MLSTest Tutorial

Contents

About MLSTest

MLSTest is a novel Windows based software for multilocus sequence data analysis. It is aimed to work with diploid sequences but it is useful for haploid sequences too. It constructs allelic profiles and makes basic clustering analysis with them, calculates discriminatory power with confidence intervals, typing efficiency and genotypic diversity. Additionally, includes tools for view, edit, convert and concatenate sequences. It uses several simple methods for tree construction (UPGMA, Neighbor Joining, Bio-Neighbor Joining, majority rule consensus) with the advantage of manage heterozygous sites. It calculates node support using bootstrap and approximated clade significance based on the Templeton test that is faster than bootstrap. Additionally, the software analyzes whether concatenation is suitable for the data set using different tests (bionj-incongruence length difference test, templeton test). It evaluates how the incongruence is distributed across the tree using a variation of the localized incongruence length difference test implementing a modified neighbor joining algorithm. MLSTest have tools to optimize and reduce MLST schemes, to select best loci for typing and analyzes whether extending the scheme will increase resolution. It calculates consensus networks and makes multidimensional scaling for visual clustering. Additionally, MLSTest looks for network structures in BURST graphs that which are evidence of recombination. Finally, it implements a user-friendly viewer for trees, BURST and multidimensional scaling graphs that allows several editing and exporting options.

Installing MLSTest

MLSTest requires a computer with Windows xp Service Pack 2 or better and Microsoft .net framework 3.5 or better.

You can download the installer from the next link:

v. 1.0.0.39:

http://mlstest.codeplex.com/downloads/get/clickOnce/MLSTest1.0.application

HOW TO INSTALL

1) Download the installer and click on the downloaded file. Follow the instructions. Installation is very easy. A warning window could be appearing saying that the publisher is unknown. Do not worry, this is because we can´t sign the app. You can uninstall the app using the control panel.
2) MLSTest will look automatically for actualizations and inform to you if you want to actualize. We will use this method to fix bugs or to include some new features.

Loading Data

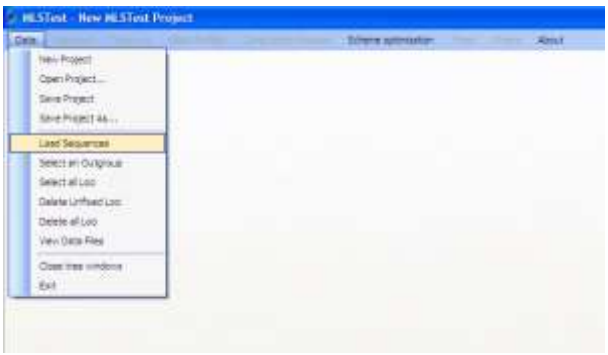
MLSTest just allows (at least for now) sequences in FASTA format:

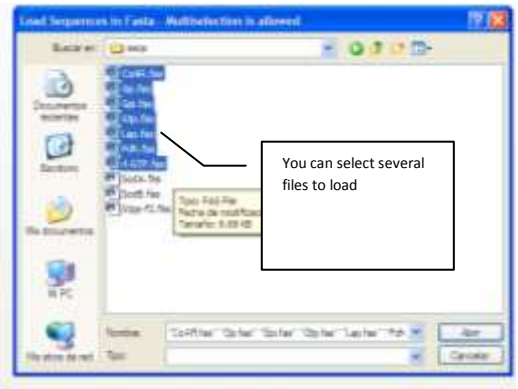
Example of the format:

>sequence1
ATATTTTTTTTTT
>sequence2
ATTATRRTTTTTT
>sequence3
AGGGGGGGGG

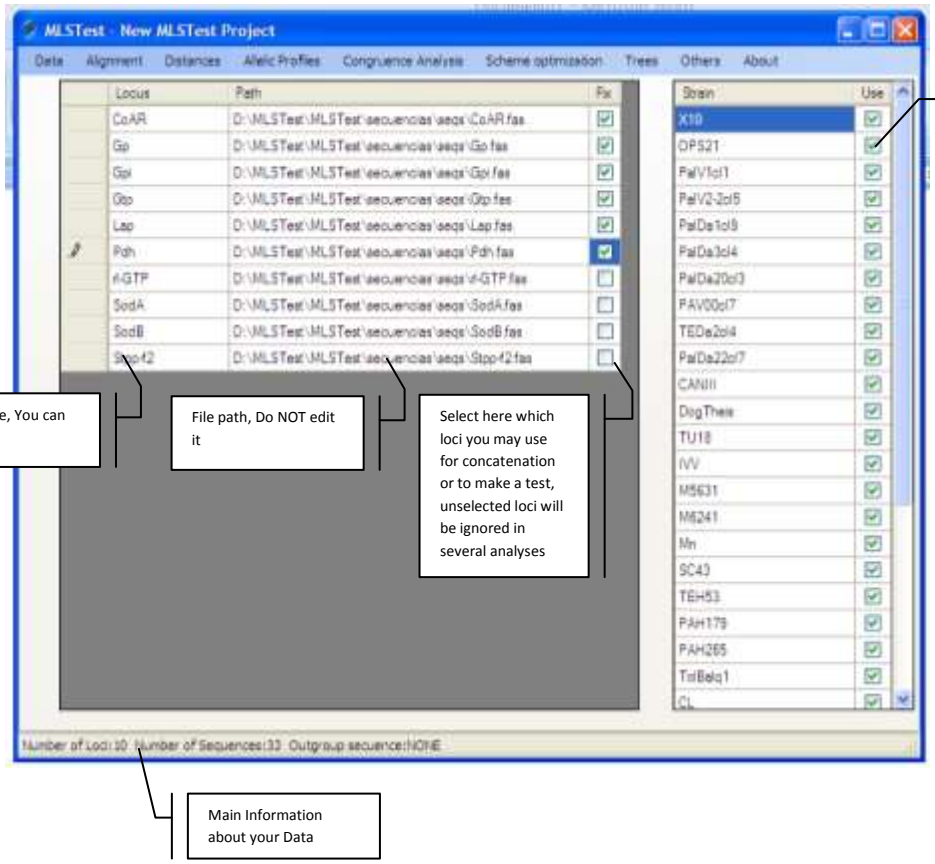
IMPORTANT!!!! Sequences must be in the same order in each alignment loaded. MLSTest ignores names and consider that two sequences in the same position in different alignments correspond to the same organism.

