

Hands on Use Clustalx and MEGA on Windows/MAC machines for sequence analysis

Part 1:

Clustal family tools include clustalw, a command line interface, clustalx, a graphical interface and clustal omega, a recent addition. Clustal was one of the earliest multiple alignment tool, published in 1988.

Higgins, D. G.; Sharp, P. M. (1988). "CLUSTAL: A package for performing multiple sequence alignment on a microcomputer". *Gene* **73** (1): 237–244. doi:10.1016/0378-1119(88)90330-7. PMID 3243435


A screenshot of a Google Scholar search for the term 'clustal'. The search bar at the top shows 'clustal' with a search button. Below the search bar, it indicates 'About 85,700 results (0.06 sec)'. On the left, there are filters for 'Articles', 'Legal documents', 'Any time' (with options: Since 2013, Since 2012, Since 2009, Custom range...), and 'Sort by relevance' (with 'Sort by date' as an option). The main results area shows two articles. The first article is 'Clustal W and Clustal X version 2.0' by MA Larkin, G Blackshields, NP Brown, R Chenna, et al., published in 2007 by Oxford Univ Press. It has a summary, is cited by 6563, and has links for related articles, BL Direct, all 40 versions, and a cite button. A '[HTML] from oxfordjoi Google Scholar' link is on the right. The second article is 'CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix ...' by JD Thompson, DG Higgins, TJ Gibson, published in 1994 by Oxford Univ Press. It has an abstract, is cited by 42191, and has links for related articles, BL Direct, all 122 versions, a cite button, and a 'More' button. A '[PDF] from nih.gov' link is on the right.

Although clustalw is not a favored tool for multiple sequence alignment, clustalx provides a nice graphical interface for visualizing sequence alignment. Here is how to use it:

1. download clustalx from <http://www.clustal.org/>


A screenshot of the Clustal website, titled 'Clustal: Multiple Sequence Alignment'. The subtitle is 'Multiple alignment of nucleic acid and protein sequences'. The website features two main download options. On the left is 'Clustal Omega', which is described as the 'Latest version of Clustal - fast and scalable (can align hundreds of thousands of sequences in hours), greater accuracy due to new HMM alignment engine' and 'Command line/web server only (GUI public beta available soon)'. On the right is 'ClustalW/ClustalX', which includes 'Classic Clustal' and 'GUI (ClustalX), command line (ClustalW), web server versions available'. The ClustalW/ClustalX option is highlighted with a red rectangular border. Logos for 'sfi' and 'UCD DUBLIN' are visible in the top left and right corners respectively.

2. click on the right side to get to the download page



ClustalW / ClustalX

Multiple alignment of nucleic acid and protein sequences



[Home](#)
[Webservers](#)
[Download](#)
[Documentation](#)
[Contact](#)
[News](#)

Webservers

You don't necessarily have to go through the hassle to install Clustal on your computer. Instead, you can run Clustal online on several servers on the web:

- [EBI web server](#)
- [Swiss Institute of Bioinformatics](#)

Download ClustalW/X

Clustal 2 comes in two flavors: the command-line version ClustalW and the graphical version ClustalX. Precompiled executables for Linux, Mac OS X and Windows (Intel, XP and Vista) of the most recent version (currently 2.1) along with the source code are [available for download here](#). You can also [browse for older versions](#) (ClustalW 1.81, ClustalX 1.81). The current version of Clustal 2 is also mirrored on the [EBI ftp site](#).

Clustal 2.1 is released under the [GNU Lesser GPL](#).

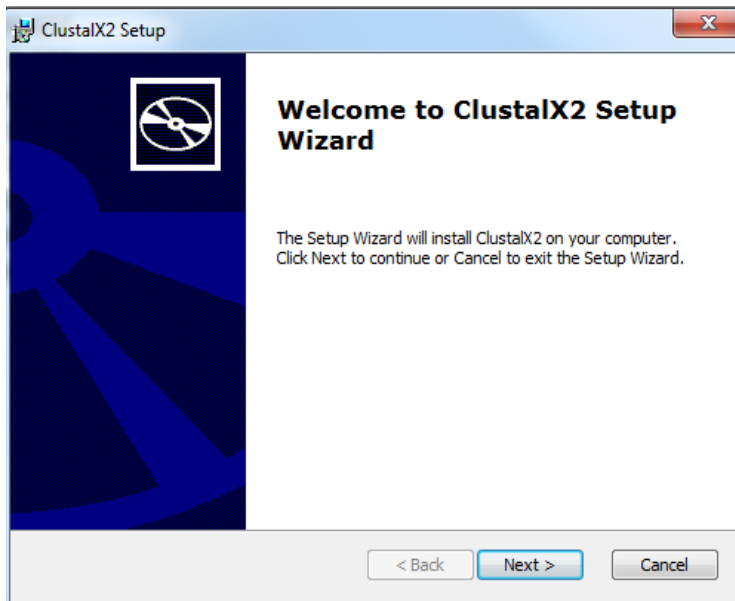
3. connect to the download page

← → ↻ www.clustal.org/download/current/

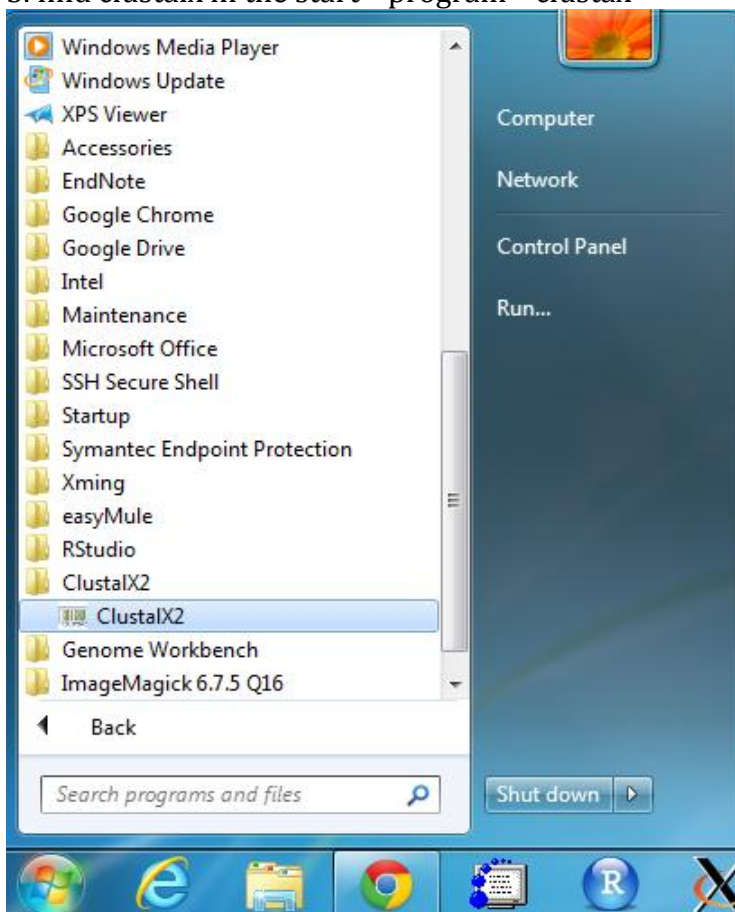
Index of /download/current

Name	Last modified	Size	Description
 Parent Directory	-		
 CHANGELOG	17-Nov-2010 11:59	9.0K	
 COPYING	17-Nov-2010 11:59	34K	
 COPYING.LESSER	17-Nov-2010 11:59	7.5K	
 Readme	17-Nov-2010 11:59	2.0K	
 clustalw-2.1-linux-x86_64-libc++static.tar.gz	17-Nov-2010 11:59	2.4M	
 clustalw-2.1-macosx.dmg	17-Nov-2010 11:59	6.5M	
 clustalw-2.1-win.msi	17-Nov-2010 11:59	1.9M	
 clustalw-2.1.tar.gz	10-Dec-2010 07:40	343K	
 clustalx-2.1-linux-i686-libc++static.tar.gz	17-Nov-2010 11:59	4.7M	
 clustalx-2.1-macosx.dmg	17-Nov-2010 11:59	12M	
 clustalx-2.1-win.msi	13-Jan-2011 09:36	4.7M	
 clustalx-2.1.tar.gz	10-Dec-2010 07:40	334K	

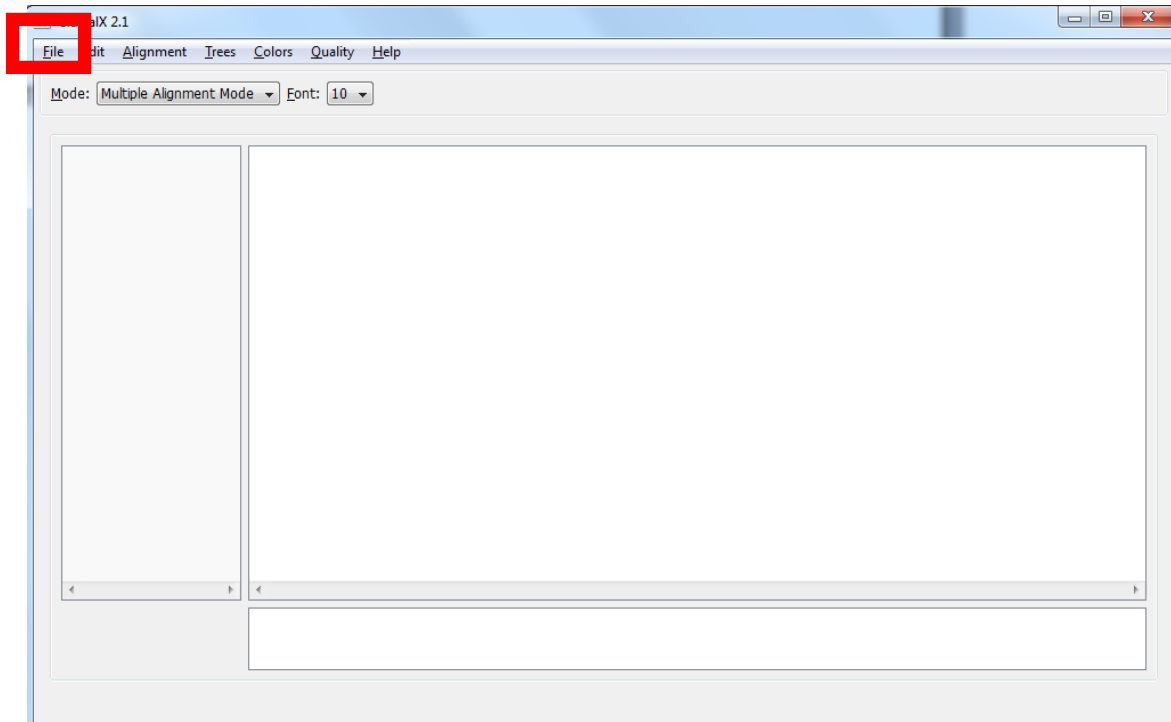
4. install on Windows



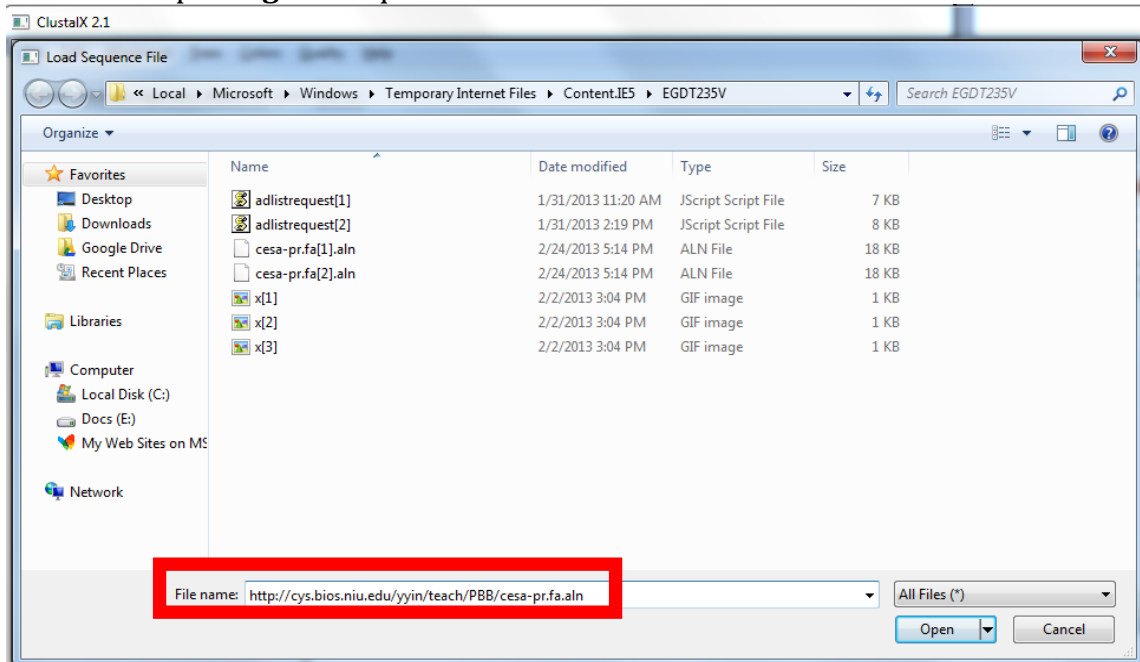
5. find clustalx in the start->program->clustax

