APPLIED MATHS
A BIOMÉRIEUX COMPANY

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 (http://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***.

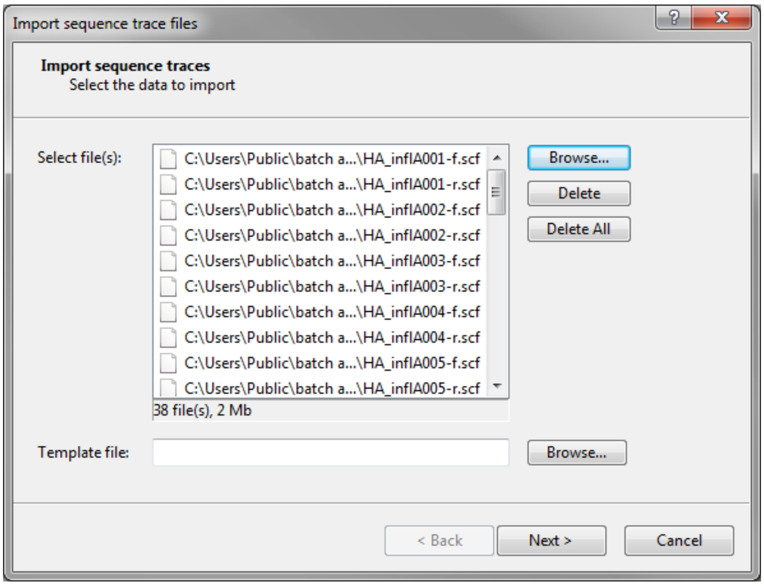
5. Select <***Create new***>.

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 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***>.

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.



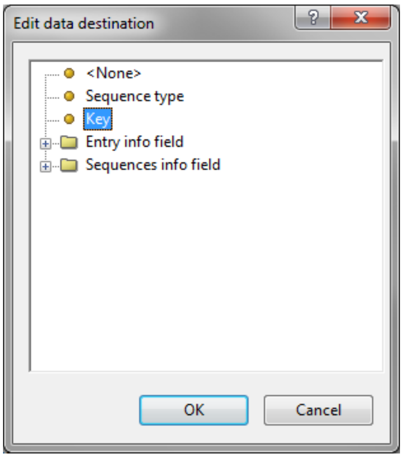
Import sequence trace files

Import sequence traces
Select the data to import

Select file(s):
C:\Users\Public\batch a...\HA_inflA001-f.scf
C:\Users\Public\batch a...\HA_inflA001-r.scf
C:\Users\Public\batch a...\HA_inflA002-f.scf
C:\Users\Public\batch a...\HA_inflA002-r.scf
C:\Users\Public\batch a...\HA_inflA003-f.scf
C:\Users\Public\batch a...\HA_inflA003-r.scf
C:\Users\Public\batch a...\HA_inflA004-f.scf
C:\Users\Public\batch a...\HA_inflA004-r.scf
C:\Users\Public\batch a...\HA_inflA005-f.scf
C:\Users\Public\batch a...\HA_inflA005-r.scf
38 file(s), 2 Mb

Template file:

Figure 1: Select all trace files.



Edit data destination

<None>
Sequence type
Key
Entry info field
Sequences info field

Figure 2: Select destination

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***>.

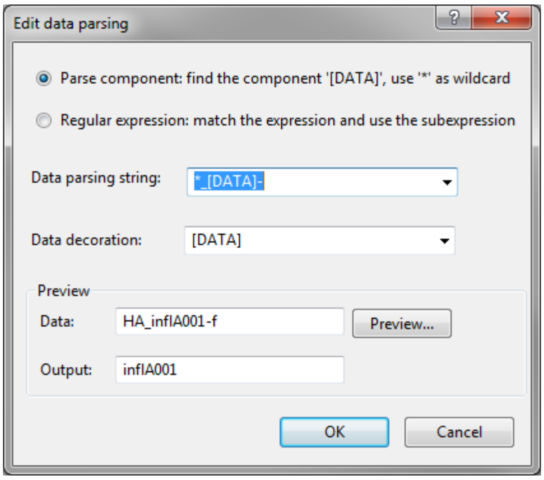
11. Select ***Sequence type*** from the list (see Figure 5) and press <***Next***> once more.

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***>.

The grid panel should now look like Figure 7.

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.


Edit data parsing

● Parse component: find the component '[DATA]', use '*' as wildcard

○ Regular expression: match the expression and use the subexpression

Data parsing string: *_[DATA]-

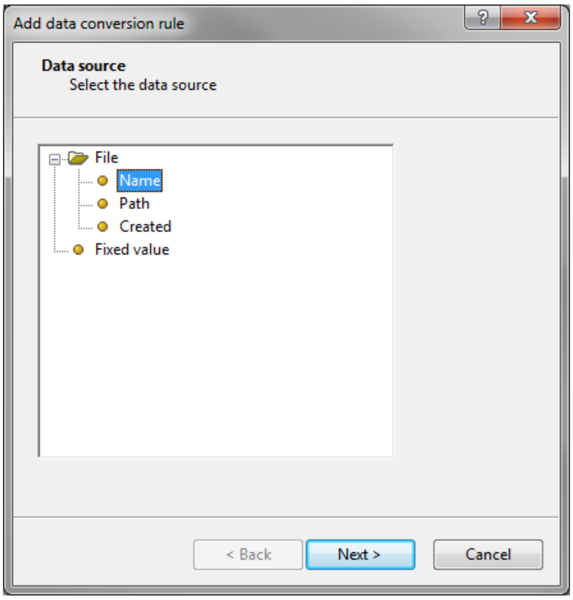
Data decoration: [DATA]

Preview

Data: HA_inflA001-f

Output: inflA001

Figure 3: Parsing string.