First step: Load sequences





Main window



Here are showed strain names in your dataset. Select which strain are included in your analysis, unselected strains will be ignored

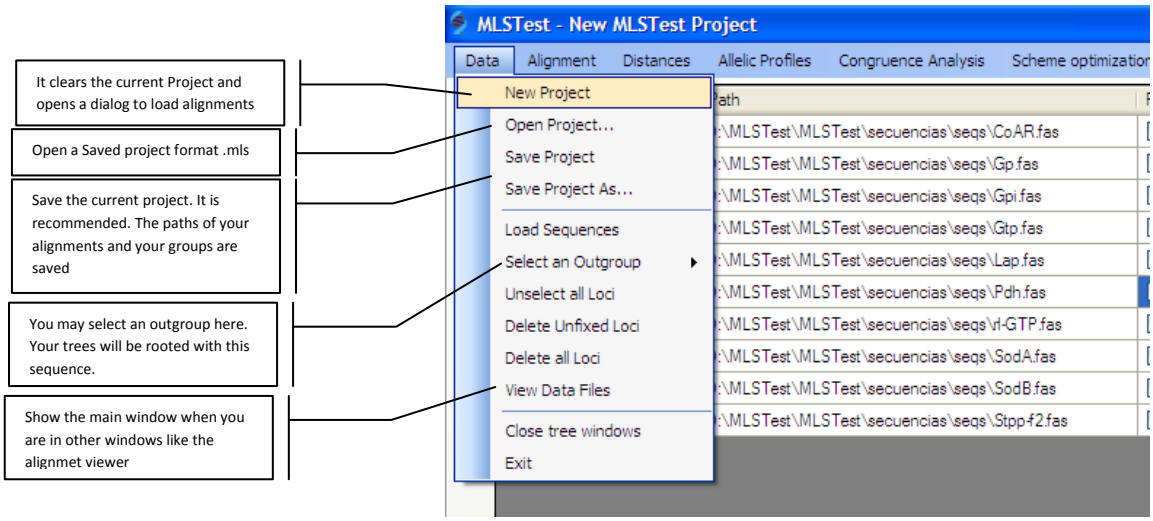
Locus Name, You can edit it

File path, Do NOT edit it

Select here which loci you may use for concatenation or to make a test, unselected loci will be ignored in several analyses