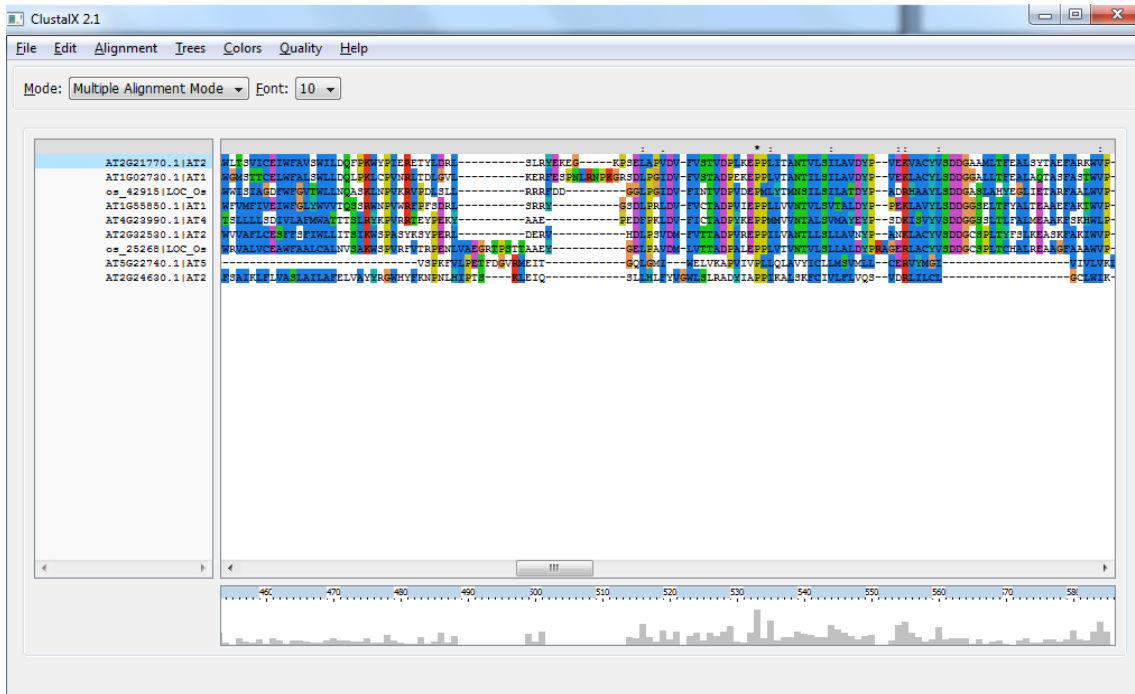


6. this is what clustalx look like

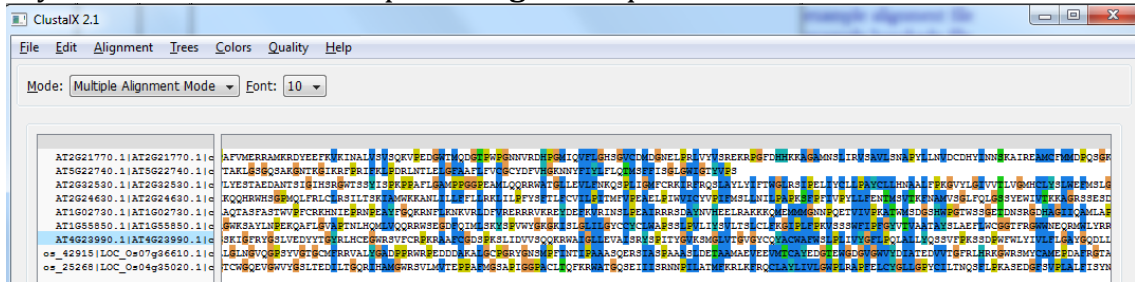


7. load multiple **aligned** sequences

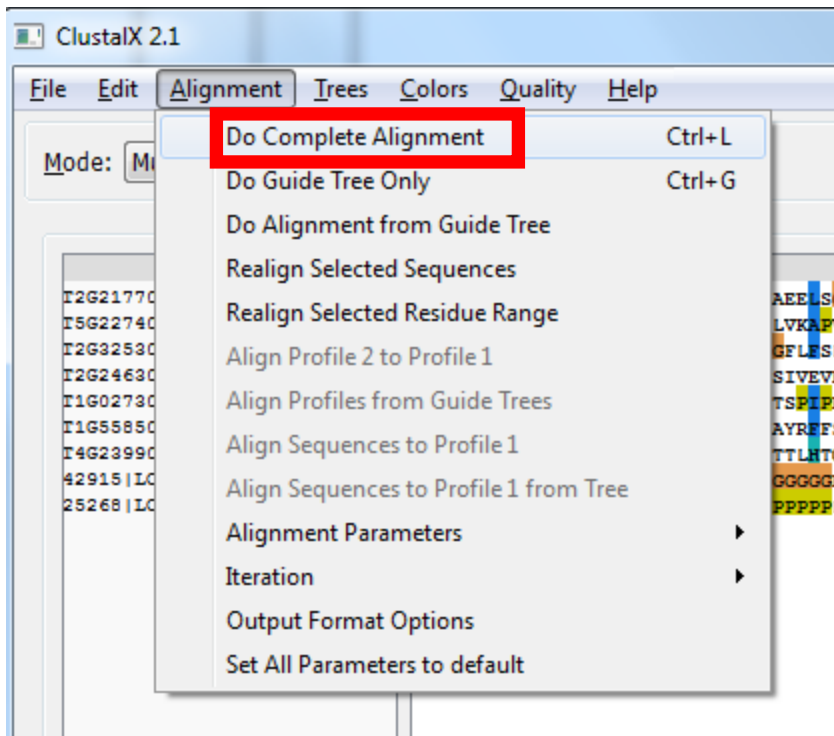




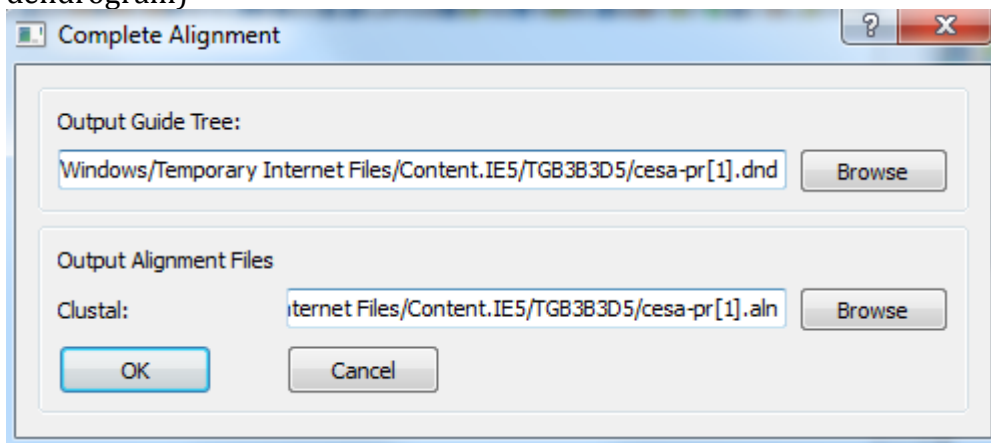
8. you can also load in multiple **unaligned** sequences



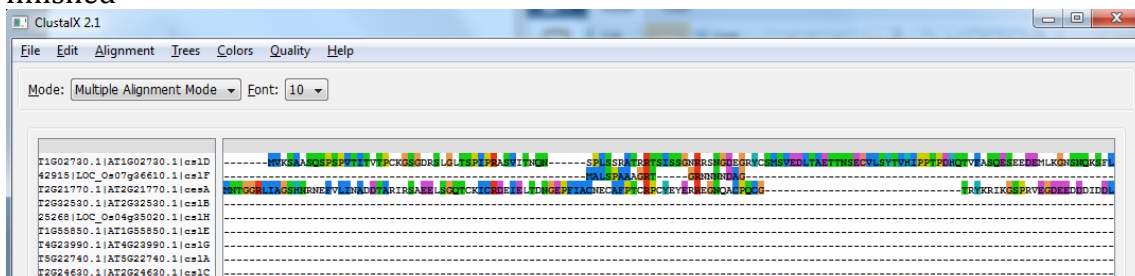
and do clustalw alignment



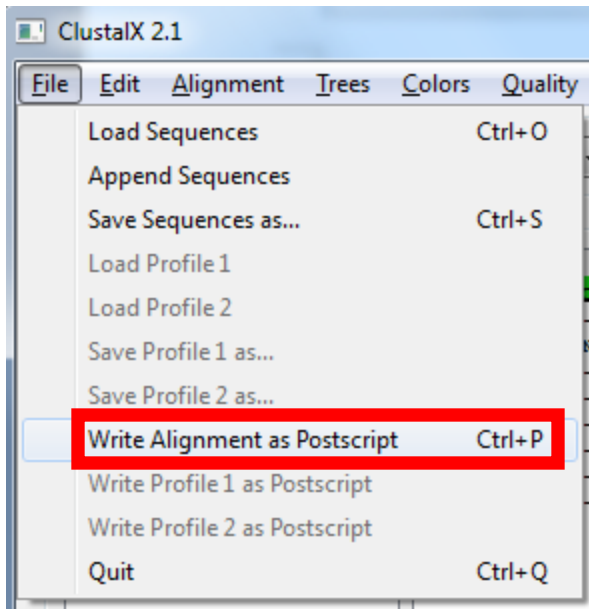
you will be asked where you want to save the alignment results (an alignment and a dendrogram)



