BIONUMERICS Tutorial:

Phylogenomics with RAxML and FastTree

1 Introduction

RAxML [2] and FastTree [1] are two open source maximum likelihood clustering methods which can be launched on aligned sequence data in BIONUMERICS. Aligned sequence data includes:

- Nucleic acid or amino acid sequences in the *Comparison* window on which a multiple alignment was calculated (see the ”Cluster analysis of sequences” tutorial).
- A SNP matrix generated by a wgSNP analysis (see the ”wgSNP with mapping performed locally on your own computer” and ”wgSNP with mapping performed on the cloud calculation engine” tutorials), stored as an aspect of a reference mapped sequence type.
- A SNP matrix generated by the CFSAN SNP pipeline (see the ”CFSAN SNP pipeline” tutorial), stored as an aspect of a sequence read set.

Both algorithms are available through the *WGS tools plugin* and can be launched on your own computer as well as on the Cloud Calculation Engine.

This tutorial will demonstrate a RAxML and FastTree analysis on a multiple alignment of concatenated MLST loci.

2 Preparing the database

The RAxML and FastTree analyses can only be performed in BIONUMERICS after installation of the *WGS tools plugin* in the BIONUMERICS database (***File*** > ***Install / remove plugins...*** ( )).

In this tutorial the **WGS demo database for the *Burkholderia cepacia* complex** will be used in which the *WGS tools plugin* is already installed. The Calculation engine option requires credits for running jobs on the Applied Maths cloud calculation engine. Credits are linked to credentials that you need to enter when installing the *WGS tools plugin*. No credits are assigned to the demo project so no RAxML or FastTree jobs can be launched on the external calculation engine. Please contact Applied Maths to obtain more information.

The demo database can be downloaded directly from the *BIONUMERICS Startup* window (see 2.1), or restored from the back-up file available on our website (see 2.2). The database contains the NCBI reference genome sequences for the *Burkholderia cepacia* complex and the genome sequences of the *Burkholderia cepacia* complex type strains if available at the time of database creation.

2.1 Option 1: Download demo database from the Startup Screen

1. Click the  button, located in the toolbar in the *BIONUMERICS Startup* window.

This calls the *Tutorial databases* window (see Figure 1).

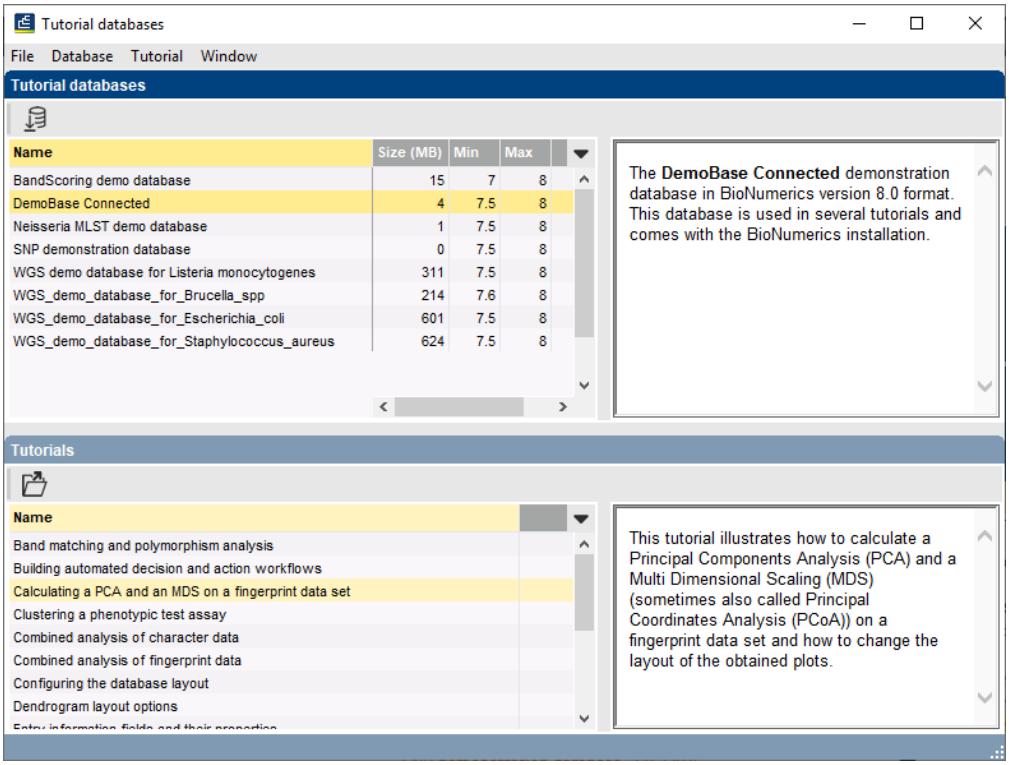

Figure 1: The *Tutorial databases* window, used to download the demonstration database.