Add data conversion rule

Data source

Select the data source

Figure 4: Add a new import rule.

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

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

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).

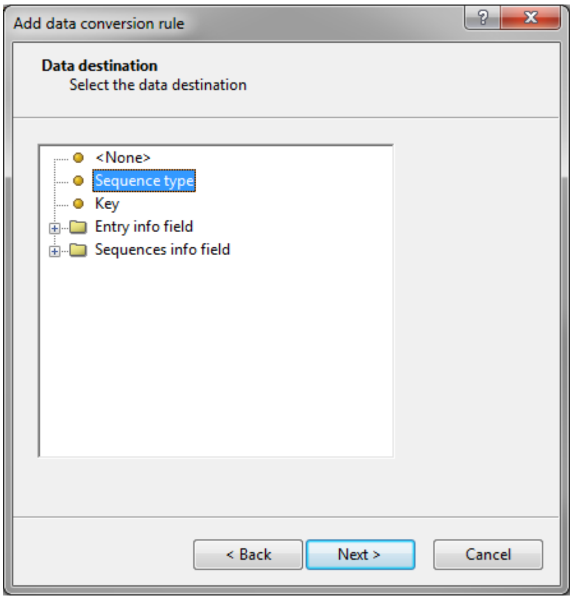
17. Press <***Yes***> twice to have the two sequence type experiments created by the software.

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).

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

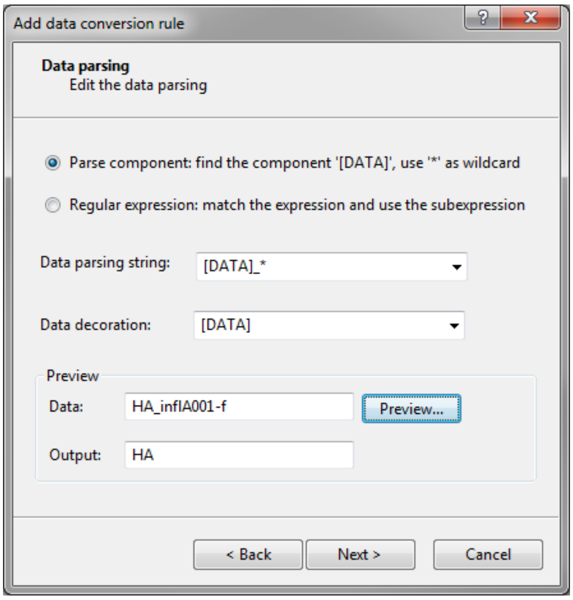


Add data conversion rule

Data destination
Select the data destination

- <None>
- Sequence type
- Key
- Entry info field
- Sequences info field

Figure 5: Link to a sequence type experiment.



Add data conversion rule

Data parsing
Edit the data parsing

● Parse component: find the component '[DATA]', use '*' as wildcard

○ Regular expression: match the expression and use the subexpression

Data parsing string: [DATA]_*

Data decoration: [DATA]

Preview

Data: HA_inflA001-f

Output: HA

Figure 6: Parsing string.

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

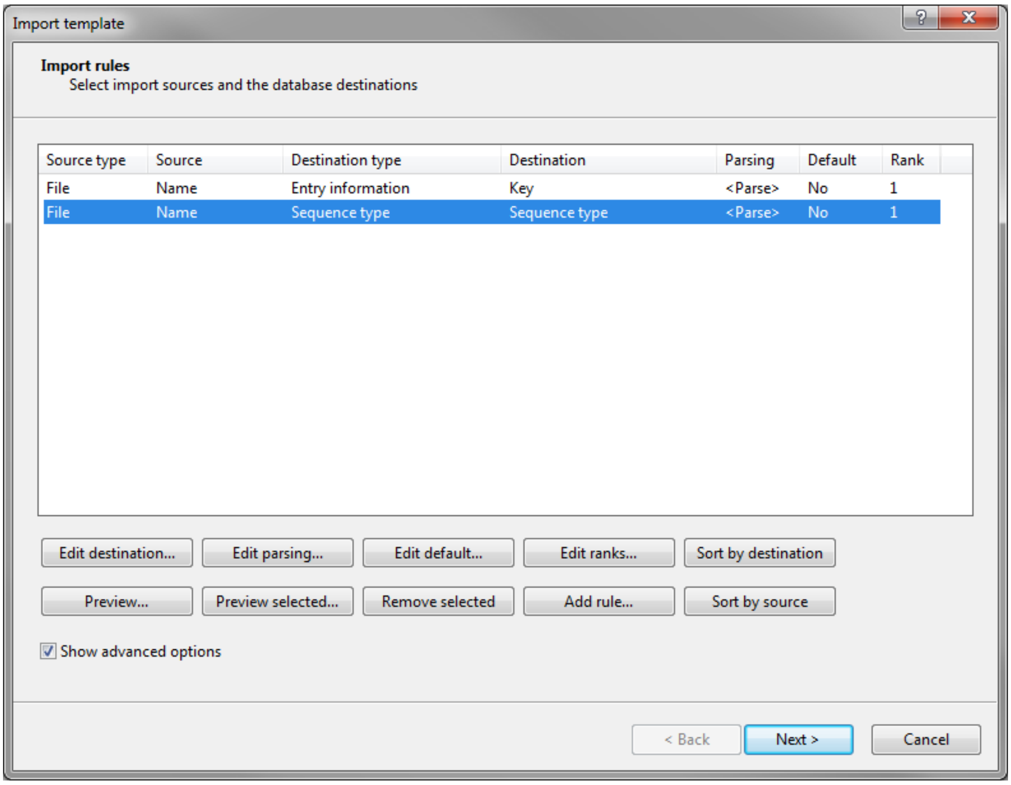
When an ***Offset*** is specified, the consensus is trimmed at that offset from the trimming target positions.

