Training course
Molecular Epidemiology
Leishmaniasis

MANUAL

IN-SILICO PROCEDURES

Leishmaniasis Epidemiology Network South America
LeishEpiNetSA

18 - 24 May 2009

Instituto Oswaldo Cruz
Rio de Janeiro
Brazil

Table of contents

C. Computer based analyses

C. COMPUTER-BASED ANALYSES

5. In-silico sequence analyses

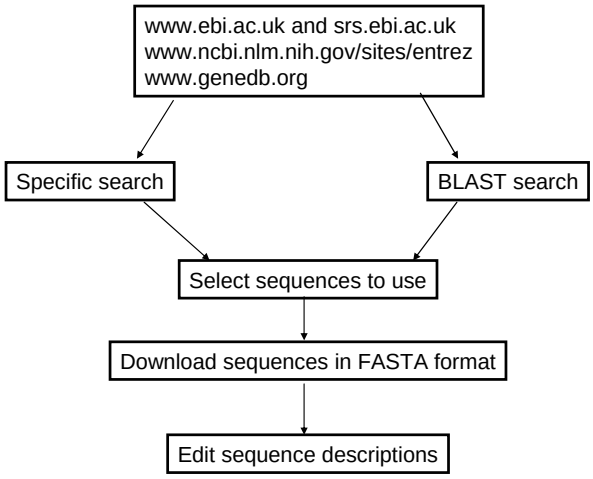
Protocol 5.1 Sequence retrieval from public domain data bases

Purpose	Retrieval of specific sequences for studying variability of *Leishmania* and for designing new assays and primers

A. Introduction:

In order to make optimal use of sequences in the public domain, retrieval of specific sequences is of great importance to study variability and to design new assays and primers. This can be done using similarity searches, or using sequence descriptions. As the workshop will deal with HSP70 RFLP, available sequences will be downloaded and used in further exercises. Also the ribosomal ITS1 region is used.

General workflow:



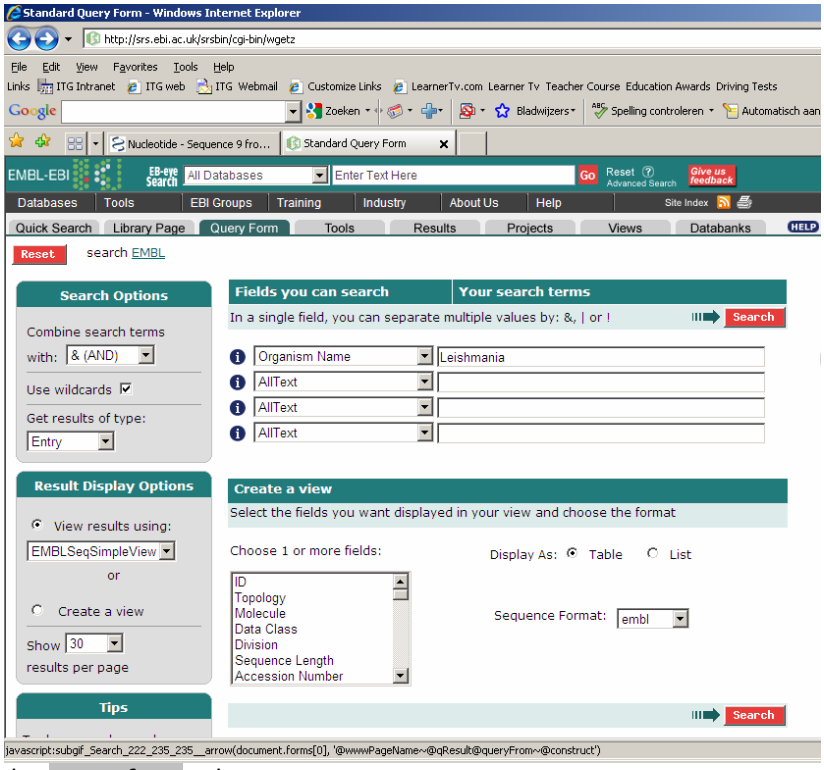
One way of collecting sequences from public domain data bases is by using searches with specific key words defined by the user.

Advantage: You will only retrieve sequences from which the submitter has indicated the name or function. In general this means that the submitter is quite sure about these, and you will in general collect only what you want. It is no guarantee for an error-free sequence.

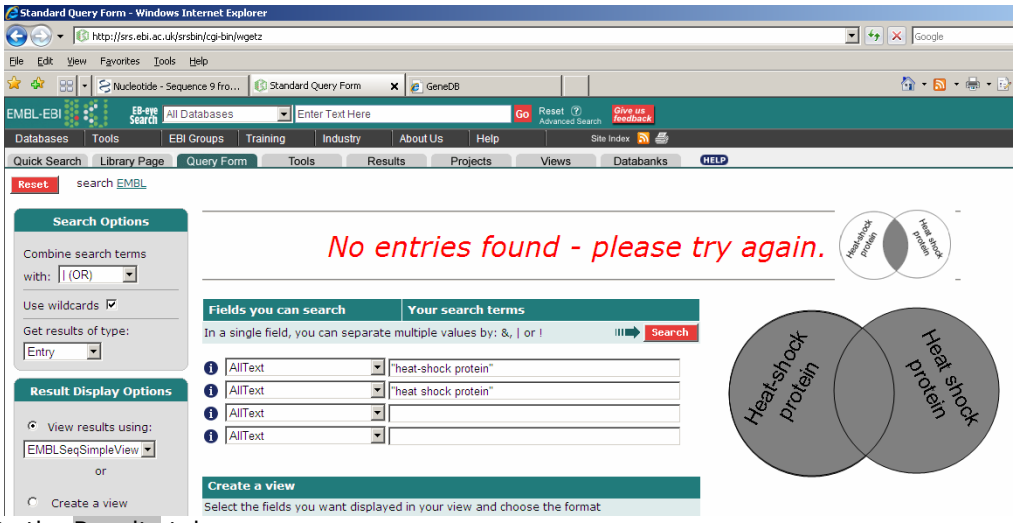
Disadvantage: Your search results depend largely on the key words used. If not properly chosen, you will miss certain sequences. In addition, if the submitter made a typographical mistake during submission, a matching key word can be missed.

Method A: Retrieval of hsp70 nucleotide sequences from EBI (joint exercise)

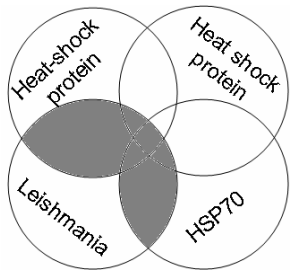
1. Open srs.ebi.ac.uk
2. Click the library page tab
3. Among nucleotide sequence databases, select EMBL
4. Click the query form tab
5. In the upper search box, select organism name from the pull-down menu
6. Type Leishmania in the search box
7. Click search



8. Click the query form tab
9. In the upper search box, select all text from the pull-down menu
10. Type HSP70 in the search box
11. Click search
12. Click the query form tab
13. Using the same technique, search for "heat-shock protein" (use the quotes)
14. Go to the query form
15. Repeat the previous search, but now enter in the second box "heat shock protein" (no dash)
16. Repeat this search, but now select Combine search terms with OR from the left pull-down menu



17. Go the Results tab.
18. Combine the previous searches to obtain the grey zone below: Q1 & (Q2 | Q4)

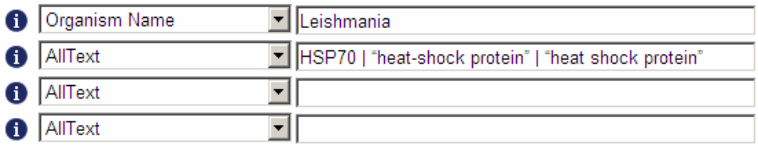


19. Select the entries you want to use.
20. Press the Save button
21. Select output of All to File using the FastaSeqs format
22. Press Save and choose a location and file name
23. Done, you have now retrieved *Leishmania* HSP70 sequences from the EMBL data base
24. Give each file entry a clear new name of maximum 40 characters. Include essential information such as the accession number and an abbreviation for the species or strain.

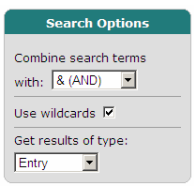
Additional remarks:

1. The above exercise is only an illustration of one possibility, it does not give a full overview of all search options. Retrieving all available sequences is a matter of proper (and sometimes inventive) selection of search terms. You can also look for particular accessions found in literature.

2. Selecting the proper set of sequences is often achieved by trial and error: some searches will result in too much junk, others will not get you all the sequences.
3. Searches can be combined in the Results page, but also already during the initial search in the Query form. The above exercise could be substituted by 1 search, when choosing Combine search terms with AND:



4. In order to select the proper sequences, it is important to critically read the sequence descriptions. A particular sequence might be listed as "putative", "similar", "precursor", "partial", or "mitochondrial". These might not always be relevant for your work.
5. Be aware that sometimes you will retrieve sequences that are the reverse complement of a coding region. In addition, not all sequences you retrieve will contain the entire fragment you are interested in, or some can contain a much larger fragment.
6. Make good use of search combinations, but be aware of their effect: "Q1 & Q2" (= "Q1 AND Q2") results in entries **common** to Q1 and Q2, in other words entries that came up in Q1 and in Q2. On the other hand, "Q1 OR Q2" (= "Q1 | Q2") takes entries that came up in Q1 or in Q2, in other words all entries from both searches. This may be counter-intuitive at times.
7. Another possible combination is "NOT" (or "!"). E.g. in the above exercise "Q4!Q3" will result in entries containing the phrase "heat shock protein", excluding the ones with the dash.
8. Always be on the lookout for missed entries, enlarge your searches if necessary. Sometimes searches do not result in what you expect because of the way search terms are checked against the entries. This you will find out only empirically.
9. You can also combine specific searches with resulting entries from BLAST.
10. Use of a wildcard "*": wildcards are used to find parts of a word in a description. Examples:
11. "shock" will find only "shock", but not "heat-shock", neither "shock-70"
12. "shock*" will find "shock-70" but not "heat-shock"
13. "*shock*" will find any word or phrase that contains shock in it, like "heat-shock-70 protein"
14. As a default, all searches are performed with a wildcard at the end of a word, as specified under search options:



15. In the Library page you will find an extensive list of libraries you can look in for finding sequences. Usually the general library used in the exercise is sufficient.
16. You can play around with different outputs besides FASTA, and search more specifically in particular entry fields. Trial-and-error is the best way of starting to deal with these options.
17. GenBank (www.ncbi.nlm.nih.gov/sites/entrez) offers the same possibilities for searching and search combinations, but sometimes comes up with quite unexpected results, e.g. a search for "donovani*" ends up with a few thousand bacteria that do not have the word donovani anywhere in their description.
18. Similarly in GeneDB (www.genedb.org) BLAST and specific search options are foreseen.

Exercises in individual groups:

1. Collect as many *Leishmania* HSP70 sequences as possible. Make sure you don't have duplicate entries by using the proper query combinations, e.g. instead of making separate FASTA files from different queries, first combine the queries and then make 1 FASTA file.

2. If the sequences are bigger than the HSP70 coding regions, delete the extra nucleotides. Info on this can be found in the entries themselves.

3. Put all sequences in one FASTA format file, which you identify as follows: "HSP70_groupID.fsa".

4. Clearly identify each sequence with "GroupID_accession_short description":

> 1_ AF291716_bra_complCDS

Sequence

> 1_ AY423868_tar_complCDS

Sequence

etc.

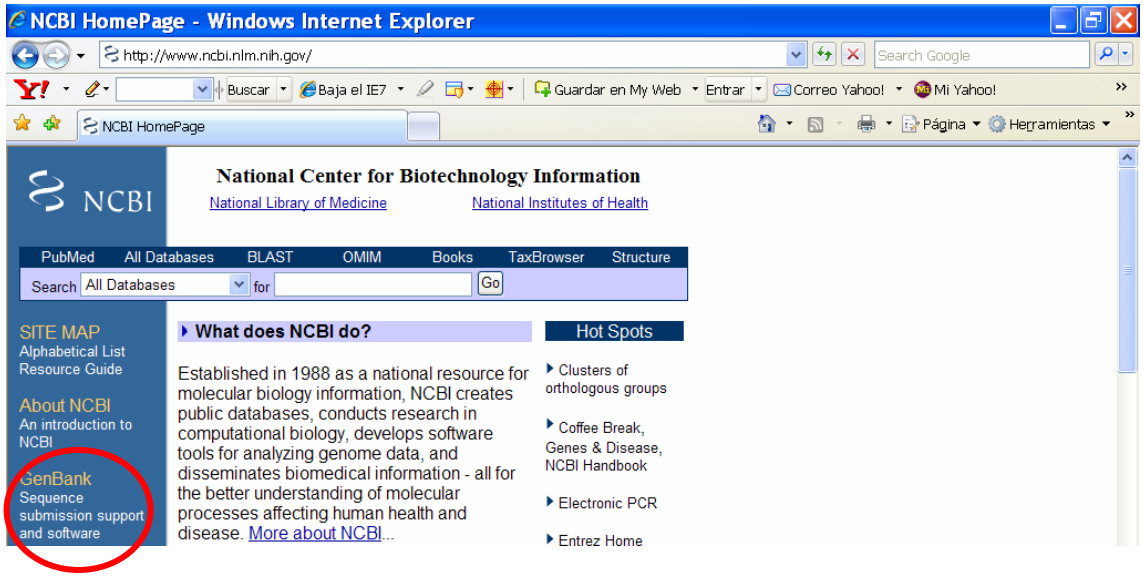
Beware: sequence names **must not exceed 40 characters** to be compatible with later programs!

5. Perform the same exercise for rDNA ITS1 sequences, but replace the "HSP70" labels by "ITS1".

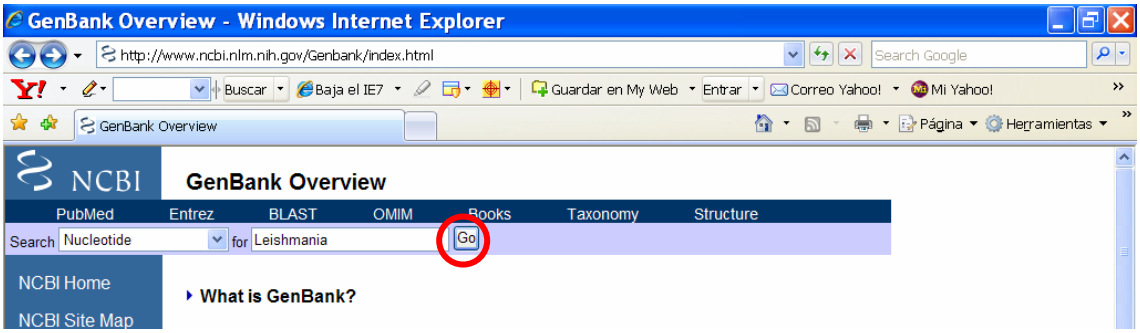
Method B*: Retrieval of hsp70 nucleotide sequences from GenBank (joint exercise)*

1. Open www.ncbi.nlm.nih.gov

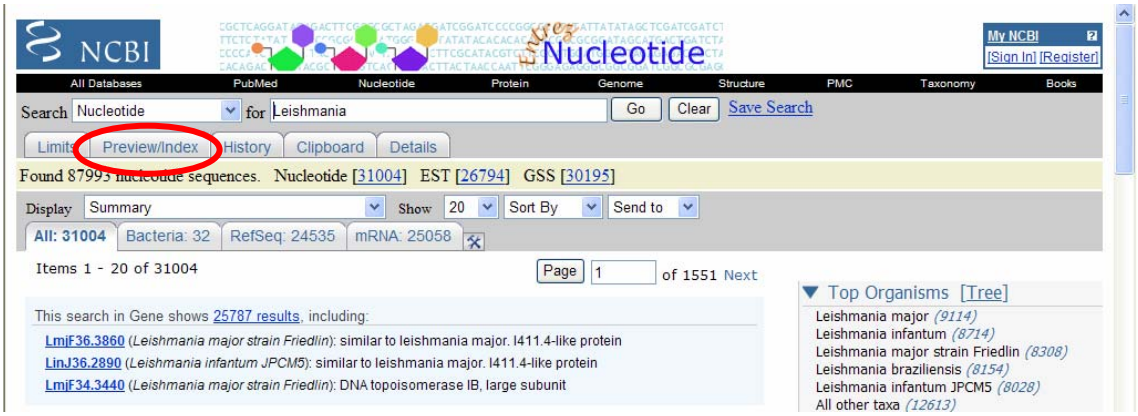
2. Click on GenBank on the left-side menu



3. Click nucleotide on the search tab and search for Leishmania, then Go

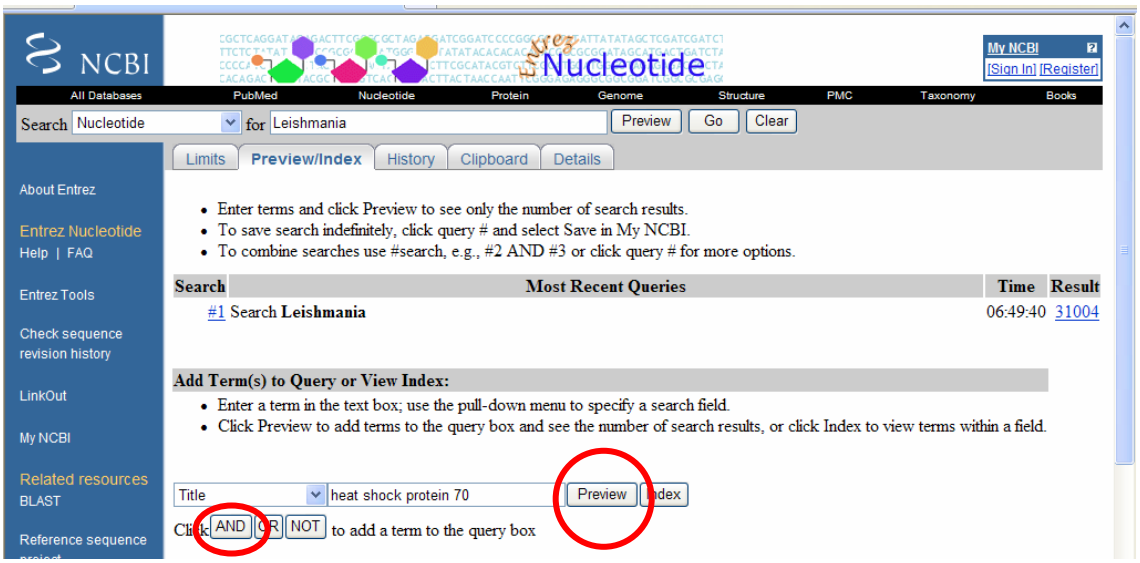


4. Now you are redirected to the Entrez nucleotide resource, where you can find all *Leishmania* related nucleotide sequences. However, you need to refine your search to *Leishmania* heat shock protein 70. Then click on the Preview/Index tab.

