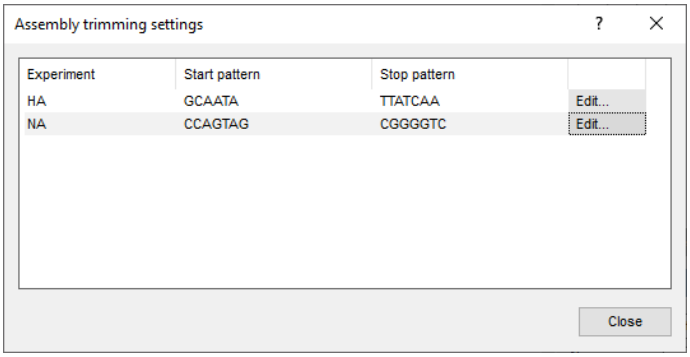

Figure 15: The *Assembly trimming settings* dialog box.

23. Press the <***Assembly settings***> button to call the *Assembly settings* dialog box (see Figure 16).

24. Double-click on the <***Edit***> button for experiment **HA** to call the *Assembly settings* dialog box (see Figure 17).

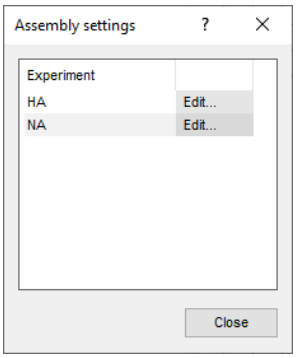

Figure 16: The *Assembly settings* dialog box.

The Assembly settings are grouped in tabs per settings dialog box in *Assembler*: ***Quality*** assignment, ***Assembly*** and ***Consensus*** determination. In the last tab the Assembly settings can be copied from or to another sequence type experiment.

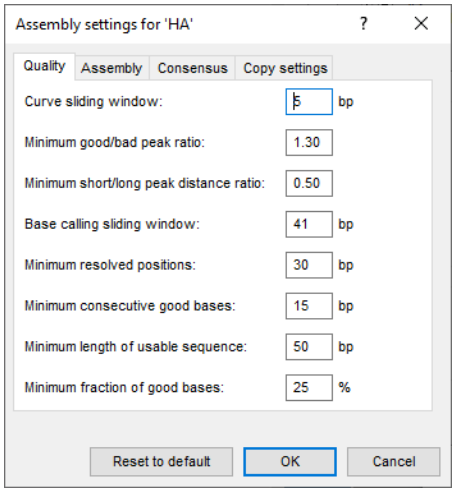

Figure 17: The Assembly settings.

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

26. Press <***Finish***> to have the 19 sequences automatically assembled.

4 Reports

The *Batch sequence assembly report* window (see Figure 18) opens when the option ***Open assembly overview report*** was checked in the *Processing* wizard page. 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. Each cell, corresponding to a key/experiment pair, provides information about the current status of the contig project. This information can be:

- **N/A**: No such experiment exists with this key.

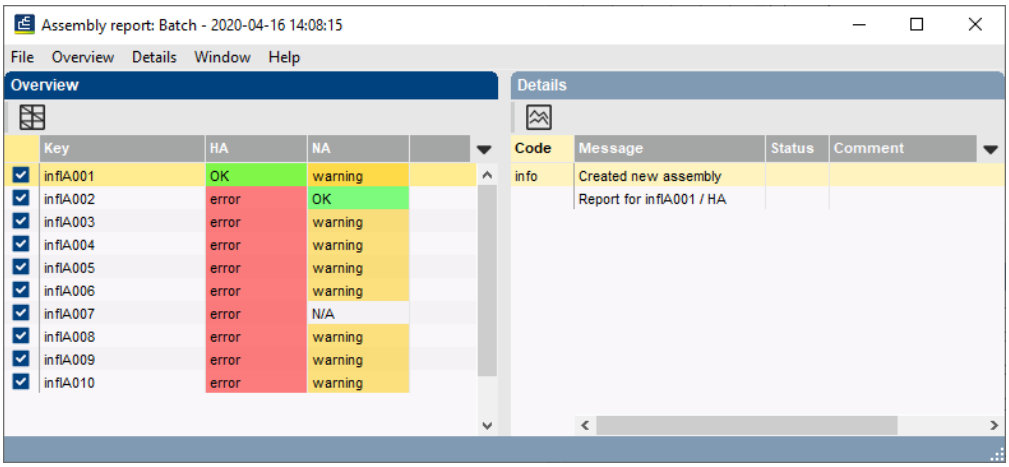

Figure 18: The *Batch sequence assembly report* window.