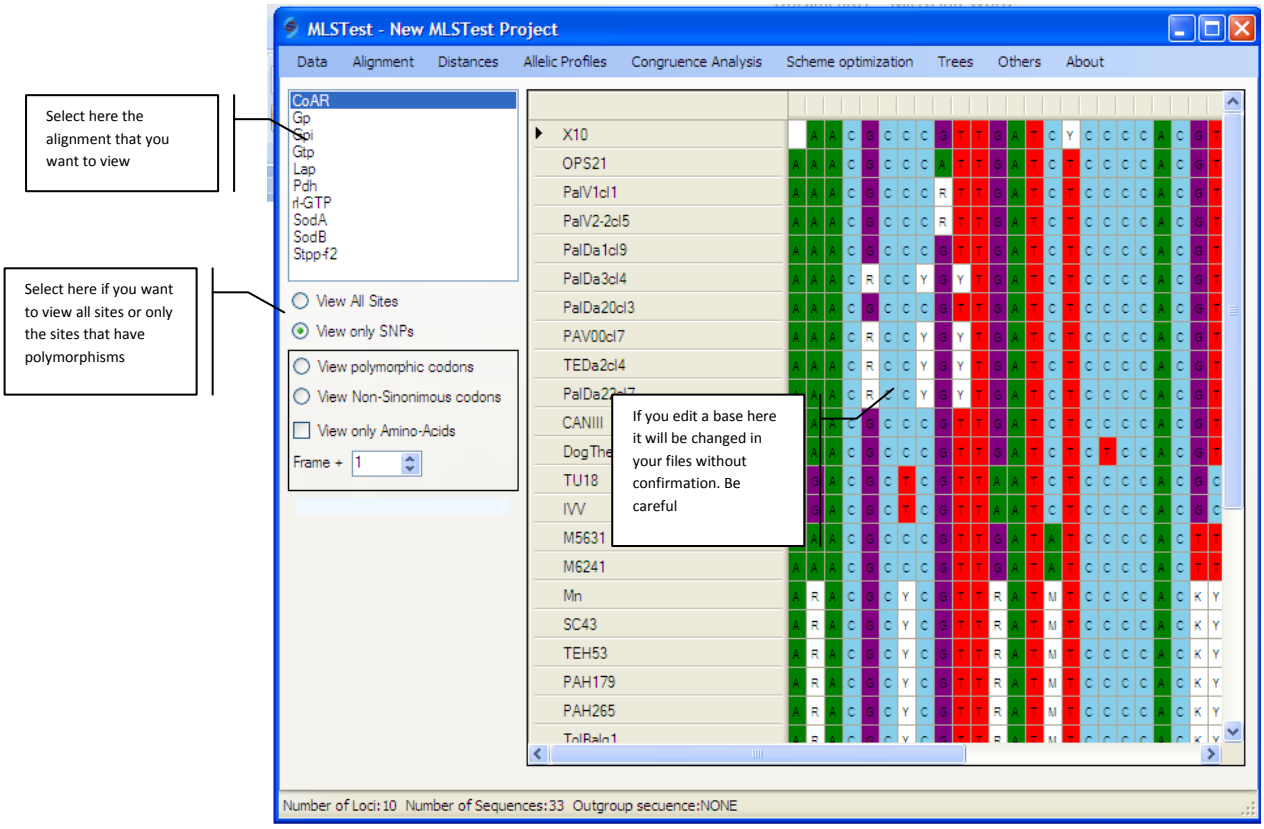
Main Information about your Data

DATA Menu

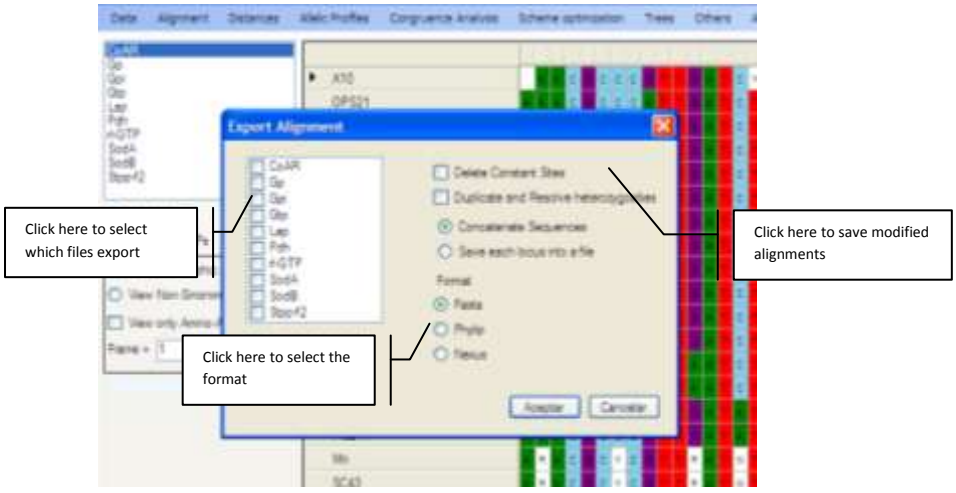

It clears the current Project and opens a dialog to load alignments
Open a Saved project format .mls
Save the current project. It is recommended. The paths of your alignments and your groups are saved
You may select an outgroup here. Your trees will be rooted with this sequence.
Show the main window when you are in other windows like the alignmet viewer
MLSTest - New MLSTest Project
Data
Alignment
Distances
Allelic Profiles
Congruence Analysis
Scheme optimizatior
New Project
Open Project...
Save Project
Save Project As...
Load Sequences
Select an Outgroup
Unselect all Loci
Delete Unfixed Loci
Delete all Loci
View Data Files
Close tree windows
Exit
Path
:\MLSTest\MLSTest\secuencias\seqs\CoAR.fas
:\MLSTest\MLSTest\secuencias\seqs\Gp.fas
:\MLSTest\MLSTest\secuencias\seqs\Gpi.fas
:\MLSTest\MLSTest\secuencias\seqs\Gtp.fas
:\MLSTest\MLSTest\secuencias\seqs\Lap.fas
:\MLSTest\MLSTest\secuencias\seqs\Pdh.fas
:\MLSTest\MLSTest\secuencias\seqs\rl-GTP.fas
:\MLSTest\MLSTest\secuencias\seqs\SodA.fas
:\MLSTest\MLSTest\secuencias\seqs\SodB.fas
:\MLSTest\MLSTest\secuencias\seqs\Stpp-f2.fas

View, modify and export your alignments

Alignment>viewer



Alignment> export



Distances

The Distances menu allows you to view and export distance matrix for alignments or allelic profiles. The file format of the matrix is compatible with phylip package for further analysis.

Managing Diploid sequences

Additionally, the user can select in "Distance>Handling Heterozygosities" how heterozygous sites are considered in the analysis. Three options are available:

1. Average States: the distance between two sites that are different is calculated as mean of distances among all the possible resolution of the heterozygosis
   Example:

$$\text{distance C-Y} = \frac{\text{Distance C-C + Distance C-T}}{\text{Number comparisons}} = \frac{0+1}{2} = 0.5$$

   Seq1: AATTY
   Seq2: ACTTC observed distance=1 + 0.5= 1.5

2. SNP duplication and heterozygosis resolution (Tavanti et al, 2005): Constant sites are removed from the alignment. Then, polymorphic site are duplicated and heterozygosis are resolved.

   Example:

   Seq1: AATTY → AY → AAYY → AATC
   Seq2: ACTTC → AC → CCCC → CCCC observed distance=3

3. Ignore in pairwise comparisons: Heterozigosities are not considered in pairwise comparisons

   Seq1: AATTY → AATT
   Seq2: ACTTC → ACTT observed distance=1.0

Only uncorrected distances are used by the software. However, considering the close relationships between strains in MLST datasets, probably this is not a problem because superimposed mutations should be rare. Additionally, this is a model-free approach, simpler and faster than corrected distances.

Working with allelic profiles

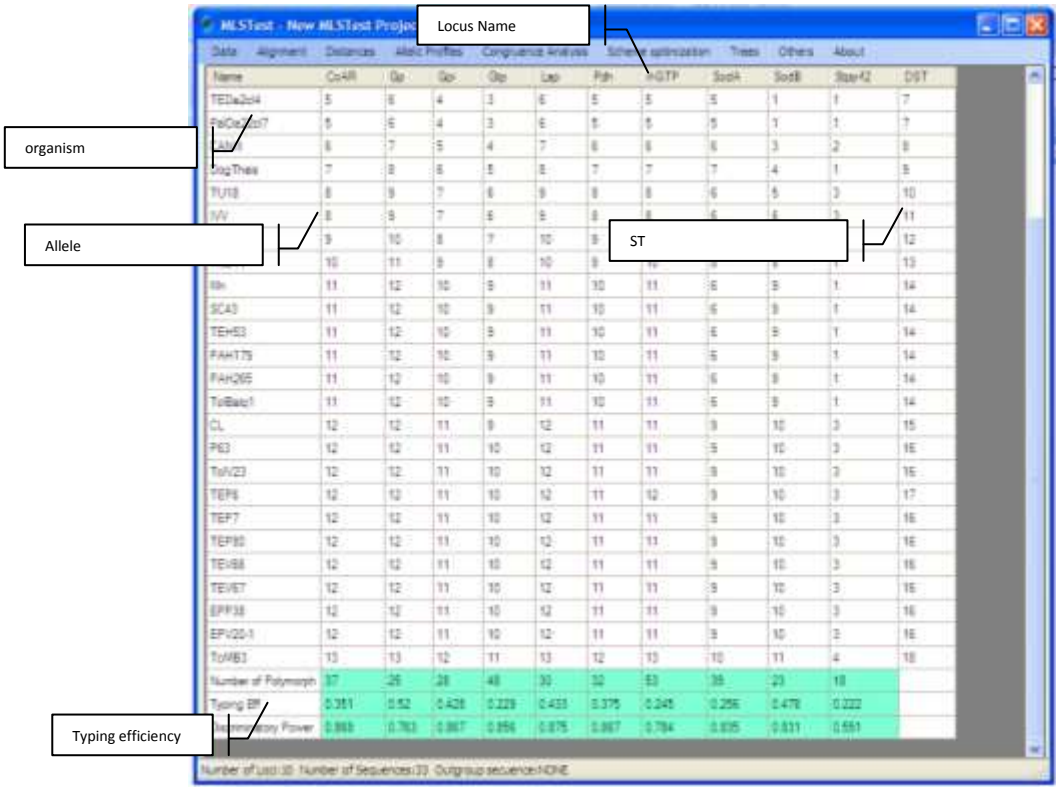
Allelic Profiles>Make/View allelic profiles

