Video Article

A Practical Guide to Phylogenetics for Nonexperts

Damien O'Halloran[1]

[1]Department of Biological Sciences and Institute for Neuroscience, The George Washington University

Correspondence to: Damien O'Halloran at damienoh@gwu.edu

URL: https://www.jove.com/video/50975
DOI: doi:10.3791/50975

Keywords: Basic Protocol, Issue 84, phylogenetics, multiple sequence alignments, phylogenetic tree, BLAST executables, basic local alignment search tool, Bayesian models

Date Published: 2/5/2014

Citation: O'Halloran, D. A Practical Guide to Phylogenetics for Nonexperts. *J. Vis. Exp.* (84), e50975, doi:10.3791/50975 (2014).

Abstract

Many researchers, across incredibly diverse foci, are applying phylogenetics to their research question(s). However, many researchers are new to this topic and so it presents inherent problems. Here we compile a practical introduction to phylogenetics for nonexperts. We outline in a step-by-step manner, a pipeline for generating reliable phylogenies from gene sequence datasets. We begin with a user-guide for similarity search tools via online interfaces as well as local executables. Next, we explore programs for generating multiple sequence alignments followed by protocols for using software to determine best-fit models of evolution. We then outline protocols for reconstructing phylogenetic relationships via maximum likelihood and Bayesian criteria and finally describe tools for visualizing phylogenetic trees. While this is not by any means an exhaustive description of phylogenetic approaches, it does provide the reader with practical starting information on key software applications commonly utilized by phylogeneticists. The vision for this article would be that it could serve as a practical training tool for researchers embarking on phylogenetic studies and also serve as an educational resource that could be incorporated into a classroom or teaching-lab.

Video Link

The video component of this article can be found at https://www.jove.com/video/50975/

Introduction

In order to understand how two (or more) species evolved, it is first necessary to obtain sequence or morphological data from each sample; these data represent quantities that we can use to measure their relationship through evolutionary space. Just like when measuring linear distance, having more data available (*e.g.* miles, inches, microns) will equate to a more accurate measurement. Ergo, the accuracy with which a researcher can deduce evolutionary distance is heavily influenced by the volume of informative data available to measure relationships. Furthermore, because different samples evolve at different rates and by different mechanisms, the method that we use to measure the relationship between two taxa also directly influences the accuracy of evolutionary measurements. Therefore, because evolutionary relationships are not directly observed but instead are extrapolated from sequence or morphological data, the problem of inferring evolutionary relationships becomes one of statistics. Phylogenetics is the branch of biology concerned with applying statistical models to patterns of evolution in order to optimally reconstruct the evolutionary history between taxa. This reconstruction between taxa is referred to as the taxa's *phylogeny*.

To help bridge the gap in expertise between molecular biologists and evolutionary biologists we describe here a step by step pipeline for inferring phylogenies from a set of sequences. Firstly, we detail the steps involved in database interrogation using the Basic Local Alignment Search Tool (BLAST[1]) algorithm through the web based interface and also by using local executables; this is often the first step in obtaining a list of similar sequences to an unidentified query, although some researchers may also be interested in gathering data for a single group via web interfaces such as Phylota (http://www.phylota.net/). BLAST is an algorithm for comparing primary amino acid or nucleotide sequence data against a database of sequences to search for "hits" that resemble the query sequence. The BLAST program was designed by Stephen Altschul *et al*. at the National Institutes of Health (NIH)[1]. The BLAST server consists of a number of different programs, and here is a list of some of the most common BLAST programs:

i) *Nucleotide-nucleotide BLAST (blastn):* This program requires a DNA sequence input and returns the most similar DNA sequences from the DNA database that the user specifies (*e.g.* for a specific organism).

ii) *Protein-protein BLAST (blastp):* Here the user inputs a protein sequence and the program returns the most similar protein sequences from the protein database that the user specifies.

iii) *Position-Specific Iterative BLAST (PSI-BLAST) (blastpgp):* The user input is a protein sequence which returns a set of closely related proteins, and from this dataset a conserved profile is generated. Next a new query is generated using only these conserved "motifs" which is used to interrogate a protein database and this returns a larger group of proteins from which a new set of conserved "motifs" are extracted and then used to interrogate a protein database until an even larger set of proteins are retuned and another profile is generated and the process repeated. By including related proteins into the query in each step this program allows the user to identify sequences that are more divergent.

iv) *Nucleotide 6-frame translation-protein (blastx):* Here the user provides a nucleotide sequence input which is converted into the six-frame conceptual translation products (*i.e.* both strands) against a protein sequence database.

v) *Nucleotide 6-frame translation-nucleotide 6-frame translation (tblastx):* This program takes a DNA nucleotide sequence input and translates the input into all six-frame conceptual translation products which it compares against the six-frame translations of a nucleotide sequence database.

vi) *Protein-nucleotide 6-frame translation (tblastn):* This program uses a protein sequence input to compare against all six reading frames of a nucleotide sequence database.

Next, we describe commonly used programs for generating a Multiple Sequence Alignment (MSA) from a sequence dataset, and this is followed by a user guide to programs that determine the best-fit models of evolution for a sequence dataset. Phylogenetic reconstruction is a statistical problem, and because of this, phylogenetic methods need to incorporate a statistical framework. This statistical framework becomes an evolutionary model that incorporates sequence change within the dataset. This evolutionary model is comprised of a set of assumptions about the process of nucleotide or amino-acid substitutions, and the best-fit model for a particular dataset can be selected through statistical testing. The fit to the data of different models can be compared via likelihood ratio tests (LRTs) or information criteria to select the best-fit model within a set of possible ones. Two common information criteria are the Akaike information criterion (AIC)[2] and the Bayesian information criterion (BIC)[3]. Once an optimal alignment is generated, there are many different methods to create a phylogeny from the aligned data. There are numerous methods of inferring evolutionary relationships; broadly, they can be divided into two categories: distance-based methods and sequence-based methods. Distance-based methods compute pairwise distances from sequences, and then use these distances to obtain the tree. Sequence-based methods use the sequence alignment directly, and usually search the tree space using an optimality criterion. We outline two sequence-based methods for reconstructing phylogenetic relationships: these are PhyML[4] which implements the maximum likelihood framework, and MrBayes[5] which uses Bayesian Markov Chain Monte Carlo inference. Likelihood and Bayesian methods provide a statistical framework for phylogenetic reconstruction. By providing user information on commonly used tree-building tools, we introduce the reader to the necessary data required to infer phylogenetic relationships.

Protocol

1. Basic Local Alignment Search Tool (BLAST): Online Interface

1. Click on this link to visit the BLAST[1] web server at the National Center for Biotechnology Information (NCBI). - http://blast.ncbi.nlm.nih.gov/Blast.cgi (**Figure 1**).
2. Input a FASTA formatted text sequence (see **Figure 2** for example) into the query box.
3. Click the appropriate BLAST program and relevant database or individual species of interest to use in the search and then click "BLAST".
   Note: FASTA formatted sequence begins with a description line indicated by a ">" sign. The description must follow immediately after the ">" sign, the sequence (*i.e.* nucleotides or amino acids) follow the description on the next line. The output from the BLAST search is viewed as HTML, plain text, XML, or hit tables (Text or csv) with the default set to HTML (**Figure 3**).

2. Basic Local Alignment Search Tool (BLAST): Local Executables

1. Download the latest BLAST command-line BLAST executables from this link:
   ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ -
2. em>For PC users: double-click the latest blast win32.exe file and accept the license agreement and click install.
   Note: The default installation directory is C:\ncbi-blast-2.2.27+.
3. Configure the PC environment variable as follows:
   1. Click the PC "start" button, and then right click "computer",
   2. Click "Properties" and in the pop-up click on the "advanced" tab
   3. Click the "Environment Variables button" and in the new pop-up click the "new" button under the "User variables for user" section
   4. In the pop-up add the variable name "Path" and variable value "C:\ncbi-blast-2.2.27+\bin.
      Note: the bin directory contains the executable (*i.e.* blastp, *etc*.).