- **N/B**: An experiment with this key exists, but (a) the assembly was not created from this batch; or (b) no assembly is present for this sequence.
- **OK** (green): A contig was assembled without any problems.
- **Warning** (orange): Align inconsistencies occurred that were resolved under the applied consensus determination settings.
- **Error** (red): At least one of several possible assembly errors occurred, e.g. a trace sequence did not meet the quality criteria, more than one contig was created, the trimming positions were not found or unresolved bases are present in the consensus.
- **Read** (red): A warning or error that was read by the user, but not solved yet.
- **Solved** (green): A warning or error that was solved by the user (see below).

1. Click a cell, e.g. **inflA002/HA** to update the *Details* panel on the right-hand side (see Figure 19).

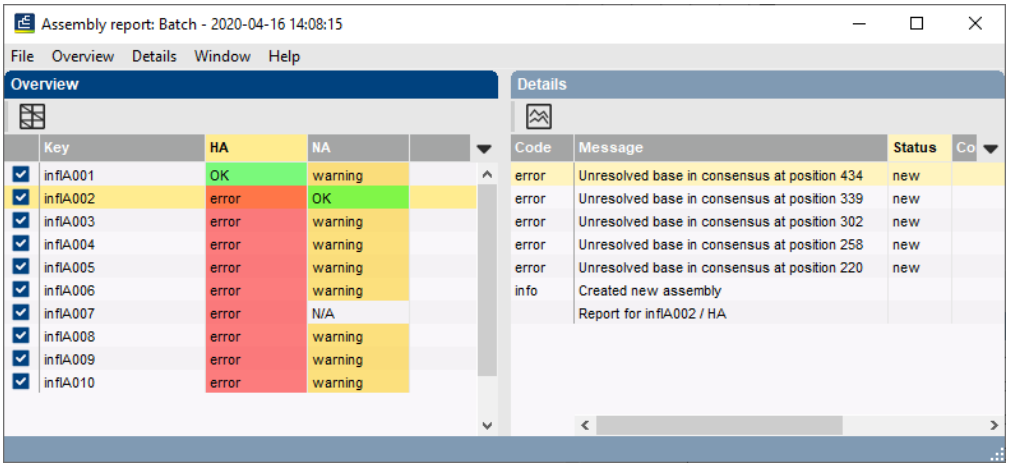

Figure 19: Details for the **inflA002/HA** assembly

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. In the *Details* panel double-click on the first error message.

This will open the sequence in the *Contig assembly* window (if not already open), with the corresponding position in focus (see Figure 20). The position can now be examined and - if needed - the base calling can be changed manually.

5 Checking assemblies

1. Select ***View*** > ***Display settings...*** to see how colors are assigned.
2. Make sure the *Aligned traces* panel is selected and use the zoom sliders or the zoom buttons to obtain an optimal view of the curves.