Allelic profiles are showed in a table. Additional information is included at the bottom of the table:

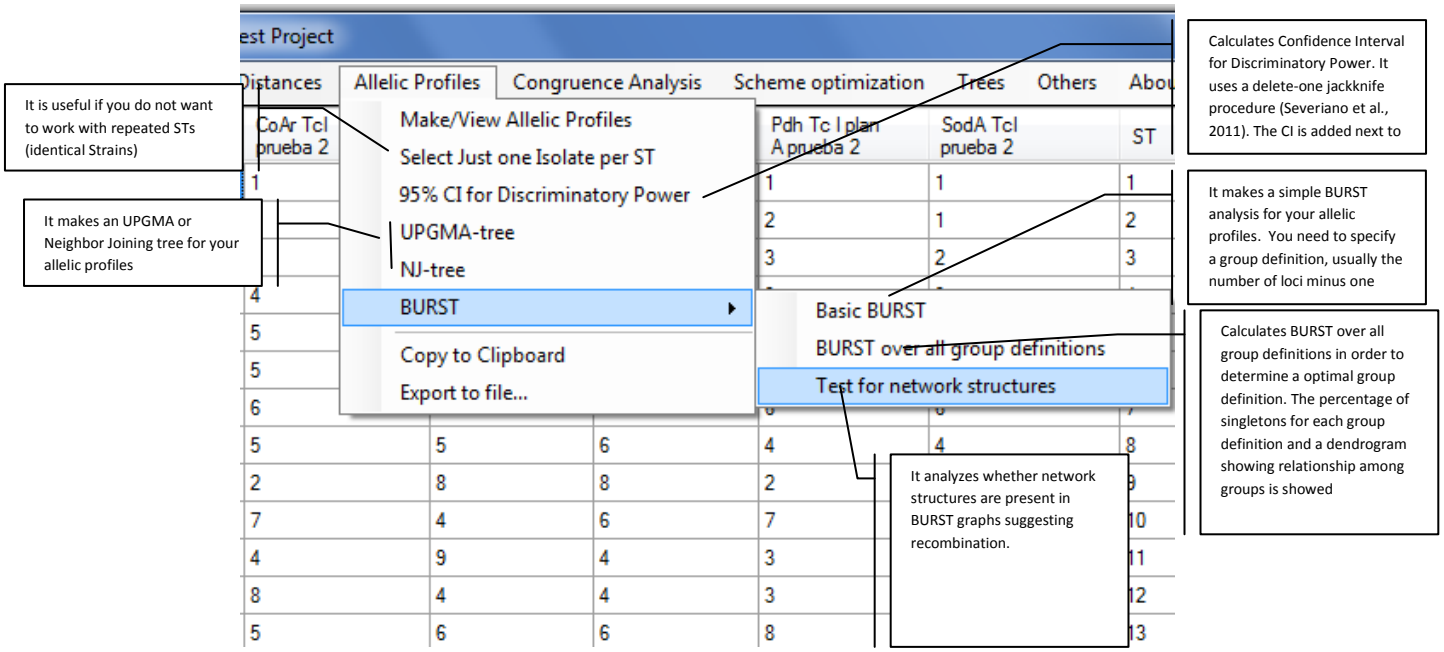
1. Number of Polymorphisms
2. Typing efficiency (TE). It is defined as the number of different genotypes (STs/DSTs) per polymorphic site.
3. Discriminatory Power: defined as the probability that two strains chosen at random from a population of unrelated strains will be distinguished by the typing method concerned (Hunter, 1990) and was calculated according to the following equation:

$$D = 1 - \frac{1}{N(N-1)} \sum_{j=1}^{s} x_j(x_j - 1)$$

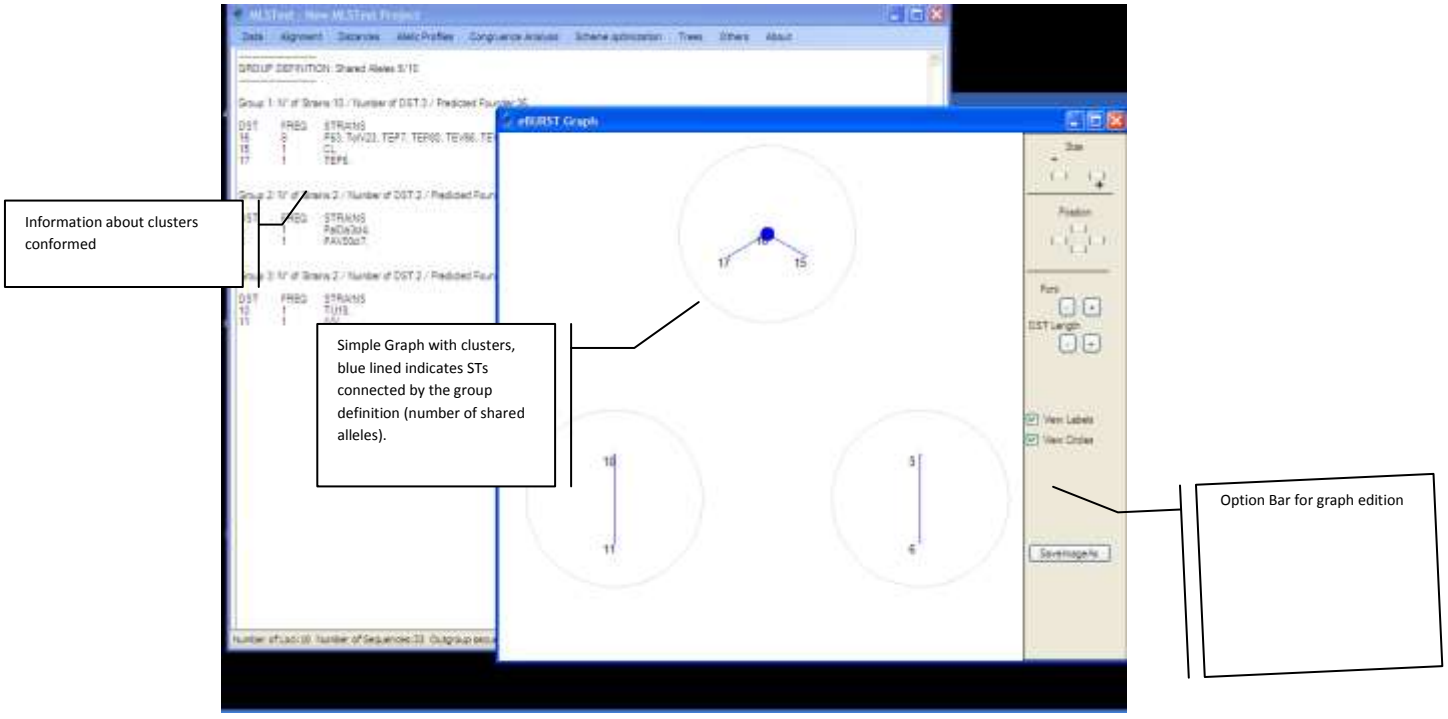
where (s) is the number of types, (xj) is the number of population members falling into the (j-th) type, and (N) is the size of the population.



