

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://o-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

mport sequenc	e trace files		?	Х
Import seque Select the	ence traces data to import.			
Select file(s):	C:\Users\Public\Docume\HA_inflA001-f.scf C:\Users\Public\Docume\HA_inflA001-r.scf C:\Users\Public\Docume\HA_inflA002-f.scf C:\Users\Public\Docume\HA_inflA002-r.scf C:\Users\Public\Docume\HA_inflA003-f.scf C:\Users\Public\Docume\HA_inflA003-f.scf C:\Users\Public\Docume\HA_inflA004-f.scf C:\Users\Public\Docume\HA_inflA004-f.scf C:\Users\Public\Docume\HA_inflA004-f.scf C:\Users\Public\Docume\HA_inflA004-f.scf C:\Users\Public\Docume\HA_inflA004-f.scf C:\Users\Public\Docume\HA_inflA004-f.scf C:\Users\Public\Docume\HA_inflA005-f.scf C:\Users\Public\Docume\HA_inflA005-f.scf C:\Users\Public\Docume\HA_inflA005-f.scf C:\Users\Public\Docume\HA_inflA005-f.scf C:\Users\Public\Docume\HA_inflA005-f.scf	Delete Delete All		
Template file:		Browse		
	< Back	Next >	Can	cel

Figure 1: Select all trace files.

to the Key field in the database:

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*>.

Edit data destination	?	×
 <none></none> Sequence type Key Entry info field Sequences info field 		
ОК	Car	icel

Figure 2: Select destination

- 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.

Edit data p	arsing				?	×
Parse	component	: find the comp	oonent '[DA	ATA]', use '*' a	s wildcar	d
🔿 Regul	ar expressi	on: match the e	expressior	n and use the	subexpre	ession
Data parsi	ng string:	*_[DATA]-*			~	
Data decor	ration:	[DATA]			~	
Preview						
Data:	HA_inflA	009-r		Preview		
Output:	in flA009					
				ОК	Can	icel

Figure 3: Parsing string.

Add data conversion rule			?	×
Data source Select the data source				
File Path Created Fixed value				
۶.				
	< Back	Next >	Can	cel

Figure 4: Add a new import rule.

Add data conversion rule			?	×
Data destination Select the data destination	I			
 <none></none> Sequence type Key Entry info field Sequences info field 				
	< Back	Next >	Can	cel

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*>.

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	Add data conversion I	uie			?)
Regular expression: match the expression and use the subexpression Data parsing string: [DATA]_* Data decoration: [DATA] Preview		sing				
Data parsing string: [DATA]_* Data decoration: [DATA] Preview Data: HA_inflA009-r Preview	Parse component	t: find the compor	ient '[DATA]', us	e '*' as wi	ldcard	
Data decoration: [DATA] Preview Data: HA_inflA009-r Preview	Regular express	ion: match the exp	pression and us	e the sub	expression	n
Preview Data: HA_inflA009-r Preview	Data parsing string:	[DATA]_*		~		
Data: HA_inflA009-r Preview	Data decoration:	[DATA]		~		
	Preview					
Output: HA	Data: HA_infl4	\009-r	Prev	/iew		
	Output: HA					
< Back Next > Cancel						

Figure 6: Parsing string.

The grid panel should now look like Figure 7.

ource type	Source	Destination type	Destination	Parsing	Default	Rank
ile	Name	Entry information	Key	<parse></parse>	No	1
ile	Name	Sequence type	Sequence type	<parse></parse>	No	1
Edit destination	Edit parsing	Edit default	Edit ranks So	rt by destination	1	

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 >.

quence type	Key inflA009 inflA010 inflA010 inflA001 inflA001			
	inflA010 inflA010 inflA001			
	in fIA010 in fIA001			
	inflA001			
	inflA001			
	inflA002			
	inflA002			
	inflA003			
	inflA003			
	inflA004			
	inflA004			
	inflA005			
	inflA005			
	inflA006			
	in flA006			
	in flA008			
	in flA008			
	in flA009			
	in flA009			_
		in fIA008 in fIA009	inflA008 inflA009	in fIA008 in fIA009

Figure 8: Preview of import.

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

Import sequence trace files		?	Х
Import template Specify how to import data into	the database.		
Import templates:			
Example import 1 Example import 2	Import my SCF trace file	Create new	
Import my SCF trace file		Edit	
		Preview	
		Сору	
Experiment type: <from import="" td="" te<=""><td>mplate> ~</td><td></td><td></td></from>	mplate> ~		
	< Back	Next > Cance	el

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).