4. em>For Mac users: Open the Terminal application (to do this just open "Finder" and search "Terminal" and this will display the "terminal" icon). Into the terminal window type:
   >ftp ftp.ncbi.nih.gov
   Note: can also type the URL used above in the example for PC
5. To access the NCBI ftp site type "anonymous" for Name and Password, and then type:
   >cd blast/executables/LATEST
6. List the executables by typing:
   >ls
7. Get the latest version by typing the following (or whatever the latest version currently is):
   >get ncbi-blast-2.2.7-macosx.tar.gz
8. Exit the NCBI ftp server site by typing "exit".
9. Decompress the downloaded files by typing:
   >tar -xzf ncbi-blast-2.2.7-macosx.tar.gz
10. Add the location of the binaries for the blast executable to your path so that the shell can search through this directory when looking for commands by typing:
    >PATH=$PATH:new_folder_location
11. Check if this added the location to your path by typing:

>echo $PATH

12. Download a preformatted BLAST databases (which are updated daily) by clicking here:
ftp://ftp.ncbi.nlm.nih.gov/blast/db/
13. Place the database into the "db" folder.
14. em>On a PC: open a MS-DOS prompt (to do this click "start" and type "cmd" in the search bar) and change the directory to the ncbi-blast folder by typing:
C:\Users>cd ..\ [moves up one folder]
C:\>cd ncbi-blast-2.2.27+
This will change the directory to:
C:\ncbi-blast-2.2.27+>
15. Create the database using the following "makedb" command:
>makedb –in db/briggsae.fasta –dbtype prot –out db/briggsae
Note: In the example below (**Figure 4**) the database is named "briggsae" and is comprised of one linkage group from the organism *Caenorhabditis briggsae.*
16. Create a query protein sequence called "test" by inserting a FASTA formatted protein text sequence into the "db" folder.
17. Interrogate the database via a blastp search by typing the following command:
>blastp –query db/test.txt –db db/briggsae –out text.txt
18. em>On a Mac: download a database for local Blast searches by accessing the NCBI ftp website as per the instructions above (step 2.4) and then type:
>lcd ../databases/
19. Download the genome or sequence of interest by typing:
>get NC_[Accession #].fna
Note: ".fna" refers to the FASTA formatted nucleotide sequence and ".faa" refers to the FASTA formatted amino acid sequences.
20. Type "quit" to exit the ftp site.
21. Make the database by typing:
>makeblastdb -in db/mouse.faa -out mouse -dbtype prot
22. Insert a FAST formatted query sequence into the "bin" folder and interrogate the database with the following command:
> blastp -query "your query.fasta" -db "your database" -out results.txt

3. Generating Multiple Sequence Alignments

1. Click on these links to access commonly used Multiple Sequence Alignment (MSA) programs:
ClustalW[6] http://www.clustal.org/
Kalign[7] http://msa.sbc.su.se/cgi-bin/msa.cgi
MAFFT[8,9] http://mafft.cbrc.jp/alignment/software/
MUSCLE[10] http://www.drive5.com/muscle/
T-Coffee[11] http://www.tcoffee.org/Projects/tcoffee/
PROBCONS[12] http://toolkit.tuebingen.mpg.de/probcons
2. Click on this link - http://tcoffee.crg.cat/apps/tcoffee/do:regular - and input FASTA formatted sequence data into the query box
Note: A sample output from T-Coffee can be seen in **Figure 5**, similar residues are color coded.
3. Download the Clustal MSA as a command line version (ClustalW) or a graphical version (ClustalX) by clicking this link: http://www.clustal.org/clustal2/ - then click on the appropriate executable (*i.e.* win, Linux, Mac OS X).
4. Upload data as FASTA formatted sequence text and align (**Figure 6**).

4. Determining Best-fit Models of Evolution

1. Click here to download the ProtTest[13] program:
http://darwin.uvigo.es/our-software/
2. Once ProtTest is downloaded, double-click on the ProtTest.jar file
3. Once ProtTest is launched, click on "select file" and load the sequence data (**Figure 7**).
4. Then click "start" and the program will begin (**Figure 8**).
Note: After completion of the run (**Figure 8**), the program will indicate the best model based on criteria *e.g.* "Best model according to AIC: WAG+I+G"

5. Inferring Sequence Based Phylogenies by Maximum Likelihood or Bayesian Inference

1. Downloaded PhyML[4] here:
https://code.google.com/p/phyml/
2. Launch the executable by double clicking the appropriate application (*i.e.* phyml Windows, phyml Linux, *etc.*) and the interface window will pop up (**Figure 9**).
3. Load the input sequence as a PHYLIP formatted sequence by typing:
>"file name".phy
Note: *To convert between sequence formats, use the "Readseq" web program available at - http://iubio.bio.indiana.edu/cgi-bin/readseq.cgi.*
4. Launch the program by typing "Y".
5. Download MrBayes[5] here:
http://mrbayes.sourceforge.net/download.php
6. To start the program click on the executable file and read NEXUS formatted sequence data into the program by typing:
>execute "file name".nex

7. Set the evolutionary model.
8. Select the number of generations to run by typing:
   >mcmcp ngen = 1000000 [this sets the number of generations to 1000000]
   >sump burnin =10000 [this sets the burnin to 10000]
9. Save the branch lengths in the results file by typing:
   >mcmcp savebrlens = yes
10. Run the analysis by typing:
    >mcmc
11. Summarize the trees using the "sumt" command.

6. Visualizing Phylogenies

1. View a list of tree viewer programs here:
   http://www.treedyn.org/overview/editors.html
2. Download the TreeView[14] program here:
   http://taxonomy.zoology.gla.ac.uk/rod/treeview.html

Representative Results

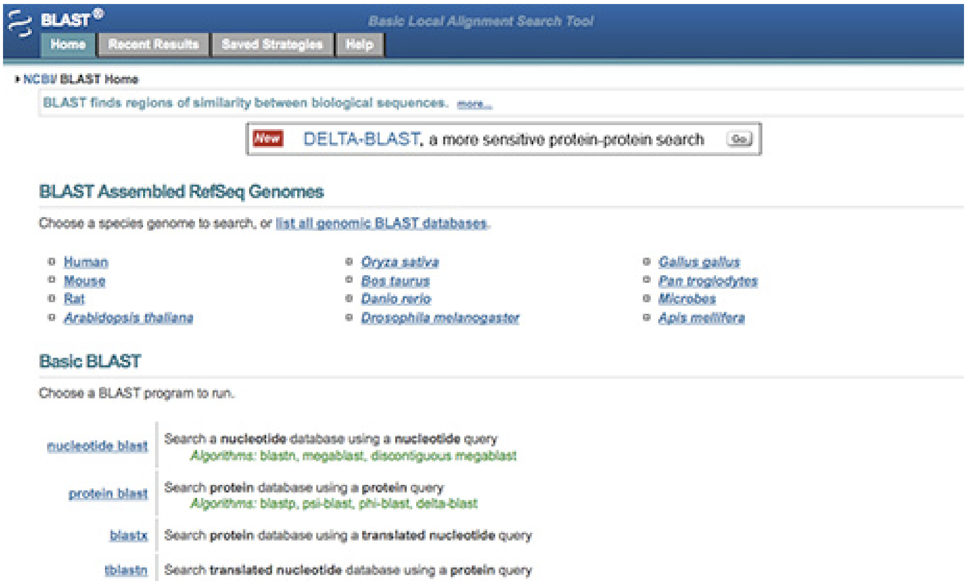
Finding similarities to a query allows researchers to ascribe a potential identity to new sequences and also infer relationships between sequences. The file input type for BLAST[1] is FASTA formatted text sequence or GenBank accession number. FASTA formatted sequence begins with a description line indicated by a ">" sign (**Figure 2**). The description must follow immediately after the ">" sign, the sequence (*i.e.* nucleotides or amino acids) follow the description on the next line. When saving and editing sequence files, it is best to use a text editor such as "Notepad" on PC or TextWrangler (http://www.barebones.com/products/textwrangler/) for Mac. The BLAST algorithm performs "local" alignments, which searches for short stretches of sequence similarity. After the algorithm has looked up all possible "stretches" from the query sequence and maximally extended these sequences, it then assembles alignments for each query sequence pair. It is then important to understand how good these matches are, and so BLAST applies statistics to each hit which comprise an expect value (E) and a bit score. The E value gives an indication of the statistical significance for a match. The lower the E-value, the more significant the hit, for example a sequence alignment with an E-value of 0.05 means that the likelihood of this match occurring by chance alone is 5 in 100. The bit score uses a specific scoring matrix to provide an indication of how good the alignment is. The higher the bit score, the better the alignment. Similar to the online version of BLAST, there are a number of parameters that can be set via commands using the local BLAST executable. A comprehensive resource describing these commands can be found here - http://www.ncbi.nlm.nih.gov/books/NBK1762/. The output of the local search is a text file just like the output from the online BLAST interface (**Figure 4**).