2. Select **WGS_demo_database_for_Bcc** from the list and select ***Database*** > ***Download*** ( ).

3. Confirm the installation of the database and press <***OK***> after successful installation of the database.

4. Close the *Tutorial databases* window with ***File*** > ***Exit***.

The **WGS_demo_database_for_Bcc** appears in the *BIONUMERICS Startup* window.

5. Double-click the **WGS_demo_database_for_Bcc** in the *BIONUMERICS Startup* window to open the database.

2.2 Option 2: Restore demo database from back-up file

A BIONUMERICS back-up file of the demo database for the *Burkholderia cepacia* complex is also available on our website. This backup can be restored to a functional database in BIONUMERICS.

6. Download the file `WGS_Bcc.bnbk` file from https://www.applied-maths.com/download/sample-data, under 'WGS_demo_database_for_Bcc'.

In contrast to other browsers, some versions of Internet Explorer rename the `WGS_Bcc.bnbk` database backup file into `WGS_Bcc.zip`. If this happens, you should manually remove the `.zip` file extension and replace with `.bnbk`. A warning will appear ("If you change a file name extension, the file might become unusable."), but you can safely confirm this action. Keep in mind that Windows might not display the `.zip` file extension if the option "Hide extensions for known file types" is checked in your Windows folder options.

7. In the *BIONUMERICS Startup* window, press the button. From the menu that appears, select ***Restore database...***.
8. Browse for the downloaded file and select ***Create copy***. Note that, if ***Overwrite*** is selected, an existing database will be overwritten.
9. Specify a new name for this demonstration database, e.g. "WGS_Bcc_demobase".
10. Click <***OK***> to start restoring the database from the backup file.
11. Once the process is complete, click <***Yes***> to open the database.

The *Main* window is displayed (see Figure 2).

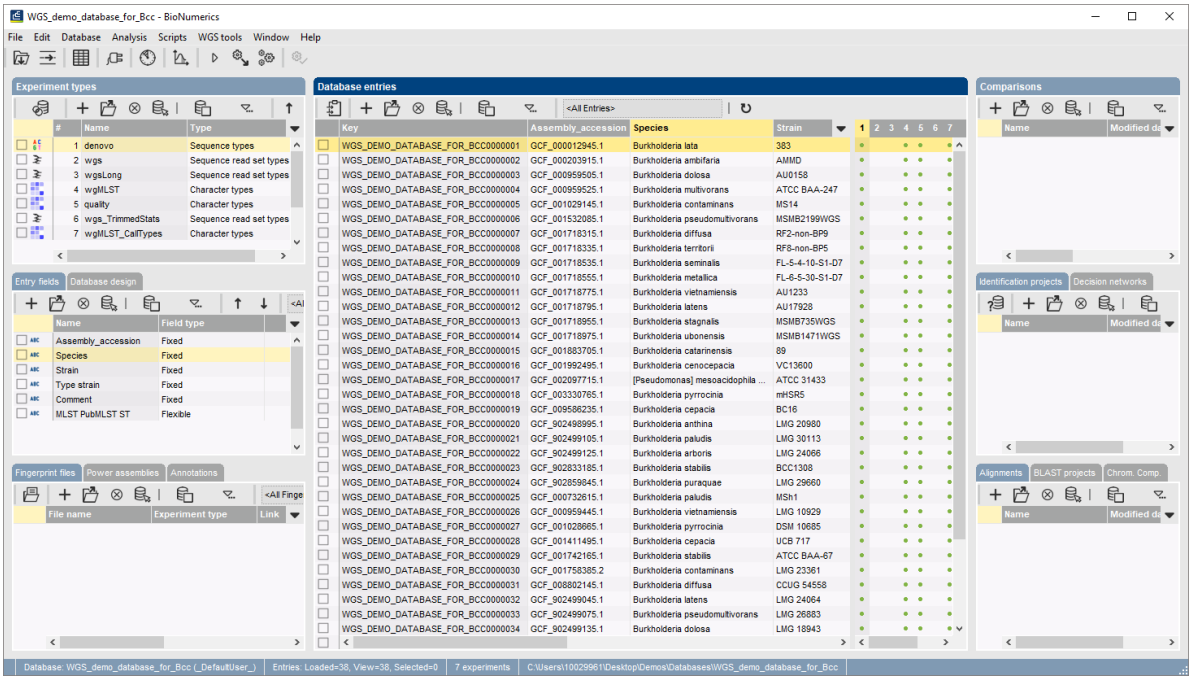

Figure 2: The *Burkholderia cepacia* complex demonstration database: the *Main* window.

3 Generate aligned sequence data

3.1 Sequence extraction

An assembly-based wgMLST analysis has already been performed on the 38 genome sequences in the database (for more information on wgMLST analysis on imported genomes in BIONUMERICS see the "wgMLST typing: routine workflow starting from imported genomes" tutorial). To

demonstrate the RAxML and FastTree maximum likelihood analyses we will extract the allele sequences of the 7 MLST loci from the 38 genome sequences and concatenate the obtained sequences.