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

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

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).



Import template

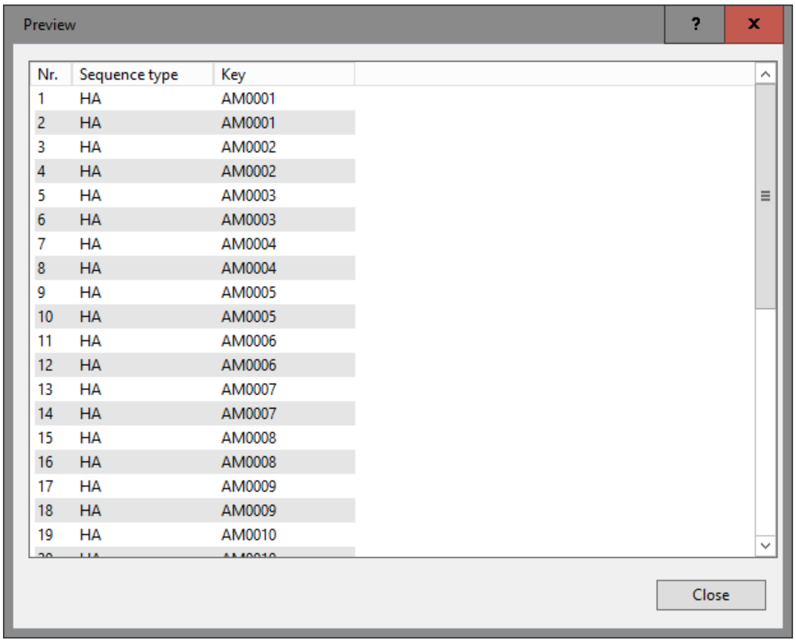
Import rules

Select import sources and the database destinations

Source type	Source	Destination type	Destination	Parsing	Default	Rank
File	Name	Entry information	Key	<Parse>	No	1
File	Name	Sequence type	Sequence type	<Parse>	No	1

☑ Show advanced options

Figure 7: Import rules.



Preview

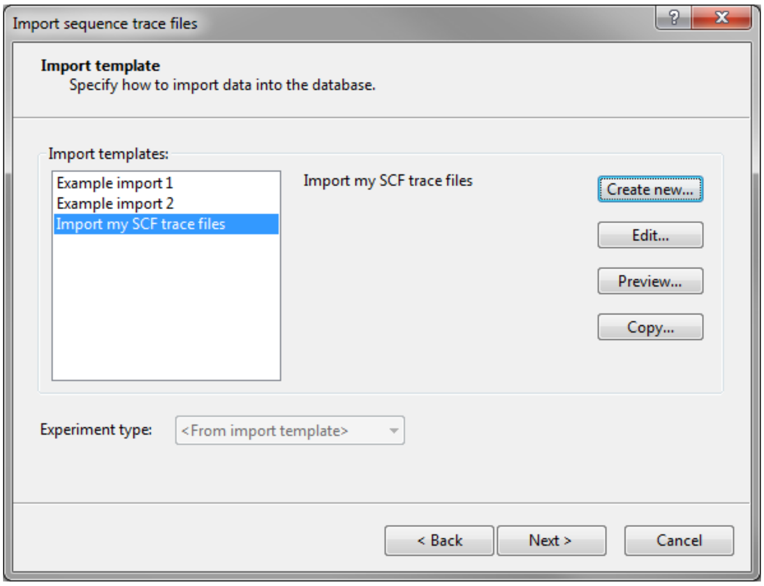
Nr.	Sequence type	Key
1	HA	AM0001
2	HA	AM0001
3	HA	AM0002
4	HA	AM0002
5	HA	AM0003
6	HA	AM0003
7	HA	AM0004
8	HA	AM0004
9	HA	AM0005
10	HA	AM0005
11	HA	AM0006
12	HA	AM0006
13	HA	AM0007
14	HA	AM0007
15	HA	AM0008
16	HA	AM0008
17	HA	AM0009
18	HA	AM0009
19	HA	AM0010

Figure 8: Preview of import.

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.

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

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



Import sequence trace files

Import template

Specify how to import data into the database.

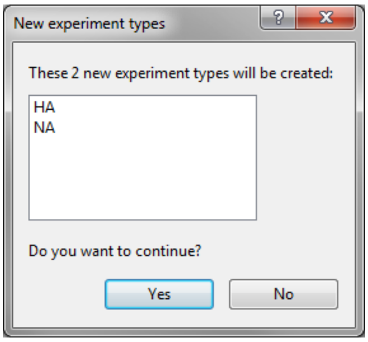
Import templates:

- Example import 1
- Example import 2
- Import my SCF trace files

Import my SCF trace files

Experiment type: <From import template>

Figure 9: My new import template.