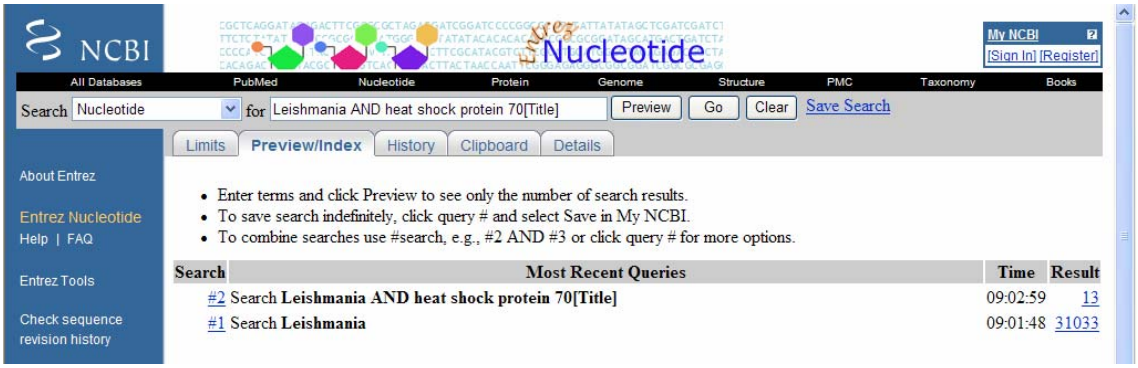


5. The result of your first search (#1) is presented, you can add new terms to refine your search:

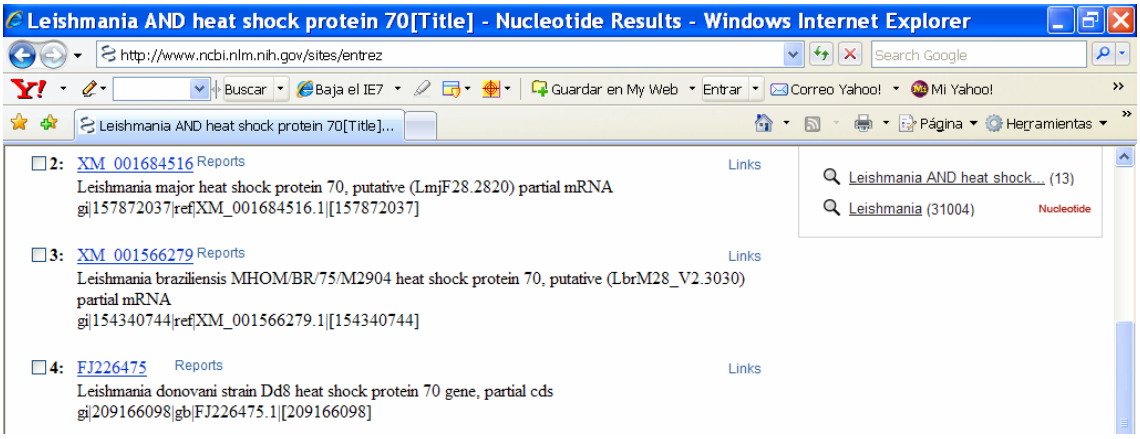
5a. Search field [title], Text box [heat shock protein 70]. Then click AND and Preview



With your new search (#2) now you have 13 results:

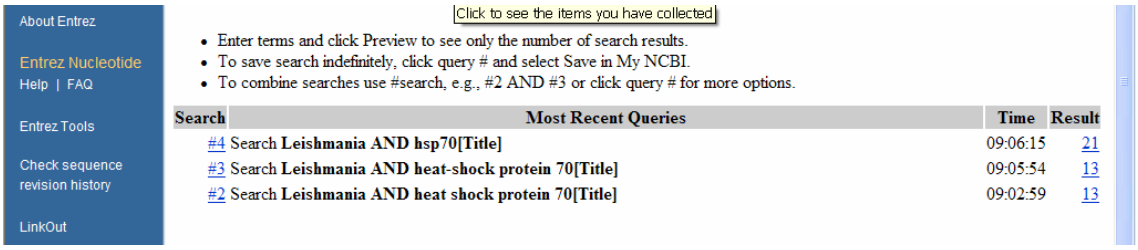


Click on results 13: And you will check that your *Leishmania* results have been filtered for heat shock protein 70.



5b. You can improve your search by changing the text in the query

- Search field [title], Text box [heat-shock protein 70]
- Search field [title], Text box [hsp70]

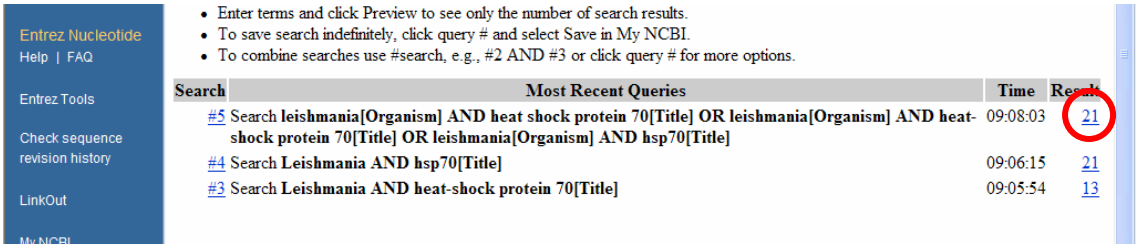


5c. Or you can also combine searches:

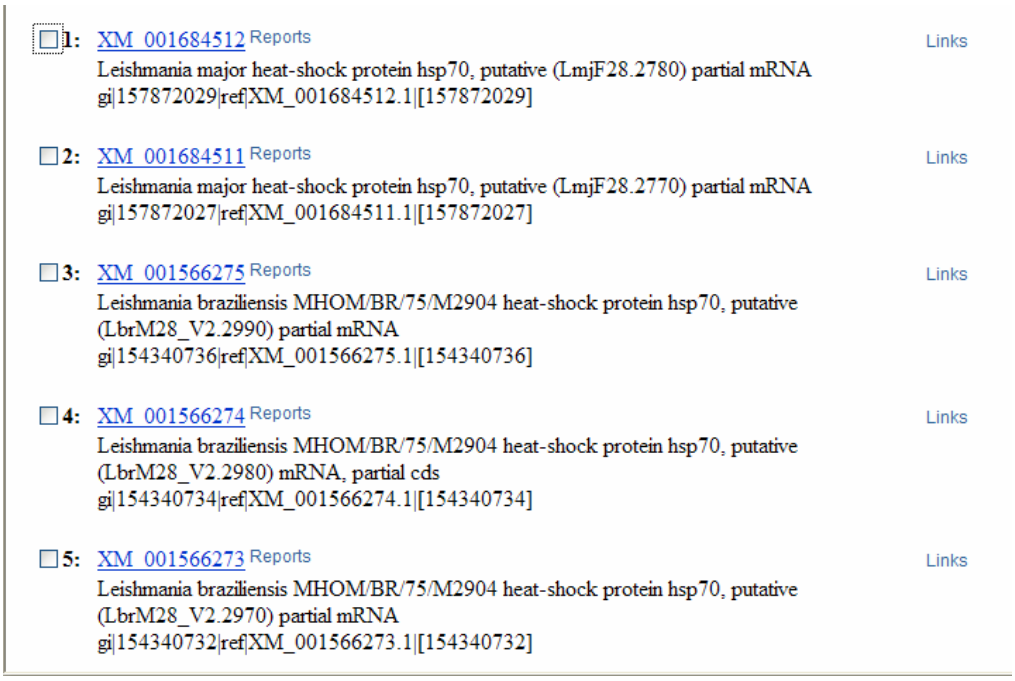
Either by writing in the upper search text box:

leishmania[Organism] AND heat shock protein 70[Title] OR leishmania[Organism] AND heat-shock protein 70[Title] OR leishmania[Organism] AND hsp70[Title]

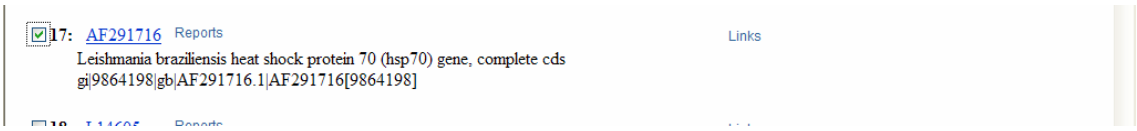
Or by combining the AND/OR buttons and the search fields and text box



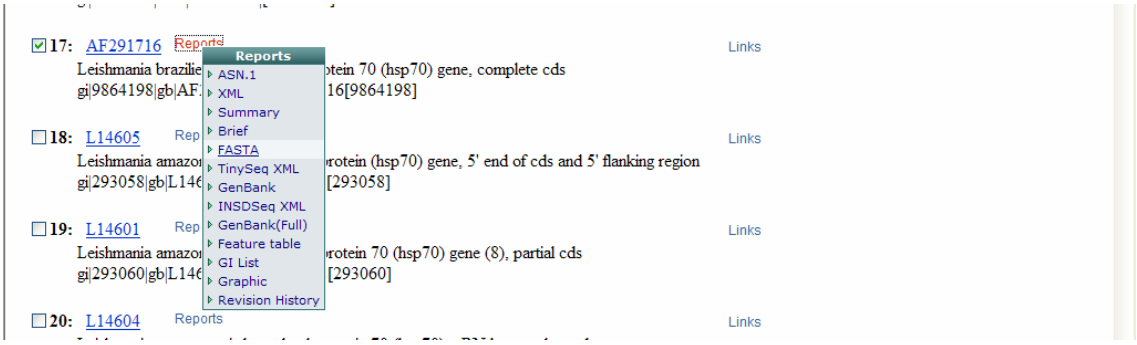
6. Check your entries by clicking the result and...:



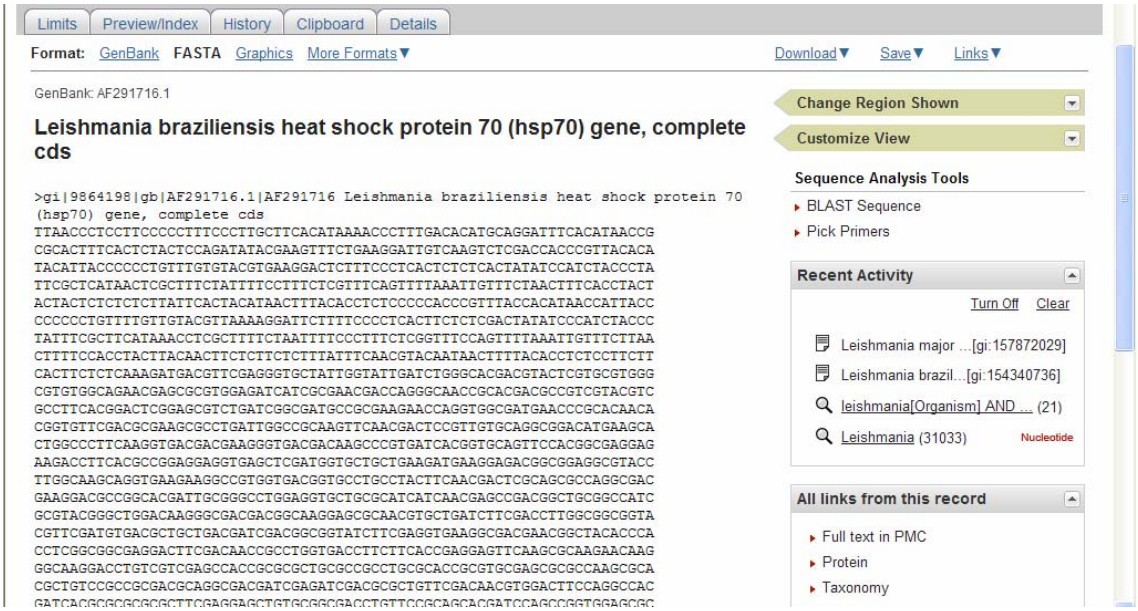
7. ...select those you want to use



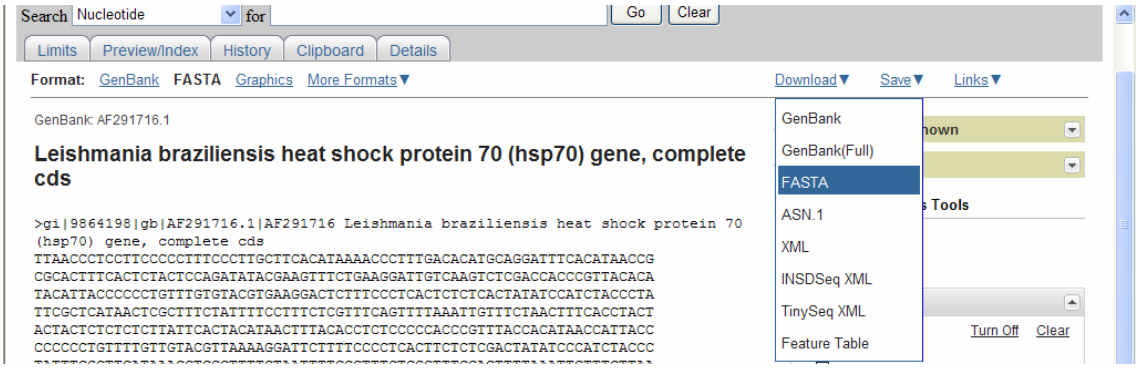
8. Click on Reports and select FASTA



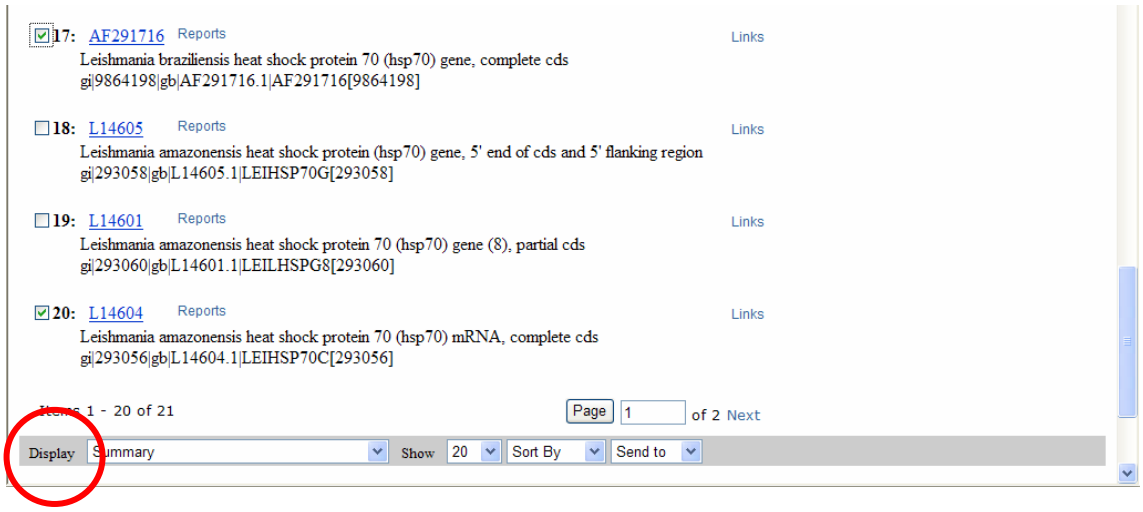
9. Now you have your sequence in FASTA format



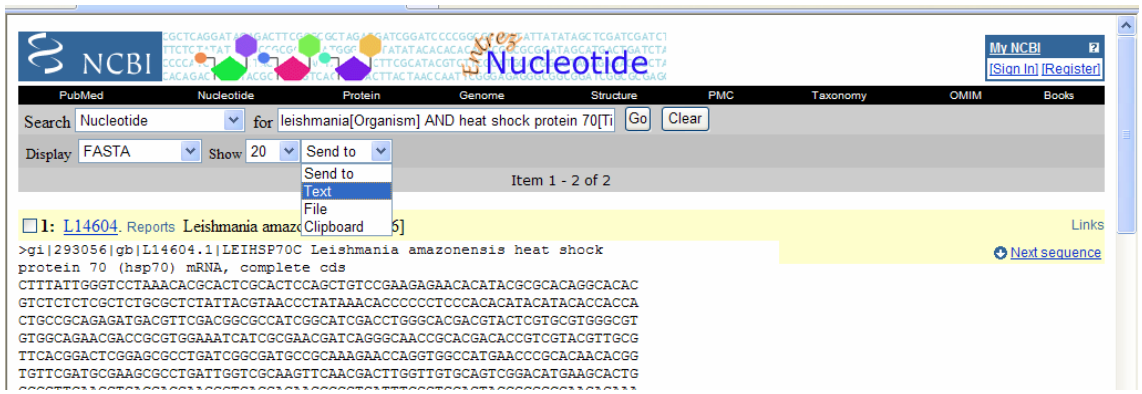
10. To save it click on Download as FASTA and save it in your sequence editor.



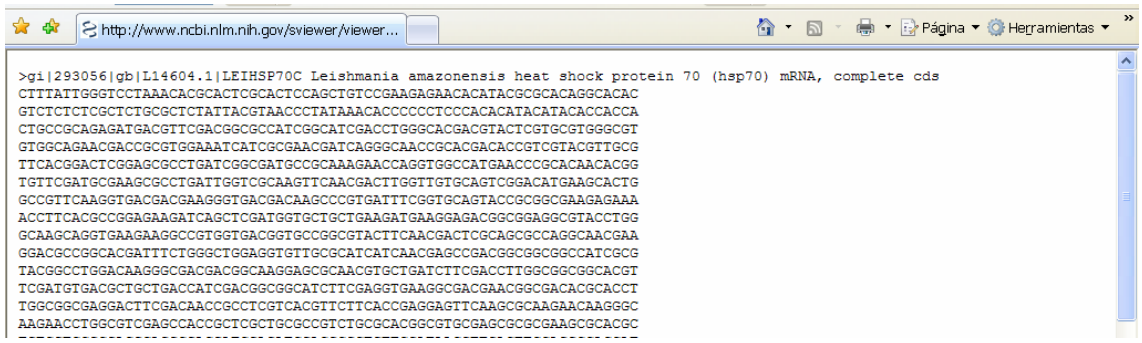
11. You can also select more than one sequence



12. Display them in FASTA format



13. and send them to Text



14. Now you can copy and paste them in a text editor.

Protocol 5.2 **Analysis of sequence chromatograms**

Purpose	To edit the chromatograms obtained from automated sequencers and resolve possible conflicts in order to obtain the final deduced sequence

A. INTRODUCTION

Output from automated sequencers is limited in size, and approaches the sequence from one direction only. It is necessary to edit the sequence chromatograms in order to extract trustworthy information, concatenate them into larger sequences, and resolve possible conflicts to obtain the final deduced sequence. To conclude, the new sequences can be submitted to the public domain sequence repositories (EBI / GenBank).

Sequences used: HSP70 and ITS1 chromatograms obtained from automated sequencing

Programs used: BioEdit (www.mbio.ncsu.edu/BioEdit/BioEdit.html)

Programs not used but excellent: MEGA (www. megasoftware.net)

Sequences used: HSP70 and ITS1 chromatograms obtained from automated sequencing.