A Multiple Sequence Alignment (MSA) is a sequence alignment of three or more primary sequences composed of amino acids, DNA, or RNA. ClustalW[6] released in 1994, is one of the most popular MSA tools for biologists. A user friendly online interface that provides one-stop access to several popular MSA tools can be found at the EMBL-EBI server here - http://www.ebi.ac.uk/Tools/msa. The input for each program can be FASTA formatted sequence data (see **Figure 2**) although many different formats are also accepted, and numerous mirror sites for each can be found online. Numerous parameters like gap penalties and output formats can be easily chosen. A sample output from the MSA T-Coffee can be seen in **Figure 5,** where similar residues are color coded. In some cases, the MSA tool can also be downloaded and executed locally. Clustal can be downloaded as a command line version (ClustalW) or a graphical version (ClustalX) from this website - http://www.clustal.org/clustal2/. To download, just click on the appropriate executable (*i.e.* win, Linux, Mac OS X). For Windows the program executable will download and a pop-up menu will require the user to click "Run", and then installation will begin. The program is very intuitive, sequences can be loaded from a text file containing sequences formatted as NBRF/PIR, FASTA, EMBL/Swiss-Prot, Clustal, GCC/MSF, GCG9 RSF, and GDE. Sequences are aligned by clicking "do complete alignment" from the "alignment" menu. A sample alignment of six protein sequences aligned using ClustalX can be seen in **Figure 6**. Various parameters such as font size and color can be easily modified, and editing of sequences is done by clicking on the "Edit" menu. Manually refined alignments are often superior to fully automated methods and because of this, MSA tool development is a very active area of research. Some common alignment editors can be found at the following links: Se-Al - http://tree.bio.ed.ac.uk/software/seal/; BSEdit - http://www.bsedit.org/; JalView - http://www.jalview.org/; SeaView - http://pbil.univ-lyon1.fr/software/seaview.html.

For amino-acid alignments the program ProtTest[13] is used to determine the selection of best-fit models of amino acid replacements within the data. ProtTest makes this selection by finding the model from the list of candidate models with the smallest Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) score, or Decision Theory Criterion (DT). The latest version of ProtTest (version 3.2) includes 15 different rate matrices that result in 120 different models. The user must have Java Runtime on their system to run ProtTest. Java Runtime is freely available here - http://www.java.com/en/download/chrome.jsp. Sequences are inputted as PHYLIP or NEXUS format. To convert between sequence formats, use the "Readseq" web program available at - http://iubio.bio.indiana.edu/cgi-bin/readseq.cgi. Click on "select file" and load the sequence data. Then click "start" and the program will begin. To modify the number of models selected, you can click the "models" button. Once the program begins it will display a progress bar at the bottom and list the models as they are being analyzed (**Figure 8**). The developer's manual provides a comprehensive user guide to ProtTest (can be found in the "manual" folder when the program is downloaded) and also provides descriptions of the models and parameters. More background information can be found here - https://code.google.com/p/prottest3/wiki/Background. There is also an online web interface for ProtTest which functions just like the downloaded version except that it can only handle a limited number of sequences. This web interface can be accessed by clicking here - http://darwin.uvigo.es/software/prottest2_server.html. For nucleotide datasets the program jModelTest[15] is used to examine the statistical selection of best-fit models of nucleotide substitutions by implementing the AIC, BIC, and DT criteria outlined above and also hierarchical and dynamical likelihood ration tests (hLRT and dLRT). jModelTest is optimized for Mac OS X. For the input, multiple formats are permitted. A clear step-by-step guide is available by the developers here - http://computing.bio.cam.ac.uk/local/doc/jmodeltest.pdf

PhyML is a program that estimates maximum likelihood phylogenies from alignments of nucleotide or amino acid sequences. PhyML will incorporate a large number of substitution models coupled to various options to search tree topology space (**Figure 10**). The program will save results into two text files. The first file will contain the ML tree in Newick format which can easily be viewed using a Tree viewer (see protocol 6), and the other file will contain the statistics (filename, model, Log-likelihood scores, *etc.*) of the analysis. All parameters are very easily set by following the Menu items. More detailed descriptions of each Menu option are explained in the PhyML manual available on the PhyML download page - https://code.google.com/p/phyml/downloads/list. MrBayes[5] is a program that utilizes Bayesian MCMC inference across a number of evolutionary models to reconstruct phylogenetic relationships. The program behaves the same on all platforms and once downloaded the installer will install the executable. To start the program, simply click on the executable. There are numerous models that can be set and details of each model and their commands can be found here - http://mrbayes.sourceforge.net/wiki/index.php/Tutorial. Another help option is to type "help lset" – this will provide details on Model setting. For example "Prset aamodelpr=mixed" will permit mixed modeling or "prset aamodelpr=fixed(wag)" will set the amino acid model to the WAG model. An outgroup can be easily set by specifying the Taxon number "outgroup 30"; the program automatically lists the sequences/Taxa by number. If an outgroup is not specified the tree will be unrooted. Once the program is running (**Figure 11**) the progress can be viewed in specific intervals which can be set using the "printfreq=X" command. More details on when to stop the analysis (*i.e.* how many generations to run for) can be found in the user's manual. Clade values on a cladogram are provided in the results alongside a phylogram which is also provided in Newick format that can easily be viewed using a tree viewer (see protocol 6).

Once a phylogenetic tree is generated, the topology needs to be visualized. There are many online tools and downloadable applications used to visualize tree topologies. A partial list of popular programs can be viewed here - http://en.wikipedia.org/wiki/List_of_phylogenetic_tree_visualization_software, and a more comprehensive list can be found here - http://www.treedyn.org/overview/editors.html. TreeView[14] and TreeDyn[16] are two popular choices. Both are very user friendly and easy to become familiar with the various options. TreeView runs on Mac and Windows, using almost identical interfaces. The input can be one of several formats including NEXUS, PHYLIP, Hennig86, MEGA, and ClustalW/X. TreeView (**Figure 12**) also includes a tree editor that allows the user to move branches, reroot trees, and rearrange the appearance of the tree.



BLAST finds regions of similarity between biological sequences. more...

DELTA-BLAST, a more sensitive protein-protein search Go

BLAST Assembled RefSeq Genomes

Choose a species genome to search, or list all genomic BLAST databases.