Analysis for Allelic Profiles

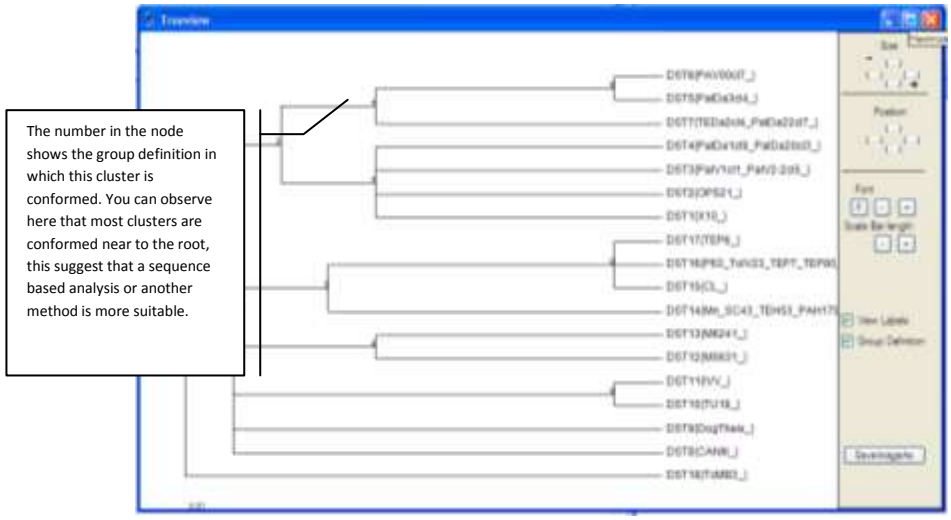

est Project
Distances
Allelic Profiles
Congruence Analysis
Scheme optimization
Trees
Others
Make/View Allelic Profiles
Select Just one Isolate per ST
95% CI for Discriminatory Power
UPGMA-tree
NJ-tree
BURST
Copy to Clipboard
Export to file...
Basic BURST
BURST over all group definitions
Test for network structures
CoAr Tcl prueba 2
Pdh Tc I plan A prueba 2
SodA Tcl prueba 2
ST
It is useful if you do not want to work with repeated STs (identical Strains)
It makes an UPGMA or Neighbor Joining tree for your allelic profiles
Calculates Confidence Interval for Discriminatory Power. It uses a delete-one jackknife procedure (Severiano et al., 2011). The CI is added next to
It makes a simple BURST analysis for your allelic profiles. You need to specify a group definition, usually the number of loci minus one
Calculates BURST over all group definitions in order to determine a optimal group definition. The percentage of singletons for each group definition and a dendrogram showing relationship among groups is showed
It analyzes whether network structures are present in BURST graphs suggesting recombination.

BURST


eBURST Graph
Information about clusters conformed
Simple Graph with clusters, blue lined indicates STs connected by the group definition (number of shared alleles).
Option Bar for graph edition

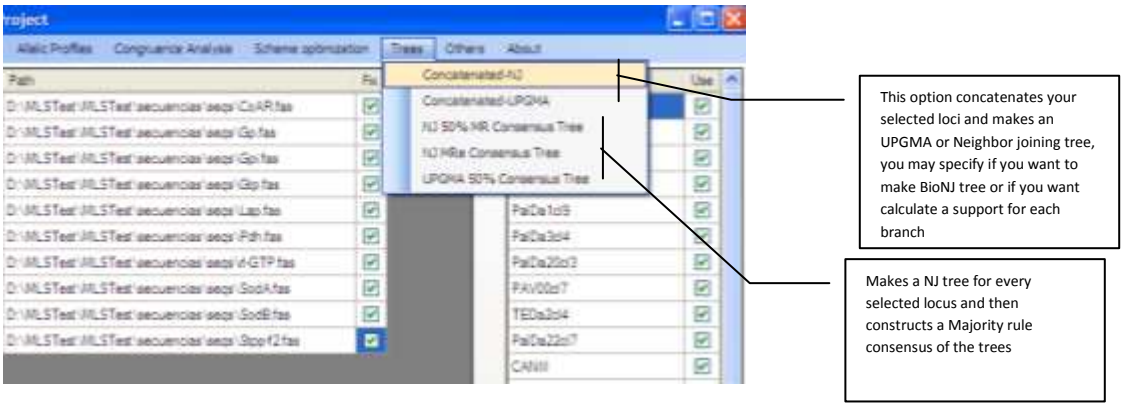
BURST over all group definitions

This analysis allow to see when groups are conformed using all possible group definitions. It is useful to select an optimal group definition or to decide if the eBURST analysis is useful for the data. For example, if the tree show the majority of groups conformed near to the root means that there are few DST/STs in common among strains. This suggests that an analysis not based on DST/STs is more suitable.