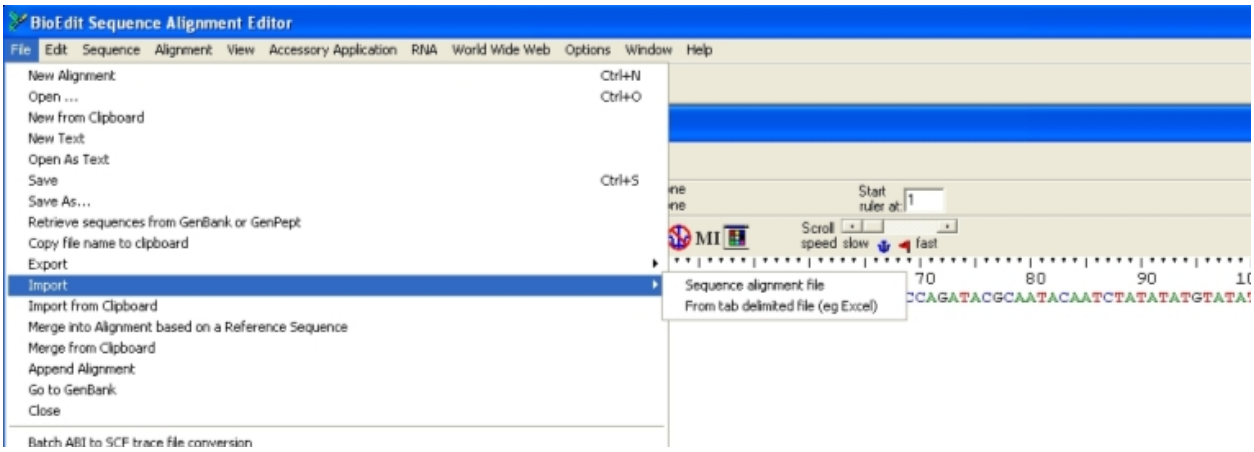
Programs used: BioEdit (www.mbio.ncsu.edu/BioEdit/BioEdit.html)

Programs not used but excellent: MEGA (www.megasoftware.net)

Method: ***Editing chromatograms for HSP70 and ITS1 sequences by using BioEdit software (joint exercise)***

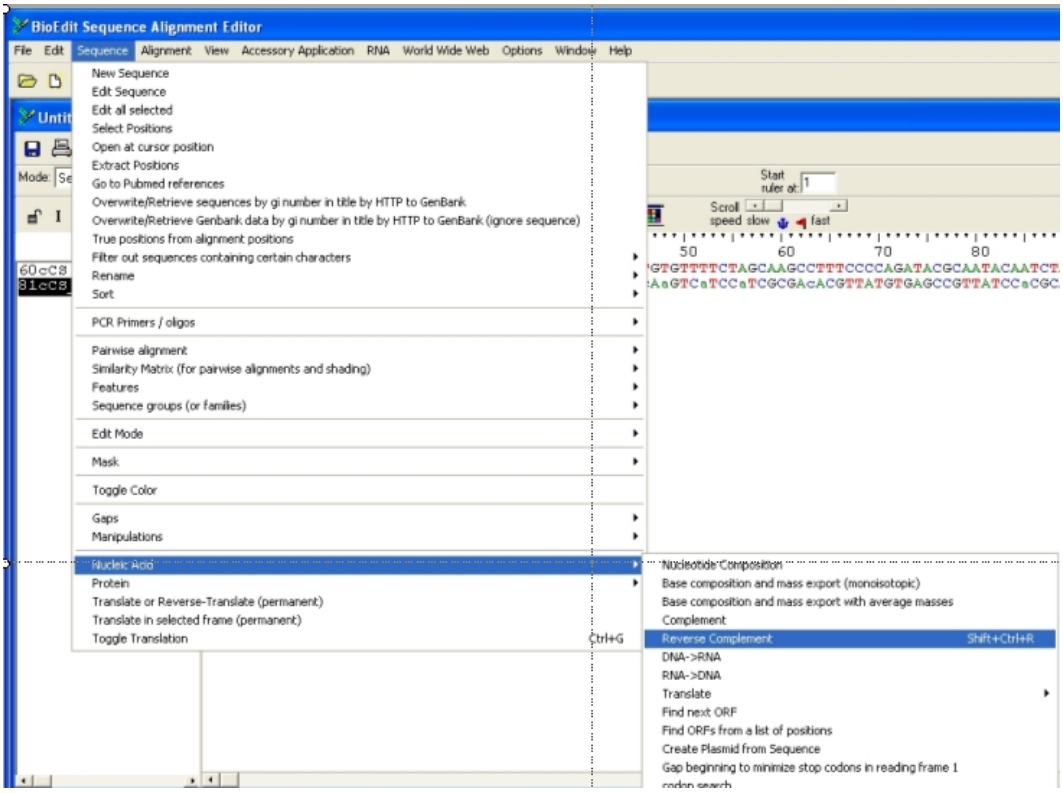
1. Start the BioEdit software from 

2. Go to File and New Alignment and Import your sequence alignment files:



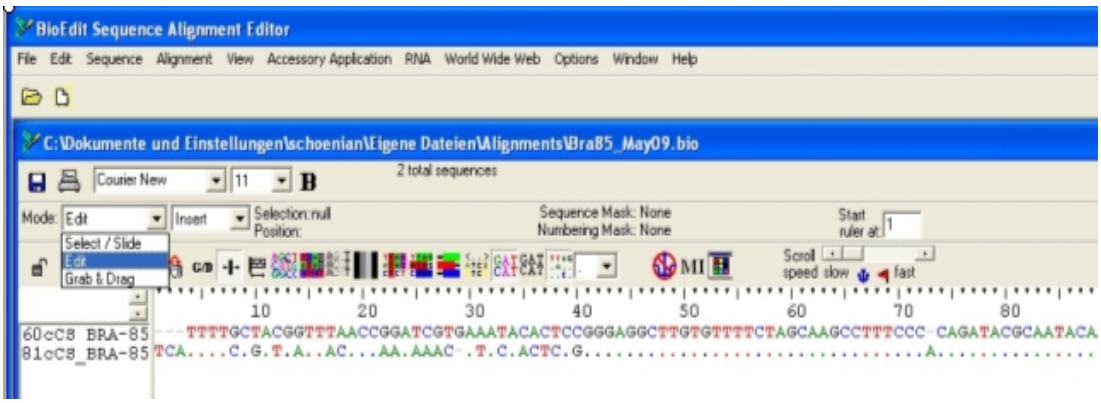
3. To see the chromatogram for a given sequence press Open. Check the quality of the sequence:

4. Compare the forward (Primer LITS for ITS1) and backward (primer L5.8S for ITS1) sequences. For this convert the backward sequence as follows: Go to Sequence, than to Nucleic Acid and than click on Reverse Complement.

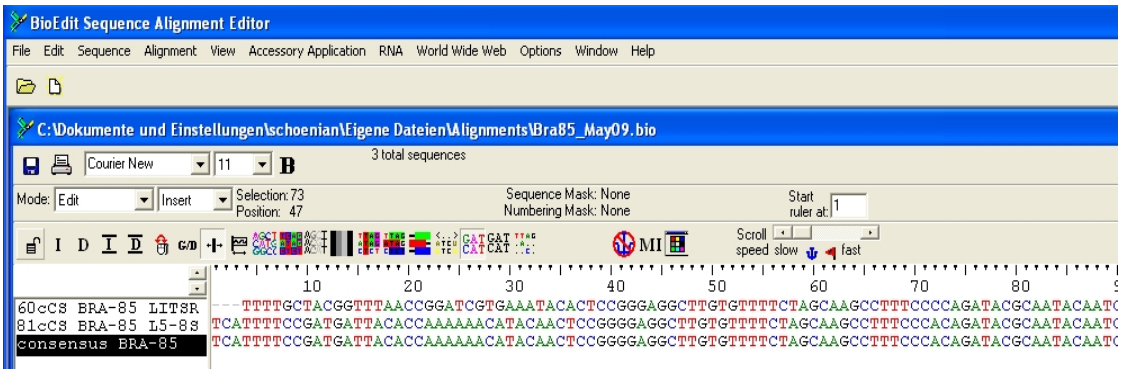


5. Align the forward and the converted backward sequence by hand or go to Sequence and choose Pairwise alignment. Save the alignment obtained and close the file.

6. For editing go to File and Open your alignment. To check again with the chromatograms Open your sequence files (see point 3!). Use Mode for editing your sequences.



It is recommended to copy one of the sequences and to do all the editing there.



7. After editing is finished, keep only the consensus sequence and delete the other two sequences by labeling them and pressing Edit and Cut. The consensus sequence will then be saved in Fasta format. This sequence can then be used for multiple alignments by either using BioEdit or MEGA softwares. In this training session we will use MEGA for multiple alignments.

Protocol 5.3 **Sequence alignments, primer design and *in-silico* RFLP**

Purpose	To align sequences for identification of homologous positions

A. INTRODUCTION

Comparing sequences requires identification of homologous positions which is done by aligning them. Bases on such alignments, primers can be designed and RFLP experiments simulated *in-silico*.

Sequences used: Edited HSP70 and ITS1 sequences from previous workshop sessions

Programs used: MEGA (www. megasoftware.net)

Primer3 (online tool, http://frodo.wi.mit.edu)

Programs not used but excellent: GeneDoc (www.nrbsc.org/gfx/genedoc/)

BioEdit (www.mbio.ncsu.edu/BioEdit/BioEdit.html)

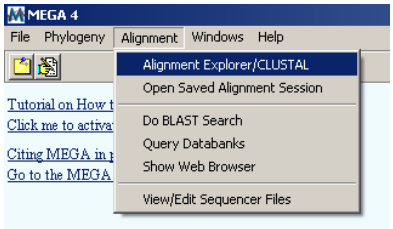
Method A: ***Aligning hsp70 nucleotide sequences (joint exercise)***

8. Prepare a FASTA file containing all sequences you want to align. This file should include sequences retrieved by SRS and BLAST, as well as the consensus sequences from the analyzed chromatograms. Each sequence should have a concise but clear labeling: "GroupID_accession_short description". Include a species abbreviation in your description.

9. Make sure you have a copy of the exercise FASTA file "Rio_Mega1.fsa".

10. Start the MEGA software.

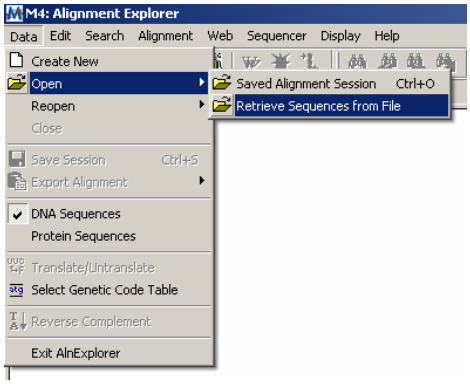
11. Open the alignment explorer from the Alignment menu:



12. Choose create a new alignment.

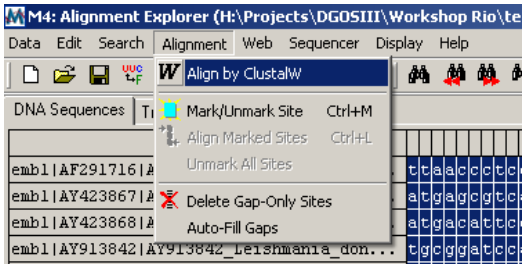
13. Select Yes for nucleic acid alignment.

14. Retrieve sequences from the FASTA file "Rio_Mega1.fsa" using the Data menu:



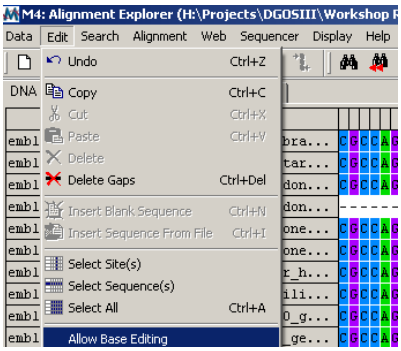
15. In the Edit menu, choose Select all, or do this by the classical Windows mouse clicking method using the SHIFT key.

16. Align by ClustalW in the Alignment menu, using the default program options:



17. Wait a bit. Perhaps take a coffee.

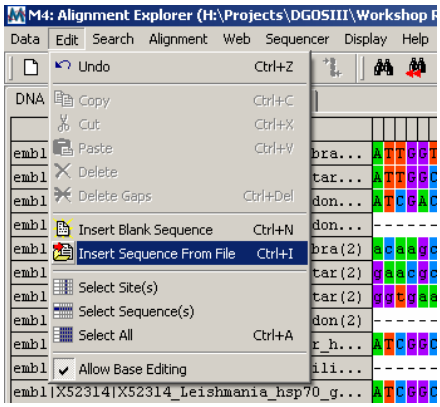
18. Once aligned, **switch off Base editing**:



This is of crucial importance when starting to edit an alignment, as it prevents you from accidentally changing, inserting, or deleting bases.

19. When not OK, modify the alignment manually by moving residues left (BACKSPACE) or right (SPACE).

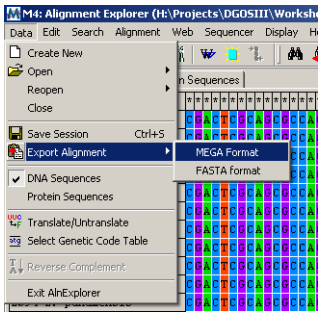
20. After finishing the alignment, add your own sequences to it. Do this from the Edit menu by allowing base editing and selecting the file option:



21. Use only sequences between the primer pair Hsp70sen (5' GACGGTGCCTGCCTACTTCAA 3') and Hsp70ant (5' CCGCCCATGCTCTGGTACATC 3').

22. Save your alignment from the Data menu.

23. Export your alignment in MEGA format, input a title and indicate that these are coding sequences:

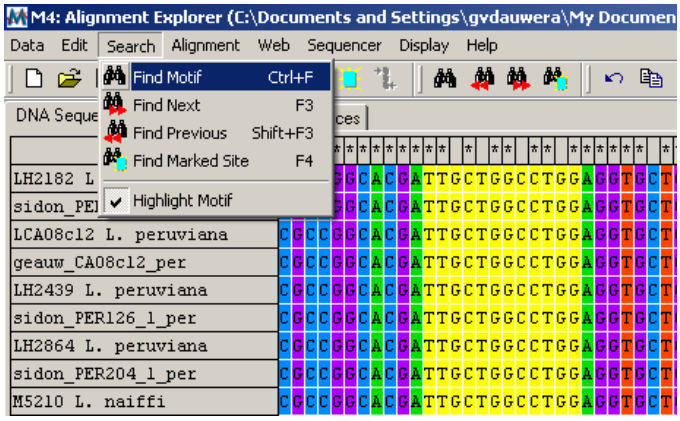


24. Done. You are now ready to start aligning your own sequences in the individual group exercise.

Additional remarks:

1. Sometimes during aligning you will get an error message that some sequence are too divergent. Just ignore this.

2. The biggest alignment problem is working with partial sequences. ClustalW does not deal with this well. There are two ways of preventing this to some extent: either manual editing after the aligning procedure, or clipping your sequences prior to aligning. The latter is possible by using the features in the GenBank or EMBL entries, or by looking for particular motives that should appear at the beginning or the end of your sequence, such as PCR primers:

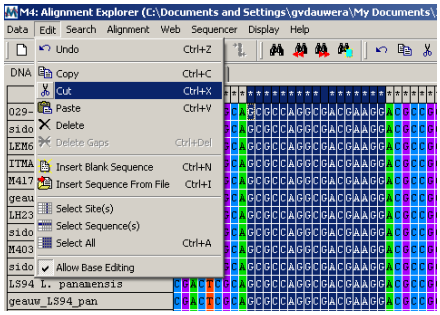


3. When working with data base retrieved sequences, you will usually have sequences in your alignment that do not match the others. These sequences have probably "escaped" your selection, and are not what you want. In such case double check the entry description in the data base you retrieved it from. Alternatively, this may be a partial sequence, part of a sequence not present in the other entries, or a reverse complement.

4. You can translate your nucleic acids into amino acids to aid in aligning in case you work with coding sequences. If your reading frame does not start from position 1, you have to select the corresponding columns first. Careful: alignment gaps are causing a frame shift in your translation, unless they are a multitude of 3.

Exercises in individual groups:

1. Align your retrieved and analyzed HSP70 sequences with the ones in the above exercise alignment. Remove any nucleotides outside the region of interest, and **remove all PCR primer sequences** by site selection and cutting:



2. Make a de novo alignment from your gathered and analyzed ITS1 sequences. When exporting into MEGA format, remember that these are not protein coding sequences. Use only sequences between primer pair LITSR (5' CTGGATCATTTTCCGATG 3') and L5.8S (5' TGATACCACTTATCGCACTT 3').

3. Use these alignments in subsequent primer design exercises.

Method B: ***PCR primers design using Primer3 software (joint exercise)***

A primer is a short synthetic oligonucleotide which is used in many molecular techniques from PCR to DNA sequencing. A pair of primers is used in most PCR variants and these are designed to have a sequence which is the reverse complement of a region of template or target DNA to which we wish the primer to anneal. When designing primers for PCR it is often necessary to make predictions about these primers, for example melting temperature (Tm) and propensity to form dimers with themselves or other primers in the reaction.

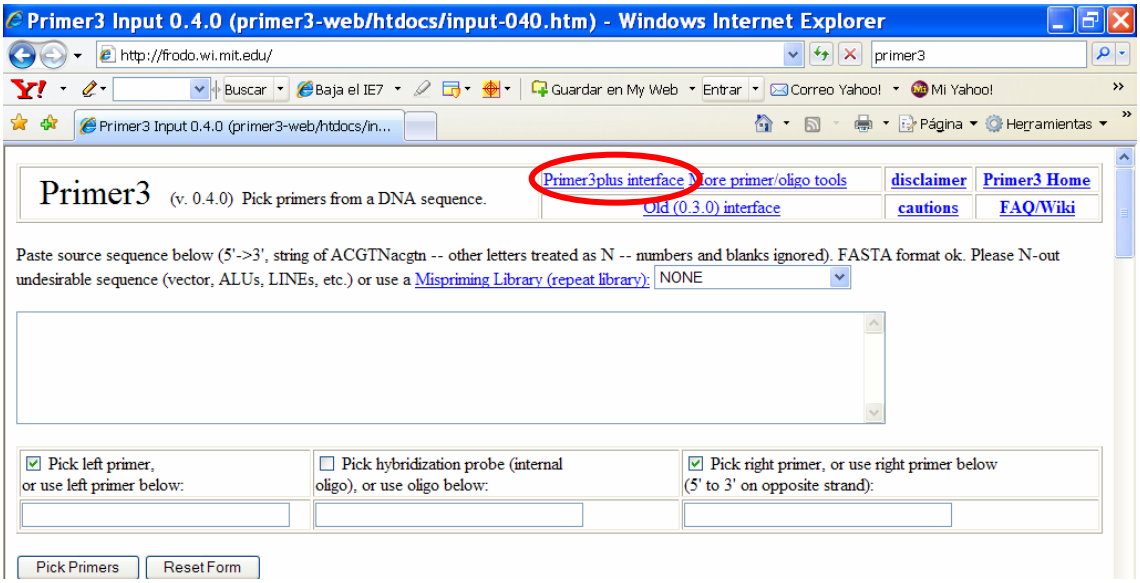
Some aspects should be taken into account when designing PCR primers:

1. primers should be 15-30 bases in length.
2. base composition should be 40-60% (G+C)
3. primers should end (3') in a G or C, or CG or GC: this prevents "breathing" of ends and increases efficiency of priming (Note: three or more Cs or Gs at the 3'- ends ofprimers may promote mispriming at G or C-rich sequences, because of stability of annealing, and should be avoided)
4. Tms between 55-70°C are preferred. Ideally, both primers should anneal at the same temperature. The annealing temperature (Ta) will ne dependent upon the primer with the lowest Tm. A simple formula to estimate the Tm of a DNA molecule is Tm= 2(A+T) + 4 (G+C); but you may be aware that this is only orientative, as different factors such as salt concentration may change the Tm value.
5. 3'-ends of primers should not be complementary, otherwise primer-dimers will be created preferentially to any other product.
6. primer self-complementarity (ability to form 2^{ary} structures such as hairpins) should be avoided.