1. Select the *Database entries* panel to make it the active panel and select all entries in the database with ***Edit*** > ***Select all*** (**Ctrl+A**).

2. Extract the allele sequences of the 7 MLST loci from the 38 genome sequences by selecting ***WGS tools*** > ***Store wgMLST locus sequences...***.

The *Store sequences* dialog box will pop up.

3. Select the **MLST PubMLST (7 loci)** subscheme and select **LocusTag** as sequence experiment type identifier (see Figure 3).

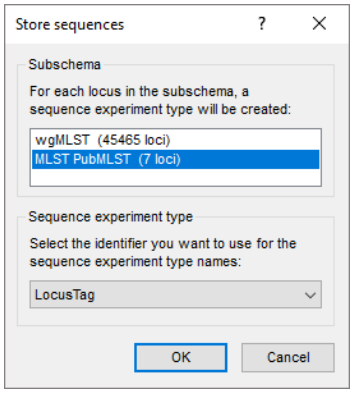

Figure 3: The *Store sequences* dialog box.

4. Press <***OK***> and <***Yes***> to confirm the creation of the 7 new sequence experiment types (see Figure 4).

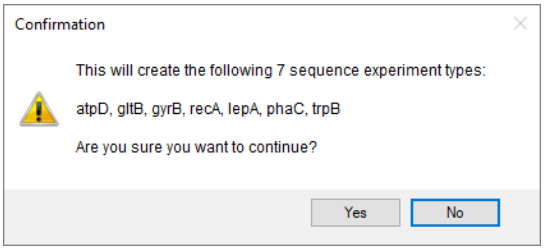

Figure 4: Confirmation dialog box.

Seven new sequence experiment types will be created and listed in the *Experiment types* panel. The 7 MLST loci sequences will be extracted from the 38 genome sequences and stored in the respective sequence experiment types. Each of these sequences can be viewed in the *Sequence editor* window by clicking on the respective green dot in the *Experiment presence* panel.

3.2 Concatenate sequences

As we want to include the sequence information from the seven loci in our phylogenetic analyses, we will first concatenate the 7 MLST loci sequences and store the concatenated sequences in a new sequence experiment type.

5. To create a new sequence experiment type, click on the *Experiment types* panel to activate it and select ***Edit*** > ***Create new object...*** (+).

6. From the *Create a new experiment type* dialog box that pops up, select ***Sequence type*** to start the *New sequence type* wizard.

7. The first page prompts you to enter a name for the new sequence experiment type. Enter a name (for example “Concatenated loci” and press <***Next***>.

8. In the second page of the *New sequence type* wizard leave the check box for ***Nucleic acid sequences*** selected and press the <***Finish***> button to complete the setup of the new sequence experiment type.

The sequence experiment type will be listed in the *Experiment types* panel of the *Main* window (see Figure 5).

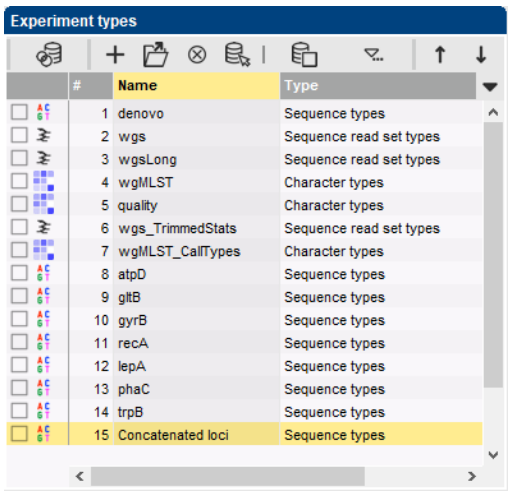

Figure 5: The *Experiment types* panel.

To concatenate the MLST loci sequences and store the concatenated sequences in the newly created sequence experiment type an online script can be used.

9. Select ***Scripts*** > ***Browse internet...*** to open the *Browse for scripts* window.

10. Select **Concatenate sequences** under **Sequence related tools**.

11. In the dialog box that pops up, select the 7 sequence experiment types containing the 7 MLST loci sequences as sequence experiment types which should be merged and select the newly created sequence experiment type as the merged sequence type (see Figure 6). Press <***OK***>.

The sequences of the 7 loci will be concatenated and stored in the merged sequence experiment type for the 38 selected entries. The *Main* window will now look like Figure 7.

3.3 Calculate a multiple alignment

A multiple sequence alignment of the concatenated sequences can be calculated in the *Comparison* window.

12. In the *Database entries* panel of the *Main* window make sure all entries are selected.

13. Click on ***Edit*** > ***Create new object...*** (+) in the *Comparisons* panel to open the *Comparison* window for the 38 selected entries.

14. In the *Experiments* panel highlight the merged sequence experiment which contains the concatenated sequences of the entries in the comparison (i.e. ”Concatenated loci”) and select ***Layout*** > ***Show image*** ( ) to visualise the concatenated sequences in the *Experiment data* panel.