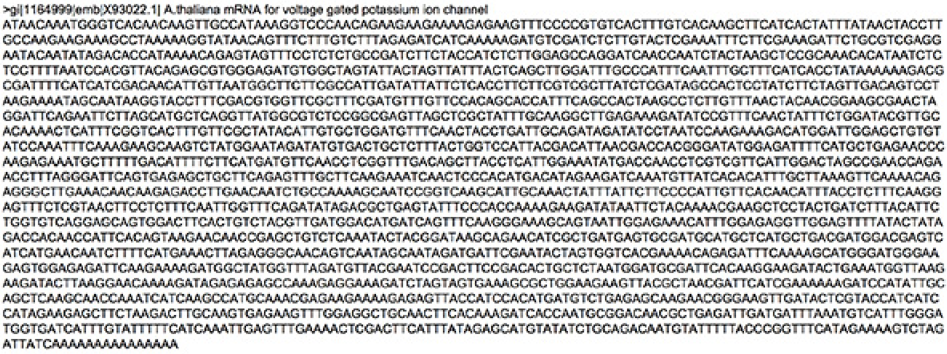
- Human
- Mouse
- Rat
- *Arabidopsis thaliana*
- *Oryza sativa*
- *Bos taurus*
- *Danio rerio*
- *Drosophila melanogaster*
- *Gallus gallus*
- *Pan troglodytes*
- *Microbes*
- *Apis mellifera*

Basic BLAST

Choose a BLAST program to run.

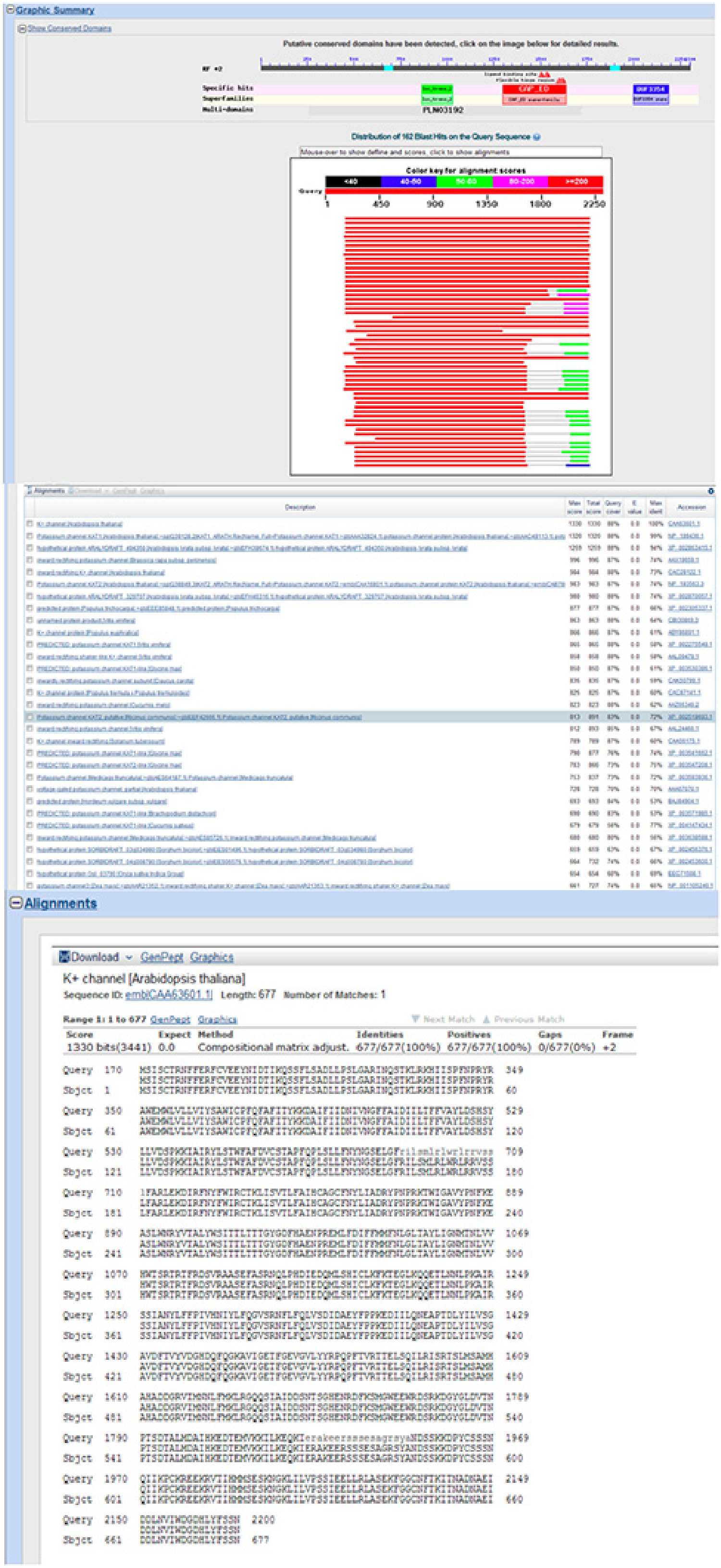
nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast, delta-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query

Figure 1. NCBI BLAST web-page. The BLAST web server contains a suite of BLAST programs and is hosted by the National Center for Biotechnology Information (NCBI). Click here to view larger image.



```
>gi|1164999|emb|X93022.1| A.thaliana mRNA for voltage gated potassium ion channel
ATAACAAATGGGTCACAACAAGTTGCCATAAAGGTCCCAACAGAAGAAGAAAAGAGAAGTTTCCCCGTGTCACTTTGTCACAAGCTTCATCACTATTTATAACTACCTT
GCCAAGAAGAAAGCCTAAAAAGGTATAACAGTTTCTTTGTCTTTAGAGATCATCAAAAAGATGTCGATCTCTTGTACTCGAAATTTCTTCGAAAGATTCTGCGTCGAGG
AATACAATATAGACACCATAAAACAGAGTAGTTTCCTCTCTGCCGATCTTCTACCATCTCTTGGAGCCAGGATCAACCAATCTACTAAGCTCCGCAAACACATAATCTC
TCCTTTTAATCCACGTTACAGAGCGTGGGAGATGTGGCTAGTATTACTAGTTATTTACTCAGCTTGGATTTGCCCATTTCAATTTGCTTTCATCACCTATAAAAAAGACG
CGATTTTCATCATCGACAACATTGTTAATGGCTTCTTCGCCATTGATATTATTCTCACCTTCTTCGTCGGCTTATCTCGATAGCCACTCCTATCTTCTAGTTGACAGTCCT
AAGAAAATAGCAATAAGGTACCTTTCGACGTGGTTCGCTTTCGATGTTTGTTCCACAGCACCATTTCAGCCACTAAGCCTCTTGTTTAACTACAACGGAAGCGAACTA
GGATTCAGAATTCTTAGCATGCTCAGGTTATGGCGTCTCCGGCGAGTTAGCTCGCTATTTGCAAGGCTTGAGAAAGATATCCGTTTCAACTATTTCTGGATACGTTGC
ACAAAACTCATTTCGGTCACTTTGTTCGCTATACATTGTGCTGGATGTTTCAACTACCTGATTGCAGATAGATATCCTAATCCAAGAAAGACATGGATTGGAGCTGTGT
ATCCAAATTTCAAAGAAGCAAGTCTATGGAATAGATATGTGACTGCTCTTTACTGGTCCATTACGACATTAACGACCACGGGATATGGAGATTTTCATGCTGAGAACCC
AAGAGAAATGCTTTTTGACATTTTCTTCATGATGTTCAACCTCGGTTTGACAGCTTACCTCATTGGAAATATGACCAACCTCGTCGTTCATTGGACTAGCCGAACCAGA
ACCTTTAGGGATTCAGTGAGAGCTGCTTCAGAGTTTGCTTCAAGAAATCAACTCCCACATGACATAGAAGATCAAATGTTATCACACATTTGCTTAAAGTTCAAAACAG
AGGGCTTGAAACAACAAGAGACCTTGAACAATCTGCCAAAAGCAATCCGGTCAAGCATTGCAAACTATTTATTCTTCCCCATTGTTCACAACATTTACCTCTTTCAAGG
AGTTTCTCGTAACTTCCTCTTTCAATTGGTTTCAGATATAGACGCTGAGTATTTCCCACCAAAAGAAGATATAATTCTACAAAACGAAGCTCCTACTGATCTTTACATTC
TGGTGTCAGGAGCAGTGGACTTCACTGTCTACGTTGATGGACATGATCAGTTTCAAGGGAAAGCAGTAATTGGAGAAACATTTGGAGAGGTTGGAGTTTTATACTATA
GACCACAACCATTCACAGTAAGAACAACCGAGCTGTCTCAAATACTACGGATAAGCAGAACATCGCTGATGAGTGCGATGCATGCTCATGCTGACGATGGACGAGTC
ATCATGAACAATCTTTTCATGAAACTTAGAGGGCAACAGTCAATAGCAATAGATGATTCGAATACTAGTGGTCACGAAAACAGAGATTTCAAAAGCATGGGATGGGAA
GAGTGGAGAGATTCAAGAAAAGATGGCTATGGTTTAGATGTTACGAATCCGACTTCCGACACTGCTCTAATGGATGCGATTCACAAGGAAGATACTGAAATGGTTAAG
AAGATACTTAAGGAACAAAAGATAGAGAGAGCCAAAGAGGAAAGATCTAGTAGTGAAAGCGCTGGAAGAAGTTACGCTAACGATTCATCGAAAAAAGATCCATATTGC
AGCTCAAGCAACCAAATCATCAAGCCATGCAAACGAGAAGAAAAGAGAGTTACCATCCACATGATGTCTGAGAGCAAGAACGGGAAGTTGATACTCGTACCATCATC
CATAGAAGAGCTTCTAAGACTTGCAAGTGAGAAGTTTGGAGGCTGCAACTTCACAAAGATCACCAATGCGGACAACGCTGAGATTGATGATTTAAATGTCATTTGGGA
TGGTGATCATTTGTATTTTTCATCAAATTGAGTTTGAAAACTCGACTTCATTTATAGAGCATGTATATCTGCAGACAATGTATTTTTACCCGGTTTCATAGAAAAGTCTAG
ATTATCAAAAAAAAAAAAAAA
```