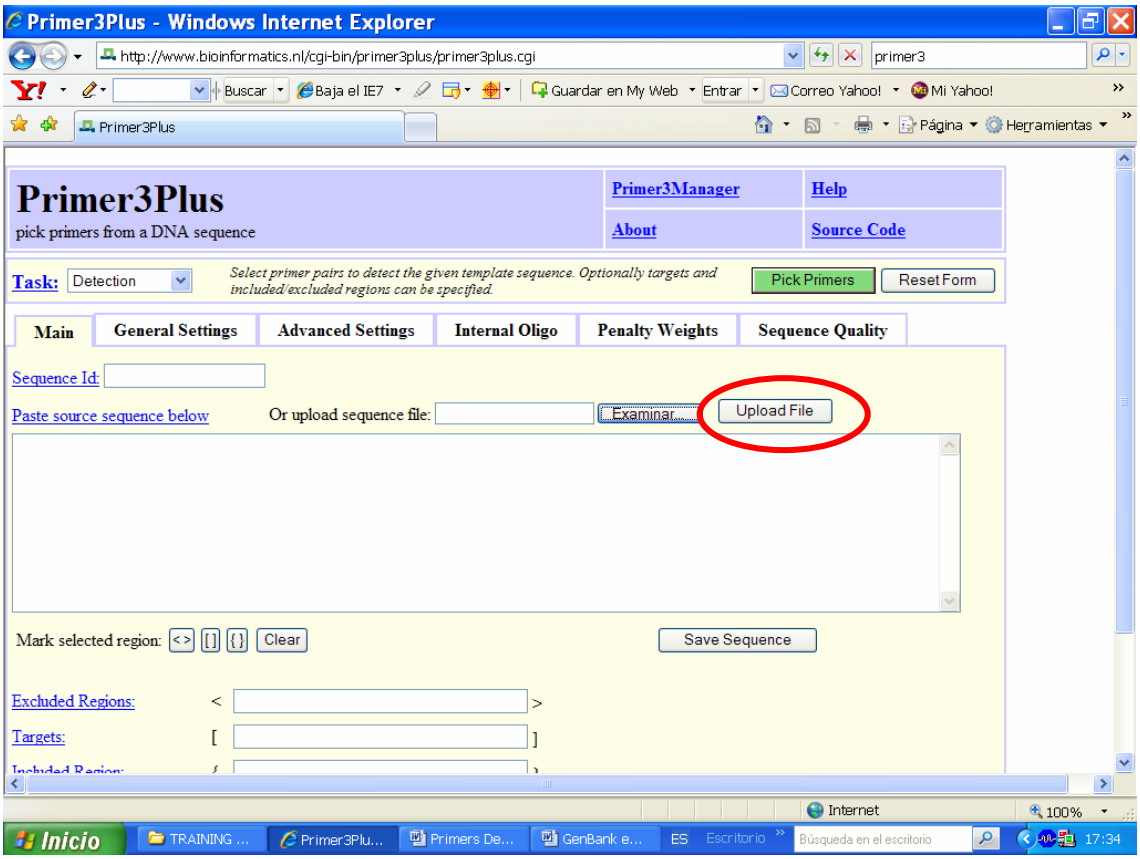
Following these rules you can design your own set of PCR primers. Nevertheless there are different software tools which can be helpful for this purpose. In this training course you will deal with Primer3 software.

Primer3 (v. 0.4.0)

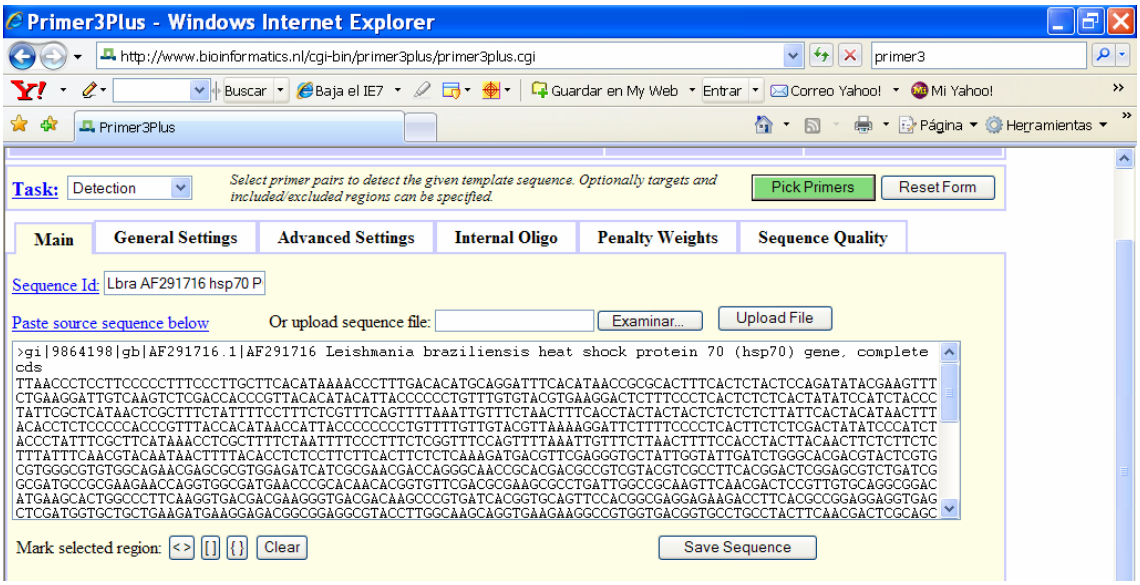
1. Open http://frodo.wi.mit.edu/ and go to Primer3plus interface



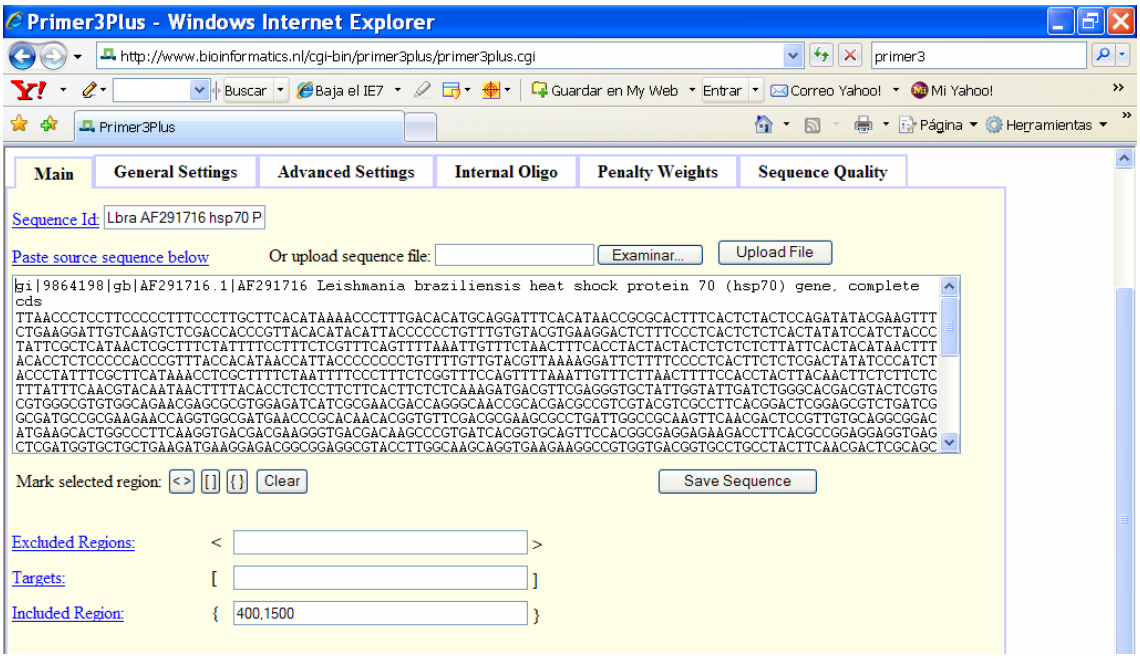
2. Be sure that you are on Task: Detection and paste your sequence (FASTA format) or Upload the File



3. We will use the *L. braziliensis* hsp70 sequence AF291716, which has been obtained in the previous exercise. Identify the sequence (Sequence ID) as Lbra AF291716 hsp70 PCR

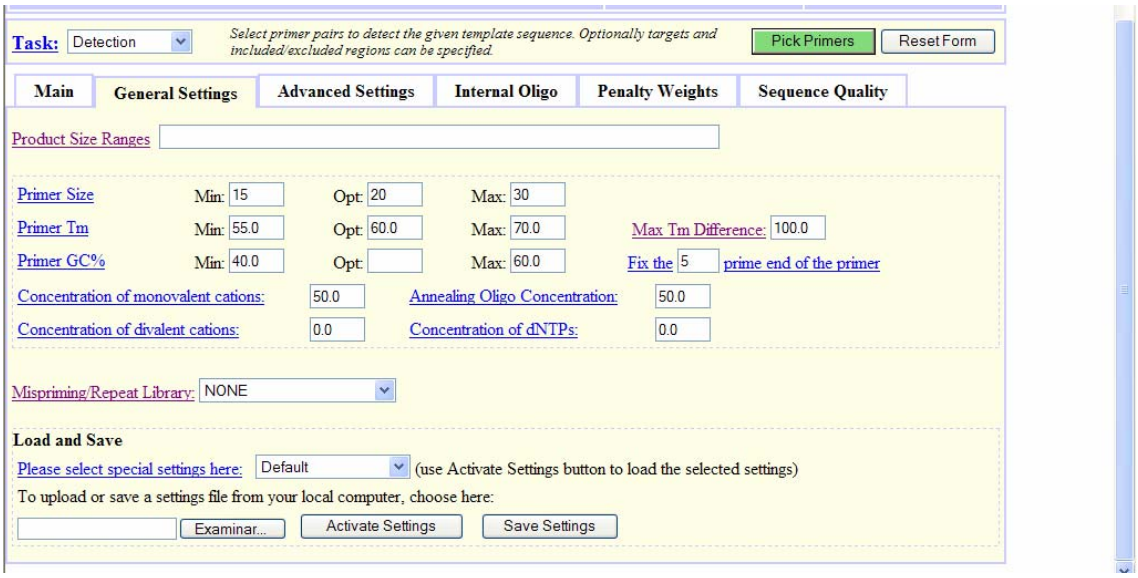


4. AF291716 sequence has 2566 bp, by alignment analysis of the sequences previously obtained we "now" that region on which we are interested is between the bases 400 and 1900. Thus in the Included Region text box we will write 400,1500, which means that we are interested on primers addressed to the region starting at base 400 and finishing at 1900 (approx. 1500 bp size).



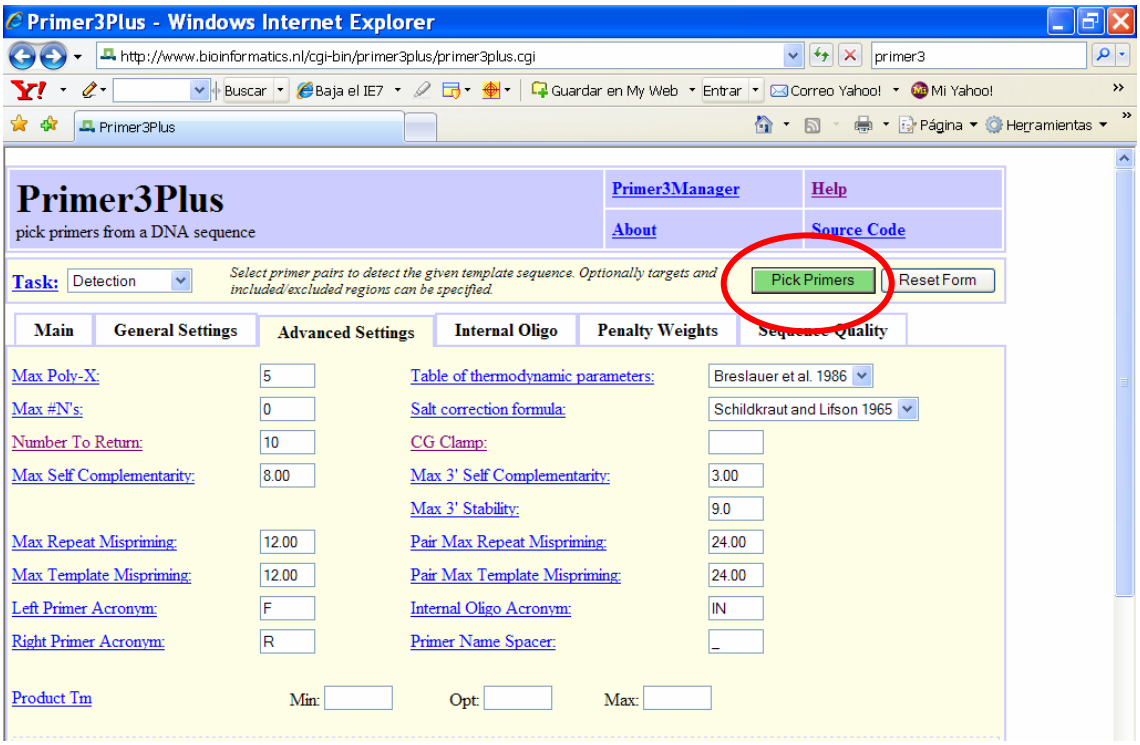
5. Now we will click on the General Settings tab and:

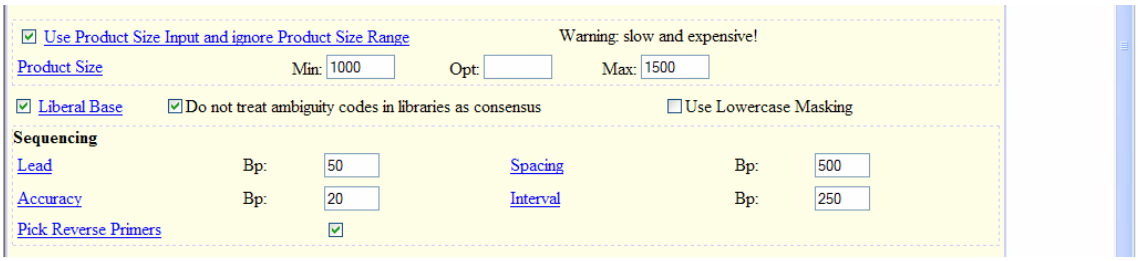
- delete Product Size Ranges
- set Primer Size Min: 15 and Max:30
- set Primer Tm Min: 55 Max: 70 Opt: 60
- set Primer GC% Min: 40 Max: 60
- Let the other settings as default by the system



6. Click on the Advanced Settings tab and:

- set Number tor Return (number of primers pairs) to 10
- delete the GC Clamp text box
- Mark Use Product Size Input and ignore Product Size Range
- set Product Size Min: 1000 Max: 1500
- let the other settings as default

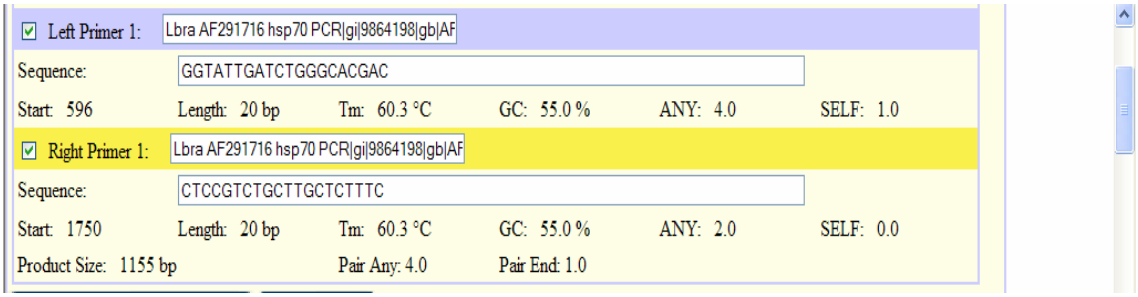




7. Click on Pick Primers.

You will obtain detailed information for the best 10 primer pairs selected by Primer3.

Carefully check which pair is most suitable for your objectives and keep in mind that you have to test their performance on the bench!!



Protocol 5.4 **Phylogeny**

Purpose	To construct and interprete phylogenetic trees

A. INTRODUCTION

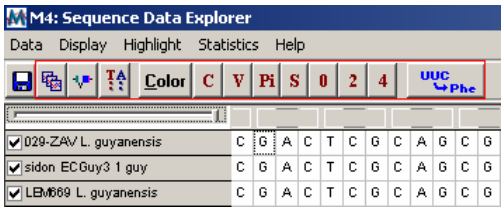
Reconstructing evolution is the primary goal of any phylogenetic analysis. Methods are numerous, and interpretation of resulting dendrograms is not trivial. This session aims at construction and interpretation of phylogenies based on the previously built HSP70 and ITS1 alignments, by using the basic and most common algorithms. Trees obtained for these two targets will be compared.

Sequences used: HSP70 and ITS1 alignments from previous workshop sessions

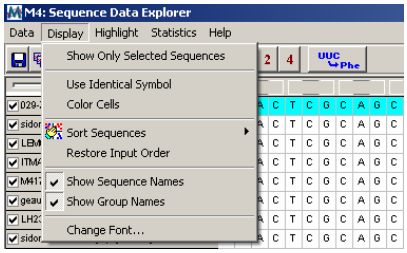
Programs used: MEGA (www. megasoftware.net)

Programs not used but excellent: PHYLIP (evolution.genetics.washington.edu/phylip.html)

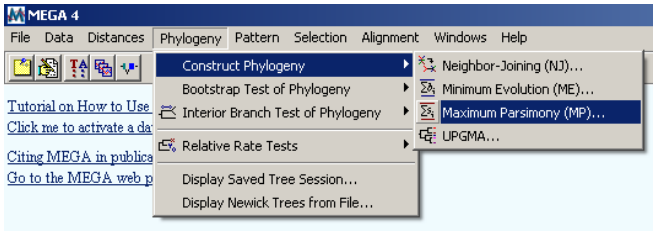
Method: ***Aligning hsp70 nucleotide sequences (joint exercise)***

1. Start MEGA 4.0
2. Open your HSP70 alignment in MEGA format from the File menu.
3. Explore the tabs
4. 

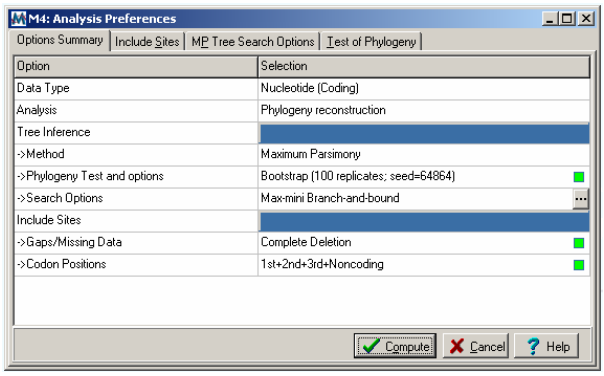
5. to familiarize yourself with data viewing, defining groups and reading frames.
6. Define the two *Leishmania* subgenera *L. (Leishmania)* and *L. (Viannia)* and add the respective sequences to these groups.
7. Make a sequence selection of the complete sequences, unchecking the boxes from the partial sequences.
8. Explore the options in the Display menu



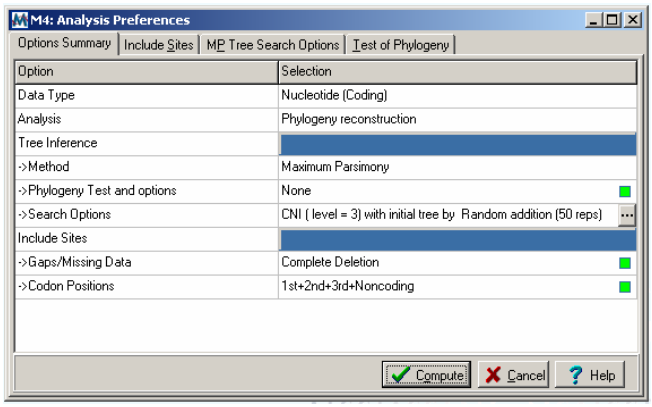
9. Identify parsimony informative sites after theoretical introduction on maximum parsimony.
10. Close the data viewer.
11. From the Phylogeny menu, select maximum parsimony trees:



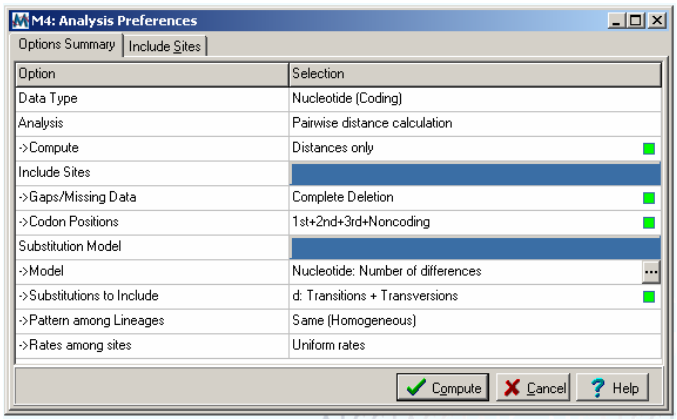
12. Use the following options and Compute:



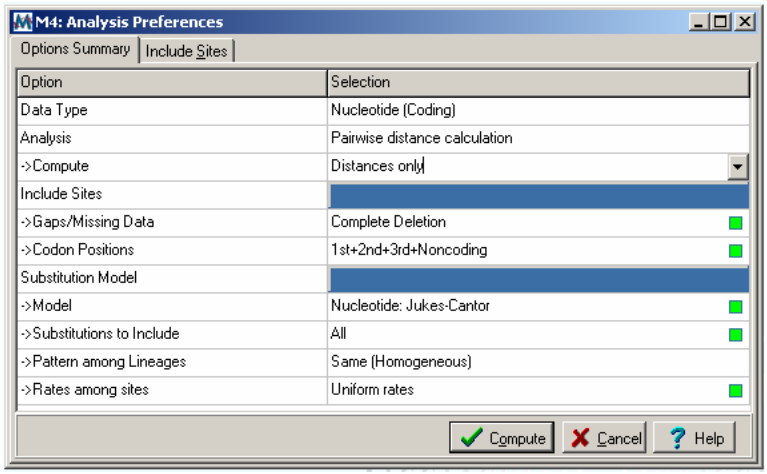
13. Wait a bit…
14. And some more…
15. Have a coffee…
16. Have another one…
17. Go powder your nose…
18. Be patient…
19. Go out for lunch…
20. Stop and abort the calculation.
21. Try again with the following options:



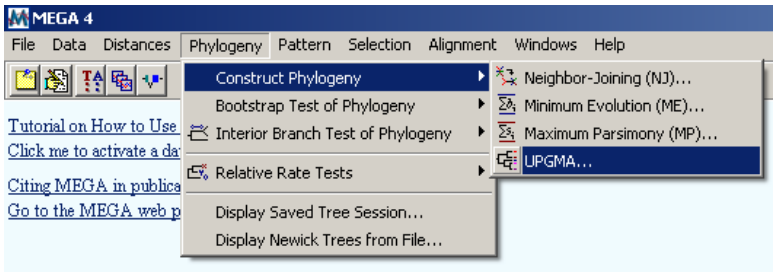
22. Look at the different output trees and compute a consensus. Play around with the different view options. Close the Tree explorer.
23. From the Distance menu, select Compute pairwise.
24. Compute with the following options:



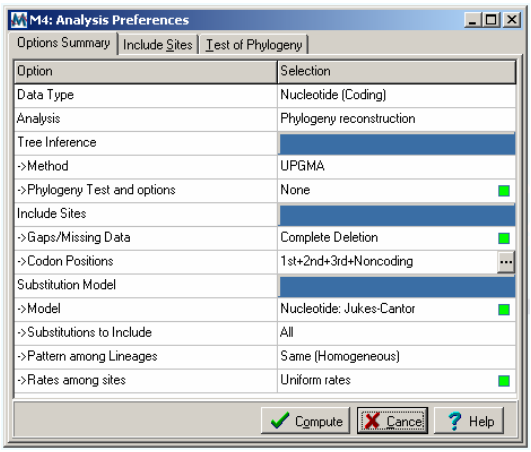
25. Repeat the calculation with the following options:



26. From the Phylogeny menu, select the following option:

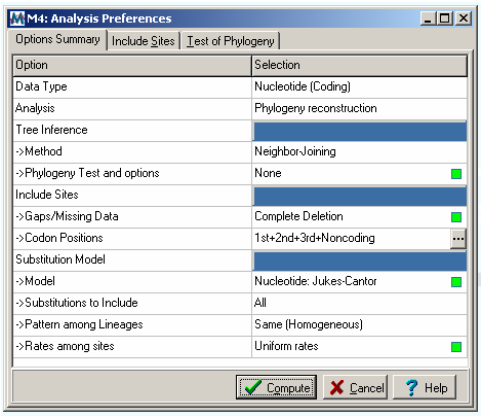


27. Calculate a UPGMA tree with the following parameters:



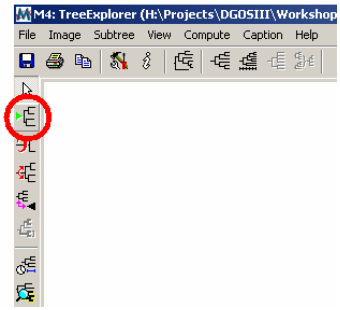
28. Observe the equal distance of all taxa to the root.

29. Calculate a Neighbor-Joining tree:

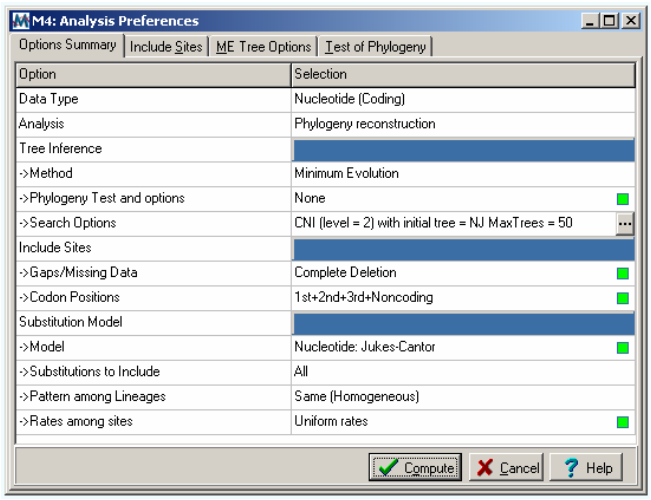


30. Observe the unequal distance to the root of all taxa.

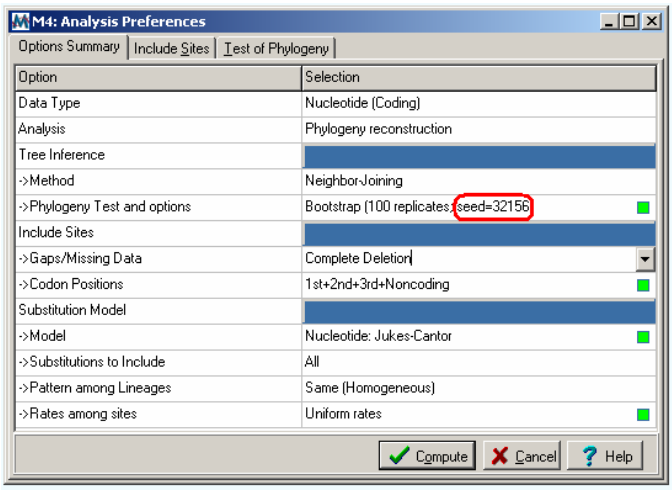
31. Re-root the tree on a branch:



32. Calculate a Minimum Evolution tree:



33. As for Maximum Parsimony, different topologies are obtained, and consensus trees can be built.
34. Build a Neighbor-Joining tree with the bootstrap option (you can pick a different seed number):



35. Re-root the tree and observe what happens with the bootstrap values.

Additional remarks:

5. A lot of different tree building options are available in MEGA, using different substitution models, tree construction algorithms, and confidence tests.

6. Trees can be built using different site selections, such as synonymous versus non-synonymous, with or without gaps, different codon positions, and also on the basis of protein sequences.

7. The better the chosen model describes the actual evolutionary mechanisms, the more reliable your phylogeny can be. It is, however, difficult to go back in time.

8. Beware of over-interpretation of confidence levels such as obtained with bootstrap: they only tell you something about the consistency in your data set with the method used, they don't tell you whether you correctly captured evolution.
9. The starting point of any analysis is a decent alignment.
10. Phylogenies start from the assumption that evolution is divergent, and works through a series of bifurcations from a common ancestor by accumulation of mutations. Unfortunately, evolution does not always treat us kindly, and the history of an organisms genome includes also horizontal gene transfers, duplications, DNA transpositions, and filtering by natural selection. It is essential to realize that the phylogeny of a piece of DNA or protein is not necessarily identical to the phylogeny of the organism it is derived from.

Exercises in individual groups:

4. Using your HSP70 alignment, build trees from protein sequences and synonymous sites only. Compare these with the trees from the joint exercise.
5. Build phylogenies from the ITS1 alignment, and compare these with the HSP70 phylogeny. In this alignment, different gap treating may have a big impact as they are numerous. As these sequences are non-coding, protein alignment is not possible to aid in identification of homologous sites.
6. From both HSP70 and ITS1 phylogenies, determine the species and if possible sub-species of the unknown sequences you received for the *in silico* analysis.

6. Multilocus sequence typing (MLST)

Protocol 6.1	**Analysis of MLST data**
Purpose	To type *Leishmania* species using the sequences of multiple housekeeping genes.

A. Introduction:

Multilocus sequence typing (MLST) is used primarily to type bacteria in the format of sequencing about 500bp of an average of 10 single copy housekeeping genes (Maiden et al 1998; Spratt 1999). You can find more information and databases for current MLST systems at http://www.mlst.net/. The same methodology has been adapted to some pathogenic fungal species and now for *Leishmania* (Mauricio et al 2006; Zemanova et al 2006). In the system being developed we have so far five genes that code for enzymes often used for enzyme electrophoresis typing, and which are fully sequenced (from 1000 to 2000 bp). Genes are amplified by PCR using primers external to the coding region and sequenced using internal primers. This methodology has several advantages over other methods, as well as drawbacks. Compared with enzyme electrophoresis, for example, it has a greatly increased discriminatory power, it can be applied to biological samples instead of cultures, it does not require reference strains and it is portable (comparable between labs and through databases). The last two features are also an advantage over PCR-RFLPs. MLST is less sensitive and has less discriminatory power than microsatellite typing, but it is more easily applicable across different species of *Leishmania* and to unknown samples. Its analysis is also potentially more straightforward. MLST has so far been used to identify genetic groups in the *L. donovani* complex and recombination.

The MLST system for the sub-genus *L. (Viannia)* is on its final stages of development. However, 5 genes have already been fully tested and chosen to integrate the fugure 10-15 system. These are: *MPI*, *ACON*, *NH2*, *MDH* and *G6PDH*. Other genes being tested are: *ASAT*, *NH2*, *C20070*, *C320290*, *GPI*, *NH1*, *ICD*, *PGM*, *6PG*, *ME* and *HK*. The final systems and conditions will be published shortly.

Technically, MLST requires knowledge of PCR and DNA sequencing, which have been covered elsewhere.

Here, participants will practice MLST analysis on 3 gene sequences that have been published for the *L. donovani* complex, which includes *L. infantum* (syn. *L. chagasi*).

B. MATERIALS:

Samples: Edited chromatograms for 10 selected strains of the *L. donovani* complex.

For PCR conditions and primers for the full 10 gene MLST system for the *L. donovani* complex, see references Mauricio et al (2006) and Zemanova el al (2007).

Equipment and software:

Computer with internet access

BioEdit (www.mbio.ncsu.edu/BioEdit/BioEdit.html)

MEGA (www.megasoftware.net)
SplitsTree4 (http://www.splitstree.org/)

PHASE (http://stephenslab.uchicago.edu/software.html)

Text editor program (Word, for example)

Spreadsheet program (Excel, for example)

C. METHODS

All programs will be available on the Desktop, as well as a folder with the initial data files.

a) Assemble consensus sequence for each gene and for each sample from partial sequences obtained from internal primers.

Step 1 – Create alignment file - Open BioEdit Sequence Alignment Editor. From the File menu, open New Alignment (or Control+N). Save it in your chosen name (Note: keep saving after each main step).

Step 2 – Import relevant raw sequences - From the File menu, open Import, then Sequence Alignment File. Select all the electropherograms for this alignment. You may need to select All files or Abi formats

Step 3 – Clean sequences - Remove stretches of NNNN from the beginning and the end.

Step 4 – Assemble sequences - Select sequences by clicking on their names on the left hand side. From the Accessory Application menu, choose CAP Contig Assembly Program. In the window click Run Application button (you can change the settings first). You may need to press Enter once the program is finished (it reads SHOW on the command prompt window). If the alignment is successful you should obtain a new file (window) with the sequences aligned and an additional sequence called "Contig-0". Rename the consensus sequence with a short easily identifiable name.

Alternative: if you are not able to obtain a single contig sequence with this method, import a reference sequence for the entire gene sequence (from GeneBank or GeneDB or previous alignments), select the reference strain and one of the raw sequences then from the Accessory Application menu, choose ClustalW Multiple Alignment. You may need to reverse some sequences prior to alignment (Select sequence name, go to the Sequence menu, choose Nucleic Acid, then Reverse Complement; or Shift+Control+R). Repeat the process for all raw sequences. You can copy the aligned raw sequence

from one window to the other with Control+F8, then paste with Control+F9. Once all aligned sequences are imported to the alignment file, you can create a consensus sequence by copying and pasting one (Select sequence name, then Control+C, then Control+V) and then adding the missing regions from the other aligned sequences.

Step 5 – Check sequences for anomalies: gaps, misalignments, incorrect multibase codes (a glitch does not reverse multibase codes in CAP). If necessary go back to the chromatograms.

Note: Instructions on how to edit raw sequences are given elsewhere.

b) Align consensus sequences from all samples for each gene and against sequences retrieved from GeneBank/EMBL

Step 1 - Create alignment file - Open BioEdit Sequence Alignment Editor. From the File menu, open New Alignment (or Control+N). Save it in your chosen name (Note: keep saving after each main step).

Step 2 – Import relevant consensus sequences. – One way is to copy them from each individual alignment file, as they are finished by copy the aligned raw sequence from one window to the other with Control+F8, then paste with Control+F9. Alternatively you can copy a sequence from the file of origin by going to the Edit menu, then Copy Sequences to Clipboard (FASTA format), and on the destination file by going to the File menu, then Import from Clipboard.

Step 3 – Align sequences – Select the names for all sequences, then from the Accessory Application menu, choose ClustalW Multiple Alignment.

Step 4 – Select coding region (remove primer sequences) – For the majority of targets, the PCR primers are outside the coding region, so by selecting the coding region their sequences are automatically eliminated. From the Edit menu, choose Search then Find Next ORF. The open reading frame (ORF, or coding region) will be highlighted in black. You can vertically select the regions outside and press the Delete button. If you need to select primer sequences, you can go to the Edit menu, choose Search then Find (Control+F).

c) Concatenate the sequences for all samples for all genes

Step 1 – All alignment files for each target should have the same samples and in the same order.

Step 2 - Save one of the alignment files with a new name. Go to the File menu, then choose Append Alignment. Choose the alignment file for the target you want to add.

Step 3 – Repeat procedure for each target at a time until all targets have been added.

d) Identify SNPs

Step 1 – Make sure all alignment files from BioEdit were saved in FASTA format.

Step 2 – Open MEGA, then from the File menu, choose Open Data (F5). Select the file.

Step 3 – Convert FASTA file to MEGA format – From the Utilities Menu on the window, choose Convert to MEGA format (Control+M). Press OK on the next two windows. You should now have a new tab with the alignment in MEGA format. Save the file. Close windows.

Step 4 – Open MEGA format file – From the File menu, choose Open Data (F5). Select the file. On the following windows, click OK/Yes for default. It should be: 1) Nucleotide sequences, 2) Protein Coding

sequence data, 3) Standard Genetic Code. You should now have a window with an alignment view.

Step 5 – Identify SNPs – From the Highlight menu, choose Variable sites (or the V button on the top toolbar).

Step 6 – Create SNP only file – From the Data menu, choose Export Data. Make sure to select Only Highlighted Sites (at the bottom of the window). It is also advisable to choose the option of writing site numbers for each site, so that you have a record of the location of each SNP. You can select different file types.

e) **Determine phase of heterozygous strains, using the program PHASE** (Stephens et al 2001)

This program implements a Bayesian statistical method for reconstructing haplotypes from population genotype data. You should be aware that it provides a probability result rather than the actual phase for sequencing or microsatellite data, so you should be very careful when interpreting results. However, it is very useful if physical linkage between SNPs or microsatellites is impossible to determine.

Step 1 – Open an MS-DOS window. Go to the Windows Start Menu, then choose Run, type COMMAND in the prompt and press OK. Alternatively, also from the Start Menu, choose Command Prompt (probably in the Programs, then Accessories menus.

Step 2 – Your PHASE program should be saved in the PHASE directory in your C:\ drive. To get there, you may need to write "cd.." at the prompt until you only see C:\. Then type "cd PHASE". You should obtain C:\PHASE\

Step 3 – Prepare the input file. Open the file you prepared with SNPs only from your sequence data for one of the gene targets. Prepare an input file in the following format:

General description:	Example:	LocusType(i):
NumberOfIindividuals NumberOfLoci P Position(1) Position(2) Position(NumberOfLoci) LocusType(1) LocusType(2) ... LocusType(NumberOfLoci) ID(1) Genotype(1) ID(2) Genotype(2) . . . ID(NumberOfIndividuals)] Genotype(NumberOfIndividuals)	3 5 P 300 1313 1500 2023 5635 MSSSM #1 12 1 0 1 3 11 0 1 0 3 #2 12 1 1 1 2 12 0 0 0 3 #3 -1 ? 0 0 2 -1 ? 1 1 13	**(a)** S for a biallelic (SNP) locus, or biallelic site in sequence data. **(b)** M for microsatellite, or other multi-allelic locus (eg tri-allelic SNP, or HLA allele). The default is that this denotes a microsatellite locus with stepwise mutation mechanism. Genotype (i) for the ith individual. Given on two rows: one allele on the 1st row, and one on the 2nd row. It does not matter which allele is entered on each row. For biallelic loci, any two characters (e.g. A/C, G/T, 0/1) can be used to represent the two alleles, and they do not need to be separated by a space. Missing alleles at SNP loci should be entered as ?. For multiallelic loci a positive integer must be used for each allele (number of repeats at microsatellite loci), and data for each locus should be separated by a space. Missing alleles at multiallelic loci should be represented by -1.

In this case, your data should all be coded as S (for a biallelic locus) and you should code it as A, C, G or T.

Step 4 – Run PHASE. At the DOS prompt, type PHASE <filename.inp> <filename.out> <number of iterations> <thinning into interval> <burn in>

For example for the gene ASAT you may prepare an ASAT.inp file, then type:

PHASE ASAT.inp ASAT.out 100 1 100 (these 3 last numbers are the default parameters, so you can miss them out)

As it runs, it keeps you informed, until you obtain a new PHASE prompt.

Step 5 – Results. You should obtain a summary results file, with the name that you specified.

In the output file, you should obtain a recreation of the input file with the estimated phase of the alleles. Uncertain phase are indicated by (), wheras uncertain genotypes are indicated by [], with p (for phase) and q (for genotypes) at the default value of 0.9. The list of probabilities for each uncertain phase call are listed at the end of the output file.

You also obtain 6 other output files with more detailed results.

Step 6 – Evaluate results. You can check the .monitor output file to examine the goodness-of-fit and the .freqs output file to check that the estimates accross each run are consistent. You can test the consistency of your results by increasing the number of iterations or the thinning interval. If the different runs keep giving different results, you can select the results from the run with the highest average value for the goodness of fit. For the purpose of this exercise, you can try another run with a higher number of iterations and compare with the results from the first run.

When you are happy with the results you can use the estimated phase to code your diplotype data as haplotypes for the next analyses, particularly for bifurcating trees.

f) Code diplotypes

Step 1 – Go to the website http://linux.mlst.net/nrdb/nrdb.htm. This opens the online program NRBD that selects non-redundant sequences.

Step 2 - Paste full alignment in FASTA or MEGA format. Click the Submit button. You should obtain an alignment of unique sequences with concatenated titles.

Step 3 - You should assign a number or letter for each unique sequence. Heterozygous sequences should be given a separate code. Where phase can be determined for heterozygous strains you can assign two codes.

Step 4 – Prepare a table listing each diplotype per sample and per target.

Step 5 – Assign an MLST code for each unique combined diplotype.

g) **Conduct network analysis of sequences on the program SplitsTree4** (Huson & Bryant, 2006)

Step 1. Make sure you have a file with aligned non-redundant sequences. SplitsTree4 accepts several formats (FASTA, Nexus, PHYLIP, Clustal, for example, as well as distance files).

Step 2. Open SplitsTree4. From the File menu, select Open. Choose your file and click Open. The program automatically produces a network. At the bottom of the window you should find information about the network produced. The default is likely to be a SplitsTree.