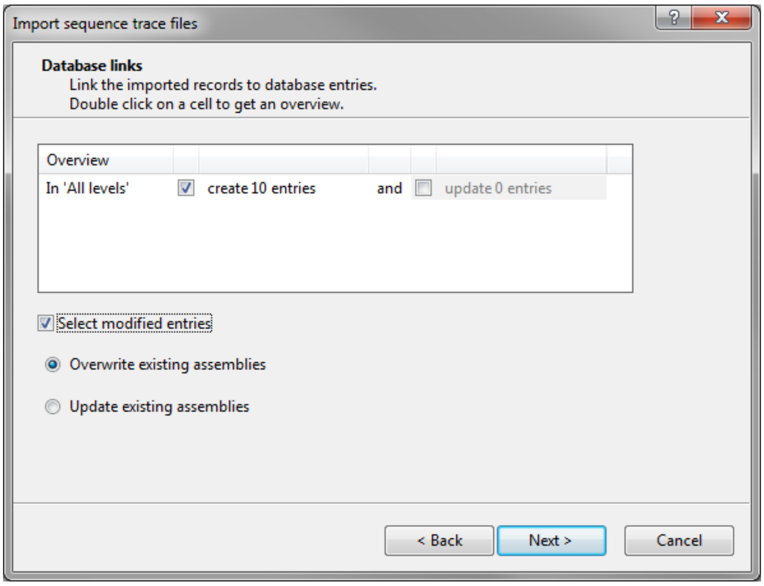
New experiment types

These 2 new experiment types will be created:

HA
NA

Do you want to continue?

Figure 10: Missing experiments in the database.



Import sequence trace files

Database links

Link the imported records to database entries.
Double click on a cell to get an overview.

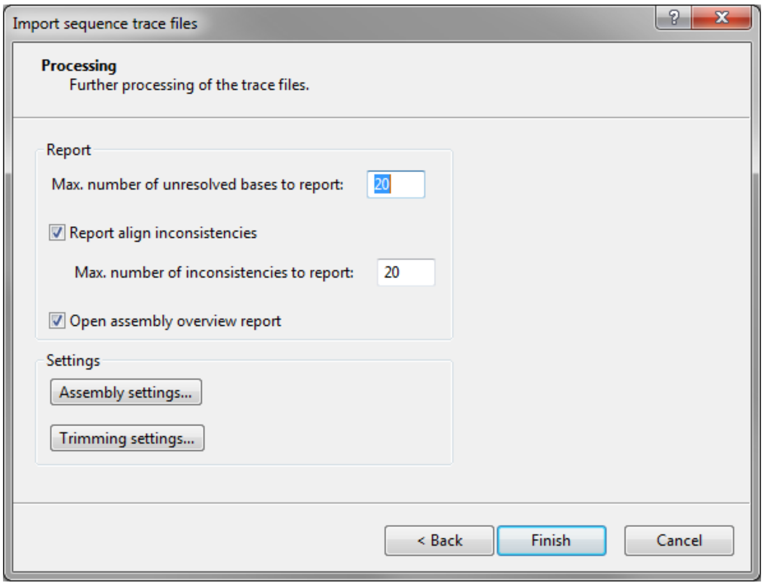
Overview				
In 'All levels'	☑	create 10 entries	and	□ update 0 entries

☑ Select modified entries

● Overwrite existing assemblies

○ Update existing assemblies

Figure 11: Create 10 new entries.



Import sequence trace files

Processing
Further processing of the trace files.

Report

Max. number of unresolved bases to report: 20

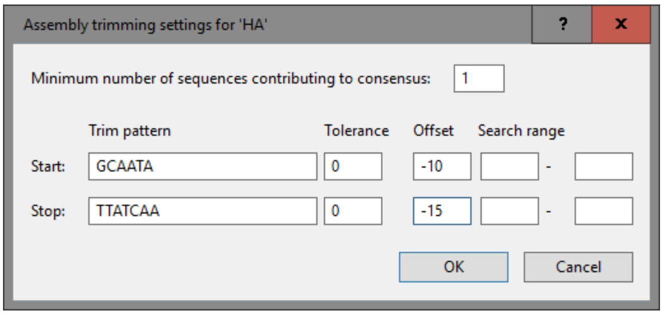
☑ Report align inconsistencies

Max. number of inconsistencies to report: 20

☑ Open assembly overview report

Settings

Figure 12: The *Processing* wizard page.