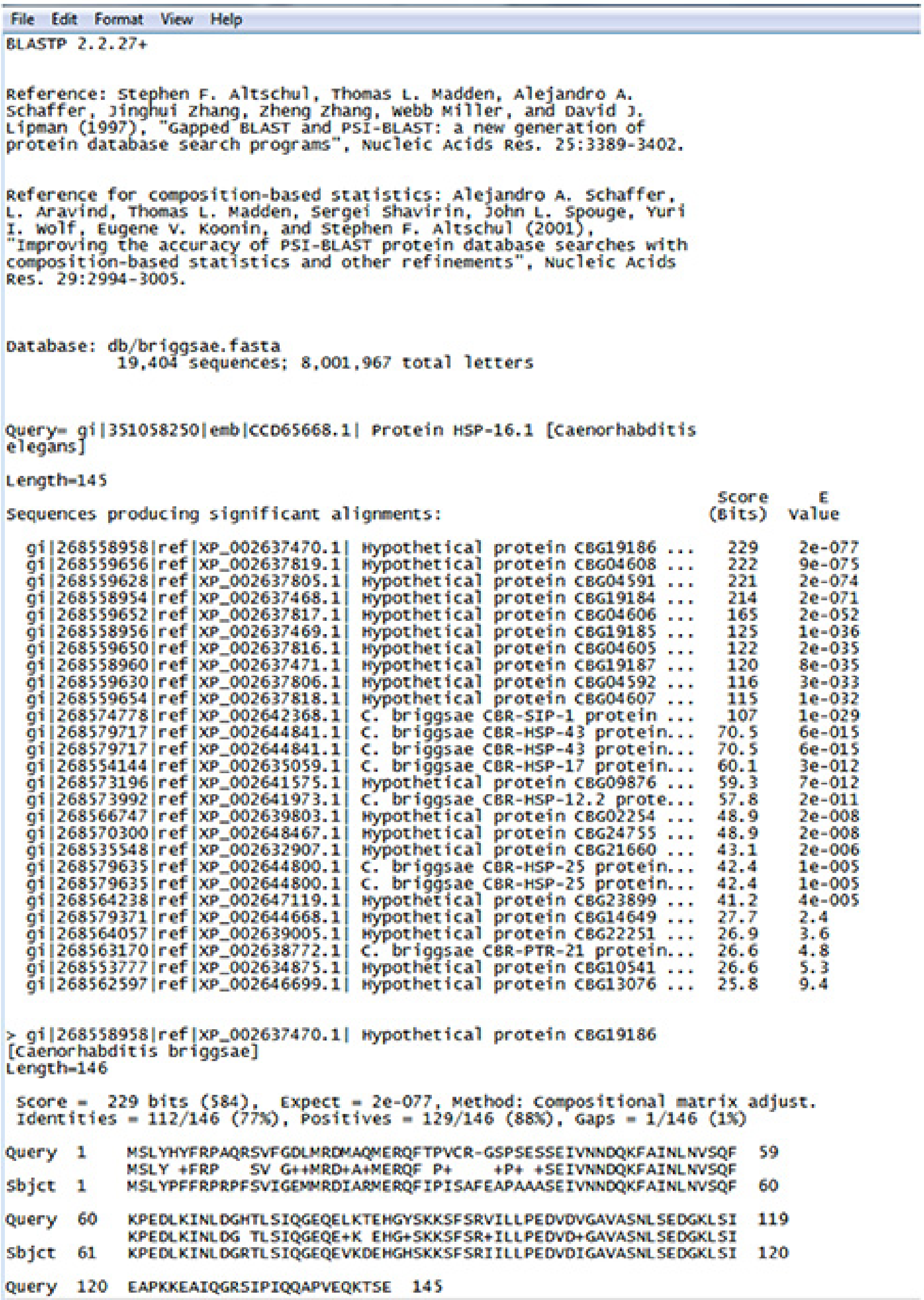
Figure 2. FASTA formatted sequence. FASTA format begins with a description line indicated by a ">". The description must follow immediately after the ">" sign, the sequence (*i.e.* nucleotides or amino acids) follow the description on the next line. Click here to view larger image.