Step 3 – Choose distance measure. From the Distances menu, you can choose different statistics. The default will be uncorrected P. Other distances incorporate factors to produce more realistic models. However, the K2P (Kimura-2-Parameter) is very robust and widely used. Try this one first. Keep the default settings shown on the new window and click the Apply button.

Step 4 – Choose Network building method. From the Networks menu, choose Neighbor-Net if not already selected, or try another method for comparison. Keep default settings and click the Apply button.

h) **Manipulate and analyze networks and trees. Explore different analyses and presentations**.

Step 1. Produce bifurcating trees. From the menu Trees, choose a tree building method. Neighbor-Joining (NJ) is recommended.

Step 2. Modify presentation of the trees and networks. From the Draw menu select a presentation style. For example, Phylogram for the NJ tree.

Step 3. Modify view settings for the trees and networks. From the View menu select, for example, Rotate Right and then Left. You can also use the arrows keys on your keyboard: Right and Left for rotating, Down for zooming in and Up for zooming out. You can also click on branches, which become highlighted in red. Then click on a node for that branch to move it around. If you click on a branch that forms part of a split, you'll notice that all splits will be highlighted. You can move splits and sequence names to make the network easier to view.

Step 4. Remove sequences from tree/network. You may want to focus on a smaller group. From the menu Data, select Filter Taxa. On the left hand side, a box lists shown taxa. Select taxa you want to remove and click the button Hide in the centre. The sequence should appear on the right hand box. Click the button Apply.

Note: you can save trees and networks by going to the menu File, then selecting Export Image.

i) **Analyze networks for evidence of recombination**.

Step 1. Notice presence of splits. A large number of splits suggests recombination. Either within the entire group or within smaller groups.

Step 2. Notice where taxa appear. Taxa at the end of long branches suggest absence of recombination, whereas taxa sitting on intermediate branches are putative hybrids. Try to identify instances of each of these cases in your dataset.

Step 3 – Test for recombination. From the Analysis menu, chose Conduct Phi Test of Recombination. You can do the analysis for the entire dataset and for subsets.

REFERENCES:

Huson D.H. & Bryant D. (2006) Application of Phylogenetic Networks in Evolutionary Studies, *Mol. Biol. Evol.*, **23**:254-267,

Maiden, M.C.J., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant,

D.A., Feavers, I.M., Achtman, M., & Spratt, B.G. (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. USA*, **95**, 3140-3145.

Mauricio IL, Yeo M, Baghaei M, Doto D, Pratlong F, Zemanova E, Dedet J-P, Lukes J, Miles MA (2006) Towards multilocus sequence typing in the *Leishmania donovani* complex: resolving genotypes and haplotypes for five polymorphic metabolic enzymes (ASAT, GPI, NH1, NH2, PGD). *Int. J. Parasitol.* **36**:757-69.

Spratt, B.G.Multilocus sequence typing: molecular typing of bacterial pathogens in an era of rapid DNA sequencing and the Internet. Curr. Opin. Microbiol. **2,** 312-316, 1999.

Stephens, M., Smith, N., & Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *Am. J. Human Genetics,* **68**, 978-989

Zemanova E, Jirku M, Mauricio IL, Horak A, Miles MA & Lukes J. (2006) Analysis of genetic polymorphism of the *Leishmania donovani* complex in five metabolic enzymes. *Int. J. Parasitol.* **37**: 149-160

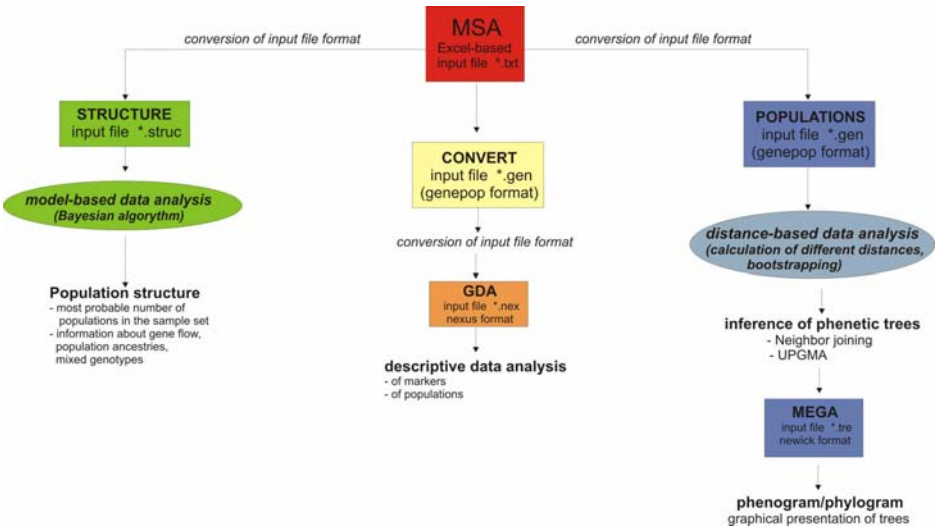
7. Analysis of MLMT data

Protocol 7.1 Population genetic analysis of *Leishmania* strains based on MLMT data

Purpose To construct phylogenetic trees based on microsatellite distance measures, to assign individuals to population by using a Bayesian statistics-based approach, to estimate the genetic isolation of the populations defined and to perform a descriptive analysis for the populations and the microsatellite loci.

A. Introduction:

Microsatellite profiles consisting of the repeat numbers for the different microsatellite loci that have been assembled for every strain under study will analysed by distance based methods and by models based on Bayesian statistics. Microsatellite genetic distances based on the "Chord distance" or "proportion of shared alleles" measures will be calculated using the program POPULATIONS and distance trees will be constructed using MEGA4. The Bayesian statistics-based method implemented in STRUCTURE uses patterns of allele frequencies for identification of distinct subpopulations and determines fractions of the genotype for each strain that belong to each subpopulation. It was shown to accurately infer individual ancestries and to provide information on population relationships and history. Fstatistics which allows for testing for the genetic variation among populations is calculated by MSA software and descriptive analysis (mean number of alleles, observed and expected heterozygosity) of microsatellite loci and of the different populations obtained is performed by GDA.



Method A. ***Preparation of data and output formats using the MSA software package.***

MSA (Microsatellite Analyzer) provides a variety of data and format outputs which are useful for analysis of microsatellite data.

Prepare an excel file with your data as following:

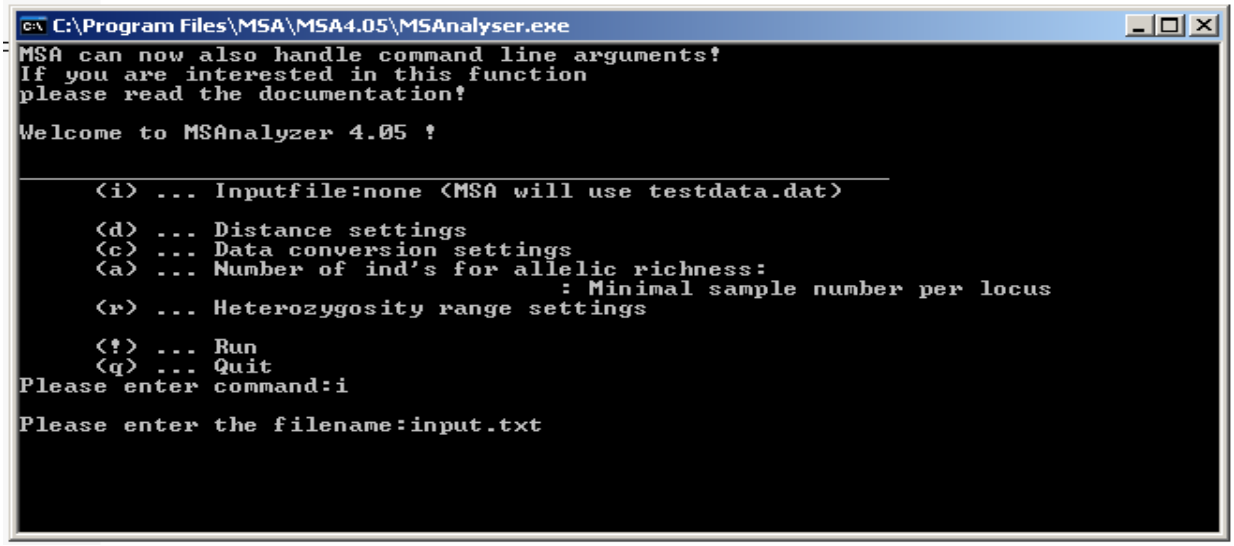
Two format column							
2			2*	2	2	2	**Repeat size**
			92*	92	62	62	**Length of flanking region**
			Lm2TG*	Lm2TG	TubCA	TubCA	**Locus names**
DON1	d	1	110*	110	82	82	**Fragment size**
DON39	d	1	110	110	82	82	
DON45	d	1	110	110	82	82	
DON28	d	1	110	110	82	82	
DON53	d	1	110	110	82	82	
DON56	d	1	110	110	82	82	

*Hint:

Marker Lm2TG contains a dinucleotide repeat (2), a the flanking region of the following has 92 bp. If the fragment size is 110 which equals to ((repeat numbers x 2) + flanking region); the repeat number in this example is 9. d=outbred individual (h=inbred), 1=group number, here we consider all individuals as belonging to one group. Missing data are indicated by empty cells, -1, nd, etc.

Convert the excel file into text file format (*.txt) before starting MSA.

Copy your input file into the folder of the MSA program, then press the MSA starting icon MSAnalyser.exe to get the following starting screen:



Type the command **(i)** and enter the name of your input file as shown. Run **(!)**, then close the command line window and go to the MSA folder where the output folder named input.txt_MSAresult00 is found This folder which contains many **input files** for other programs and results **as allele counts and allele frequencies**.

Method B. ***Construction of distance matrices using POPULATIONS***

Open the MSA folder input.txt_MSAresult00 and select the following files:

Formats&Data → Genepop.gen

To start the program POPULATIONS open the program folder and copy the genepop.gen file to it. Open the command line window using populations.exe :

```
\\Mikro_cd\VOL1\Domainen_Benutzer\amro\an_diesen_Benutzer\All programs from Katrin\Program...

*************************************
*  Populations 1.2.28   CNRS UPR9034  *
*       langella@pge.cnrs-gif.fr      *
*      http://www.cnrs-gif.fr/pge     *
*************************************



Main menu

0) Exit
1) Compute individuals distances + tree
2) Compute populations distances + tree
3) Allelic frequencies, Fstats
4) Build phylogenetic tree with a distance matrix
5) Formated output for other softwares
6) Choose the output format of matrix
7) Structured populations

Your choice:
```

type **(1)** for computing individual distances, press enter. Type the name of the input file (in this example genepop.gen), press enter to get to this screen:

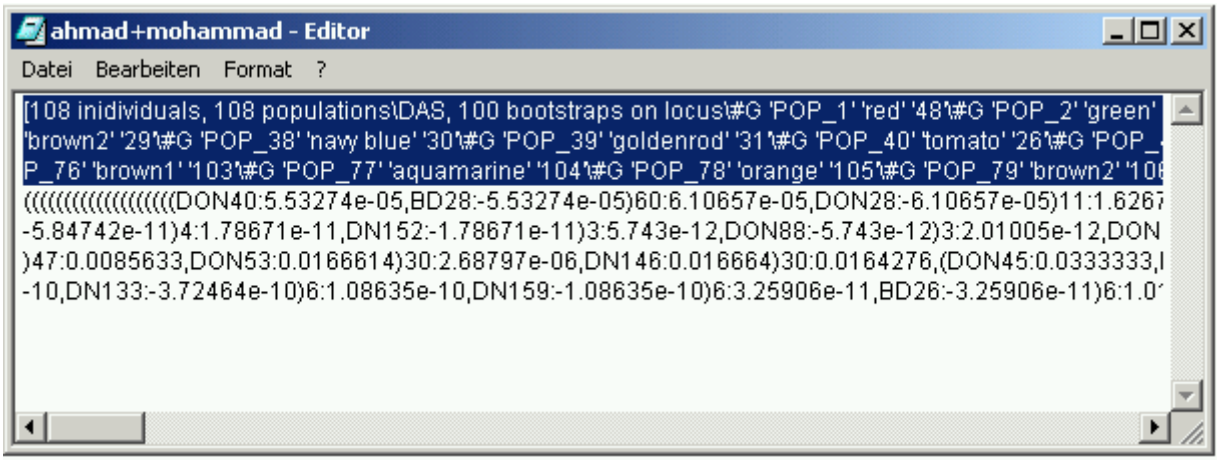
```
\\Mikro_cd\VOL1\Domainen_Benutzer\amro\an_diesen_Benutzer\All programs from Katrin\Program...
1) Compute individuals distances + tree
2) Compute populations distances + tree
3) Allelic frequencies, Fstats
4) Build phylogenetic tree with a distance matrix
5) Formated output for other softwares
6) Choose the output format of matrix
7) Structured populations

Your choice:
1
Name of input file (Populations or Genepop format) ?
genepop.gen


Reconstructing phylogenetic trees with individuals


0) back
1) Get distance matrix between individuals
2) Phylogenetic tree of individuals
3) Phylogenetic tree of individuals with bootstraps on locus
4) Choose locus to compute distances (default: all)

Your choice:
```

Type **(2)** to calculate the phylogenetic tree of individuals without bootstrap, or **(3)** for the phylogenic tree of individuals with bootstraps on locus. Enter and choose on the following screen **(3)** for Chord distance Cavalli-Sforza and Edwards, Dc, or **(5)** for "proportion of shared alleles" distance measure DAS (also called Dps). Here we select the Dc measure.

Enter and select either **(1)** UPGMA or **(2)** Neighbor Joining for tree calculations in the next screen. We will calculate a neighbor-joining tree in this session. If the bootstrap option is used it is needed to type the number bootstraps wanted. In this example we selected 100.

Type an output file name. The tree calculation process will take about 15 seconds depending on to your sample size.

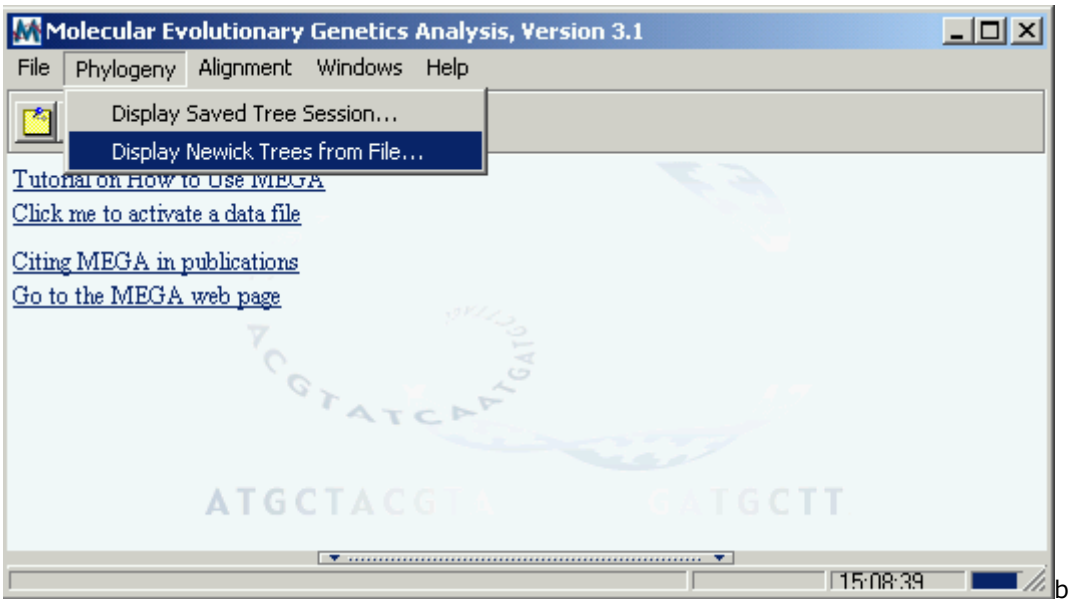
Close the command line screen and go to the program folder of **POPULATIONS** where the output file can be found. Open this file using the program **EDITOR** and delete the lines before the brackets as shown in the following screen:



Then save the file as text file and switch to the program **MEGA 4** in order to draw your tree.

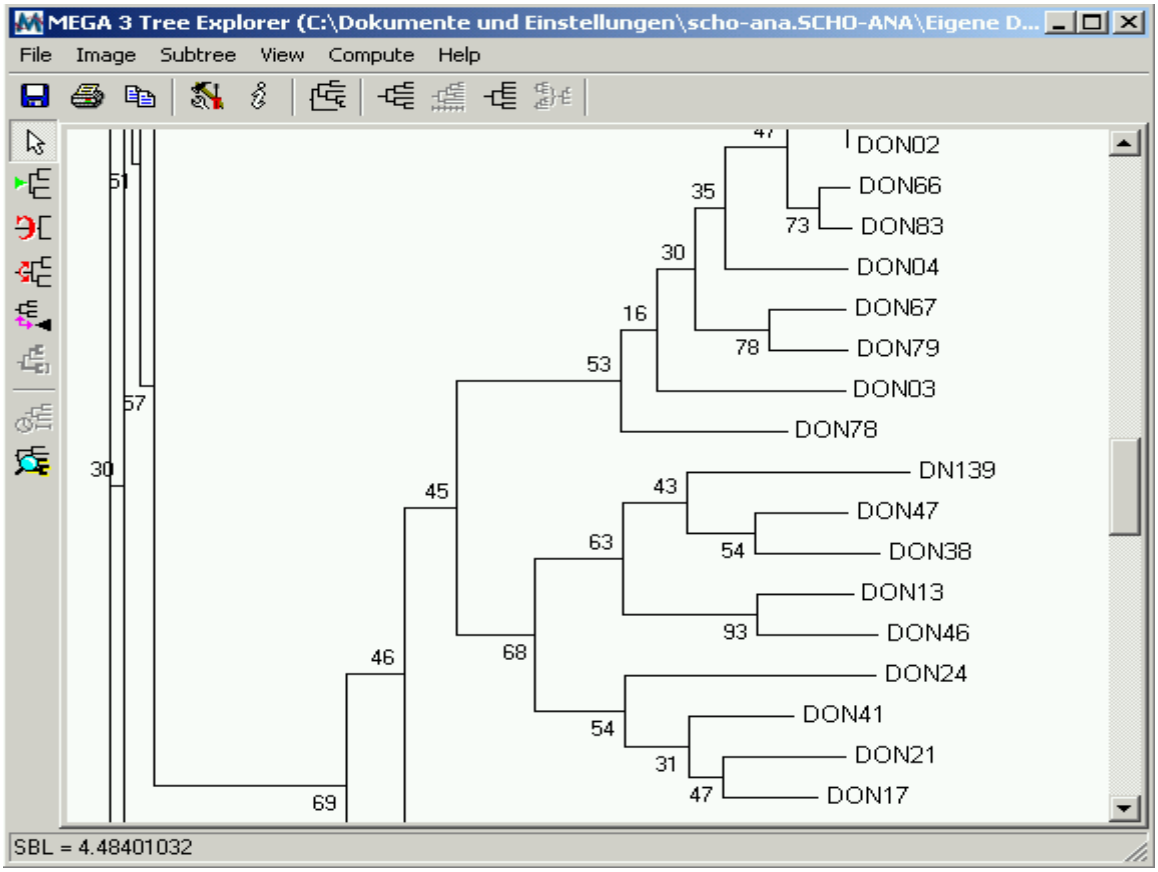
Method C: ***Drawing distance trees using MEGA 4***

Open **MEGA 4** and select 'Phylogeny', then select the second option 'Display Newick Trees from File':



b

Go to the folder of **POPULATIONS** and select your outputfile.The tree will appear as follows:

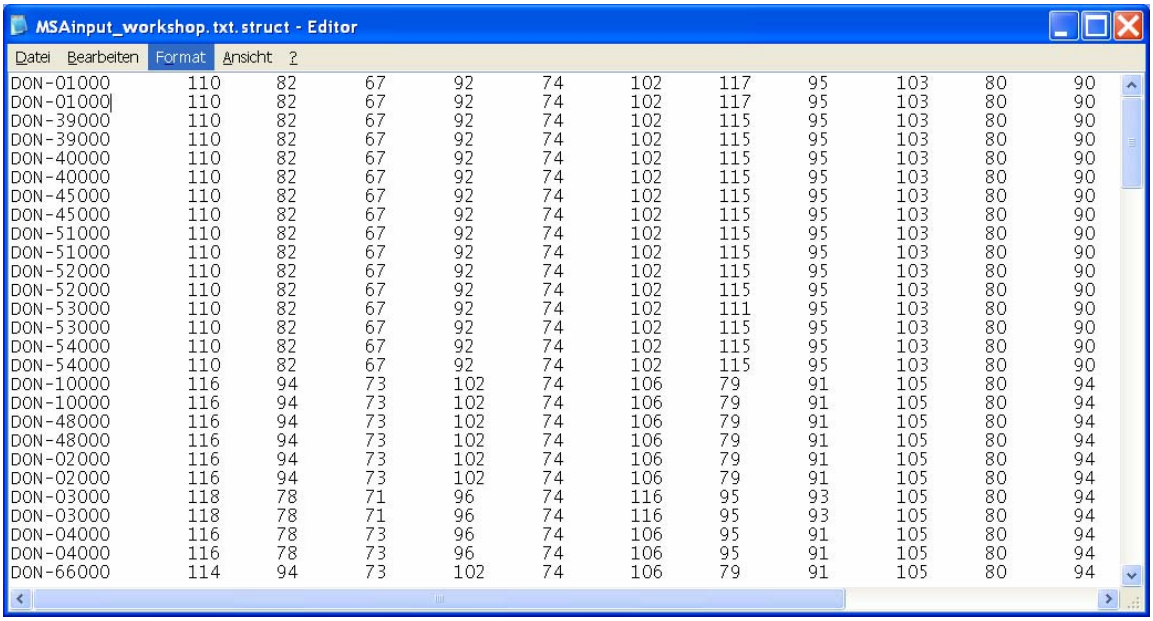


MEGA 4 allows for rooting the trees and changing the format, colors, appearance etc. by pressing the small icons at the top or at the left side of the screen. For further graphical applications and changes you can export the tree as graphic in *.emf (enhanced metafile) format.