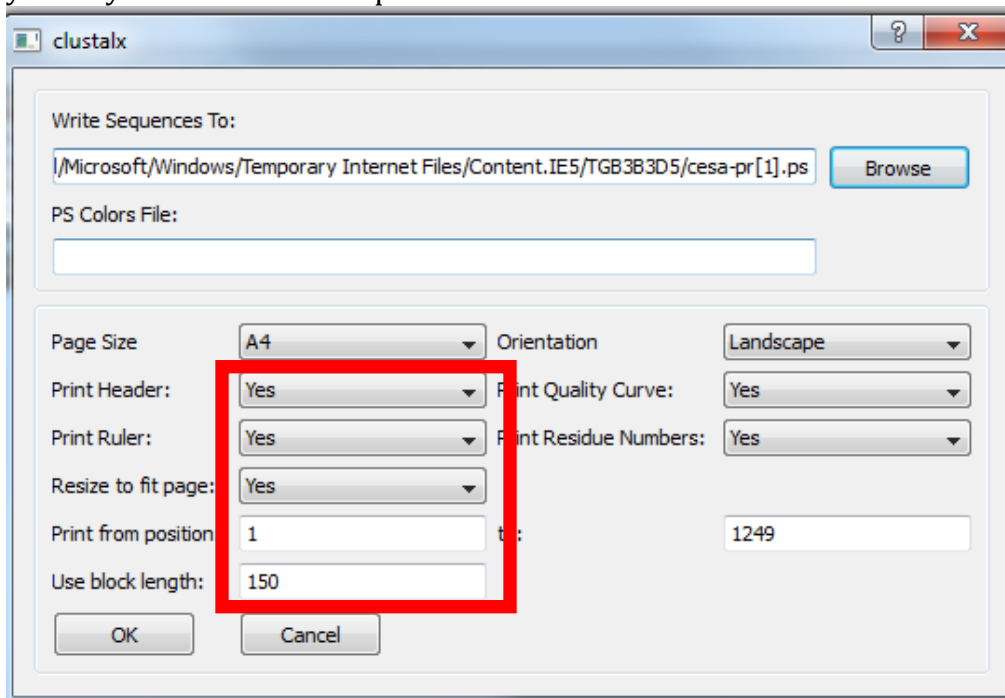
finished



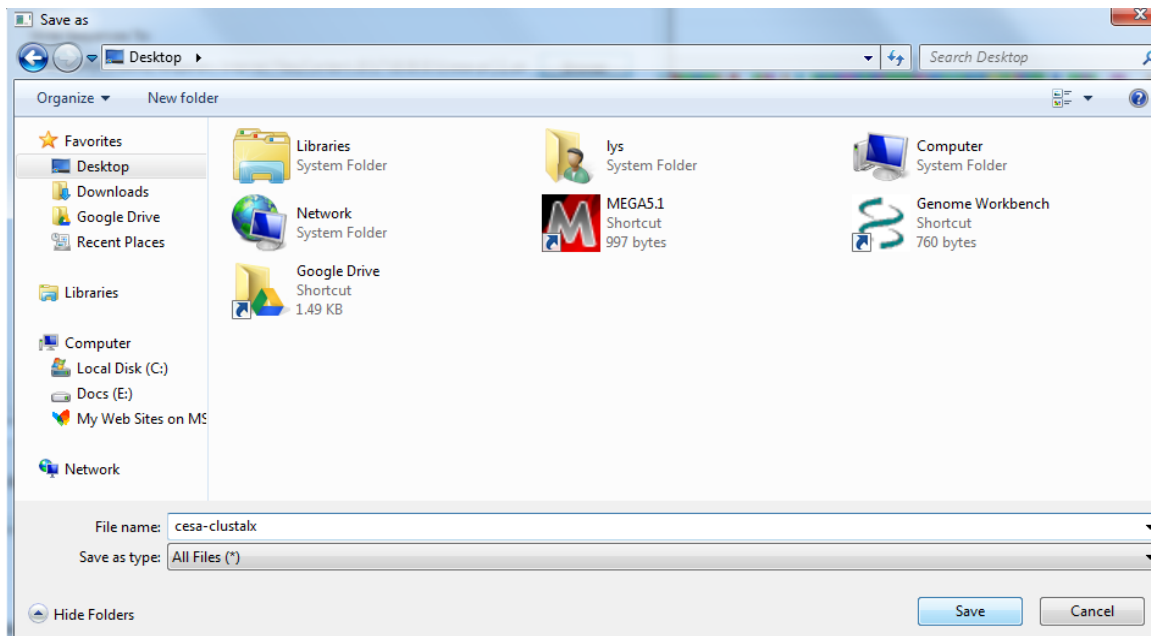
9. you can save the alignment as a postscript format file



you may choose different options



I am gonna save it in the desktop folder



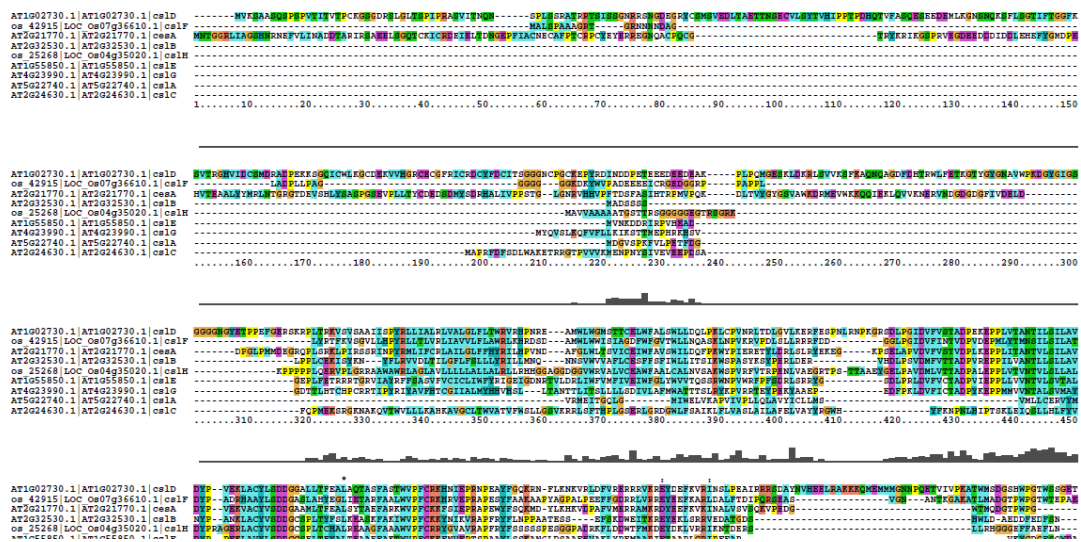
10. you may use **acrobat pro** to convert the ps file to a pdf file

CLUSTAL 2.1 MULTIPLE SEQUENCE ALIGNMENT

File: C:/Users/lys/Desktop/cesa-clustalx

Date: Sun Feb 24 17:28:50 2013

Page 1 of 3



Part 2:

MEGA5: Molecular Evolutionary Genetics Analysis version 5

MEGA is an integrated tool for conducting sequence alignment, inferring phylogenetic trees, mining web-based databases, estimating rates of molecular evolution, inferring ancestral sequences, and testing evolutionary hypotheses. MEGA is used by biologists in a large number of laboratories for reconstructing the evolutionary histories of species and inferring the extent and nature of selective forces shaping the evolution of genes and species

Mega was developed as a software with GUI, but recently it released a command line version to facilitate large scale analyses using terminals:

http://www.kumarlab.net/pdf_new/KumarTamura12a.pdf

Scholar About 20,400 results (0.11 sec) My Citations

Articles **MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0** [HTML] from 9med.net Google Scholar
K Tamura, J Dudley, M Nei, S Kumar - *Molecular biology and evolution*, 2007 - SMOE

Legal documents Abstract We announce the release of the fourth version of **MEGA software**, which expands on the existing facilities for editing DNA sequence data from autosequencers, mining Web-databases, performing automatic and manual sequence alignment, analyzing sequence ...
Cited by 17875 Related articles BL Direct All 15 versions Cite

Any time
Since 2013
Since 2012
Since 2009
Custom range...

Sort by relevance
Sort by date

MEGA5: **molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods** [HTML] from oxfordjourn Google Scholar
K Tamura, D Peterson, N Peterson, G Stecher... - *Molecular biology and evolution*, 2011 - SMOE
... Here, we announce the release of **Molecular Evolutionary Genetics Analysis version 5 (MEGA5)**, which is a user ... ancestral states and sequences (along with probabilities), and estimating evolutionary rates site-by ... This version of **MEGA** is intended for the Windows platform, and it ...
Cited by 4162 Related articles All 14 versions Cite

☒ include patents
☒ include citations

MEGA: **a biologist-centric software for evolutionary analysis of DNA and protein sequences** [HTML] from oxfordjourn Google Scholar
S Kumar, M Nei, J Dudley, K Tamura - *Briefings in bioinformatics*, 2008 - Oxford Univ Press
... It was made available over the Internet (<http://www.megasoftware.net>) and was downloaded by ...
In **MEGA 4**, we expanded the transparency of choices and assumptions by adding a new Caption ...
MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0 ...
Cited by 1431 Related articles BL Direct All 27 versions Cite More

☐ Create alert

1. find MEGA at <http://www.megasoftware.net/>

www.megasoftware.net

MEGA MOLECULAR EVOLUTIONARY GENETICS ANALYSIS
Authors: Koichiro Tamura, Daniel Peterson, Nicholas Peterson, Glen Stecher, Masatoshi Nei and Sudhir Kumar

Version 5.10 Follow @iluvmega

Windows Download v5.1 Updated: Oct 19 2012 Build: 5121019

Mac OS Download v5.1 Updated: Dec 19 2012 Build: 5121218

Computational Core Download v5.1 Updated: May 31 2012 Build: 5120531

Older Versions MEGA 4 MEGA for DOS

Alignments & Data
Data Types
Web Data Acquisition
Manual & Automated Alignments

Major Analyses
Models and Parameters
Infer Phylogenies
Compute Distances
Tests of Selection
Ancestral Sequences
Clocks and Rates

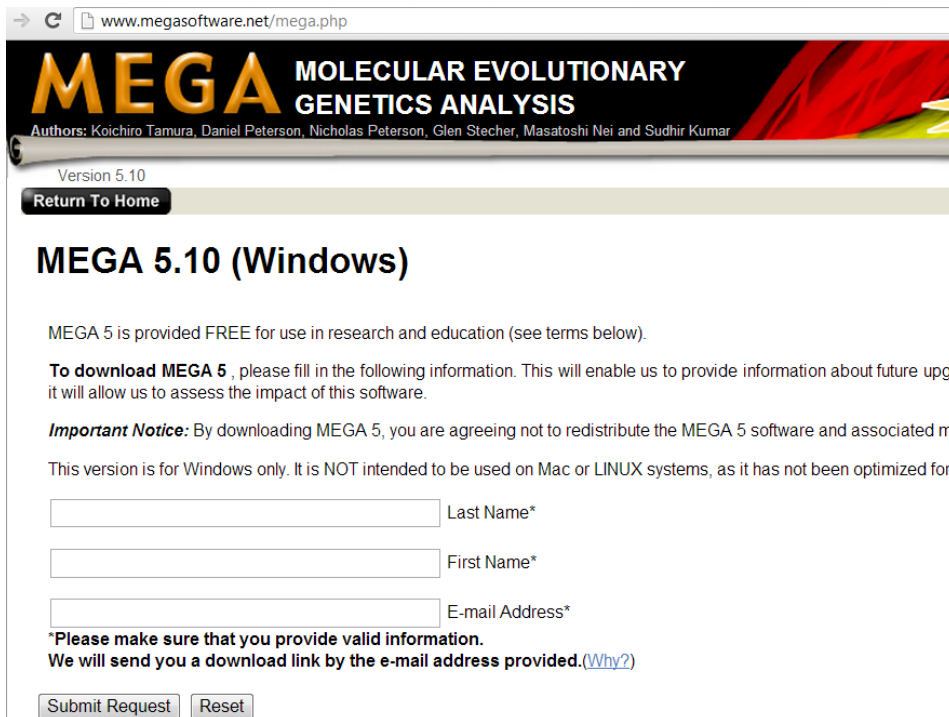
Substitution Models
DNA/RNA
Codon
Protein
Rates & Composition