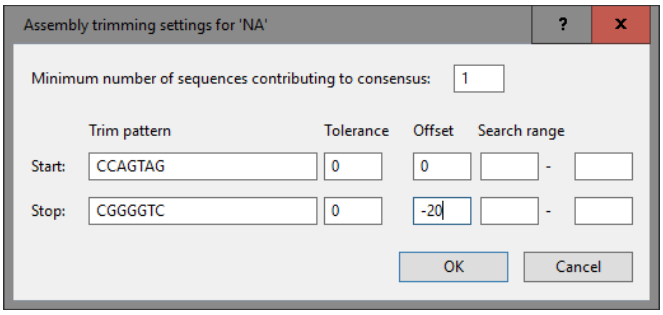


Assembly trimming settings for 'HA'

Minimum number of sequences contributing to consensus: 1

	Trim pattern	Tolerance	Offset	Search range	
Start:	GCAATA	0	-10	-	
Stop:	TTATCAA	0	-15	-	

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



Assembly trimming settings for 'NA'

Minimum number of sequences contributing to consensus: 1

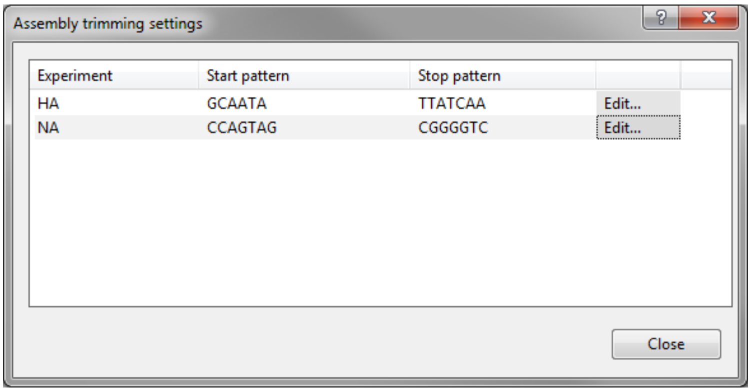
	Trim pattern	Tolerance	Offset	Search range	
Start:	CCAGTAG	0	0	-	
Stop:	CGGGGTC	0	-20	-	

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

4 Reports

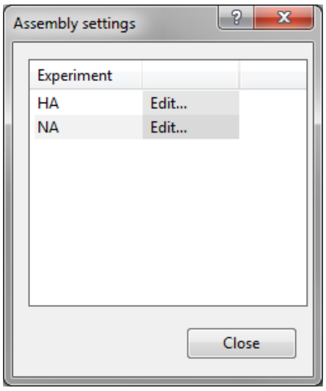
The *Batch sequence assembly report* window (see Figure 18) opens when the option ***Open assembly overview report*** was checked in the *Processing* wizard page. This window can also be displayed from the *Main* window with ***Analysis*** > ***Sequence types*** > ***Batch assembly reports...***.

The *Overview* panel displays the entries (keys) as rows and the experiments as columns. Each cell, corre-



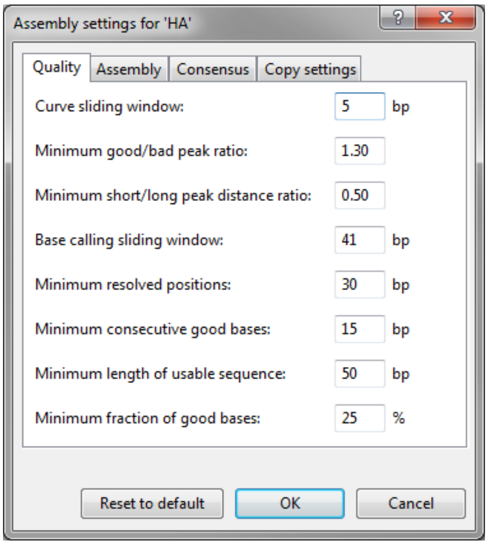
Experiment	Start pattern	Stop pattern	
HA	GCAATA	TTATCAA	Edit...
NA	CCAGTAG	CGGGGTC	Edit...

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



Experiment	
HA	Edit...
NA	Edit...

Figure 16: The *Assembly settings* dialog box.



Curve sliding window: 5 bp

Minimum good/bad peak ratio: 1.30

Minimum short/long peak distance ratio: 0.50

Base calling sliding window: 41 bp

Minimum resolved positions: 30 bp

Minimum consecutive good bases: 15 bp

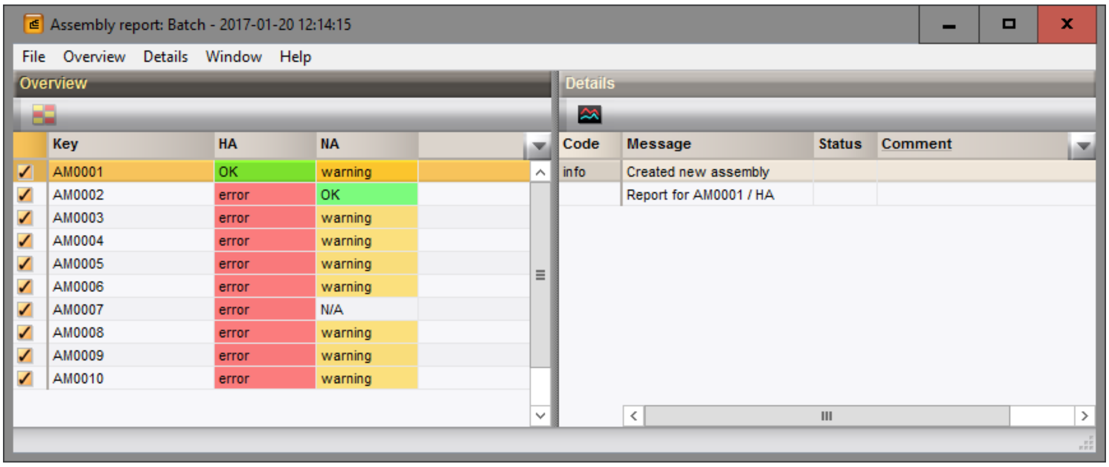
Minimum length of usable sequence: 50 bp