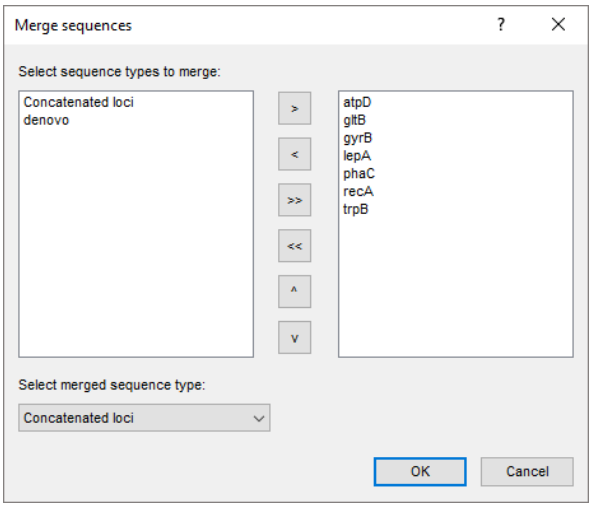

Figure 6: The ***Merge sequences*** dialog box.

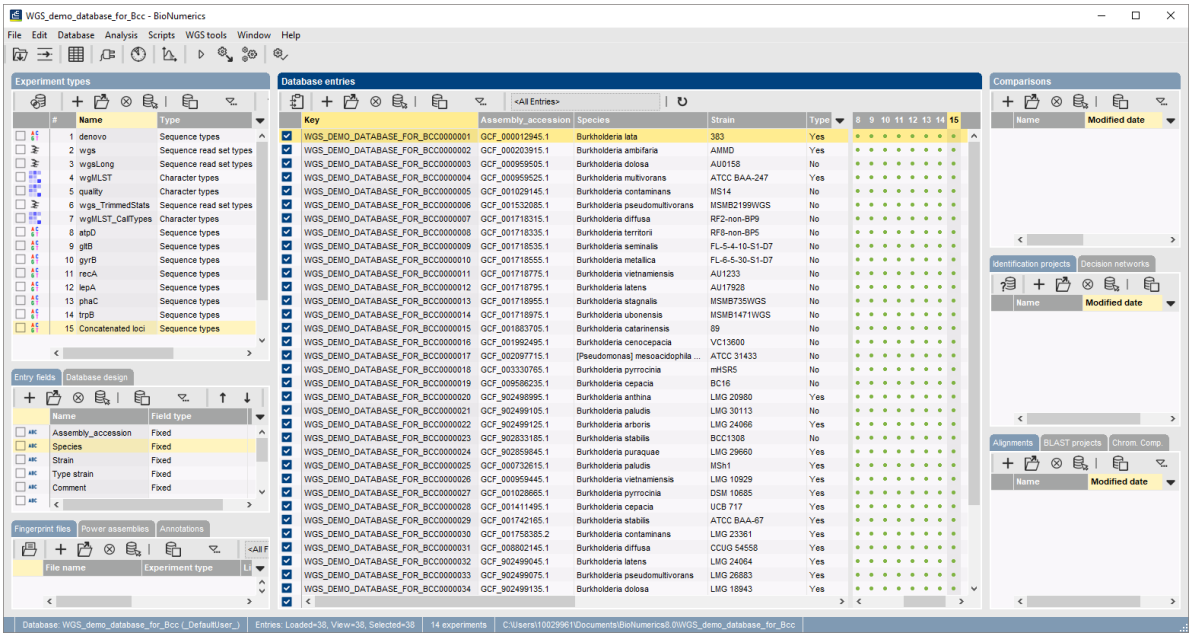

Figure 7: The *Main* window.

15. Select ***Sequence*** > ***Multiple alignment...*** ( ) to open the *Multiple alignment* dialog box (see Figure 8).

16. Leave the settings as default and press <***OK***>.

A multiple alignment is calculated. The *Comparison* window now looks like Figure 9.

For more information on multiple sequence alignment in BIONUMERICS see the "Cluster analysis of sequences" tutorial. Note that for MLST loci a good multiple alignment can be obtained on the concatenated sequences. However, when non-MLST loci are considered it is good practice to concatenate the multiple sequence alignments of the loci instead to reduce the amount of misaligned positions.

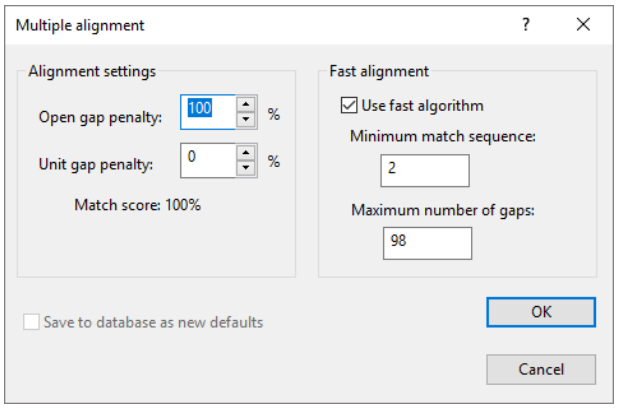

Figure 8: The *Multiple alignment* dialog box.