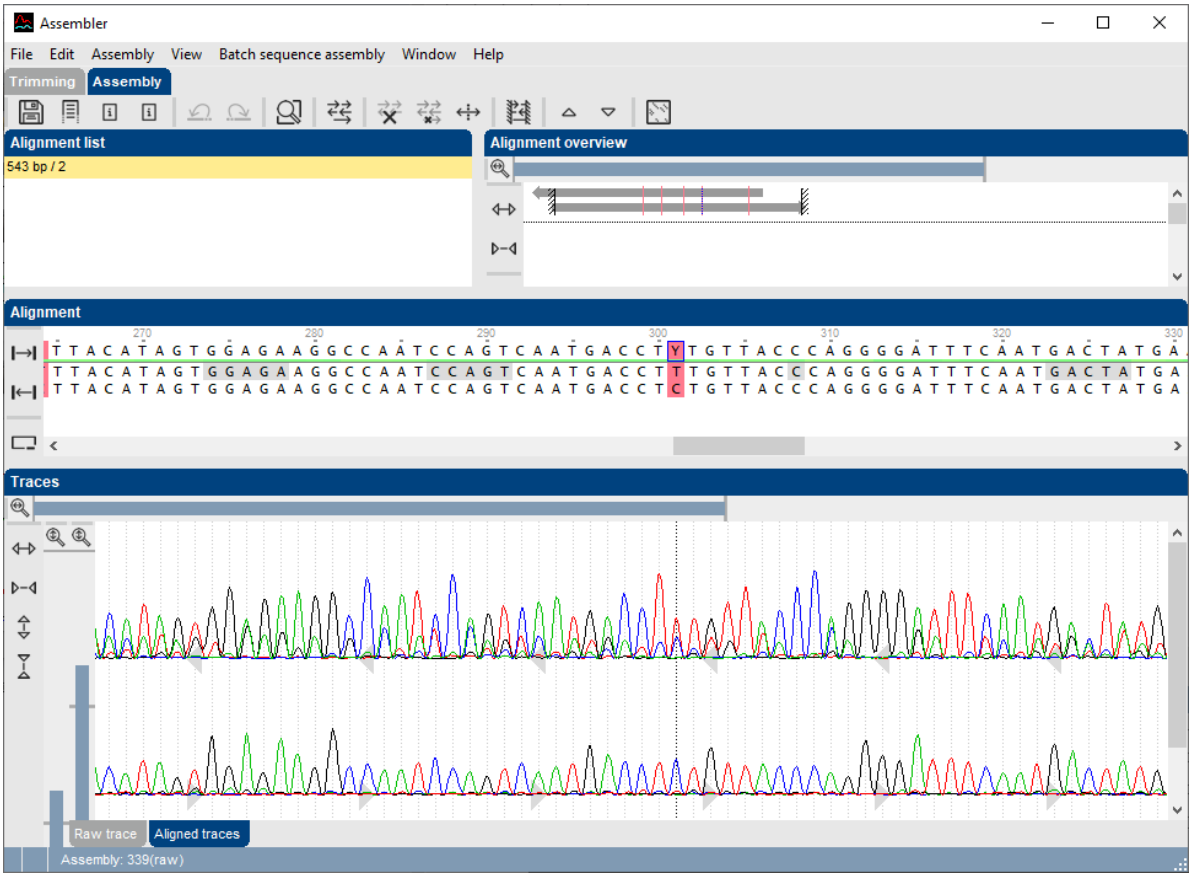

Figure 20: The *Contig assembly* window as called from the detailed report by double-clicking an error message. The window shows the contig project with the unresolved base in focus.

In case of the unresolved base highlighted in Figure 20, the ”T” needs to be changed into a ”y”.

3. Change the ”T” into a ”y”.

The base is now resolved under the default assembly settings and is no longer highlighted in red.

4. Check and resolve all other error/warning messages.

5. Select ***Batch sequence assembly*** > ***Set report to solved, save and close*** (**Ctrl+Shift+S**) in Assembler.

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 (see Figure 21).

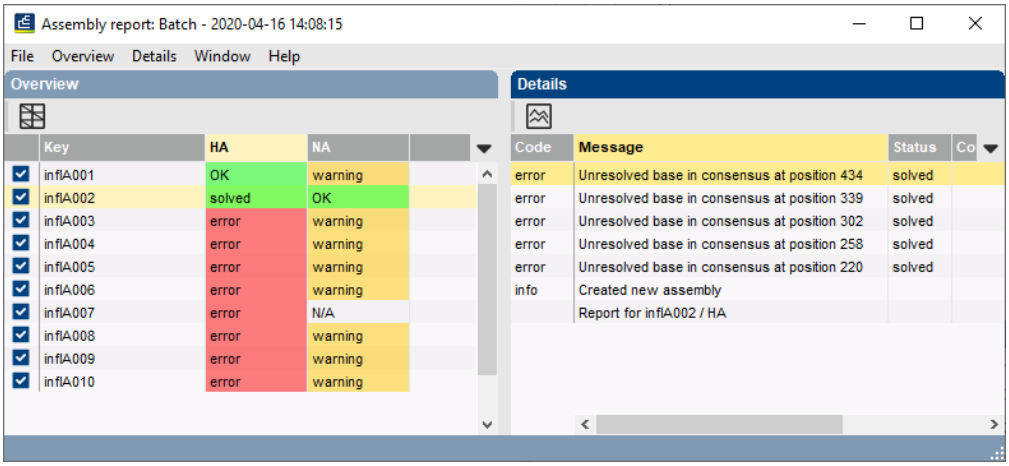

Figure 21: Solved status.

6. Close the *Batch sequence assembly report* window.

6 Open Assembler

The *Experiment presence* panel in the *Main* window shows for each database entry whether an experiment is available (colored dot) or not (see Figure 22).