Method D: ***Inference of population structure using STRUCTURE***

The program **STRUCTURE 2.1** is used for studying the nature and extent of genetic variation within and between populations.

To obtain the **STRUCTURE** input file open the MSA folder again and got to the folder input.txt_MSAresult00. There you will find a folder Formats&Data from where you can select the *.struct file which should look as follows:

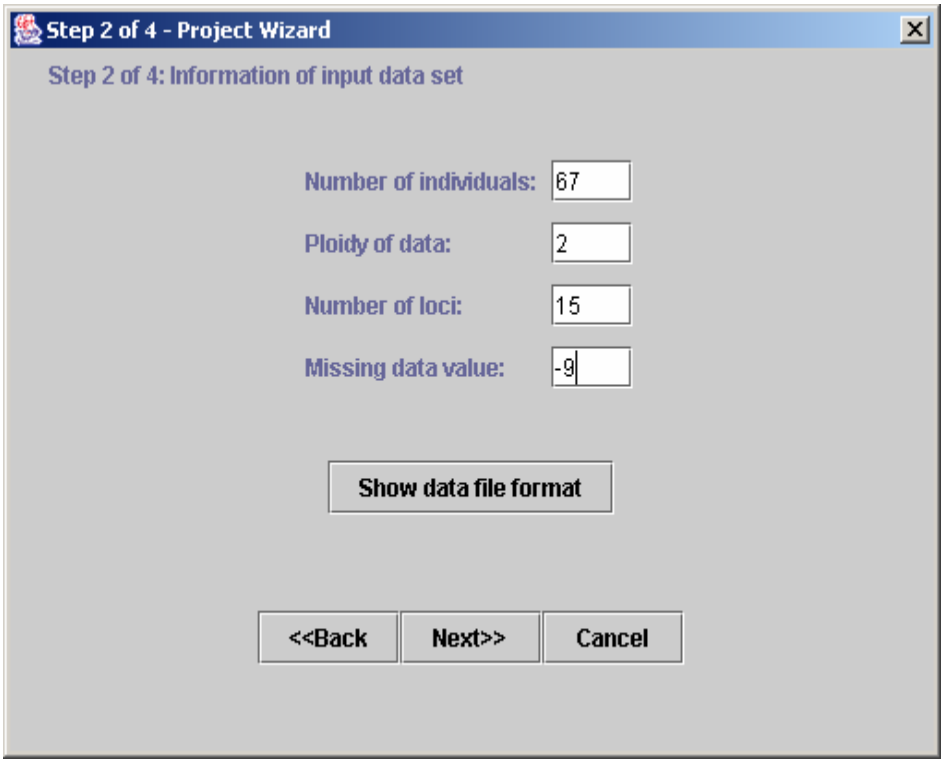


MSAinput_workshop.txt.struct - Editor

Datei Bearbeiten Format Ansicht ?

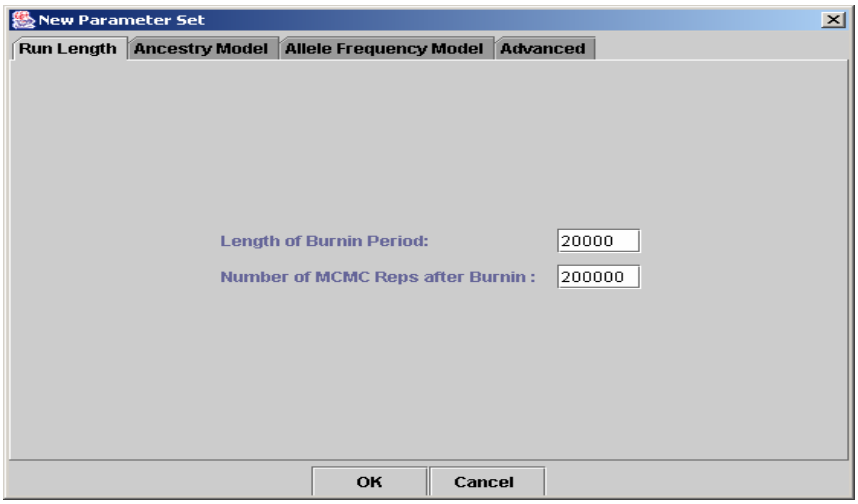
DON-01000	110	82	67	92	74	102	117	95	103	80	90
DON-01000	110	82	67	92	74	102	117	95	103	80	90
DON-39000	110	82	67	92	74	102	115	95	103	80	90
DON-39000	110	82	67	92	74	102	115	95	103	80	90
DON-40000	110	82	67	92	74	102	115	95	103	80	90
DON-40000	110	82	67	92	74	102	115	95	103	80	90
DON-45000	110	82	67	92	74	102	115	95	103	80	90
DON-45000	110	82	67	92	74	102	115	95	103	80	90
DON-51000	110	82	67	92	74	102	115	95	103	80	90
DON-51000	110	82	67	92	74	102	115	95	103	80	90
DON-52000	110	82	67	92	74	102	115	95	103	80	90
DON-52000	110	82	67	92	74	102	115	95	103	80	90
DON-53000	110	82	67	92	74	102	111	95	103	80	90
DON-53000	110	82	67	92	74	102	115	95	103	80	90
DON-54000	110	82	67	92	74	102	115	95	103	80	90
DON-54000	110	82	67	92	74	102	115	95	103	80	90
DON-10000	116	94	73	102	74	106	79	91	105	80	94
DON-10000	116	94	73	102	74	106	79	91	105	80	94
DON-48000	116	94	73	102	74	106	79	91	105	80	94
DON-48000	116	94	73	102	74	106	79	91	105	80	94
DON-02000	116	94	73	102	74	106	79	91	105	80	94
DON-02000	116	94	73	102	74	106	79	91	105	80	94
DON-03000	118	78	71	96	74	116	95	93	105	80	94
DON-03000	118	78	71	96	74	116	95	93	105	80	94
DON-04000	116	78	73	96	74	106	95	91	105	80	94
DON-04000	116	78	73	96	74	106	95	91	105	80	94
DON-66000	114	94	73	102	74	106	79	91	105	80	94

Open **STRUCTURE 2.1**, select '**File**' and then **'New Project'**. Enter the name of your new project and select your input file. Click **'Next'** and enter the number of individuals, the ploidy (2), the number of loci and the code for missing values (-9) as shown on this screen:



Click **'Next'**. In the '*step 3 screen'* don't add anything. Click **'Next'** again to open the '*step 4 screen'* where you should select Individual ID for each individual, then press Finish to get to a new screen and there on Proceed.

After this the small screens will disappear and a big table will show up on the main screen of the program. There click on **'Parameter set'** on the upper tool bar and select **'new parameter set'**:

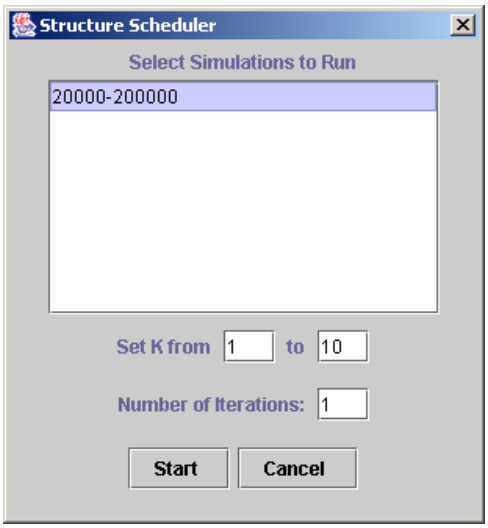


Type in the length of burn-in (usually between 10000-20000) and the number of MCMC repetitions after burn-in (usually between 100000-200000). Choose also the appropriate 'Ancestry model', usually 'admixture model'. All other values are default values.

Press on **OK**, and type the new parameter set (20000-200000) on the following small screen:

Press **OK,** go back to the main tool bar, select **'Project'** and then **'Start Job'.** Define the numbers of K and of iterations. K represents the number of populations that will be simulated and tested. It depends on the nature and origin of organisms tested and usually varies between 1 and 10. For a first analysis use one iteration.

If you wish to calculate delta K (see below) you have to run at least 10 iterations per each K.



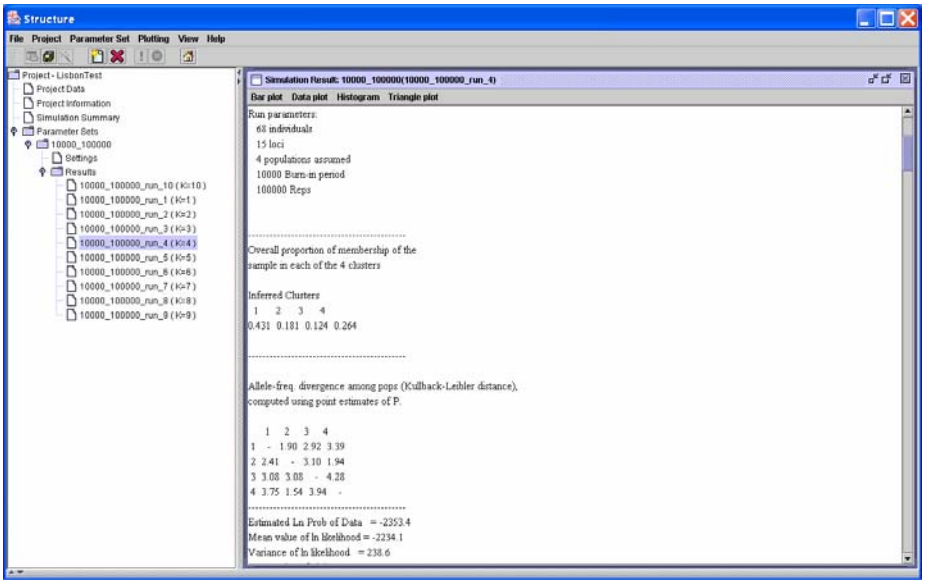
K represents the number of populations that will be simulated and tested. It depends on the nature and origin of organisms tested and usually varies between 1 and 10. For a first analysis use one iteration.

If you wish to calculate delta K,s ee below, you have to run at least 10 iterations per K.

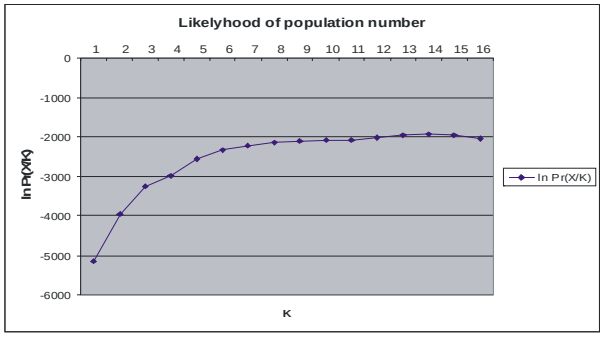
Press **'Start',** the job may take a long time depending on your data size and the number of iterations.

Press **'Start',** the job may take a long time depending on your data size and the number of iterations.

To show **STRUCTURE** results for each tested K value click on the result folder for the respective K value in the left part of the screen.



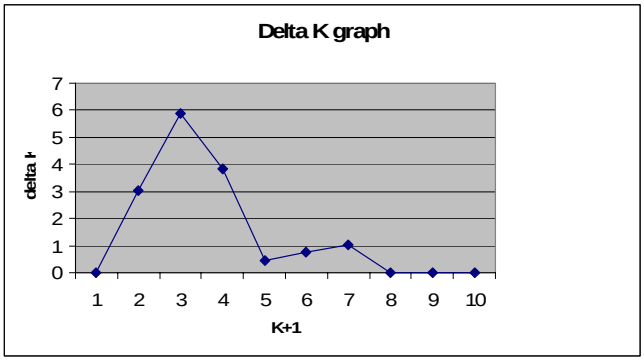
In this example the results for K = 4 are shown, including the values for the estimated ln probability and the mean value of ln likelihood. The latter are used for each K to draw the "likelihood of population number graph" by **EXCEL**. When the derived Gaussian graph reaches a plateau, the value of *K* captures the major population structure in the data set.



The number of population can be obtained more precisely by calculating the delta (Δ) K (first derivative of K) values according to *Evanno et al. 2005*. For this 10 iterations are needed for each K to calculate Mean and Standard deviation values. ΔK is estimated as the mean of the absolute values of $L''(K)$ averaged over 10 runs divided by the standard deviation of $L(K)$, $DK = m(|L''(K)|)/s[L(K)]$, which expands to

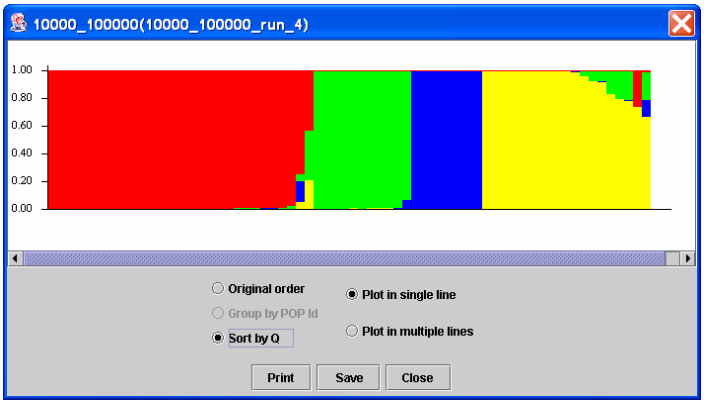
$\Delta K = m(|L(K + 1) - 2\,L(K) + L(K - 1)|)/s[L(K)]$.

For our example (K=4) the following graph will be obtained:



The maximum of the graph shows the most likely number of populations in the data set.

To see the barplots for the data analysed click on **`bar plot'** and then on **`show'**.

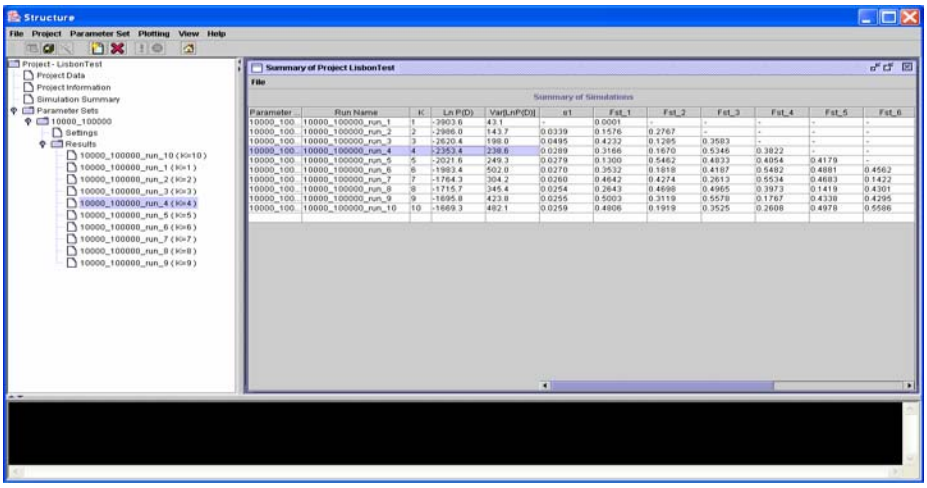


(A) Barplots for K = 4 using order according to Q (population assignment) plotted as single lines.

If you want to see the barplots in input order press on the respective key.

(b) Barplots for K = 4 using order according to Q (population assignment) plotted as multiple lines with strain input numbers – assignment of individual strains is possible (see order of numbers in the input file!).

To see the summary of the **STRUCTURE** run click on **'view'** and then **'simulation summary'**:



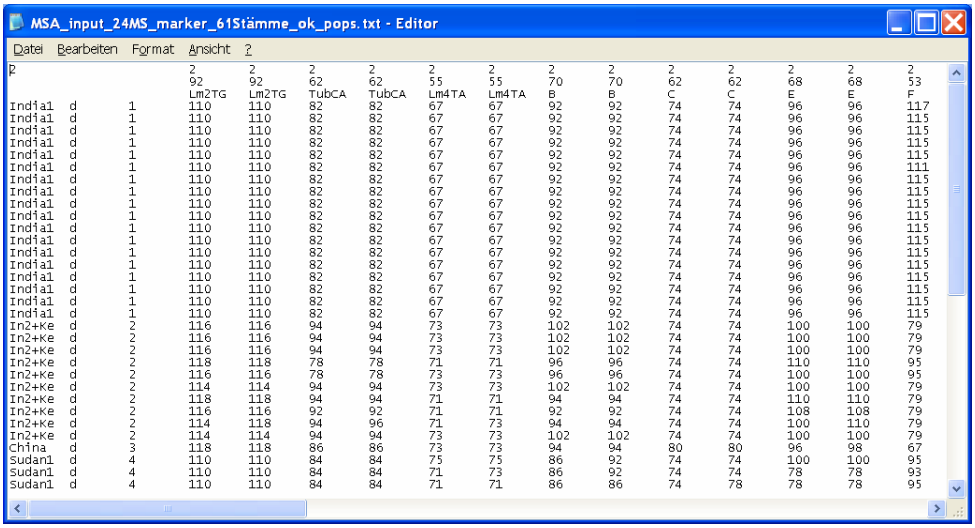
Summary of Simulations

Parameter ...	Run Name	K	Ln P(D)	Var[LnP(D)]	α1	Fst_1	Fst_2	Fst_3	Fst_4	Fst_5	Fst_6
10000_100...	10000_100000_run_1	1	-3903.6	43.1	-	0.0001	-	-	-	-	-
10000_100...	10000_100000_run_2	2	-2986.0	143.7	0.0339	0.1576	0.2767	-	-	-	-
10000_100...	10000_100000_run_3	3	-2620.4	198.0	0.0495	0.4232	0.1285	0.3583	-	-	-
10000_100...	10000_100000_run_4	4	-2353.4	238.6	0.0289	0.3166	0.1670	0.5346	0.3822	-	-
10000_100...	10000_100000_run_5	5	-2021.6	249.3	0.0279	0.1300	0.5462	0.4833	0.4054	0.4179	-
10000_100...	10000_100000_run_6	6	-1983.4	502.0	0.0270	0.3532	0.1818	0.4187	0.5482	0.4881	0.4562
10000_100...	10000_100000_run_7	7	-1764.3	304.2	0.0260	0.4642	0.4274	0.2613	0.5534	0.4683	0.1422
10000_100...	10000_100000_run_8	8	-1715.7	345.4	0.0254	0.2643	0.4698	0.4965	0.3973	0.1419	0.4301
10000_100...	10000_100000_run_9	9	-1695.8	423.8	0.0255	0.5003	0.3119	0.5578	0.1767	0.4338	0.4295
10000_100...	10000_100000_run_10	10	-1669.3	482.1	0.0259	0.4806	0.1919	0.3525	0.2608	0.4978	0.5586

Method E. Measurement of Fst values using MSA

*F*statistics allows for testing the genetic variation among populations.