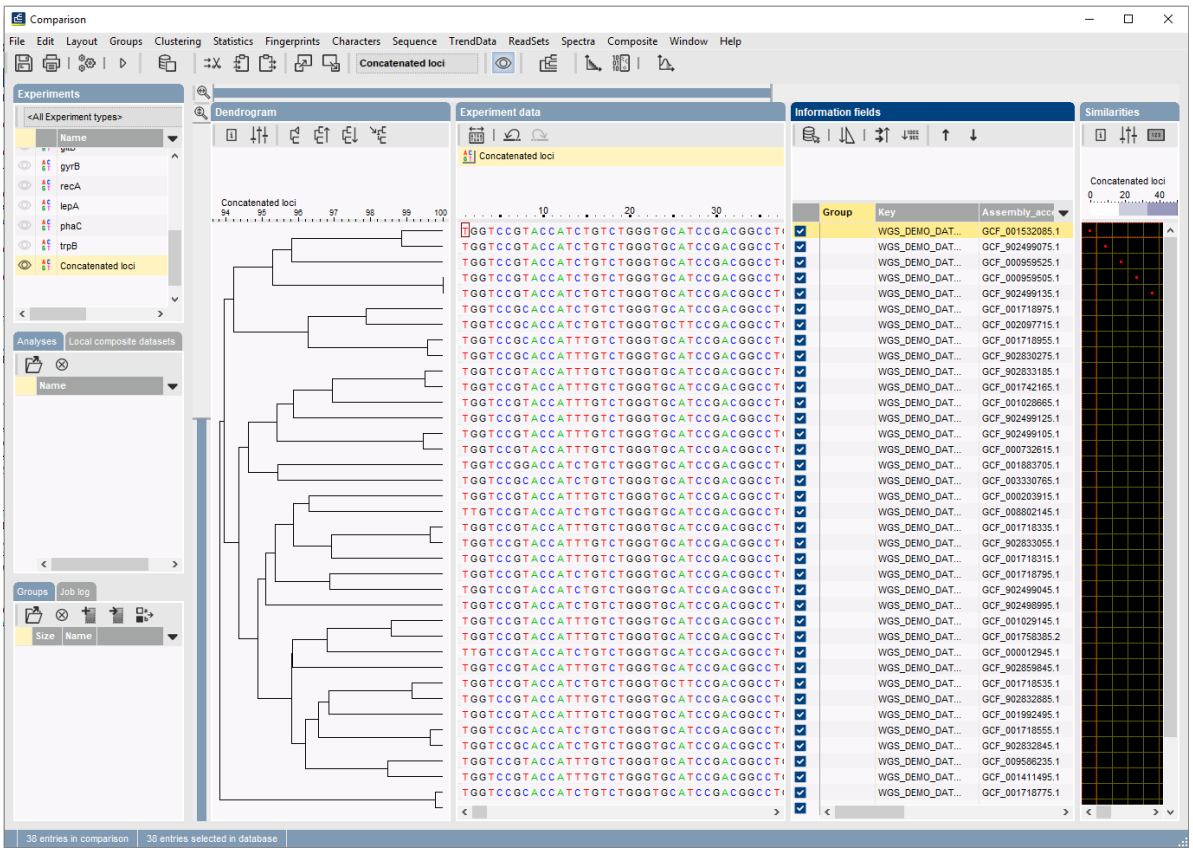

Figure 9: The *Comparison* window.

4 Launch a maximum likelihood analysis

A maximum likelihood analysis with RAxML and FastTree can be launched on your own computer or on the calculation engine for the entries in the created comparison. Following the same principles as cluster analyses in comparisons, comparison jobs apply on the whole comparison and the data stored in the *active* experiment and (where applicable) the active aspect in the *Experiments* panel.

1. Save the comparison by selecting ***File*** > ***Save*** (, **Ctrl+S**). Specify a name for the comparison for example ”ML analyses for Bcc strains” and press <***OK***>.

2. To launch a maximum likelihood comparison job, select ***File*** > ***Launch comparison jobs...*** ( ) from the menu in the *Comparison* window.

In case the comparison was not saved to the database yet, you will be prompted to save first. Comparisons should be saved before any jobs can be launched.

Subsequently, the *Submit comparison jobs* dialog box will open (see Figure 10).

If the active experiment does not support any comparison jobs, an error message is shown and the *Submit comparison jobs* dialog box will not appear.