	Description	Max score	Total score	Query cover	E value	Max ident	Accession
□	K+ channel [Arabidopsis thaliana]	1330	1330	88%	0.0	100%	CAA63601.1
□	Potassium channel KAT1 [Arabidopsis thaliana] >sp|Q39128.2|KAT1_ARATH RecName: Full=Potassium channel KAT1 >gb|AAA32824.1| potassium channel protein [Arabidopsis thaliana] >gb|AAC49113.1| pota	1320	1320	88%	0.0	99%	NP_199436.1
□	hypothetical protein ARALYDRAFT_494350 [Arabidopsis lyrata subsp. lyrata] >gb|EFH39574.1| hypothetical protein ARALYDRAFT_494350 [Arabidopsis lyrata subsp. lyrata]	1259	1259	88%	0.0	94%	XP_002863415.1
□	inward rectifying potassium channel [Brassica rapa subsp. pekinensis]	996	996	87%	0.0	74%	AAV13659.1
□	inward rectifying K+ channel [Arabidopsis thaliana]	984	984	88%	0.0	77%	CAC05122.1
□	Potassium channel KAT2 [Arabidopsis thaliana] >sp|Q38849.2|KAT2_ARATH RecName: Full=Potassium channel KAT2 >emb|CAA16801.1| potassium channel protein KAT2 [Arabidopsis thaliana] >emb|CAB78	983	983	87%	0.0	74%	NP_193563.3
□	hypothetical protein ARALYDRAFT_329797 [Arabidopsis lyrata subsp. lyrata] >gb|EFH45316.1| hypothetical protein ARALYDRAFT_329797 [Arabidopsis lyrata subsp. lyrata]	980	980	88%	0.0	74%	XP_002870057.1
□	predicted protein [Populus trichocarpa] >gb|EEE85848.1| predicted protein [Populus trichocarpa]	877	877	87%	0.0	66%	XP_002305337.1
□	unnamed protein product [Vitis vinifera]	863	863	88%	0.0	64%	CBI30893.3
□	K+ channel protein [Populus euphratica]	866	866	87%	0.0	61%	ADY69691.1
□	PREDICTED: potassium channel KAT1 [Vitis vinifera]	865	865	88%	0.0	58%	XP_002270549.1
□	inward rectifying shaker-like K+ channel [Vitis vinifera]	858	858	88%	0.0	58%	AAL09479.1
□	PREDICTED: potassium channel KAT1-like [Glycine max]	850	850	87%	0.0	61%	XP_003530300.1
□	inwardly rectifying potassium channel subunit [Daucus carota]	836	836	87%	0.0	59%	CAA05790.1
□	K+ channel protein [Populus tremula x Populus tremuloides]	825	825	87%	0.0	60%	CAC87141.1
□	inward rectifying potassium channel [Cucumis melo]	823	823	88%	0.0	62%	AAZ95349.2
□	Potassium channel KAT2, putative [Ricinus communis] >gb|EEF42666.1| Potassium channel KAT2, putative [Ricinus communis]	813	891	82%	0.0	72%	XP_002518693.1
□	inward rectifying potassium channel [Vitis vinifera]	812	893	85%	0.0	67%	AAL24465.1
□	K+ channel inward rectifying [Solanum tuberosum]	789	789	87%	0.0	60%	CAA60175.1
□	PREDICTED: potassium channel KAT1-like [Glycine max]	790	877	76%	0.0	74%	XP_003541662.1
□	PREDICTED: potassium channel KAT2-like [Glycine max]	783	866	73%	0.0	75%	XP_003547208.1
□	Potassium channel [Medicago truncatula] >gb|AES64187.1| Potassium channel [Medicago truncatula]	753	837	73%	0.0	72%	XP_003593606.1
□	voltage-gated potassium channel, partial [Arabidopsis thaliana]	728	728	70%	0.0	70%	AAA67979.1
□	predicted protein [Hordeum vulgare subsp. vulgare]	693	693	84%	0.0	53%	BAJ84904.1
□	PREDICTED: potassium channel KAT1-like [Brachypodium distachyon]	690	690	83%	0.0	53%	XP_003571980.1
□	PREDICTED: potassium channel KAT1-like [Cucumis sativus]	679	679	58%	0.0	77%	XP_004147434.1
□	inward rectifying potassium channel [Medicago truncatula] >gb|AES85725.1| inward rectifying potassium channel [Medicago truncatula]	680	680	80%	0.0	56%	XP_003638598.1
□	hypothetical protein SORBIDRAFT_03g034980 [Sorghum bicolor] >gb|EES01495.1| hypothetical protein SORBIDRAFT_03g034980 [Sorghum bicolor]	659	659	67%	0.0	67%	XP_002456379.1
□	hypothetical protein SORBIDRAFT_04g006790 [Sorghum bicolor] >gb|EES06075.1| hypothetical protein SORBIDRAFT_04g006790 [Sorghum bicolor]	664	732	74%	0.0	66%	XP_002452600.1
□	hypothetical protein OsI_03790 [Oryza sativa Indica Group]	654	654	60%	0.0	69%	EEC71595.1
□	potassium channel3 [Zea mays] >gb|ABR21352.1| inward rectifying shaker K+ channel [Zea mays] >gb|ABR21353.1| inward rectifying shaker K+ channel [Zea mays]	661	727	74%	0.0	66%	NP_001105240.1