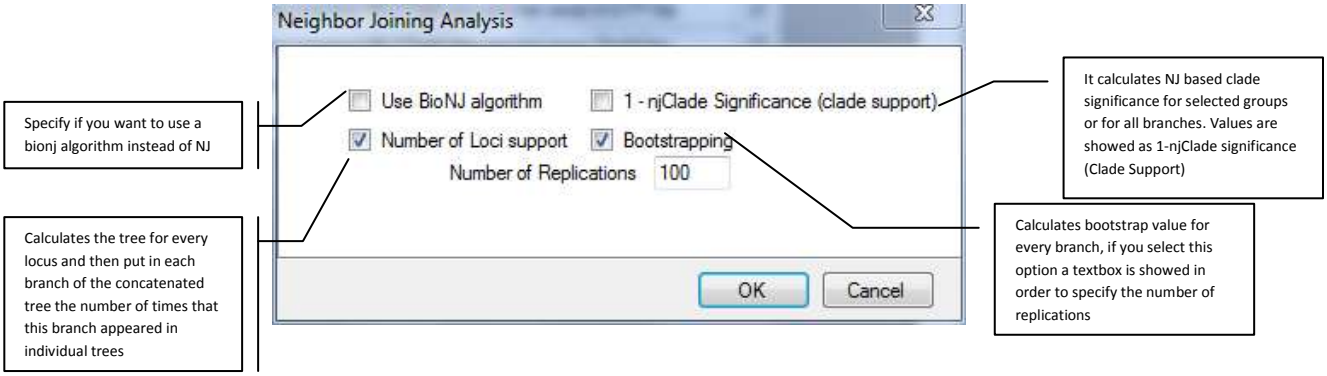


Making Trees for sequences

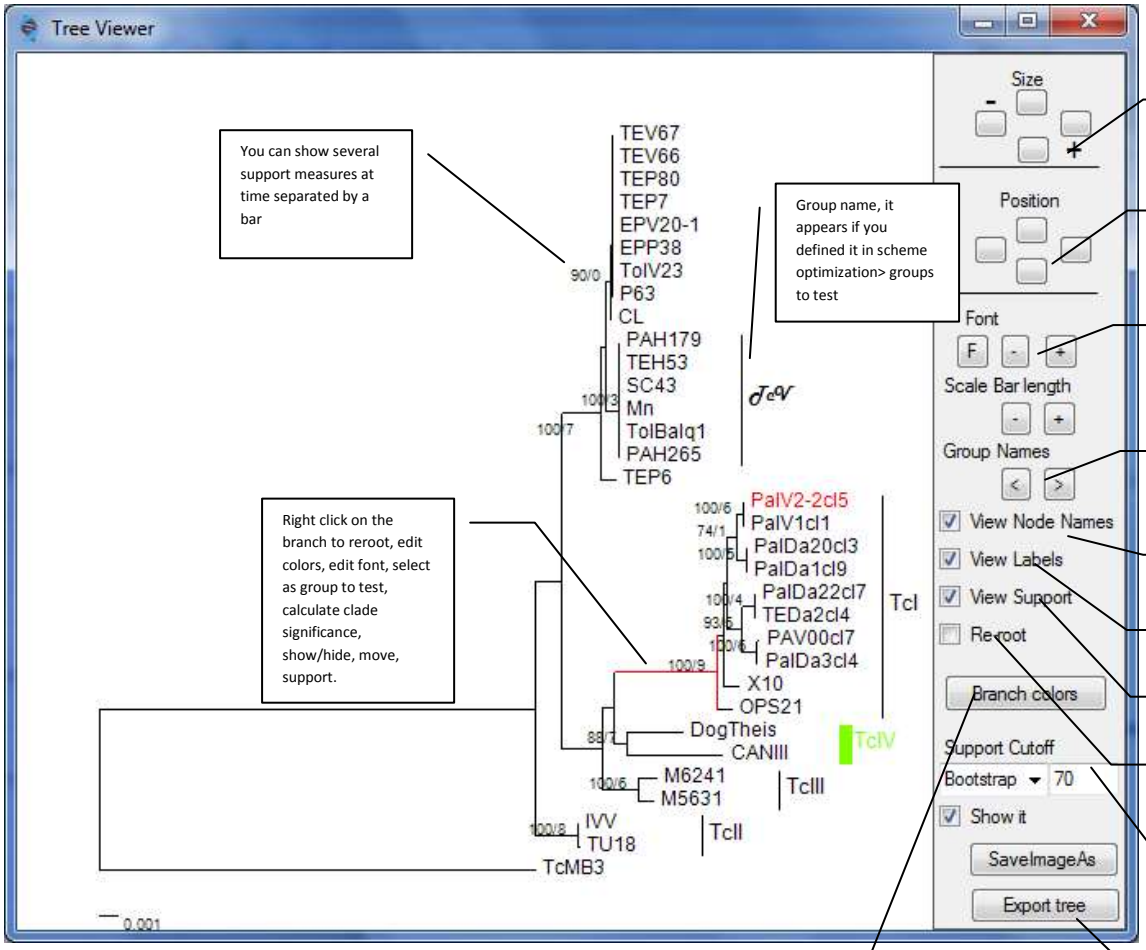
1) First, you need to specify how heterozygosities are treated in the analysis. Click on Distances>Handling heterozygosis and select one option. The default is Average States
2) Making a tree



Neighbor Joining Analysis (Concatenated NJ option)



The Tree Viewer



You can show several support measures at time separated by a bar

Group name, it appears if you defined it in scheme optimization> groups to test

Right click on the branch to reroot, edit colors, edit font, select as group to test, calculate clade significance, show/hide, move, support.

This menu allows to color branches according their support value.