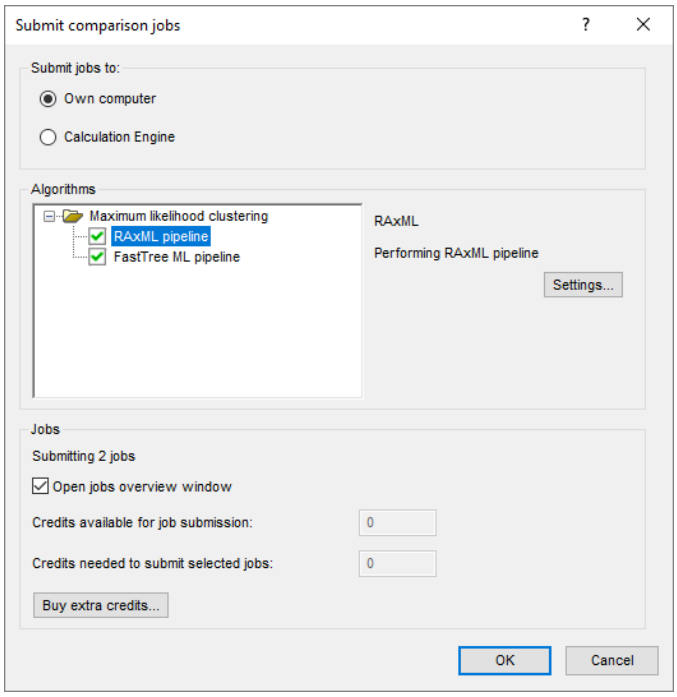

Figure 10: The *Submit comparison jobs* dialog box.

The RAxML and FastTree algorithms are available on the local calculation engine and on the Cloud Calculation Engine. In this tutorial the jobs will be launched on the local calculation engine (i.e. will be launched on your own computer).

3. In the ***Submit jobs to*** panel select ***Own computer*** and in the ***Algorithms*** panel select both the ***RAxML*** and ***FastTree*** algorithm (see Figure 10).

4. With the ***RAxML*** algorithm highlighted, press the <***Settings...***> to open the *RAxML pipeline settings* dialog box in which the evolutionary model and other settings for the RAxML job can be defined (see Figure 11).

In the ***Evolutionary models*** panel an evolutionary model for maximum likelihood analysis can be selected. The available models in the drop-down list next to the ***Model*** option consist of a combination of a basic evolutionary model (e.g. GTR, DAYHOFF,. . .) with a model for rate heterogeneity among sites (+GAMMA or +CAT) and an allowance for the presence of invariant sites (+I).

The GTR model is the most common substitution model for nucleotide sequence analysis and the only model available through RAxML. For amino acid sequence analysis twelve general amino acid substitution models can be selected, i.e. BLOSUM62, DAYHOFF, DCMUT, GTR, JTT, JTTDCMUT, LG, LG4M, LG4X, PMB, VT and WAG.

The GAMMA and CAT models account for variable rates of evolution across sites. The GAMMA model assumes a gamma distribution of rates across sites with four discrete rate categories. The CAT model optimizes the individual per-site substitution rates and classifies these individual rates

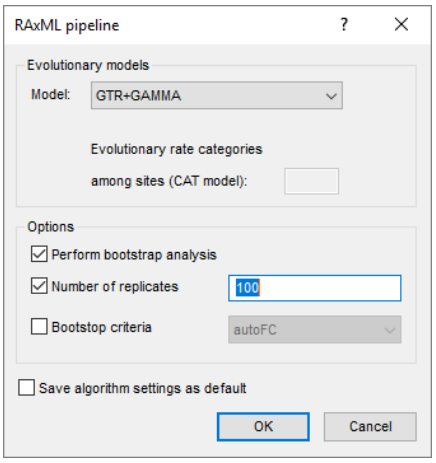

Figure 11: The *RAxML pipeline settings* dialog box.

into the number of rate categories specified by the ***Evolutionary rate categories among sites*** option. The default number of rate categories is set to 25. Note that this option is specific for the CAT model and is therefore not available when the GAMMA model has been chosen from the drop-down list.

The ***Options*** panel allows you to set the criteria for bootstrap analysis. When the ***Perform bootstrap analysis*** option is checked, bootstrap analysis is enabled and two options for bootstrap analysis become available:

- ***Number of replicates***: This option allows the user to set the number of bootstrap replicates. Default the number of bootstrap replicates is set to 100.
- ***Bootstop criteria***: When this option is checked the optimal number of bootstrap replicates to obtain stable support values will be determined automatically. In the drop-down list a RAxML bootstop criterion (i.e. threshold to decide when enough replicates have been computed) can be selected to determine convergence of the bootstrap estimates.

The bootstrap support values are drawn on the best-scoring maximum likelihood tree.

When altering the RAxML settings, one can save the updated settings as defaults to the database with ***Save algorithm settings as default***.

5. Leave the settings as default and press <***OK***>.
6. In the *Submit comparison jobs* dialog box highlight the ***FastTree*** algorithm and press the <***Settings...***> to open the *FastTree ML pipeline settings* dialog box in which the evolutionary model and other settings for the FastTree job can be defined (see Figure 12).