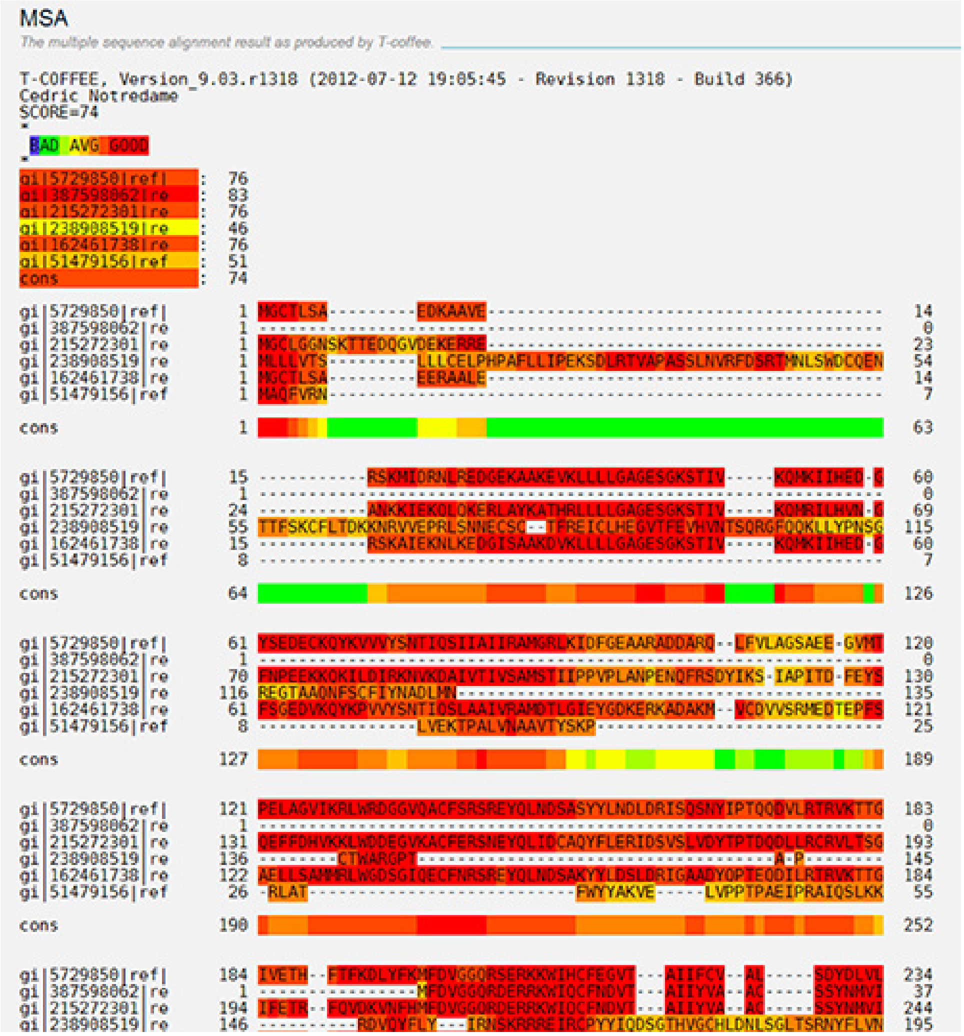
```
K+ channel [Arabidopsis thaliana]
Sequence ID: emb|CAA63601.1|  Length: 677  Number of Matches: 1

Range 1: 1 to 677 GenPept Graphics                    ▼ Next Match  ▲ Previous Match
 Score             Expect  Method                       Identities      Positives       Gaps       Frame
 1330 bits(3441)   0.0     Compositional matrix adjust. 677/677(100%)   677/677(100%)   0/677(0%)  +2

Query  170   MSISCTRNFFERFCVEEYNIDTIKQSSFLSADLLPSLGARINQSTKLRKHIISPFNPRYR  349
             MSISCTRNFFERFCVEEYNIDTIKQSSFLSADLLPSLGARINQSTKLRKHIISPFNPRYR
Sbjct  1     MSISCTRNFFERFCVEEYNIDTIKQSSFLSADLLPSLGARINQSTKLRKHIISPFNPRYR  60

Query  350   ANEMWLVLLVIYSAWICPFQFAFITYKKDAIFIIDNIVNGFFAIDIILTFFVAYLDSHSY  529
             ANEMWLVLLVIYSAWICPFQFAFITYKKDAIFIIDNIVNGFFAIDIILTFFVAYLDSHSY
Sbjct  61    ANEMWLVLLVIYSAWICPFQFAFITYKKDAIFIIDNIVNGFFAIDIILTFFVAYLDSHSY  120

Query  530   LLVDSPKKIAIRYLSTWFAFDVCSTAPFQPLSLLFNYNGSELGFrilsmlrlwrlrrvss  709
             LLVDSPKKIAIRYLSTWFAFDVCSTAPFQPLSLLFNYNGSELGFRILSMLRLWRLRRVSS
Sbjct  121   LLVDSPKKIAIRYLSTWFAFDVCSTAPFQPLSLLFNYNGSELGFRILSMLRLWRLRRVSS  180

Query  710   lFARLEKDIRFNYFWIRCTKLISVTLFAIHCAGCFNYLIADRYPNPRKTWIGAVYPNFKE  889
             LFARLEKDIRFNYFWIRCTKLISVTLFAIHCAGCFNYLIADRYPNPRKTWIGAVYPNFKE
Sbjct  181   LFARLEKDIRFNYFWIRCTKLISVTLFAIHCAGCFNYLIADRYPNPRKTWIGAVYPNFKE  240

Query  890   ASLWNRYVTALYWSITTLTTTGYGDFHAENPREMLFDIFFMMFNLGLTAYLIGNMTNLVV  1069
             ASLWNRYVTALYWSITTLTTTGYGDFHAENPREMLFDIFFMMFNLGLTAYLIGNMTNLVV
Sbjct  241   ASLWNRYVTALYWSITTLTTTGYGDFHAENPREMLFDIFFMMFNLGLTAYLIGNMTNLVV  300

Query  1070  HWTSRTRTFRDSVRAASEFASRNQLPHDIEDQMLSHICLKFKTEGLKQQETLNNLPKAIR  1249
             HWTSRTRTFRDSVRAASEFASRNQLPHDIEDQMLSHICLKFKTEGLKQQETLNNLPKAIR
Sbjct  301   HWTSRTRTFRDSVRAASEFASRNQLPHDIEDQMLSHICLKFKTEGLKQQETLNNLPKAIR  360

Query  1250  SSIANYLFFPIVHNIYLFQGVSRNFLFQLVSDIDAEYFPPKEDIILQNEAPTDLYILVSG  1429
             SSIANYLFFPIVHNIYLFQGVSRNFLFQLVSDIDAEYFPPKEDIILQNEAPTDLYILVSG
Sbjct  361   SSIANYLFFPIVHNIYLFQGVSRNFLFQLVSDIDAEYFPPKEDIILQNEAPTDLYILVSG  420

Query  1430  AVDFTVYVDGHDQFQGKAVIGETFGEVGVLYYRPQPFTVRITELSQILRISRTSLMSAMH  1609
             AVDFTVYVDGHDQFQGKAVIGETFGEVGVLYYRPQPFTVRITELSQILRISRTSLMSAMH
Sbjct  421   AVDFTVYVDGHDQFQGKAVIGETFGEVGVLYYRPQPFTVRITELSQILRISRTSLMSAMH  480

Query  1610  AHADDGRVIMNNLFMKLRGQQSIAIDDSNTSGHENRDFKSMGWEEWRDSRKDGYGLDVTN  1789
             AHADDGRVIMNNLFMKLRGQQSIAIDDSNTSGHENRDFKSMGWEEWRDSRKDGYGLDVTN
Sbjct  481   AHADDGRVIMNNLFMKLRGQQSIAIDDSNTSGHENRDFKSMGWEEWRDSRKDGYGLDVTN  540

Query  1790  PTSDTALMDAIHKEDTEMVKKILKEQKIerakeerssesagrsyaNDSSKKDPYCSSSN  1969
             PTSDTALMDAIHKEDTEMVKKILKEQKIERAKEERSSSESAGRSYANDSSKKDPYCSSSN
Sbjct  541   PTSDTALMDAIHKEDTEMVKKILKEQKIERAKEERSSSESAGRSYANDSSKKDPYCSSSN  600

Query  1970  QIIKPCKREEKRVTIHMMSESKNGKLILVPSSIEELLRLASEKFGGCNFTKITNADNAEI  2149
             QIIKPCKREEKRVTIHMMSESKNGKLILVPSSIEELLRLASEKFGGCNFTKITNADNAEI
Sbjct  601   QIIKPCKREEKRVTIHMMSESKNGKLILVPSSIEELLRLASEKFGGCNFTKITNADNAEI  660

Query  2150  DDLNVIWDGDHLYFSSN  2200
             DDLNVIWDGDHLYFSSN
Sbjct  661   DDLNVIWDGDHLYFSSN  677
```