Changes the size of the tree

Moves the tree in the window or a selected element

Changes the font size and type

It moves the all the Groups names and bars

Show/hide Group Names

Show/hide Taxa

Show/hide Support for all branches

If it is selected, when you click on a branch the tree is re-roted there

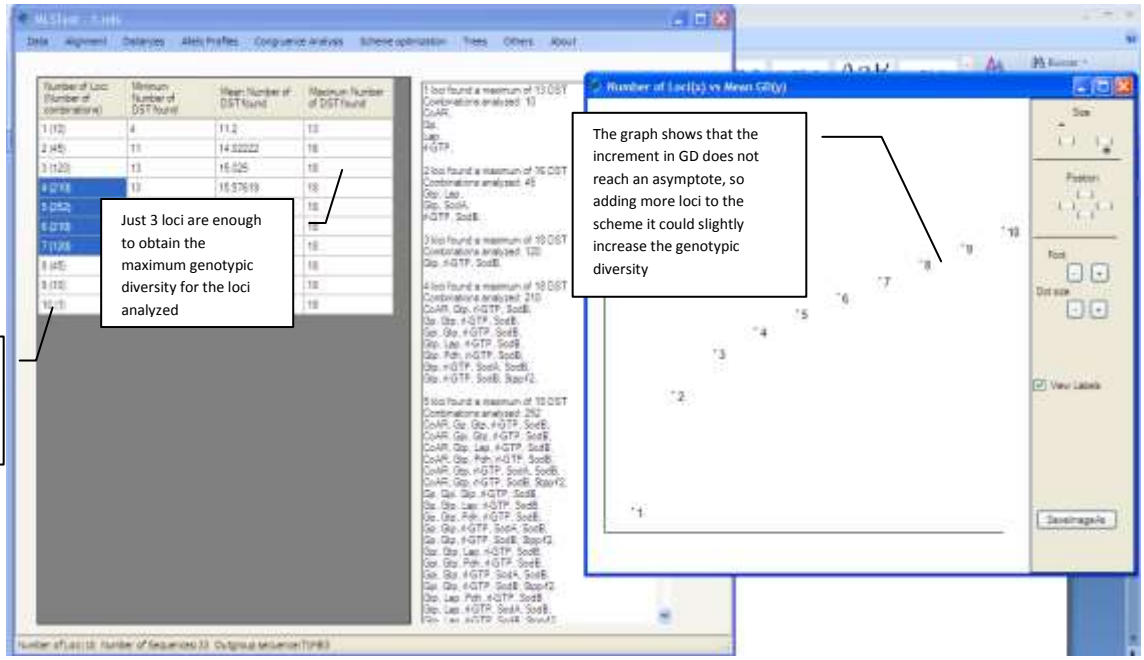
hide Support lower that the cutoff

Save the tree in Newick Format or MLSTest Format

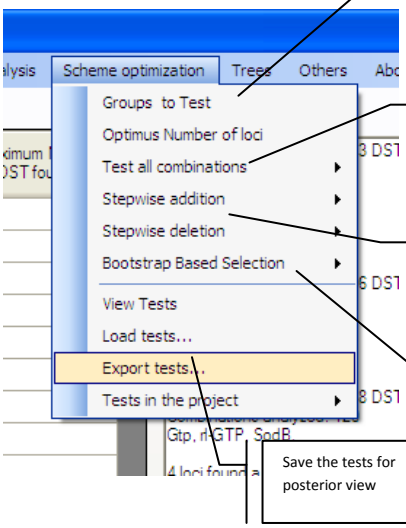
Scheme optimization

>Optimus Number of Loci

This option is useful to reduce the number of Loci analyze and select a good number of loci in order to obtain a certain genotypic diversity. Additionally, it is useful to determine if the genotypic diversity will increase by adding more loci to the scheme.



>Selecting a good combination of loci



Click to select groups to test if you previously know the phylogeny of your strains or based on the concatenated tree topology

It makes the NJ trees for all possible combinations for a determined number of loci and evaluates and select the combination with better support and genotypic diversity

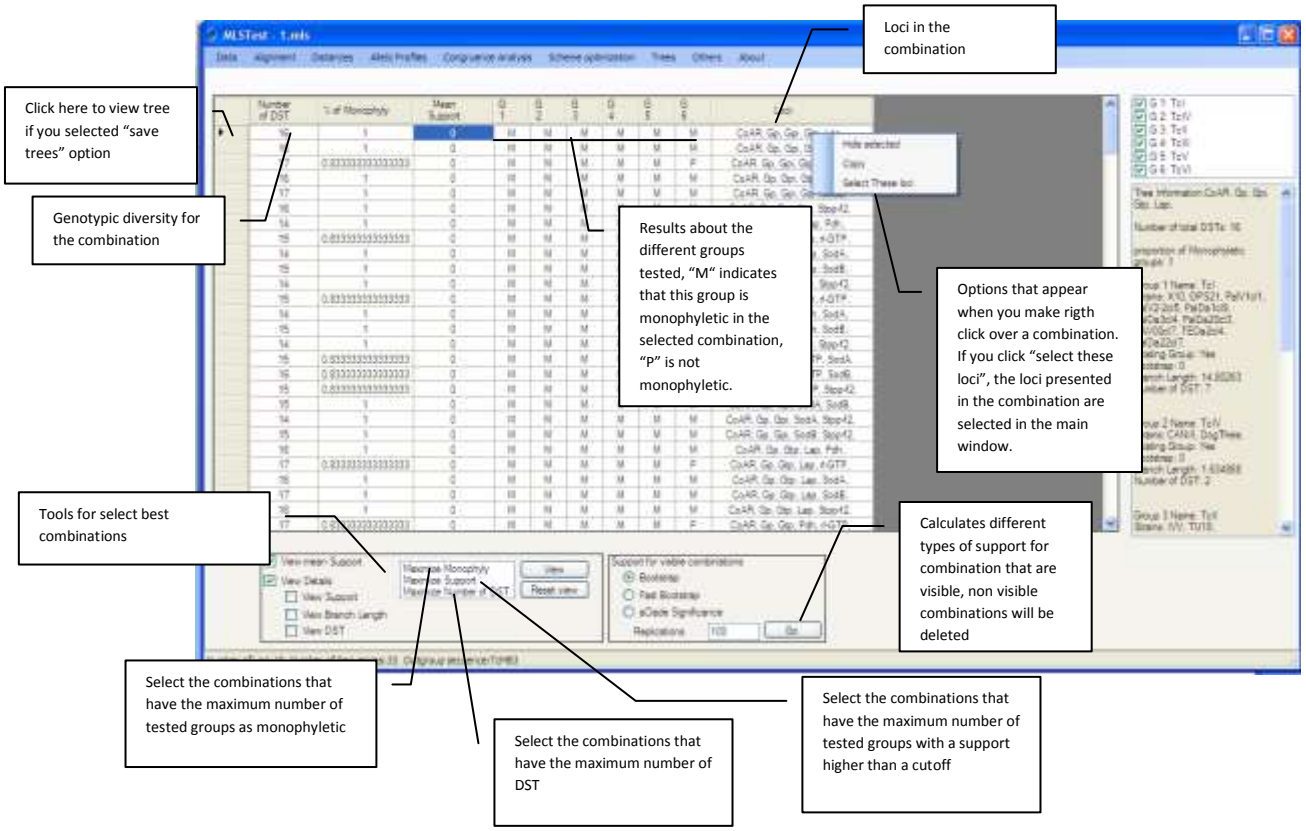
Add or delete one locus to a selected group of loci in order to test if there is a improvement or impairment in support, genotypic diversity or support.