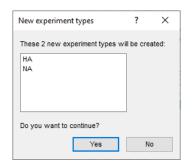


Figure 10: Missing experiments in the database.

Import sequence trace files		?	×
Database links Link the imported records to database ent Double click on a cell to get an overview.	ries.		
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: Report align inconsistencies Max. number of inconsistencies to report: Open assembly overview report Settings Assembly settings Trimming settings	20				
		< Back	Finish	Can	cel

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.

Assemb	bly trimming settings for 'HA			?	×
Minimu	m number of sequences contrib	uting to consensu	us: 1		
	Trim pattern	Tolerance	Offset Searc	h range	
Start:	GCAATA	0	-10	-	
Stop:	TTATCAA	0	-15	-	
			ОК	Car	ncel

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

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

Assemb	ly trimming settings for 'NA'			?	×
Minimun	n number of sequences contributing	g to consensi	is: 1		
	Trim pattern	Tolerance	Offset Search r	ange	
Start:	CCAGTAG	0	0]-	
Stop:	CGGGGTC	0	-20	- [
			ОК	Can	icel

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).

ssembly trimmin	g settings		? >
Experiment	Start pattern	Stop pattern	
HA	GCAATA	TTATCAA	Edit
NA	CCAGTAG	CGGGGTC	Edit

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).

ssembly settings	?	×
Experiment		
HA	Edit	
NA	Edit	
		se

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.

Assembly settings for 'HA'	?	×
Quality Assembly Consensus Copy settings		
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	%	
Reset to default OK		Cancel

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.

Ē	Assembly report: Ba	tch - 2020-04-1	6 14:08:15						-		×
ile	Overview Details	s Window H	lelp								
)ve	rview				D)etails					
Ħ	3					\approx					
	Key	HA	NA		C	ode	Message	Status	Comme	ent	
~	inflA001	ок	warning	^	in	fo	Created new assembly				
~	inflA002	error	ОК				Report for inflA001 / HA				
~	inflA003	error	warning								
~	inflA004	error	warning								
~	inflA005	error	warning								
~	inflA006	error	warning								
~	in flA007	error	N/A								
~	inflA008	error	warning								
~	in flA009	error	warning								
~	inflA010	error	warning								
				~			<				

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).

ile	Overview Detai	ls Window H	lelp					
ve	rview		а 		Details	;		
Ð	3				\approx			
	Кеу	HA	NA	-	Code	Message	Status	Co
~	inflA001	ок	warning	^	error	Unresolved base in consensus at position 434	new	
~	in flA002	error	ок		error	Unresolved base in consensus at position 339	new	
~	in flA003	error	warning		error	Unresolved base in consensus at position 302	new	
~	in flA004	error	warning		error	Unresolved base in consensus at position 258	new	
~	in flA005	error	warning		error	Unresolved base in consensus at position 220	new	
~	in flA006	error	warning		info	Created new assembly		
~	in flA007	error	N/A			Report for inflA002 / HA		
~	in flA008	error	warning					
~	in flA009	error	warning					
~	in flA010	error	warning					
				~		<		

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.

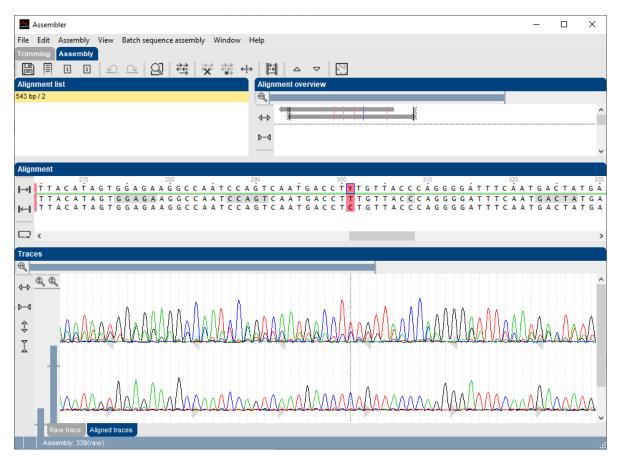


Figure 20: The *Contig assembly* window as called from the detailed report by double-clicking an error message. The window shows the contig project with the unresolved base in focus.