In the ***Evolutionary models*** panel an evolutionary model for maximum likelihood analysis can be selected. Two models are available in the drop-down list for nucleotide sequence analysis (i.e. ***Generalized time-reversible*** and ***Jukes-Cantor***), while three models are available for amino acid analysis (i.e. ***Jones-Taylor-Thorton***, ***Le-Gascuel (2008)*** and ***Whelan-And-Goldman (2001)***).

In the ***Options*** panel four additional options for FastTree analysis can be enabled:

- ***Speed up***: This option can be checked to speed up the neighbor joining phase and to reduce memory usage.

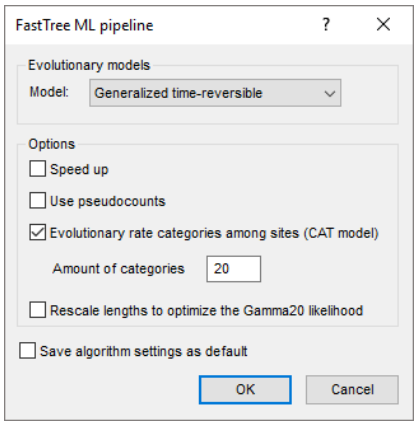

Figure 12: The *FastTree ML pipeline settings* dialog box.

- ***Use pseudocounts***: This option is recommended if the alignment has sequences with little or no overlap. A pseudocount (weight of 1.0) will be used to estimate the distances between these sequences.
- ***Evolutionary rate categories among sites (CAT model)***: If this option is checked the evolutionary model will be run under the CAT model for rate heterogeneity among sites. The CAT model optimizes the individual per-site substitution rates and classifies these individual rates into the number of rate categories specified by ***Amount of categories***. The default amount of categories is set to 20.
- ***Rescale lengths to optimize the Gamma20 likelihood***: After the final round of optimizing branch lengths with the CAT model, report the likelihood under the discrete gamma model with the same number of categories. FastTree uses the same branch lengths but optimizes the gamma shape parameter and the scale of the lengths. The final tree will have rescaled lengths.

Local support values computed with the Shimodaira-Hasegawa test are drawn on the resulting maximum likelihood tree and provide an estimate of the reliability of each split in the tree. The FastTree support values range from 0 to 1 but are multiplied by 100 in BIONUMERICS.

When altering the FastTree settings, one can save the updated settings as defaults to the database with ***Save algorithm settings as default***.

7. Leave the settings as default and press <***OK***>.

8. Press <***OK***> in the *Submit comparison jobs* dialog box to launch the RAxML and FastTree jobs on your computer.

The jobs are submitted to your computer and the *Job overview* window opens. In the *Job overview* window, the job type, job name, time of submission, job status, a description of the job, its progress and much more can be monitored.

5 Import and analyse the maximum likelihood job results

Once the jobs have been finished (see Figure 13), the results can be imported in the database by selecting ***Jobs*** > ***Get results*** ( ) from the *Job overview* window.

1. When the jobs are finished, select the jobs and select ***Jobs*** > ***Get results*** ( ) to import the results in the comparison window.

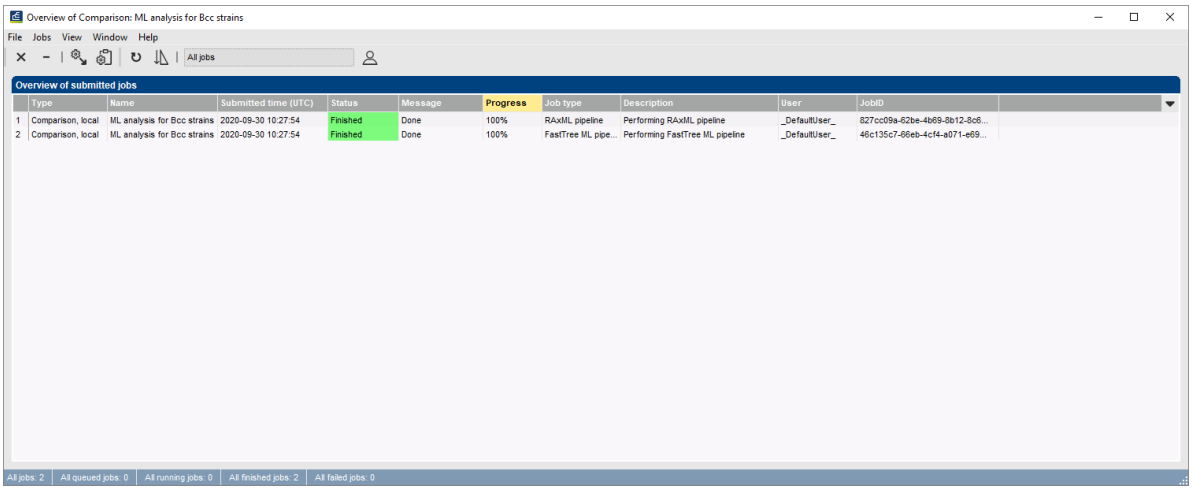

Figure 13: The *Job overview* window listing a finished RAxML and FastTree pipeline job.