1)Discard low supported loci, 2)calculate if a bootstrap for a branch in a combination is significantly better than the bootstrap of the same branch in another combination according to the number of replications used

Save the tests for posterior view

Tests window

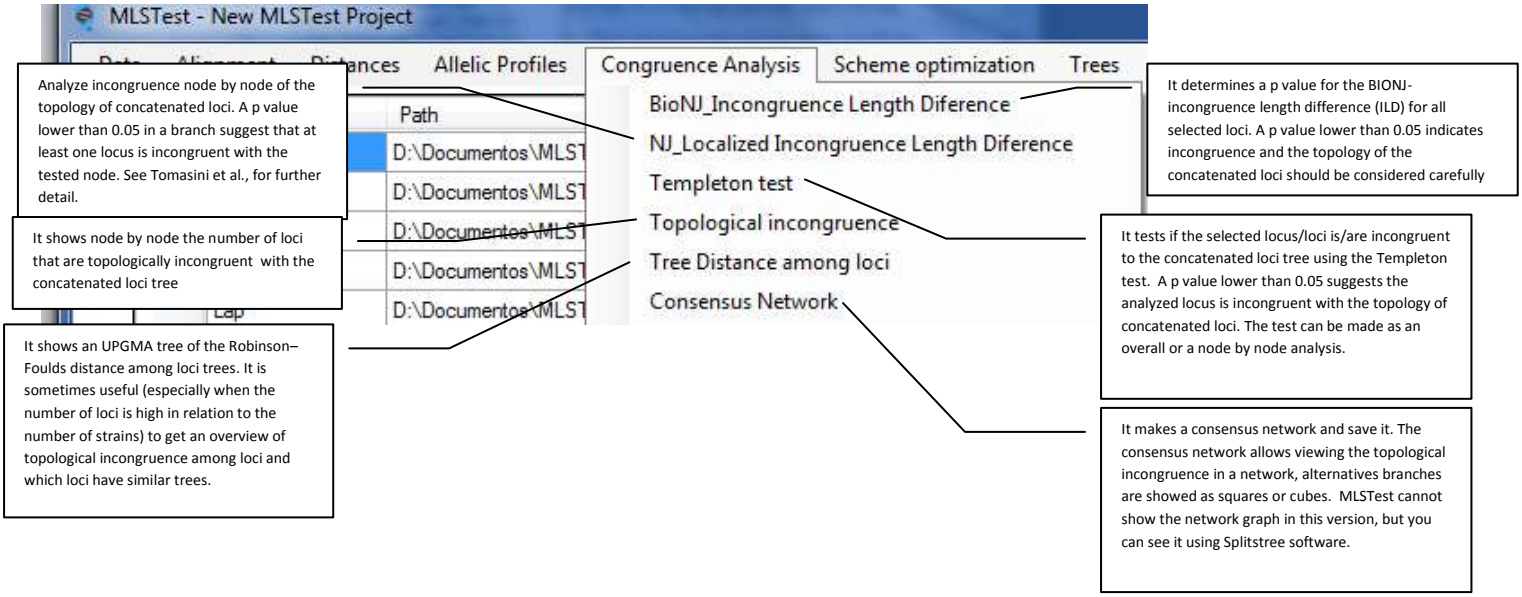
When you test all combinations of loci for a determined number of loci a window like this is showed. All possible combination of loci are showed in a table.



How to select a good combination:

1) Define your selection criteria: i.e. maximize number of monophyletic groups and then Maximize observed diversity
2) Select the groups to test (scheme optimization menu>select group to test)
3) Run the analysis (scheme optimization menu>test all combinatios>test user determined groups. The tests window is showed
4) Apply the criteria clicking on "Maximize Monophyly", then click on "view" button. MLSTest applied this criterion and combinations that maximize the criteria, then click on maximize number of DST (the second criteria is applied over the first). So, combinations that maximize your criteria are showed

Analyzing congruence among loci



About the Templeton test

Two tests are included based on the Templeton test. The first one, tests if the topology of the concatenated tree is not significantly different than the "optimum" topology for a selected partition based on the NJ method. It means that the data for this partition is incongruent with the topology of concatenated loci. This is made by evaluating both topologies site by site on the partition. The length of each topology for a site is calculated using the Pauplin formula (Pauplin, 2000). Then, the rank order of the differences is evaluated using the Wilcoxon sum-Rank test

The second test analyze node by node. It allows seeing how incongruence is distributed across the tree. Each branch in the topology of the concatenated is forced to exist in the data partition using the same algorithm described for NJ-LILD. Then, Templeton test is used to compare the optimal topology with the constrained topology. A p value lower than 0.05 indicates the partition is significantly incongruent with the evaluated node.

Recommended steps for analyzing incongruence

1) Look for topological incongruence (if there is topological incongruence go to the next step)
2) Make a BIONJ-ILD test, if the p-value is significant or near to the significance make a more detailed test as the NJ-LILD to find possible branches affected by the incongruence.
3) Test each locus to determine which is incongruent by using the Templeton test. You have two options here, to test the locus of interest against the concatenated topology or to test all the loci less the analyzed locus against the concatenated

topology. The first says if the data of the selected locus is incongruent with the combined topology. The second says if excluding certain locus, the remaining loci are poorly explained by the topology of the combined analysis. This means that this locus introduce significant topological modifications and sometimes it is powerful than the first approach. Sometimes, the incongruence is not detected in any single locus because it is distributed among several loci.