Minimum fraction of good bases: 25 %

Figure 17: The Assembly settings.

sponding 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.
- **N/B**: An experiment with this key exists, but (a) the assembly was not created from this batch; or (b)


Assembly report: Batch - 2017-01-20 12:14:15
File Overview Details Window Help
Overview
Details

	Key	HA	NA
☑	AM0001	OK	warning
☑	AM0002	error	OK
☑	AM0003	error	warning
☑	AM0004	error	warning
☑	AM0005	error	warning
☑	AM0006	error	warning
☑	AM0007	error	N/A
☑	AM0008	error	warning
☑	AM0009	error	warning
☑	AM0010	error	warning

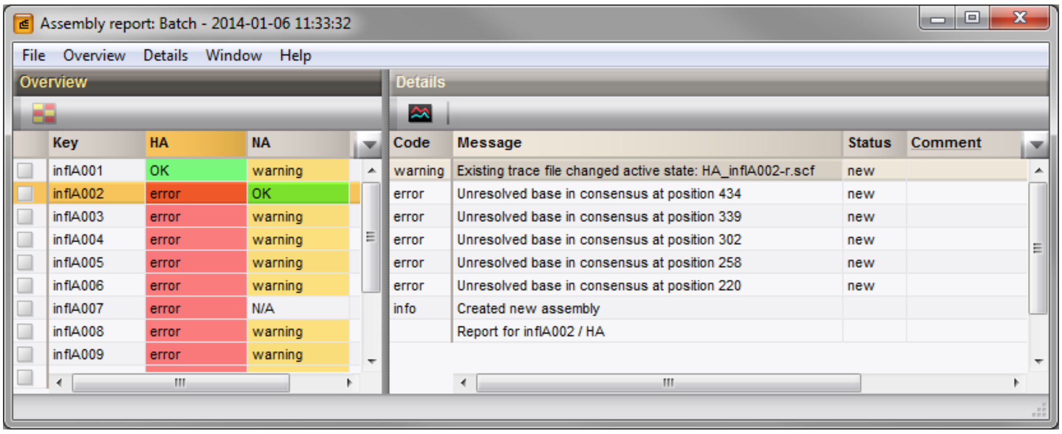
Code	Message	Status	Comment
info	Created new assembly		
	Report for AM0001 / HA		

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

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).


Assembly report: Batch - 2014-01-06 11:33:32
File Overview Details Window Help
Overview
Details

	Key	HA	NA
□	inflA001	OK	warning
□	inflA002	error	OK
□	inflA003	error	warning
□	inflA004	error	warning
□	inflA005	error	warning
□	inflA006	error	warning
□	inflA007	error	N/A
□	inflA008	error	warning
□	inflA009	error	warning

Code	Message	Status	Comment
warning	Existing trace file changed active state: HA_inflA002-r.scf	new	
error	Unresolved base in consensus at position 434	new	
error	Unresolved base in consensus at position 339	new	
error	Unresolved base in consensus at position 302	new	
error	Unresolved base in consensus at position 258	new	
error	Unresolved base in consensus at position 220	new	
info	Created new assembly		
	Report for inflA002 / HA		

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. Use the zoom sliders in the *Traces* panel or use the zoom buttons to obtain an optimal view of the curves (see Figure 20).