About MEGA
MEGA is an integrated tool for conducting sequence alignment, inferring phylogenetic trees, mining web-based databases, estimating rates of molecular evolution, inferring ancestral sequences, and testing evolutionary hypotheses. MEGA is used by biologists in a large number of laboratories for reconstructing the evolutionary histories of species and inferring the extent and nature of selective forces shaping the evolution of genes and species. [Download PDF](#)

About MEGA Computational Core (MEGA-CC)

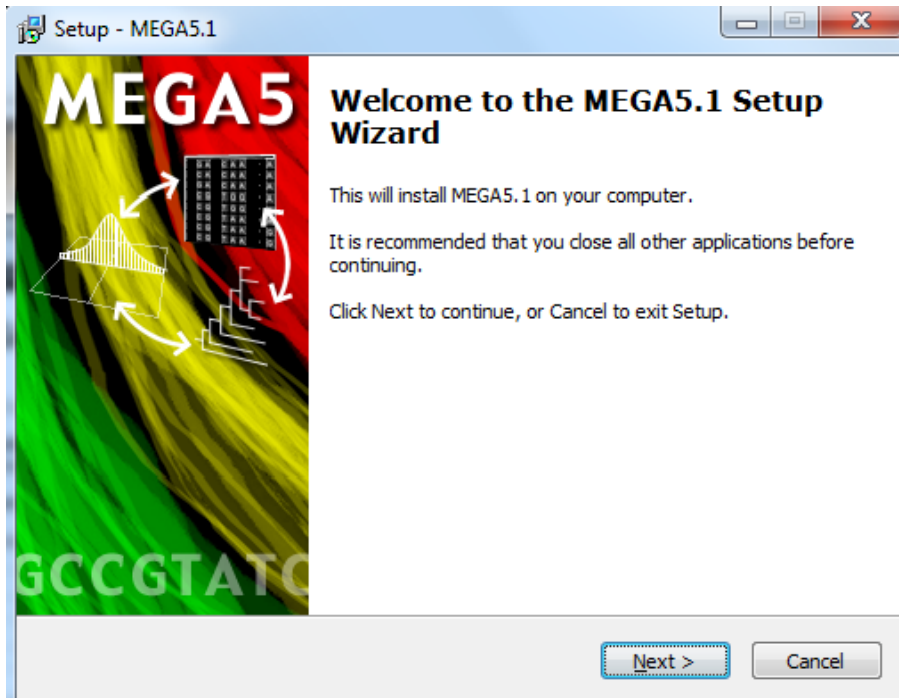
2. it's free, but you need to fill out an on-line form to download

I also put a recently downloaded file at
http://cys.bios.niu.edu/yyin/teach/MEGA5.10_Setup.exe

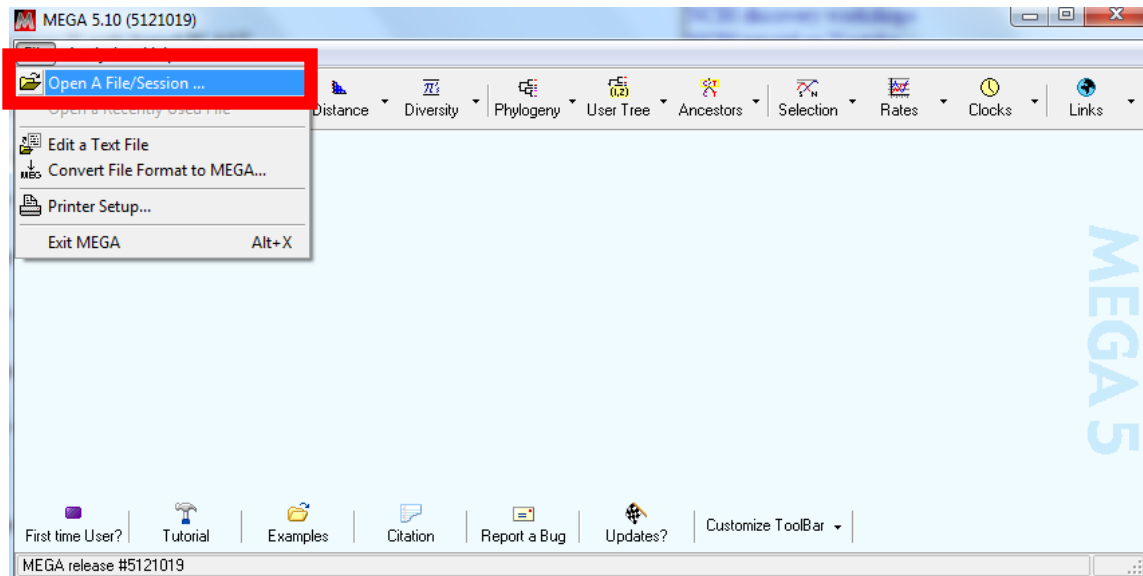


The screenshot shows the MEGA 5.10 website. At the top, there is a banner with the text "MEGA MOLECULAR EVOLUTIONARY GENETICS ANALYSIS" and the authors' names: Koichiro Tamura, Daniel Peterson, Nicholas Peterson, Glen Stecher, Masatoshi Nei and Sudhir Kumar. Below the banner, it says "Version 5.10" and has a "Return To Home" button. The main heading is "MEGA 5.10 (Windows)". The text states that MEGA 5 is provided FREE for use in research and education. It asks users to fill in information to enable future updates. An important notice states that by downloading MEGA 5, users agree not to redistribute the software. It also mentions that this version is for Windows only. There are three input fields for "Last Name*", "First Name*", and "E-mail Address*". Below these fields, there is a note: "Please make sure that you provide valid information. We will send you a download link by the e-mail address provided. (Why?)". At the bottom, there are two buttons: "Submit Request" and "Reset".

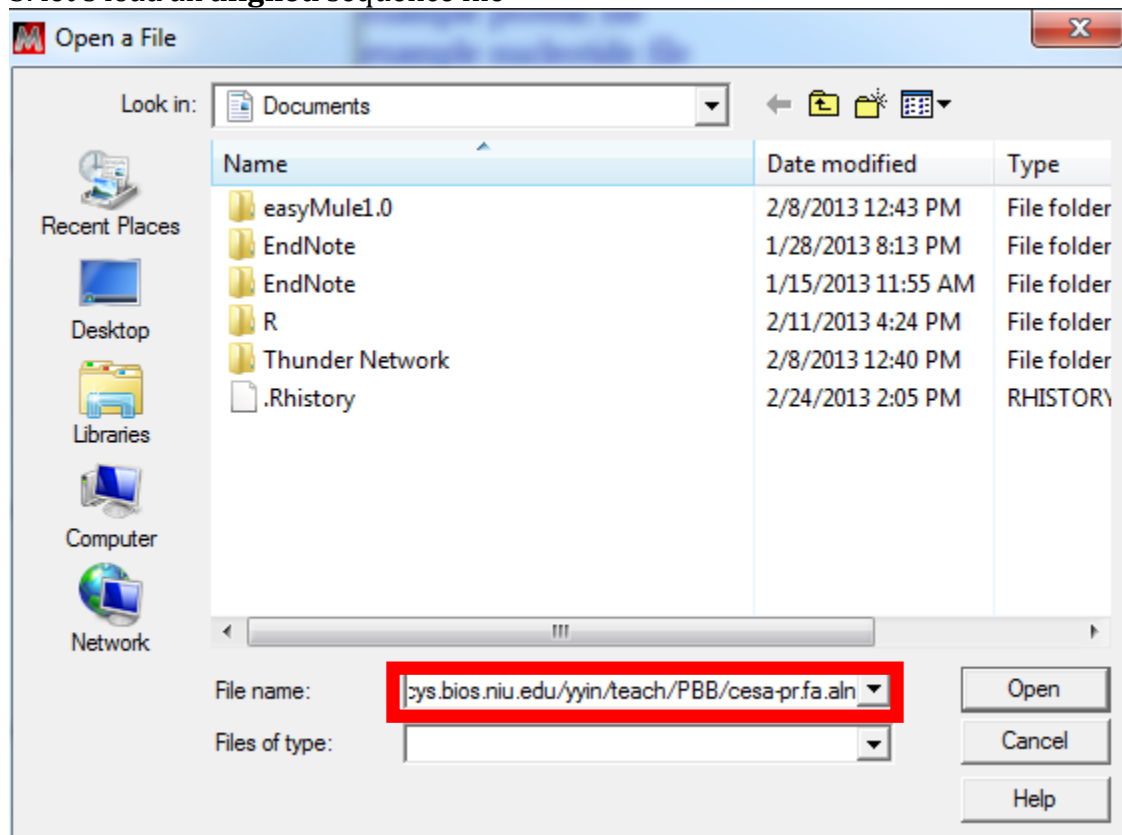
3. follow the instructions to install



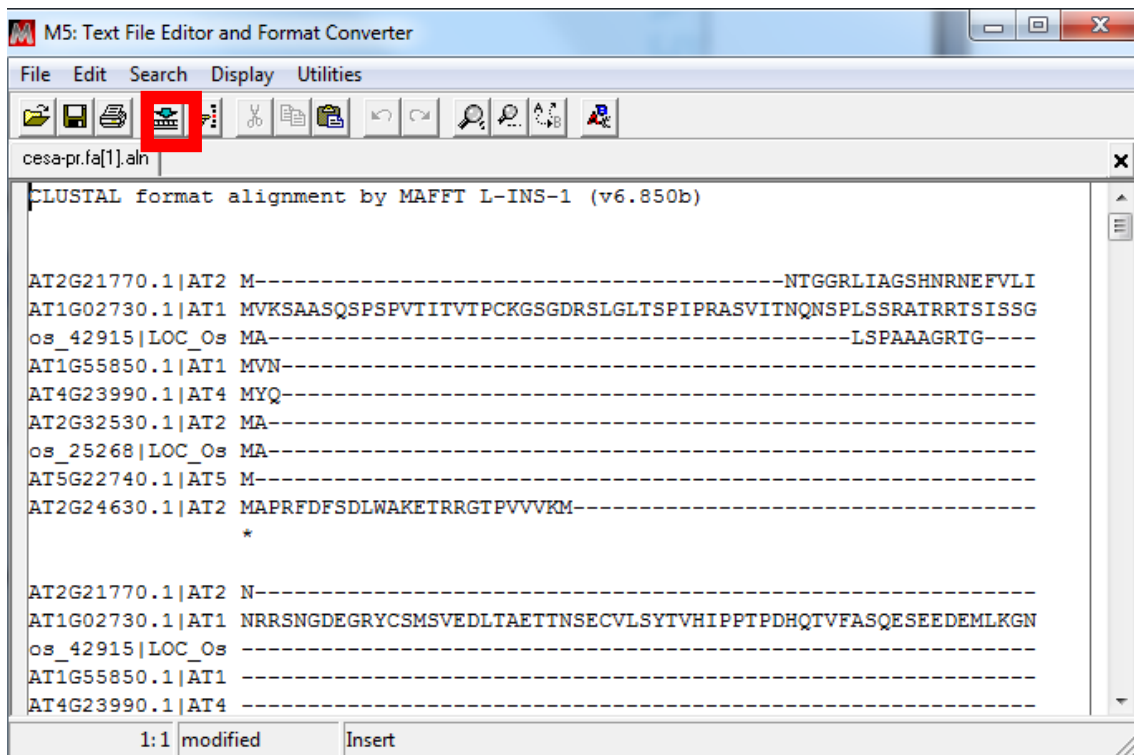
4. this is what MEGA looks like



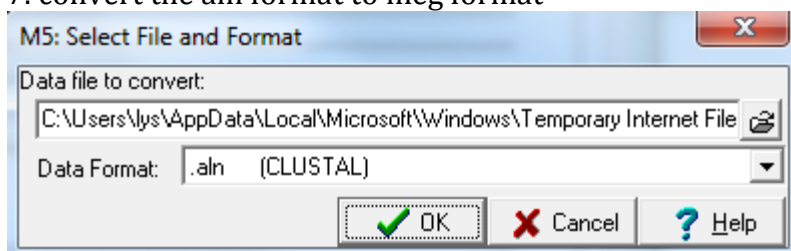
5. let's load an **aligned** sequence file