2. Close the *Job overview* window.

The *Comparison* window now looks like Figure 14.

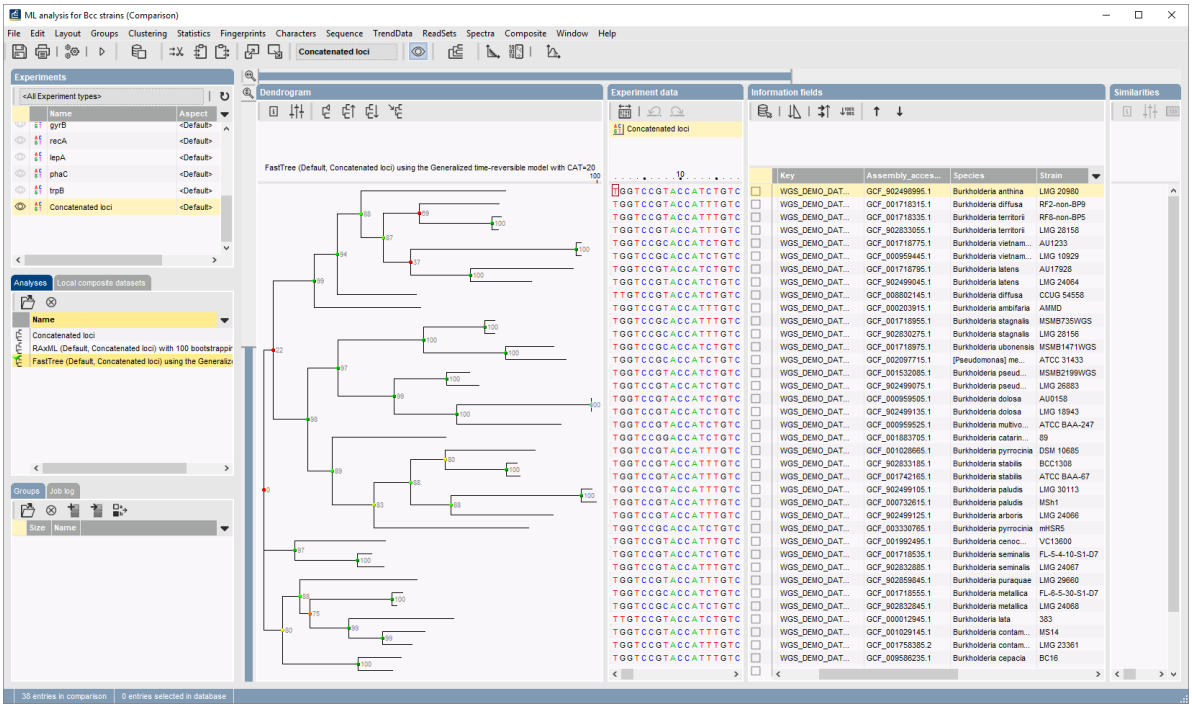

Figure 14: The *Comparison* window after import of the RAxML and FastTree pipeline job results.

The *Analyses* panel lists the performed maximum likelihood analyses. The name of the analysis includes the analysis type (RAxML or FastTree), the aspect and the experiment type on which the analysis was performed and the selected maximum likelihood settings. If the dendrogram is not automatically displayed in the *Dendrogram* panel, double-click on the analysis or use ***File*** > ***Analysis components*** > ***Open*** ( ) to display.

The toolbar of the *Dendrogram* panel (see Figure 15) provides different options to manipulate the visualised dendrogram (i.e. changing the display settings, rearranging the branches, collapsing/-expanding the branches and rerooting the tree).

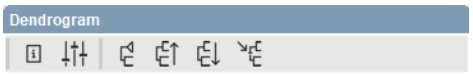

Figure 15: The toolbar of the *Dendrogram* panel.

The log file of the maximum likelihood pipeline jobs can be consulted in the Job log panel (see Figure 16).

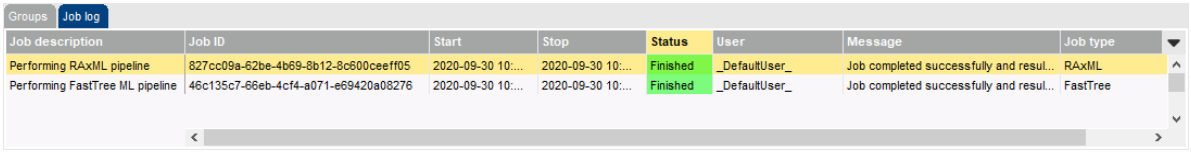

Figure 16: The Job log panel.

Bibliography

[1] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. Fasttree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, 5(3), 2010.

[2] Alexandros Stamatakis. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.