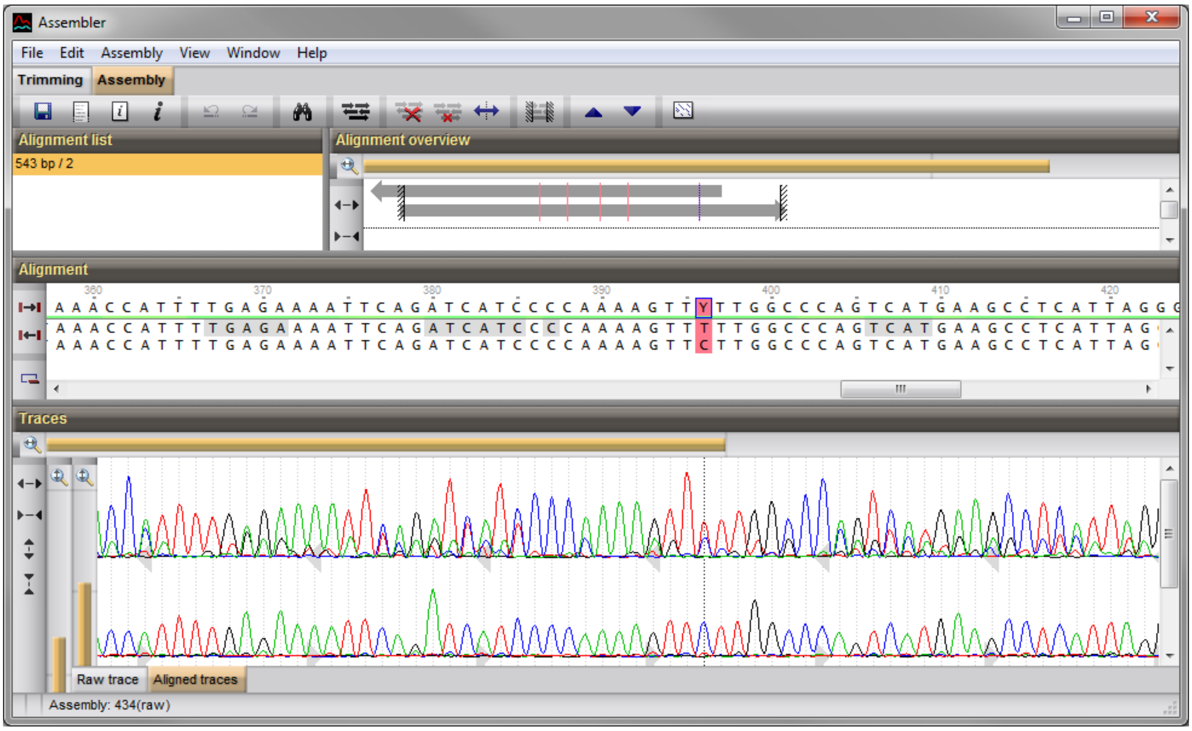

Assembler
File Edit Assembly View Window Help
Trimming Assembly
Alignment list
543 bp / 2
Alignment overview
Alignment
360 370 380 390 400 410 420
AAACCATTTTGAGAAAATTCAGATCATCCCCAAAAGTTYTTGGCCCAGTCATGAAGCCTCATTAGGG
AAACCATTTTGAGAAAATTCAGATCATCCCCAAAAGTTTTTGGCCCAGTCATGAAGCCTCATTAG
AAACCATTTTGAGAAAATTCAGATCATCCCCAAAAGTTCTTGGCCCAGTCATGAAGCCTCATTAG
Traces
Raw trace
Aligned traces
Assembly: 434(raw)

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”.

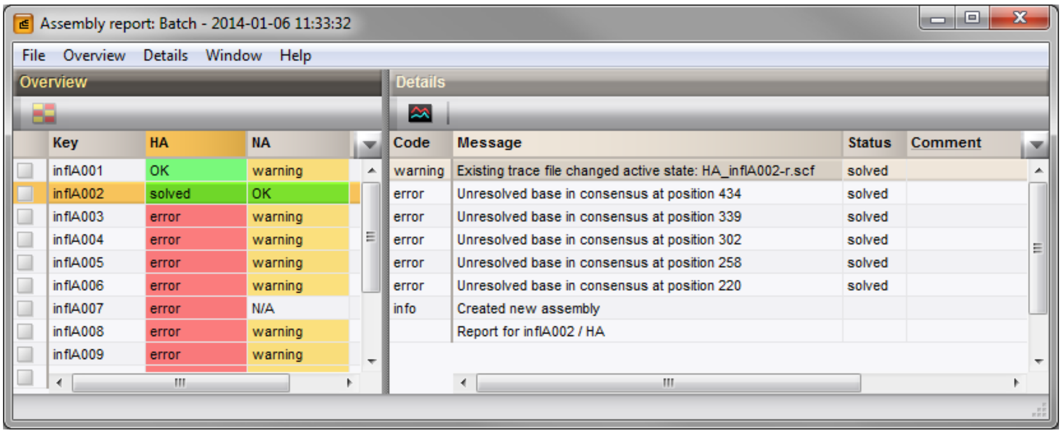
2. Change the ”T” into a ”y”.

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

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

4. 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).


Assembly report: Batch - 2014-01-06 11:33:32
File Overview Details Window Help
Overview
Details

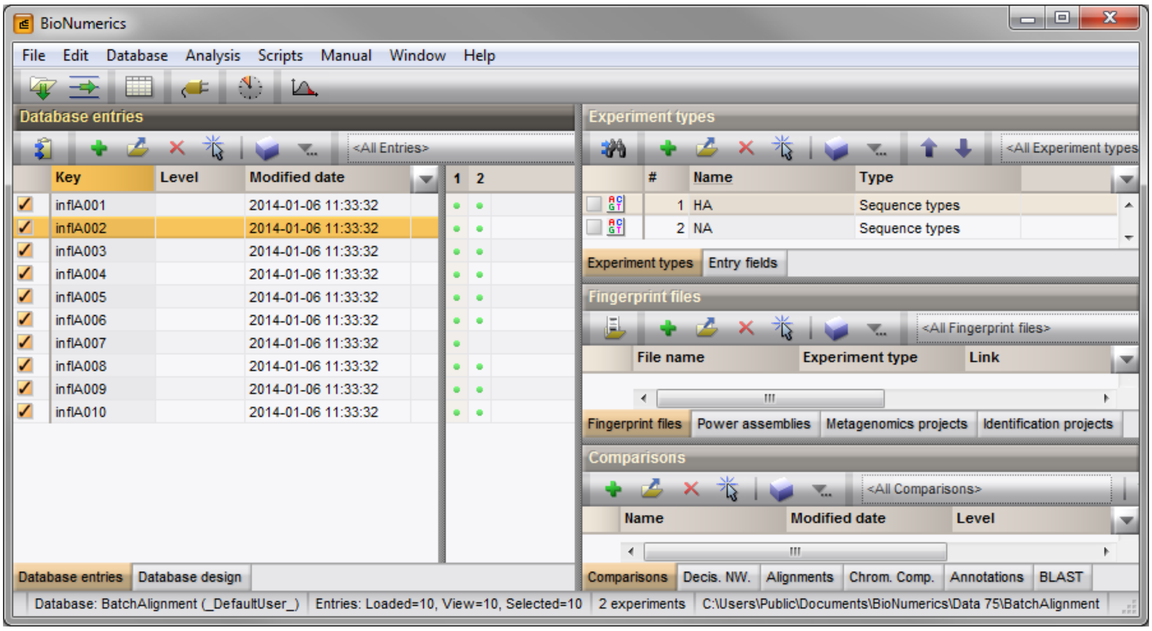
	Key	HA	NA
□	inflA001	OK	warning
□	inflA002	solved	OK
□	inflA003	error	warning
□	inflA004	error	warning
□	inflA005	error	warning
□	inflA006	error	warning
□	inflA007	error	N/A
□	inflA008	error	warning
□	inflA009	error	warning