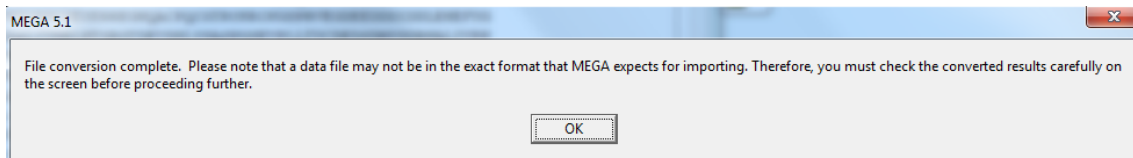
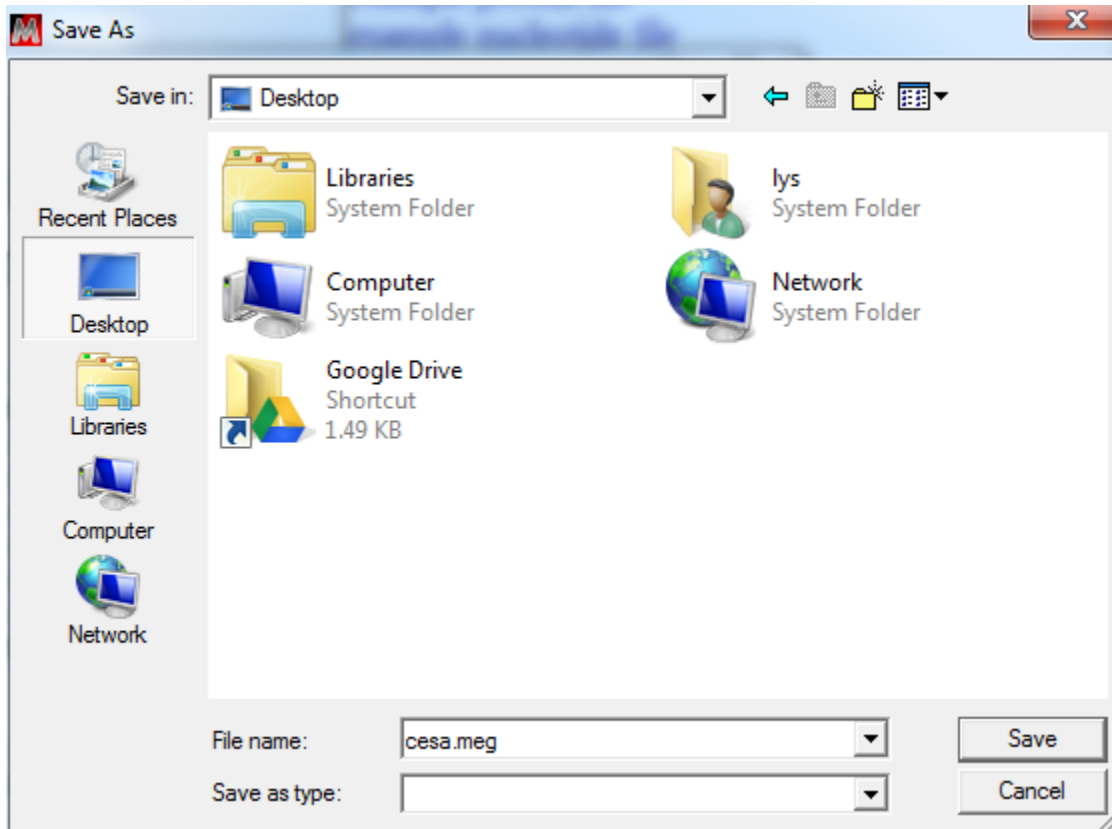
6. a new window called sequence editor will appear



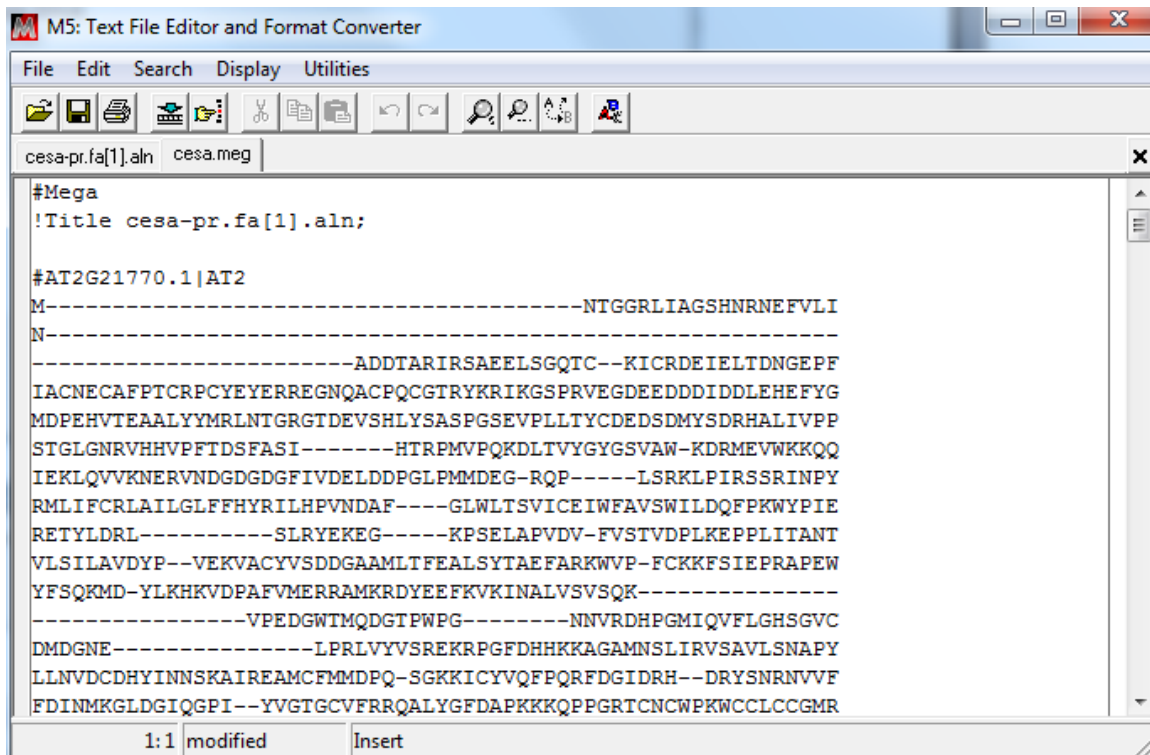
7. convert the aln format to meg format



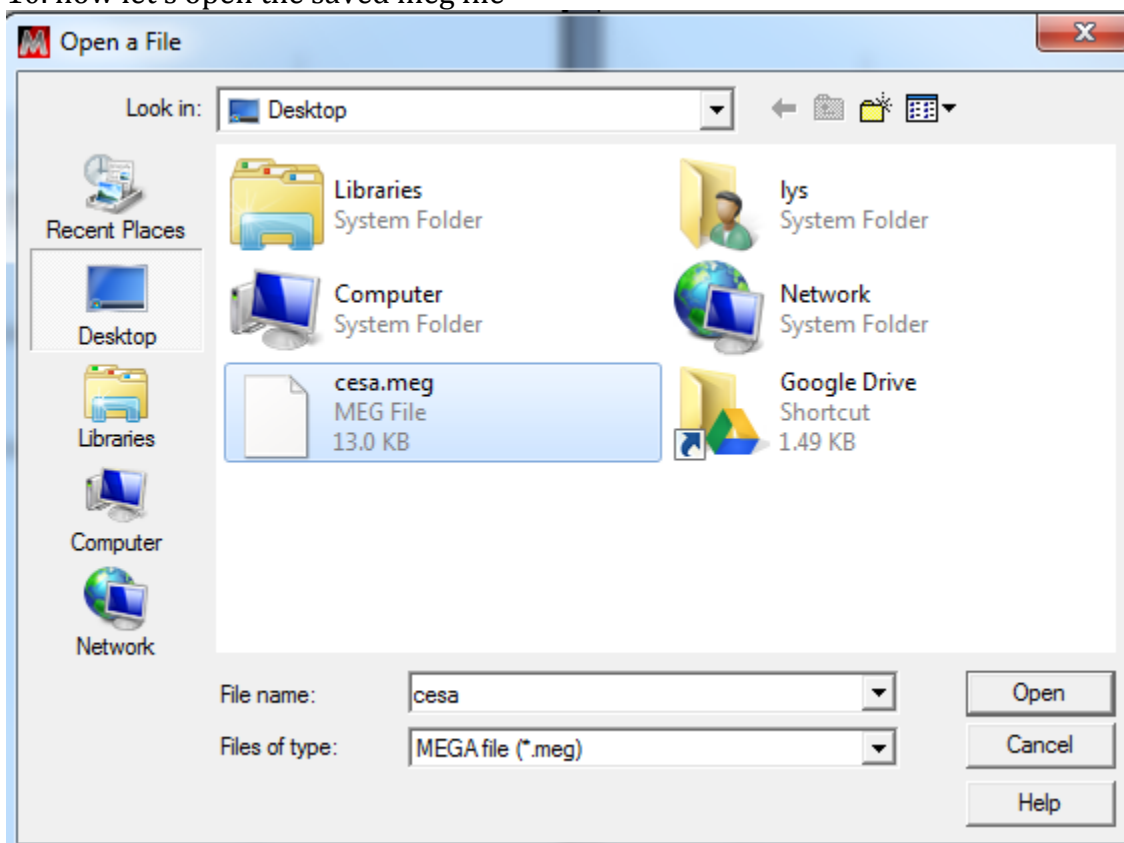
8. save it as a meg file



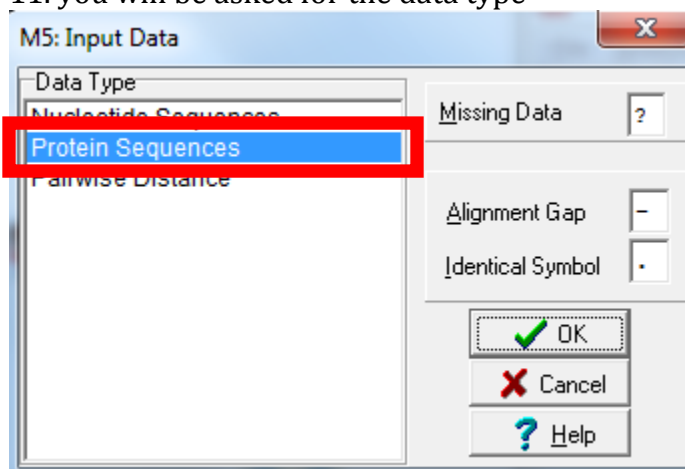
9. what meg format look like:



10. now let's open the saved meg file



11. you will be asked for the data type

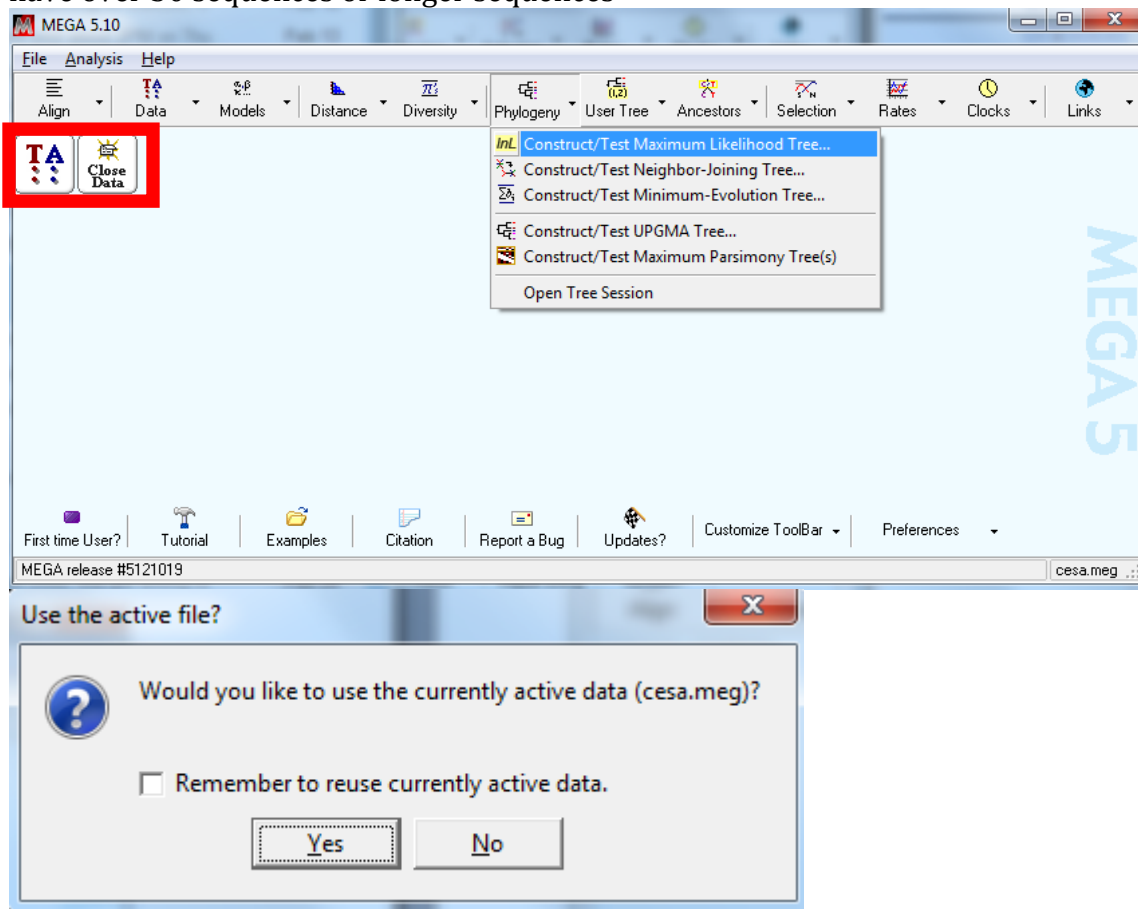


12. the data is loaded; we can build the tree now

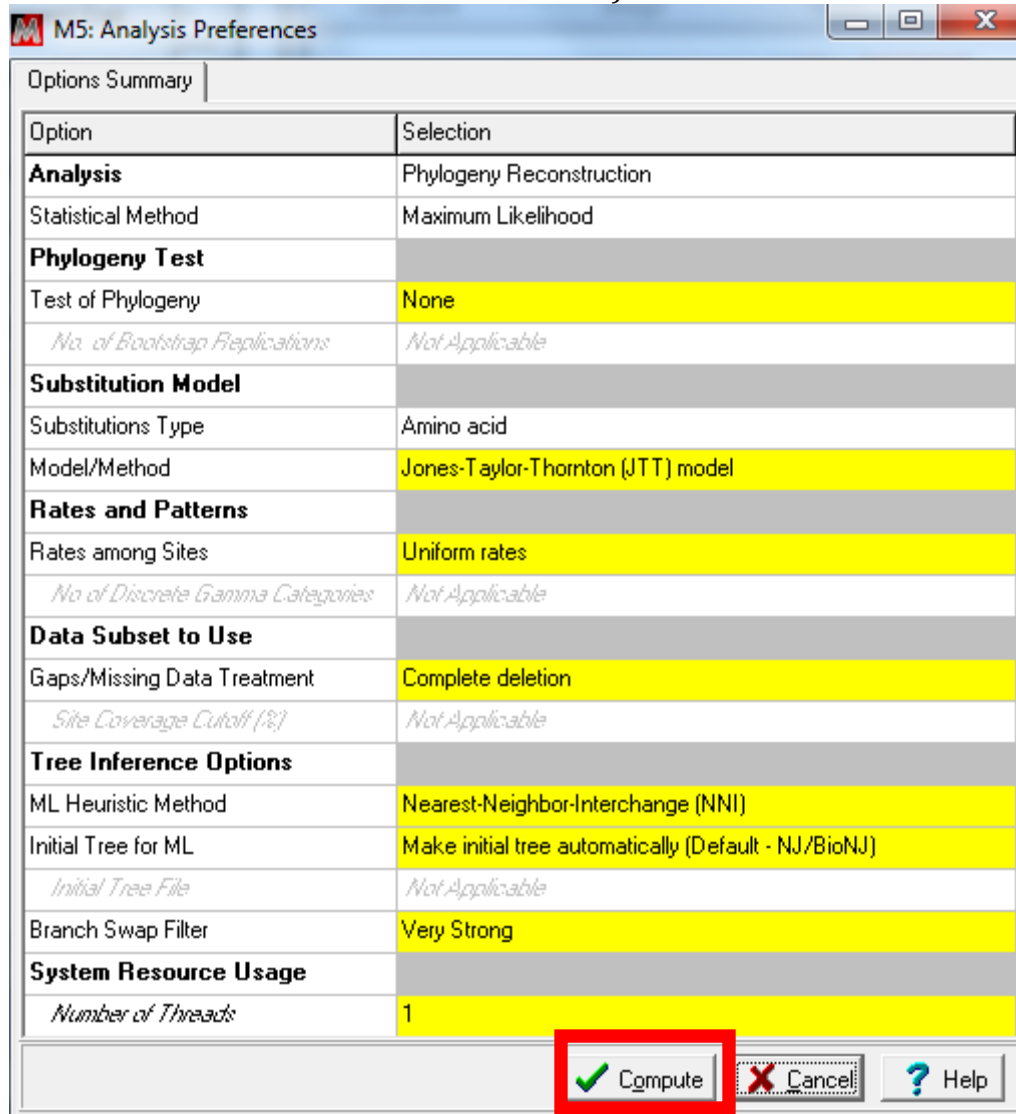
You may choose from a list of different building algorithms

Basically, maximum likelihood is the most accurate but also the slowest

Neighbor-joining and maximum parsimony are also very popular and faster if you have over 50 sequences or longer sequences



13. you may choose parameters for tree building, but let's try default first (the default ones are the fastest but less accurate)

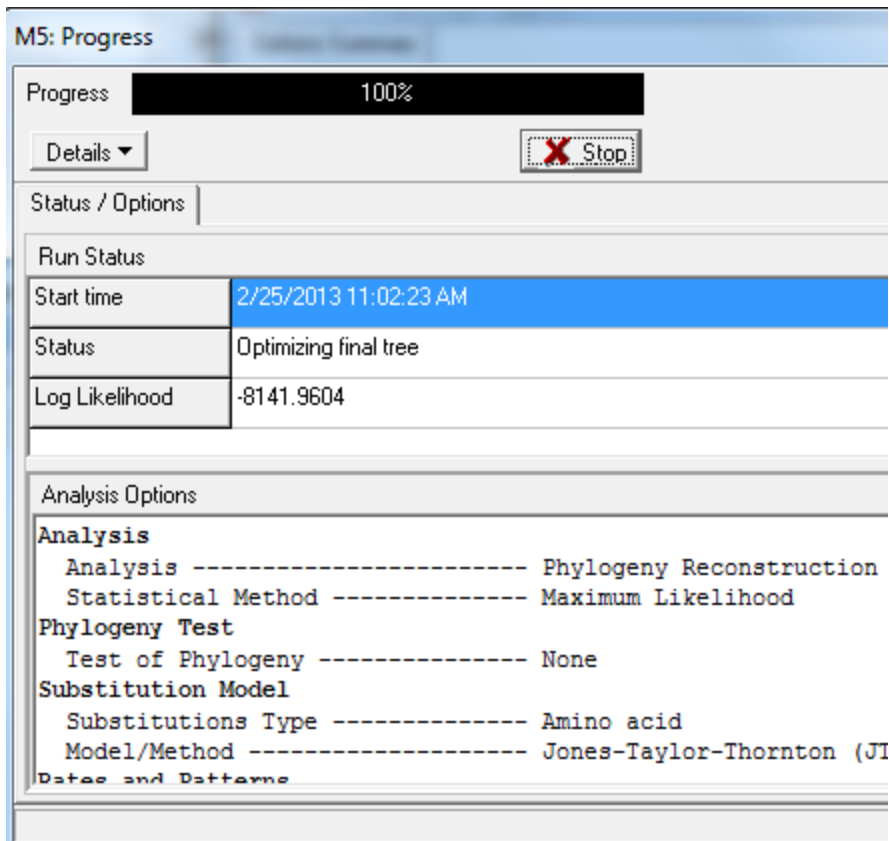


M5: Analysis Preferences

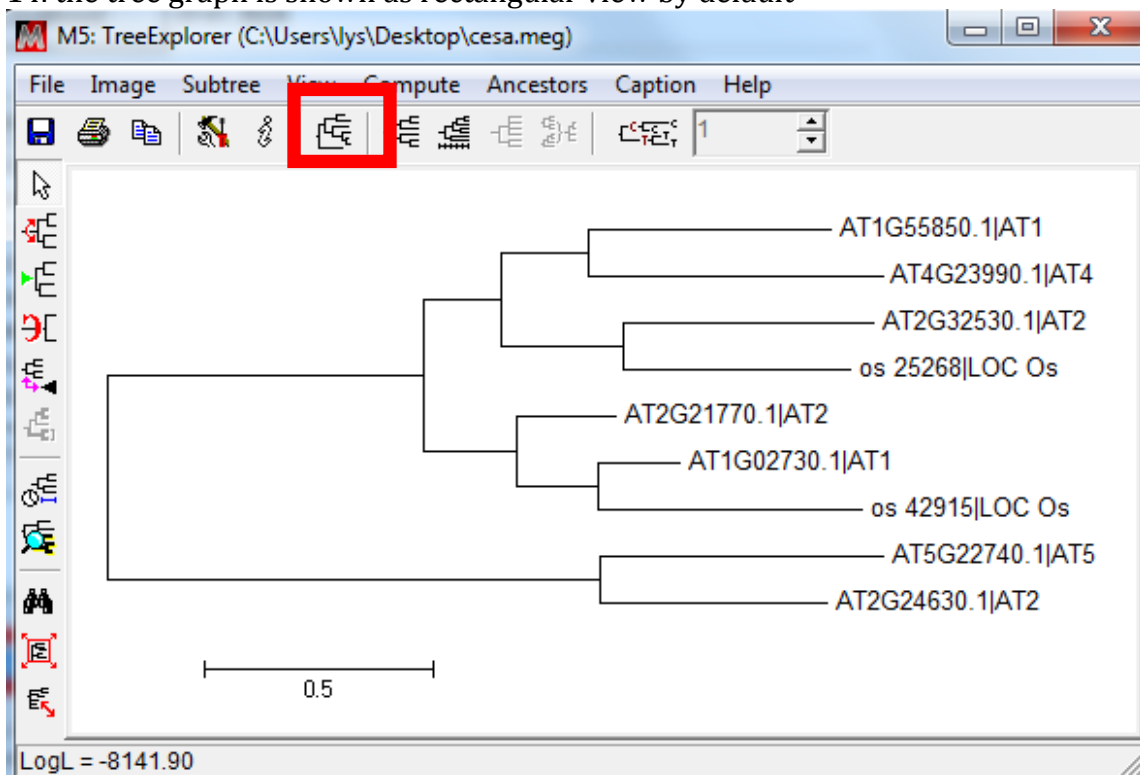
Options Summary

Option	Selection
Analysis	Phylogeny Reconstruction
Statistical Method	Maximum Likelihood
Phylogeny Test	
Test of Phylogeny	None
<i>No. of Bootstrap Replications</i>	<i>Not Applicable</i>
Substitution Model	
Substitutions Type	Amino acid
Model/Method	Jones-Taylor-Thornton (JTT) model
Rates and Patterns	
Rates among Sites	Uniform rates
<i>No. of Discrete Gamma Categories</i>	<i>Not Applicable</i>
Data Subset to Use	
Gaps/Missing Data Treatment	Complete deletion
<i>Site Coverage Cutoff (%)</i>	<i>Not Applicable</i>
Tree Inference Options	
ML Heuristic Method	Nearest-Neighbor-Interchange (NNI)
Initial Tree for ML	Make initial tree automatically (Default - NJ/BioNJ)
<i>Initial Tree File</i>	<i>Not Applicable</i>
Branch Swap Filter	Very Strong
System Resource Usage	
<i>Number of Threads</i>	1