Make an input file by using the **same MSA input file** that has been applied before. The populations as inferred by STRUCTURE should be arranged in the first column of the input file. Indicate the population number in the third column. Save this file as Text file **(*.txt).** The input file will look as follows:



```
2                         2      2      2      2      2      2      2      2      2      2      2      2      2
                          92     92     62     62     55     55     70     70     62     62     68     68     53
                          Lm2TG  Lm2TG  TubCA  TubCA  Lm4TA  Lm4TA  B      B      C      C      E      E      F
India1  d      1          110    110    82     82     67     67     92     92     74     74     96     96     117
India1  d      1          110    110    82     82     67     67     92     92     74     74     96     96     115
India1  d      1          110    110    82     82     67     67     92     92     74     74     96     96     115
India1  d      1          110    110    82     82     67     67     92     92     74     74     96     96     115
India1  d      1          110    110    82     82     67     67     92     92     74     74     96     96     115
India1  d      1          110    110    82     82     67     67     92     92     74     74     96     96     111
India1  d      1          110    110    82     82     67     67     92     92     74     74     96     96     115
India1  d      1          110    110    82     82     67     67     92     92     74     74     96     96     115
India1  d      1          110    110    82     82     67     67     92     92     74     74     96     96     115
India1  d      1          110    110    82     82     67     67     92     92     74     74     96     96     115
India1  d      1          110    110    82     82     67     67     92     92     74     74     96     96     115
India1  d      1          110    110    82     82     67     67     92     92     74     74     96     96     115
India1  d      1          110    110    82     82     67     67     92     92     74     74     96     96     115
India1  d      1          110    110    82     82     67     67     92     92     74     74     96     96     115
India1  d      1          110    110    82     82     67     67     92     92     74     74     96     96     115
India1  d      1          110    110    82     82     67     67     92     92     74     74     96     96     115
India1  d      1          110    110    82     82     67     67     92     92     74     74     96     96     115
India1  d      1          110    110    82     82     67     67     92     92     74     74     96     96     115
In2+Ke  d      2          116    116    94     94     73     73     102    102    74     74     100    100    79
In2+Ke  d      2          116    116    94     94     73     73     102    102    74     74     100    100    79
In2+Ke  d      2          116    116    94     94     73     73     102    102    74     74     100    100    79
In2+Ke  d      2          118    118    78     78     71     71     96     96     74     74     110    110    95
In2+Ke  d      2          116    116    78     78     73     73     96     96     74     74     100    100    95
In2+Ke  d      2          114    114    94     94     73     73     102    102    74     74     100    100    79
In2+Ke  d      2          118    118    94     94     71     71     94     94     74     74     110    110    79
In2+Ke  d      2          116    116    92     92     71     71     92     92     74     74     108    108    79
In2+Ke  d      2          114    118    94     96     71     73     94     94     74     74     100    110    79
In2+Ke  d      2          114    114    94     94     73     73     102    102    74     74     100    100    79
China   d      3          118    118    86     86     73     73     94     94     80     80     96     98     67
Sudan1  d      4          110    110    84     84     75     75     86     92     74     74     100    100    95
Sudan1  d      4          110    110    84     84     71     73     86     92     74     74     78     78     93
Sudan1  d      4          110    110    84     84     71     71     86     86     74     78     78     78     95
```

Copy this input text file to the folder of the **MSA** program.

Open the **MSA**nalyser.exe, type **(i)** and type your input file name in the command line of file name:

```
C:\Program Files\MSA\MSAnalyser.exe
MSA can now also handle command line arguments!
If you are interested in this function
please read the documentation!

Welcome to MSAnalyzer 4.00 !

_____________________________________________________
     (i) ... Inputfile:none (MSA will use testdata.dat)

     (d) ... Distance settings
     (c) ... Data conversion settings
     (r) ... Heterozygosity range settings

     (!) ... Run
     (q) ... Quit
Please enter command:i

Please enter the filename:MSA_Nabila_ok_61_popsnwithhybrids.txt_
```

Enter, type (d) in the command line and enter:

```
C:\Program Files\MSA\MSAnalyser.exe
     (p) ... Distances                          POP_OFF IND_OFF
     (s) ... Fst,Fit,Fis                        OFF
     (m) ... back to main menu

     (!) ... Run
     (q) ... Quit
Please enter command:s

_____________________________________________________
     (c) ... Fst/Fis/Fit distances calc: OFF
     (g) Fst calculated global         : ON
     (g) Fst calculated pairwise       : OFF

     (1) Calculation based on          : Hobs
     (4) Assume Hardy-Weinberg         : No
     (7) Heterogeneity among loci      : No

     (n) ... Number of permutations    : 10000

     (b) ... back to distance menu
     (m) ... back to main menu

     (!) ... Run
     (q) ... Quit
Please enter command:_
```

Type (s) in the command line to activate **(ON)** the Fst, Fit, Fis option, and enter. Then type (c) in the command line and enter, then type (g) and enter again.

Type (b) and press enter, to see that Fst, Fit, Fis are ON. Now type (!) in the command line and press enter to run the program. It will take approximately 30 seconds.

After that close the window and open the output folder, which is deposited after analysis in the MSA program folder as *.txt_MSAresult00.

Open the output folder and find the F-Statistic folder:

Now open the **'FST_WC84-pValue…'** data file:

FIS_WC84.txt	1 KB	Textdatei	16.07.2007 11:22
FIT_WC84.txt	1 KB	Textdatei	16.07.2007 11:22
FST_WC84.txt	1 KB	Textdatei	16.07.2007 11:22
FST_WC84-pValu...	1 KB	Microsoft Excel-Arb...	16.07.2007 11:22
GlobFst.xls	2 KB	Microsoft Excel-Arb...	16.07.2007 11:22

The *F*st-values and *p*-values as shown in the following screen:

	A	B	C	D
1	Fst-values:			
2	pop1MON24	0.000000	0.276766	0.393683
3	pop2MON1	0.276766	0.000000	0.340363
4	pop3MON1	0.393683	0.340363	0.000000
5				
6				
7	P-values:			
8	pop1MON24	0.000000	0.000100	0.000100
9	pop2MON1	0.000300	0.000000	0.000100
10	pop3MON1	0.000300	0.000300	0.000000
11				

The p-value indicates the statistical significance of *F*st- value ($p<0.05$: significant, $p>0.05$: not significant).

Guideline for interpretation of F_{st}– values according to Wright (1978):

< 0.05 - little genetic differentiation

0.05-0.15 - moderate genetic differentiation

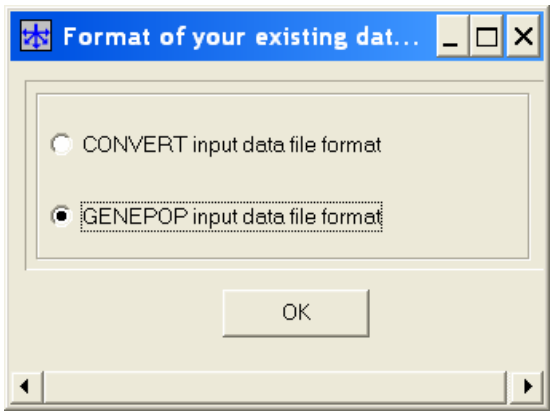
0.15-0.25 - great genetic differentiation

> 0.25 - very great genetic differentiation

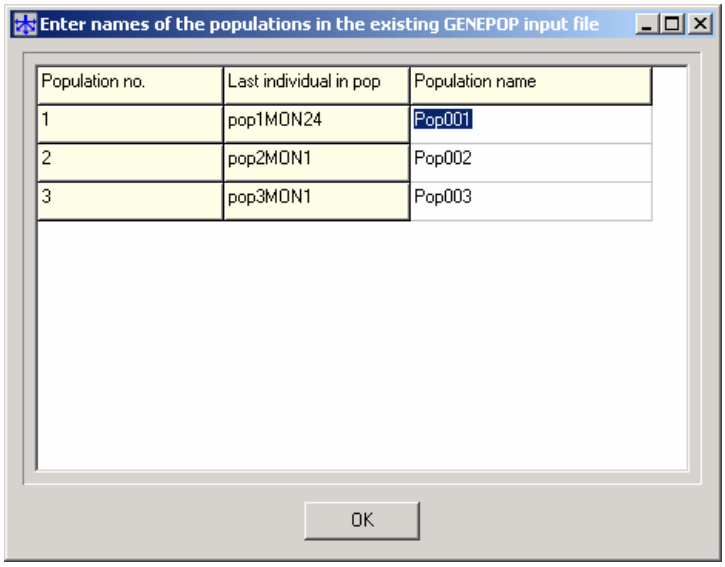
Method F: ***Descriptive statistics analysis by using GDA***

GDA software allows calculation of mean number of alleles (MNA), observed heterozygosity (H_o) and expected heterozygosity (H_e)

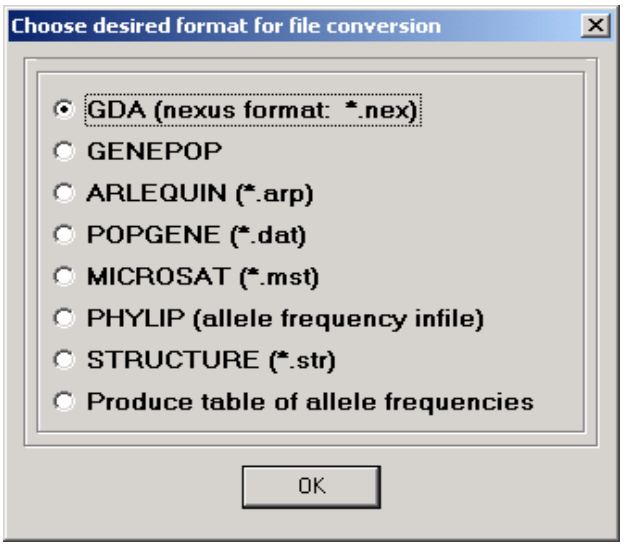
As input file use the Genepop.gen file already generated by **MSA** analysis for calculation of *F*st values. This file is located inside the folder of Formats and Data**.** Convert the Genepop.gen file into the GDA input file which is a nexus file (*.nex)**.** For this copy of the Genepop.gen file into the folder of the **CONVERT131** program. Open the CONVERT.exe, click on file**,** after that on Load data file**,** select GENEPOP input data file format and click on **OK.**



Select your file from the **CONVERT131** folder and click on open to see the following screen:

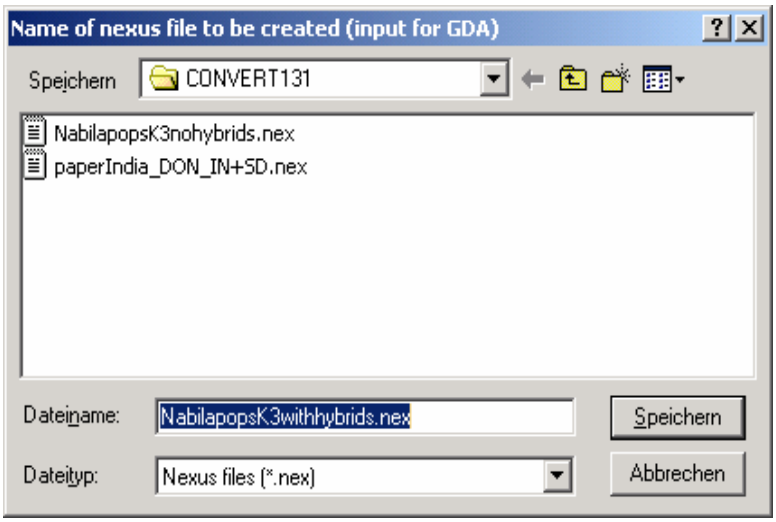


Click on OK to open the following screen. Select GDA (nexus format).



GDA can also be used to create input files for other programs as shown above.

Click on OK and then on Save in the following screen:



A message of successful conversion will be received!

Select the file converted to nexus format from the **CONVERT131** folder and copy it **(*.nex)** into the folder of the **GDA** program. Open the file:

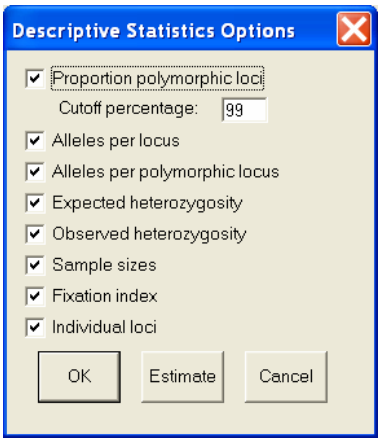
Example of a nexus file:

```
NabilapopsK3withhybrids.nex - Editor
Datei  Bearbeiten  Format  ?
#nexus

begin gdadata;  [!MSA-Output from file: MSA_Nabila_ok_61_popsnohybrids.txt
  dimensions npops=   3  nloci=  14;
  format missing=? separator=/;
  locusallelelabels
    1   Lm2TG,
    2   TubCA,
    3   Lm4TA,
    4   B,
    5   C,
    6   E,
    7   F,
    8   G,
    9   P,
    10  Q,
    11  R,
    12  CS20,
    13  7031,
    14  7039
    ;


matrix

Pop001 :
   pop1MON24      7/   7      1/   1      7/   7      3/   3      4/   4      6/   6
   pop1MON24      9/   9      1/   1      7/   7      3/   3      4/   4      6/   6
   pop1MON24      6/   8      1/   4      2/   3      2/   3      1/   4      2/   8
   pop1MON24      9/   9      1/   1     10/  10      3/   3      4/   4      5/   6
   pop1MON24      8/   9      1/   1      7/   9      2/   3      4/   4      5/   6
   pop1MON24      9/   9      1/   1      6/  10      3/   3      4/   4      5/   6
```

Cave: marker names should have more than one letter otherwise they have to be edited, e.g by adding a second letter!

The GDA nexus file can also be created manually!

Now run **GDA** by opening the **gda.exe**, click on the file menu and after that on open. Select the desired file and open. Go to the file, click on Log to save the file and type a file name. Then go to Descr and option and choose the values to be calculated:



P - proportion of polymorphic loci

n- number of individuals

A- alleles per locus (mean number of alleles-MNA)

Ap- alleles per polymorphic locus

He- expected heterozygosity (genetic diversity)

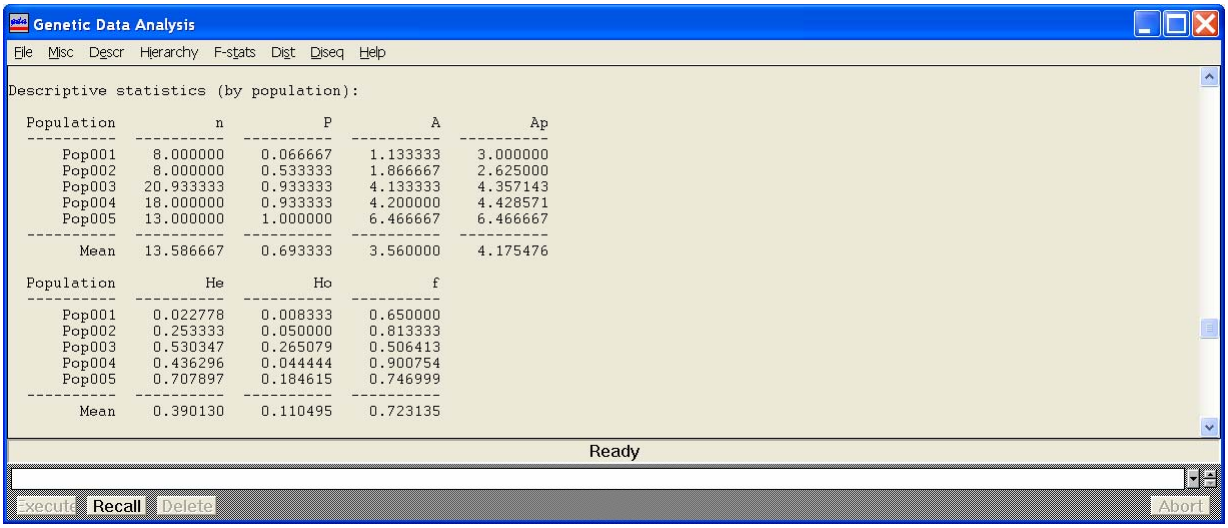
Ho- observed heterozygosity

f- fixation index

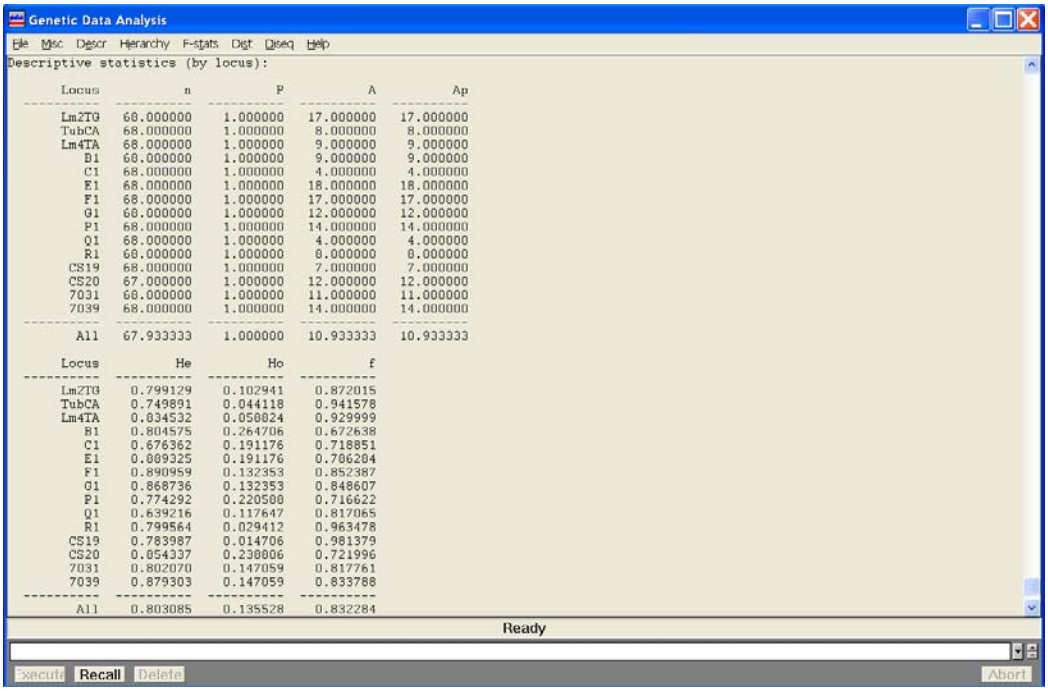
Click on Estimate to start the calculation and go back to the file menu to make the file unlog.

The results of descriptive statistics are shown on the following screen:

Results per population for all loci:



Descriptive data of the microsatellite loci:



References:

Manuals for the softwares used

Evanno,G., S. Regnaut, J. Goudet, Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study, Mol. Ecol. 14 (2005) 2611-2620