Code	Message	Status	Comment
warning	Existing trace file changed active state: HA_inflA002-r.scf	solved	
error	Unresolved base in consensus at position 434	solved	
error	Unresolved base in consensus at position 339	solved	
error	Unresolved base in consensus at position 302	solved	
error	Unresolved base in consensus at position 258	solved	
error	Unresolved base in consensus at position 220	solved	
info	Created new assembly		
	Report for inflA002 / HA		

Figure 21: Solved status.

6 Open Assembler

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



	Key	Level	Modified date	1	2
☑	inflA001		2014-01-06 11:33:32	●	●
☑	inflA002		2014-01-06 11:33:32	●	●
☑	inflA003		2014-01-06 11:33:32	●	●
☑	inflA004		2014-01-06 11:33:32	●	●
☑	inflA005		2014-01-06 11:33:32	●	●
☑	inflA006		2014-01-06 11:33:32	●	●
☑	inflA007		2014-01-06 11:33:32	●	
☑	inflA008		2014-01-06 11:33:32	●	●
☑	inflA009		2014-01-06 11:33:32	●	●
☑	inflA010		2014-01-06 11:33:32	●	●

	#	Name	Type
□	1	HA	Sequence types
□	2	NA	Sequence types

Database: BatchAlignment (_DefaultUser_) | Entries: Loaded=10, View=10, Selected=10 | 2 experiments | C:\Users\Public\Documents\BioNumerics\Data 75\BatchAlignment

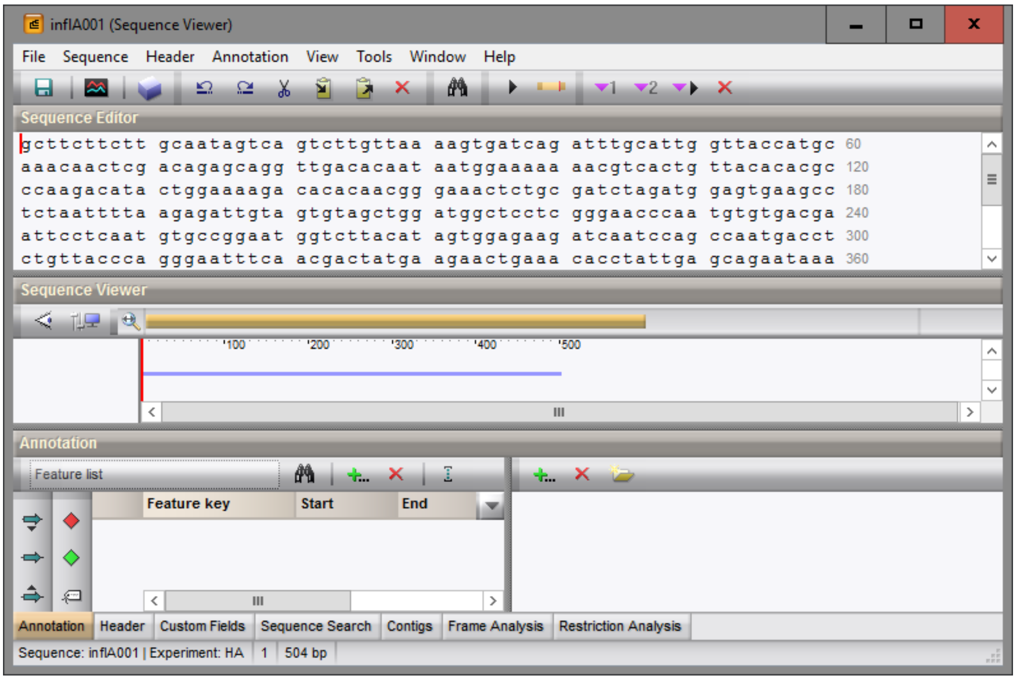
Figure 22: The *Main* window.

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

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

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***.



```
gcttcttctt gcaatagtca gtcttgttaa aagtgatcag atttgcattg gttaccatgc 60
aaacaactcg acagagcagg ttgacacaat aatggaaaaa aacgtcactg ttacacacgc 120
ccaagacata ctggaaaaga cacacaacgg gaaactctgc gatctagatg gagtgaagcc 180
tctaatttta agagattgta gtgtagctgg atggctcctc gggaacccaa tgtgtgacga 240
attcctcaat gtgccggaat ggtcttacat agtggagaag atcaatccag ccaatgacct 300
ctgttaccca gggaatttca acgactatga agaactgaaa cacctattga gcagaataaa 360
```

Sequence: infIA001 | Experiment: HA | 1 | 504 bp

Figure 23: The *Sequence editor* window.

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.