☒ Compute
 ☐ Cancel

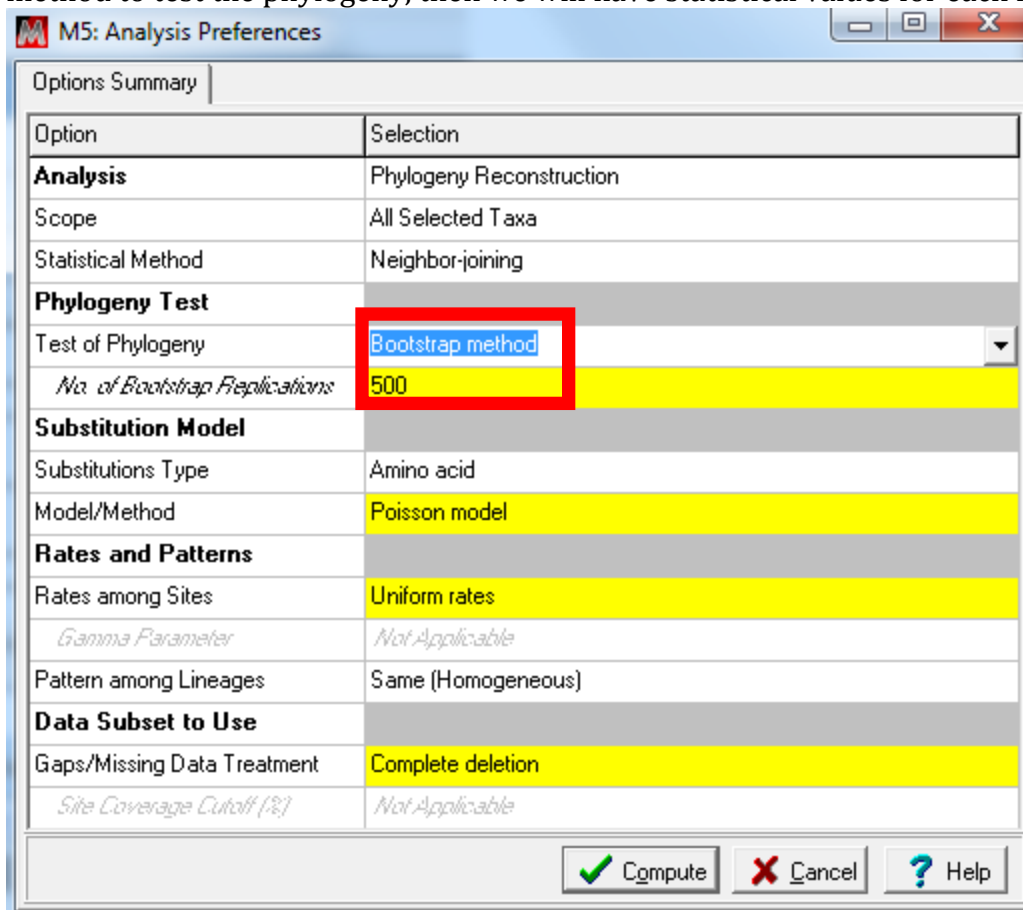


14. the tree graph is shown as rectangular view by default

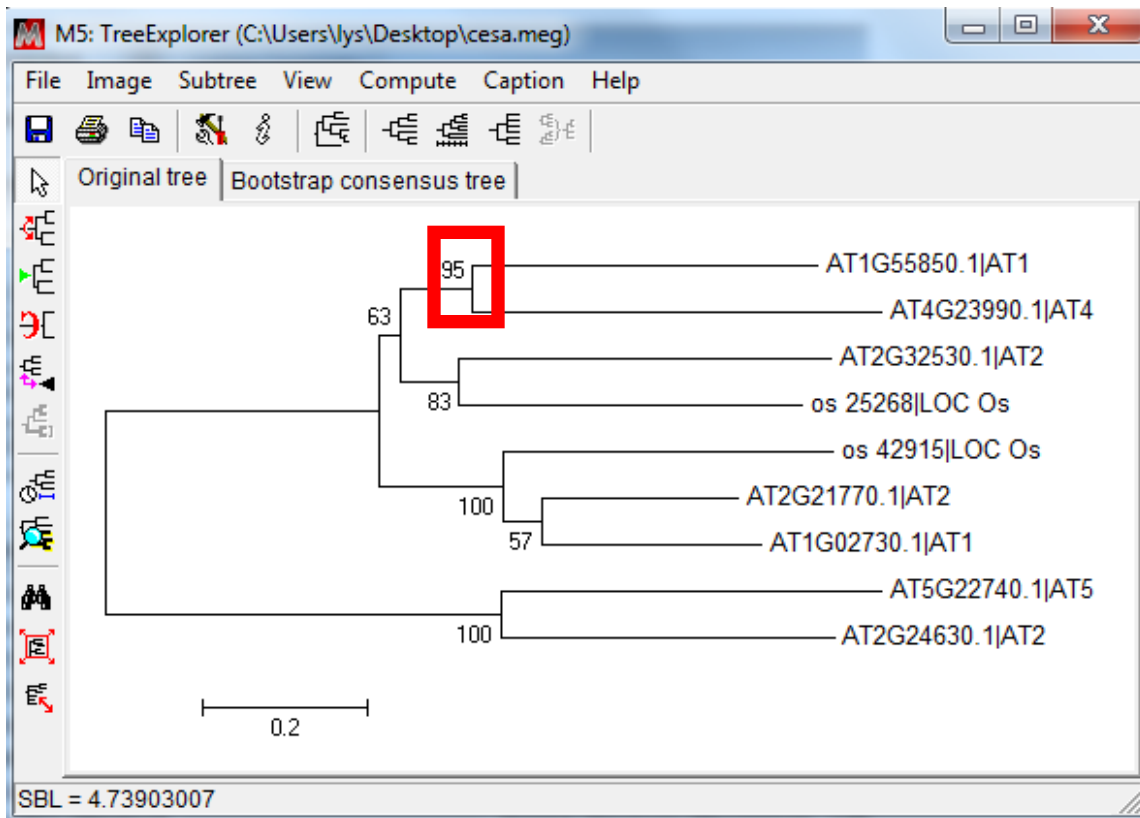


15. how do we have statistical values on the internal nodes?

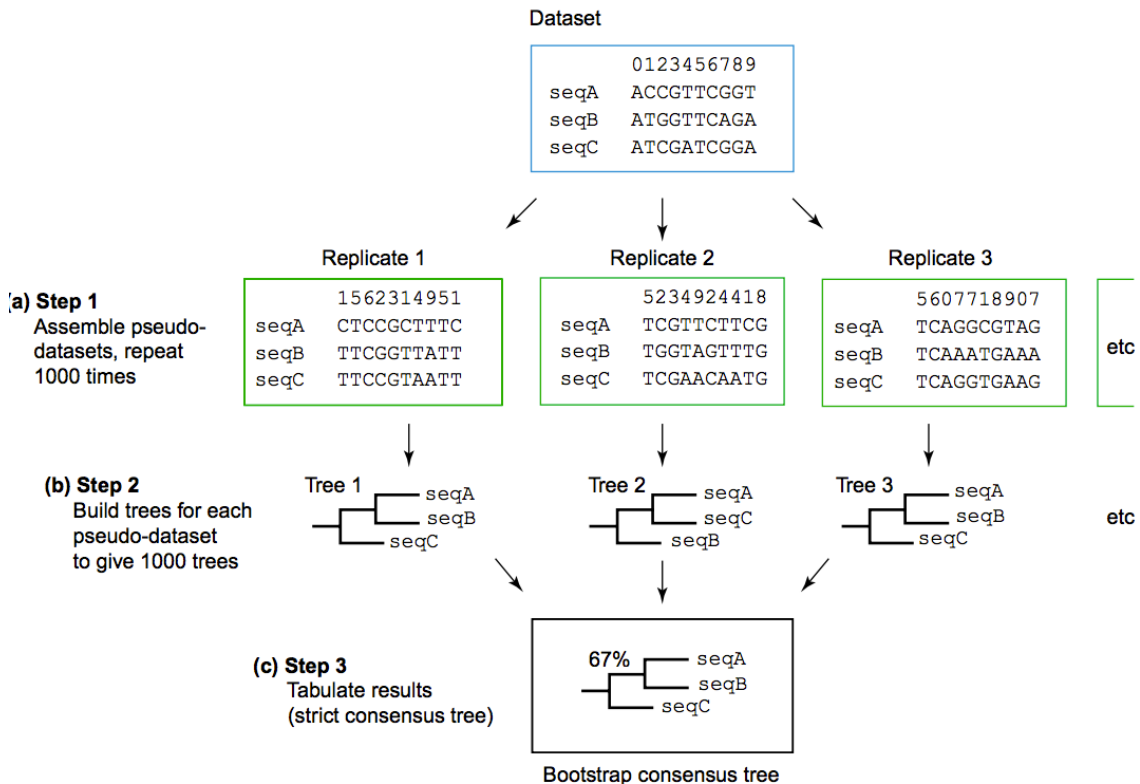
This time we want to go back to choose neighbor-joining algorithm because it is much faster than maximum likelihood. Here we also want to choose bootstrap method to test the phylogeny, then we will have statistical values for each node.



16. now we have the bootstrapped neighbor-joining tree.



How are bootstrap tests performed?