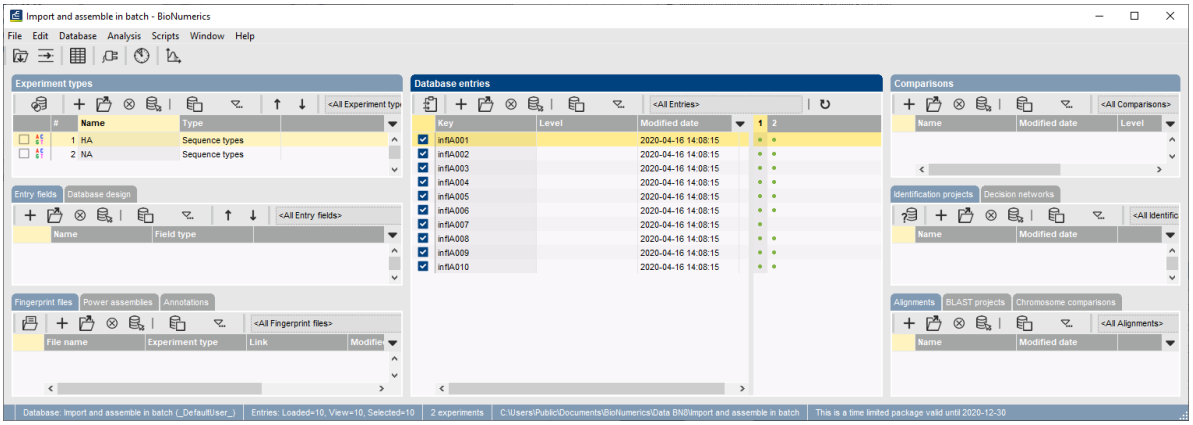

Figure 22: The *Main* window.

1. Click on a colored dot of a linked sequence type.

This action opens the *Sequence editor* window (see Figure 23).

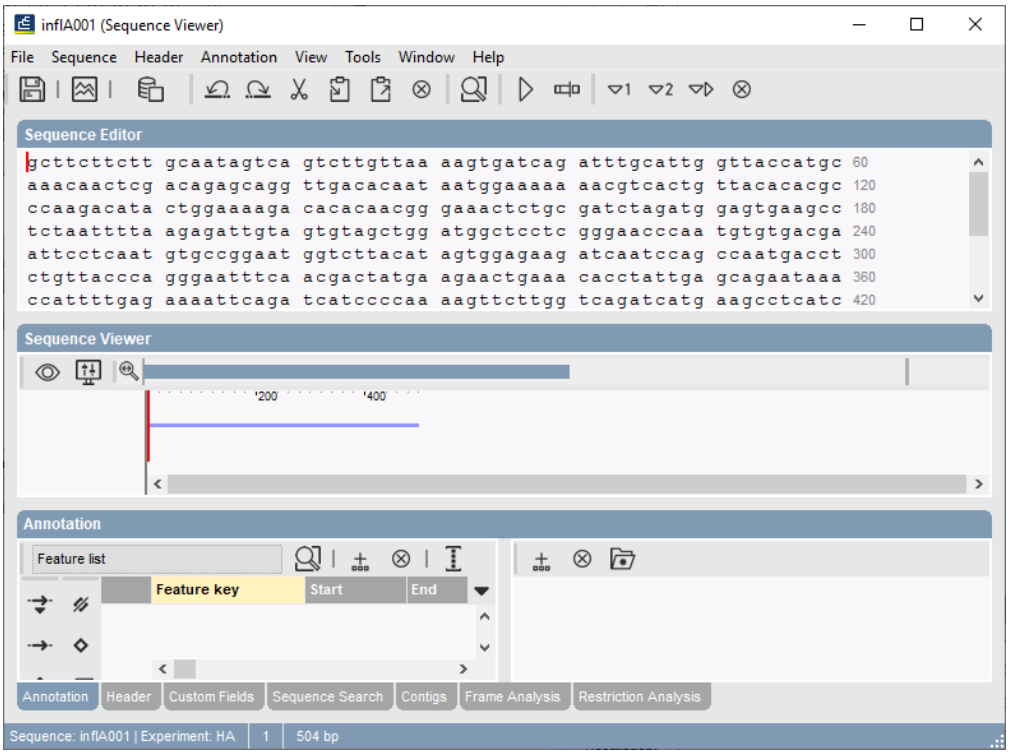

Figure 23: The *Sequence editor* window.

2. Press the button to launch Assembler to open the contig project associated with this sequence.

Alternatively, Assembler can be called from the Batch Overview reports, which are displayed from the *Main* window with ***Analysis*** > ***Sequence types*** > ***Batch assembly reports...***.

7 Conclusion

In this tutorial you have seen how to import and assemble trace files in batch. The sequences can now be analyzed in BIONUMERICS (aligning, clustering, mutation search, etc.). More information about these tools can found in the analysis tutorials on our website.