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).

ile	Overview Details	Window H	elp					
)ve	rview				Details			
đ					\approx			
	Кеу	HA	NA	-	Code	Message	Status	Co
~	in flA001	ок	warning	^	error	Unresolved base in consensus at position 434	solved	
/	inflA002	solved	ОК		error	Unresolved base in consensus at position 339	solved	
/	inflA003	error	warning		error	Unresolved base in consensus at position 302	solved	
1	inflA004	error	warning		error	Unresolved base in consensus at position 258	solved	
1	inflA005	error	warning		error	Unresolved base in consensus at position 220	solved	
1	in flA006	error	warning		in fo	Created new assembly		
1	inflA007	error	N/A			Report for inflA002 / HA		
1	inflA008	error	warning					
/	in flA009	error	warning					
/	in flA010	error	warning					
				~		<		

Figure 21: Solved status.

6. Close the Batch sequence assembly report window.

Open Assembler 6

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).

Simport and assemble in batch - BioNumerics			- 🗆 X
File Edit Database Analysis Scripts Window Help			
☞ 코 ▦ ☞ ♥ ኴ			
Experiment types	Database entries		Comparisons
🖓 🕂 💾 🛞 🕄 I 🔂 🧟 🕇 🕹 «All Experiment type	\$ + ₺ ⊗ ₪ ₪ ~	<all entries=""> し</all>	+ 🐴 🛞 🗟 🛍 🖳 <all comparisons=""></all>
# Name Type 🔻	Key Level	Modified date 🛛 🔽 1	Name Modified date Level 🔫
1 HA Sequence types	inflA001	2020-04-16 14:08:15 • •	^
C 2 NA Sequence types	inflA002	2020-04-16 14:08:15 • •	~
v	inflA003	2020-04-16 14:08:15 • •	< >
Entry fields Database design	inflA004	2020-04-16 14:08:15 • •	Identification projects Decision networks
	inflA005	2020-04-16 14:08:15 • •	
+ 💾 ⊗ 🗟 🔂 🤄 ↑ ↓ «All Entry fields»	inflA006	2020-04-16 14:08:15 • •	2월 🕂 💾 🛇 🕄 🔂 🗸. <all identific<="" td=""></all>
Name Field type	 ✓ inflA007 ✓ inflA008 	2020-04-16 14:08:15 • 2020-04-16 14:08:15 •	Name Modified date 🗸
	✓ inflA008 ✓ inflA009	2020-04-16 14:08:15 • • 2020-04-16 14:08:15 • •	
	InflA010	2020-04-16 14:08:15	
v	MINAUTO	2020-04-16 14:00:15	~
Fingerprint files Power assemblies Annotations			Alignments BLAST projects Chromosome comparisons
P + P ⊗ B F マ. <all files="" fingerprint=""></all>			+ 🐴 🛞 🗟 🕴 🛱 🗸 - <all alignments=""></all>
File name Experiment type Link Modifier			Name Modified date
			Marine Modified date
< · · · · · · · · · · · · · · · · · · ·	4	>	
	· ·	,	
Database: Import and assemble in batch (_DefaultUser_) Entries: Loaded=10, View=10, Selected=	10 2 experiments C:\Users\Public\Documents\BioNume	erics\Data BN8\Import and assemble in batch This is a time limi	ted package valid until 2020-12-30

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).

11

inflA001 (Sequence Viewer)		×
		^
ile Sequence Header Annotation View Tools Window Help		
┣║図 Ē ⊻♀♀∦ŸŸŸ⊗ Q ▷ œ <□ <2 >> ⊗		
Sequence Editor		
gcttcttctt gcaatagtca gtcttgttaa aagtgatcag atttgcattg gttaccatg	: 60	^
aaacaactcg acagagcagg ttgacacaat aatggaaaaa aacgtcactg ttacacacg	: 120	
ccaagacata ctggaaaaga cacacaacgg gaaactctgc gatctagatg gagtgaagc	: 180	
tctaatttta agagattgta gtgtagctgg atggctcctc gggaacccaa tgtgtgacga	a 240	
atteeteaat gtgeeggaat ggtettaeat agtggagaag ateaateeag eeaatgaee	: 300	
ctgttaccca gggaatttca acgactatga agaactgaaa cacctattga gcagaataaa		
ccattttgag aaaattcaga tcatccccaa aagttcttgg tcagatcatg aagcctcat	420	~
		>
Annotation		
Feature list Q ± ⊗ ± Feature key Start End ▼		
Annotation Header Custom Fields Sequence Search Contigs Frame Analysis Restriction Analysis		
Sequence: inflA001 Experiment: HA 1 504 bp		

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.