TRENDS in Genetics

17. if we want to have more accurate tree, we can tune more parameters to use sophisticated statistical models to estimate the tree

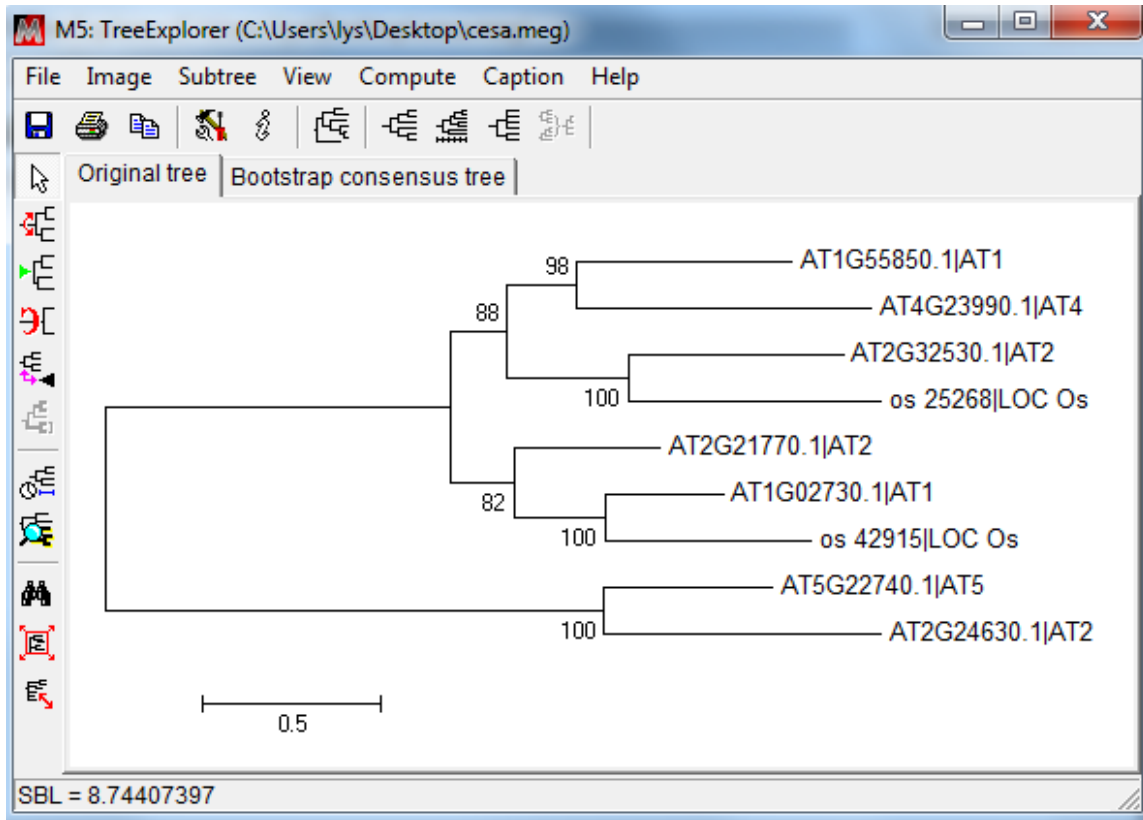
M5: Analysis Preferences

Options Summary

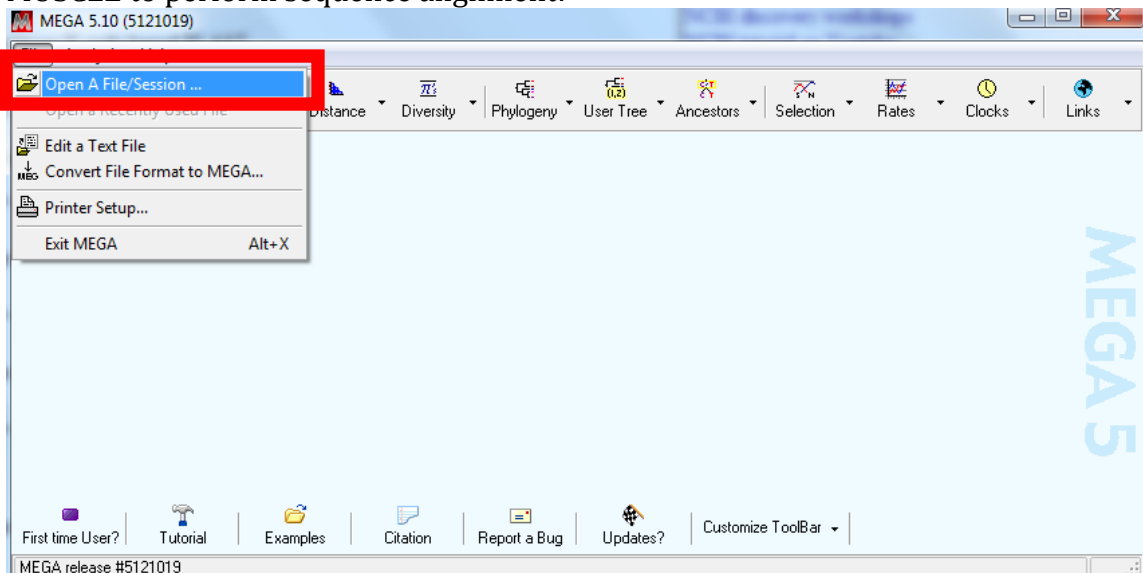
Option	Selection
Analysis	Phylogeny Reconstruction
Scope	All Selected Taxa
Statistical Method	Neighbor-joining
Phylogeny Test	
Test of Phylogeny	Bootstrap method
<i>No. of Bootstrap Replications</i>	500
Substitution Model	
Substitutions Type	Amino acid
Model/Method	Jones-Taylor-Thornton (JTT) model
Rates and Patterns	
Rates among Sites	Gamma Distributed (G)
<i>Gamma Parameter</i>	4
Pattern among Lineages	Same (Homogeneous)
Data Subset to Use	
Gaps/Missing Data Treatment	Pairwise deletion
<i>Site Coverage Cutoff (%)</i>	Not Applicable

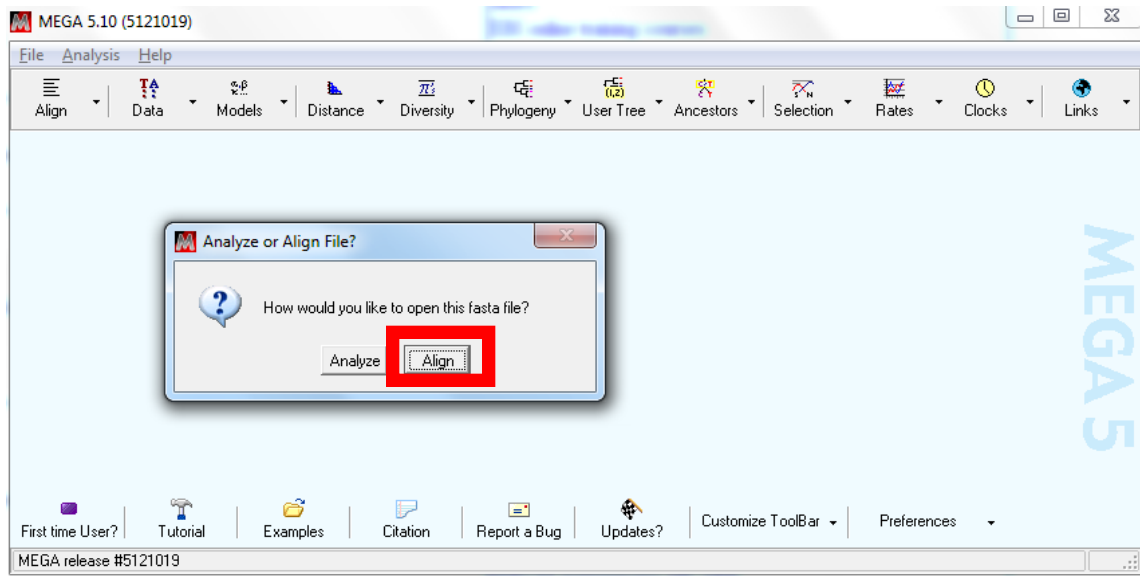
✓ Compute ✗ Cancel ? Help

Note the tree topology and bootstrap supports changed?

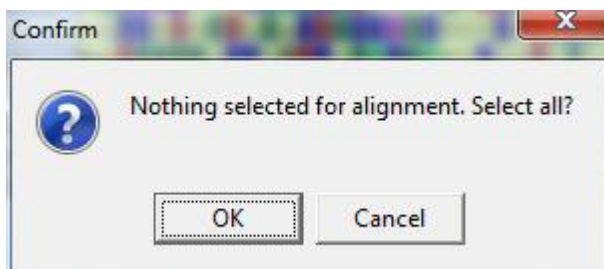
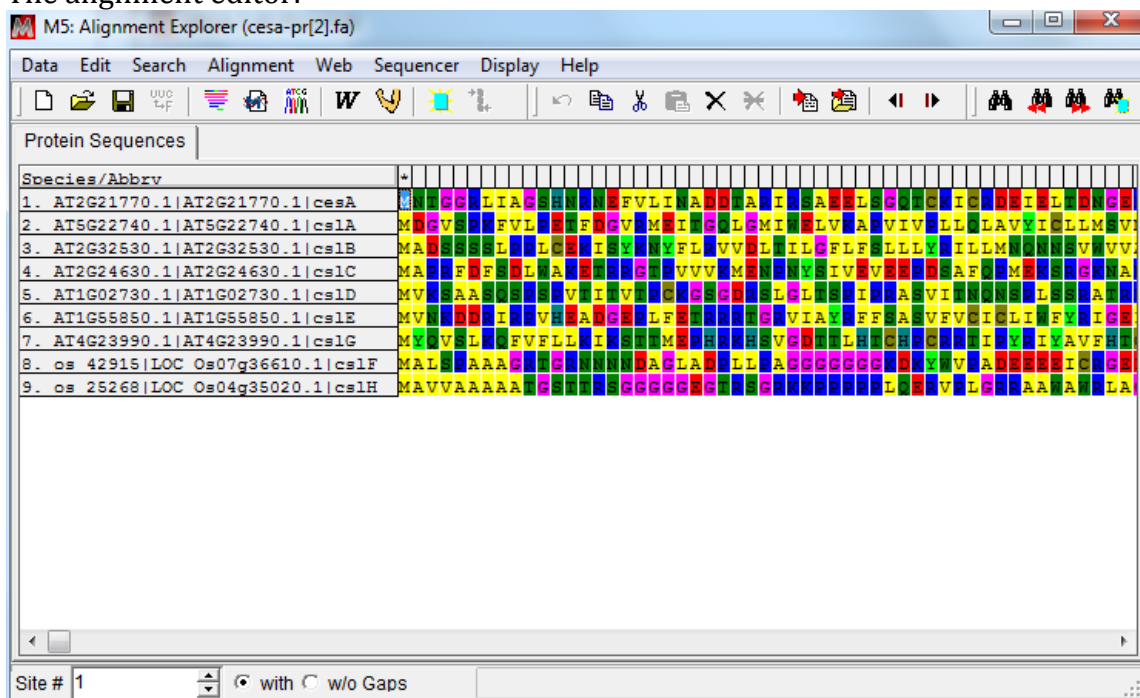


18. If we want to start from unaligned sequences and use built-in clustalw or MUSCLE to perform sequence alignment:





The alignment editor:



Select MUSCLE alignment

M5: MUSCLE - AppLink

Option	Selection
<input type="checkbox"/> Presets	None
Gap Penalties	
Gap Open	-2.9
Gap Extend	0
Hydrophobicity Multiplier	1.2
Memory/Iterations	
Max Memory in MB	808
Max Iterations	8
More Advanced Options	
Clustering Method (Iteration 1,2)	UPGMB
Clustering Method (Other Iterations)	UPGMB
Min Diag Length (lambda)	24
<input type="checkbox"/> Genetic Code (when using cDNA)	Standard
Alignment Info	MUSCLE Citation: Edgar, Robert C. (2004), MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Research 32(5), 1792-1797.

M5: Alignment Explorer (cesa-pr[2].fa)

Data Edit Search Alignment Web Sequencer Display Help

Protein Sequences

Species/Abbrv	
1. AT1G02730.1 AT1G02730.1 cs1D	MVFAA...IIV...C...L...I...A...VI...L...A...I
2. AT1G5850.1 AT1G5850.1 cs1E	-----MV...D...I
3. AT2G21770.1 AT2G21770.1 csA	-----M...G...L...I...A...F...V...L...I...A...D...A
4. AT2G24630.1 AT2G24630.1 cs1C	-----M...A...F...F...L...A...S...I
5. AT2G32530.1 AT2G32530.1 cs1B	-----M...A...S...I
6. AT4G23990.1 AT4G23990.1 cs1G	-----M...V...V...L...X...F...V...F...L...I...I...I
7. AT5G22740.1 AT5G22740.1 cs1A	-----M...C...V
8. os 25268 LOC Os04g35020.1 cs1H	-----M...A...V...V...A...A...A...A...C
9. os 42915 LOC Os07g36610.1 cs1F	-----M...A...L...L...A...A...A...C

Site # 1129 ☐ with ☐ w/o Gaps