Figure 3. HTML output from a BLAST search. The output from the BLAST search illustrates the areas of identity within the query sequence, and also provides bit-scores, expect values and pairwise alignments with each match. Click here to view larger image.



```
BLASTP 2.2.27+


Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro A.
Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J.
Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of
protein database search programs", Nucleic Acids Res. 25:3389-3402.


Reference for composition-based statistics: Alejandro A. Schaffer,
L. Aravind, Thomas L. Madden, Sergei Shavirin, John L. Spouge, Yuri
I. Wolf, Eugene V. Koonin, and Stephen F. Altschul (2001),
"Improving the accuracy of PSI-BLAST protein database searches with
composition-based statistics and other refinements", Nucleic Acids
Res. 29:2994-3005.



Database: db/briggsae.fasta
           19,404 sequences; 8,001,967 total letters



Query= gi|351058250|emb|CCD65668.1| Protein HSP-16.1 [Caenorhabditis
elegans]

Length=145
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

  gi|268558958|ref|XP_002637470.1| Hypothetical protein CBG19186 ...   229     2e-077
  gi|268559656|ref|XP_002637819.1| Hypothetical protein CBG04608 ...   222     9e-075
  gi|268559628|ref|XP_002637805.1| Hypothetical protein CBG04591 ...   221     2e-074
  gi|268558954|ref|XP_002637468.1| Hypothetical protein CBG19184 ...   214     2e-071
  gi|268559652|ref|XP_002637817.1| Hypothetical protein CBG04606 ...   165     2e-052
  gi|268558956|ref|XP_002637469.1| Hypothetical protein CBG19185 ...   125     1e-036
  gi|268559650|ref|XP_002637816.1| Hypothetical protein CBG04605 ...   122     2e-035
  gi|268558960|ref|XP_002637471.1| Hypothetical protein CBG19187 ...   120     8e-035
  gi|268559630|ref|XP_002637806.1| Hypothetical protein CBG04592 ...   116     3e-033
  gi|268559654|ref|XP_002637818.1| Hypothetical protein CBG04607 ...   115     1e-032
  gi|268574778|ref|XP_002642368.1| C. briggsae CBR-SIP-1 protein ...   107     1e-029
  gi|268579717|ref|XP_002644841.1| C. briggsae CBR-HSP-43 protein...  70.5     6e-015
  gi|268579717|ref|XP_002644841.1| C. briggsae CBR-HSP-43 protein...  70.5     6e-015
  gi|268554144|ref|XP_002635059.1| C. briggsae CBR-HSP-17 protein...  60.1     3e-012
  gi|268573196|ref|XP_002641575.1| Hypothetical protein CBG09876 ...  59.3     7e-012
  gi|268573992|ref|XP_002641973.1| C. briggsae CBR-HSP-12.2 prote...  57.8     2e-011
  gi|268566747|ref|XP_002639803.1| Hypothetical protein CBG02254 ...  48.9     2e-008
  gi|268570300|ref|XP_002648467.1| Hypothetical protein CBG24755 ...  48.9     2e-008
  gi|268535548|ref|XP_002632907.1| Hypothetical protein CBG21660 ...  43.1     2e-006
  gi|268579635|ref|XP_002644800.1| C. briggsae CBR-HSP-25 protein...  42.4     1e-005
  gi|268579635|ref|XP_002644800.1| C. briggsae CBR-HSP-25 protein...  42.4     1e-005
  gi|268564238|ref|XP_002647119.1| Hypothetical protein CBG23899 ...  41.2     4e-005
  gi|268579371|ref|XP_002644668.1| Hypothetical protein CBG14649 ...  27.7     2.4
  gi|268564057|ref|XP_002639005.1| Hypothetical protein CBG22251 ...  26.9     3.6
  gi|268563170|ref|XP_002638772.1| C. briggsae CBR-PTR-21 protein...  26.6     4.8
  gi|268553777|ref|XP_002634875.1| Hypothetical protein CBG10541 ...  26.6     5.3
  gi|268562597|ref|XP_002646699.1| Hypothetical protein CBG13076 ...  25.8     9.4


> gi|268558958|ref|XP_002637470.1| Hypothetical protein CBG19186
[Caenorhabditis briggsae]
Length=146

 Score =  229 bits (584),  Expect = 2e-077, Method: Compositional matrix adjust.
 Identities = 112/146 (77%), Positives = 129/146 (88%), Gaps = 1/146 (1%)

Query  1    MSLYHYFRPAQRSVFGDLMRDMAQMERQFTPVCR-GSPSESSEIVNNDQKFAINLNVSQF  59
            MSLY +FRP   SV G++MRD+A+MERQF P+     +P+ +SEIVNNDQKFAINLNVSQF
Sbjct  1    MSLYPFFRPRPFSVIGEMMRDIARMERQFIPISAFEAPAAASEIVNNDQKFAINLNVSQF  60

Query  60   KPEDLKINLDGHTLSIQGEQELKTEHGYSKKSFSRVILLPEDVDVGAVASNLSEDGKLSI  119
            KPEDLKINLDG TLSIQGEQE+K EHG+SKKSFSR+ILLPEDVD+GAVASNLSEDGKLSI
Sbjct  61   KPEDLKINLDGRTLSIQGEQEVKDEHGHSKKSFSRIILLPEDVDIGAVASNLSEDGKLSI  120

Query  120  EAPKKEAIQGRSIPIQQAPVEQKTSE  145
```

Figure 4. A sample output from a local BLAST executable search. The output of this search is a text file just like the output from the online BLAST interface, that include the expect value and bit score, as well as match description. Click here to view larger image.



MSA

The multiple sequence alignment result as produced by T-coffee.

```
T-COFFEE, Version_9.03.r1318 (2012-07-12 19:05:45 - Revision 1318 - Build 366)
Cedric Notredame
SCORE=74
*
 BAD AVG GOOD
*
gi|5729850|ref|   :  76
gi|387598062|re   :  83
gi|215272301|re   :  76
gi|238908519|re   :  46
gi|162461738|re   :  76
gi|51479156|ref   :  51
cons              :  74

gi|5729850|ref|     1 MGCTLSA---------EDKAAVE----------------------------------------  14
gi|387598062|re     1 ---------------------------------------------------------------   0
gi|215272301|re     1 MGCLGGNSKTTEDQGVDEKERRE----------------------------------------  23
gi|238908519|re     1 MLLLVTS---------LLLCELPHPAFLLIPEKSDLRTVAPASSLNVRFDSRTMNLSWDCQEN  54
gi|162461738|re     1 MGCTLSA---------EERAALE----------------------------------------  14
gi|51479156|ref     1 MAQFVRN--------------------------------------------------------   7

cons                1                                                                  63

gi|5729850|ref|    15 -----------RSKMIDRNLREDGEKAAKEVKLLLLGAGESGKSTIV-----KQMKIIHED-G  60
gi|387598062|re     1 ---------------------------------------------------------------   0
gi|215272301|re    24 -----------ANKKIEKQLQKERLAYKATHRLLLLGAGESGKSTIV-----KQMRILHVN-G  69
gi|238908519|re    55 TTFSKCFLTDKKNRVVEPRLSNNECSC--TFREICLHEGVTFEVHVNTSQRGFQQKLLYPNSG 115
gi|162461738|re    15 -----------RSKAIEKNLKEDGISAAKDVKLLLLGAGESGKSTIV-----KQMKIIHED-G  60
gi|51479156|ref     8 ---------------------------------------------------------------   7

cons               64                                                                 126

gi|5729850|ref|    61 YSEDECKQYKVVVYSNTIQSIIAIIRAMGRLKIDFGEAARADDARQ--LFVLAGSAEE-GVMT 120
gi|387598062|re     1 ---------------------------------------------------------------   0
gi|215272301|re    70 FNPEEKKQKILDIRKNVKDAIVTIVSAMSTIIPPVPLANPENQFRSDYIKS-IAPITD-FEYS 130
gi|238908519|re   116 REGTAAQNFSCFIYNADLMN------------------------------------------- 135
gi|162461738|re    61 FSGEDVKQYKPVVYSNTIQSLAAIVRAMDTLGIEYGDKERKADAKM--VCDVVSRMEDTEPFS 121
gi|51479156|ref     8 ----------------LVEKTPALVNAAVTYSKP-----------------------------  25

cons              127                                                                 189

gi|5729850|ref|   121 PELAGVIKRLWRDGGVQACFSRSREYQLNDSASYYLNDLDRISQSNYIPTQQDVLRTRVKTTG 183
gi|387598062|re     1 ---------------------------------------------------------------   0
gi|215272301|re   131 QEFFDHVKKLWDDEGVKACFERSNEYQLIDCAQYFLERIDSVSLVDYTPTDQDLLRCRVLTSG 193
gi|238908519|re   136 --------CTWARGPT------------------------------------A-P-------- 145
gi|162461738|re   122 AELLSAMMRLWGDSGIQECFNRSREYQLNDSAKYYLDSLDRIGAADYQPTEQDILRTRVKTTG 184
gi|51479156|ref    26 -RLAT---------------------------FWYYAKVE-----LVPPTPAEIPRAIQSLKK  55

cons              190                                                                 252

gi|5729850|ref|   184 IVETH--FTFKDLYFKMFDVGGQRSERKKWIHCFEGVT---AIIFCV--AL-----SDYDLVL 234
gi|387598062|re     1 ----------------MFDVGGQRDERRKWIQCFNDVT---AIIYVA--AC-----SSYNMVI  37
gi|215272301|re   194 IFETR--FQVDKVNFHMFDVGGQRDERRKWIQCFNDVT---AIIYVA--AC-----SSYNMVI 244
gi|238908519|re   146 ----------RDVQYFLY---IRNSKRRREIRCPYYIQDSGTHVGCHLDNLSGLTSRNYFLVN 195
```

Figure 5. Output of a MSA using T-Coffee. The output highlights similar sites and weights the match by color. Gaps are inserted as "-" signs and the residue or nucleotide position is preserved for each taxon. Click here to view larger image.

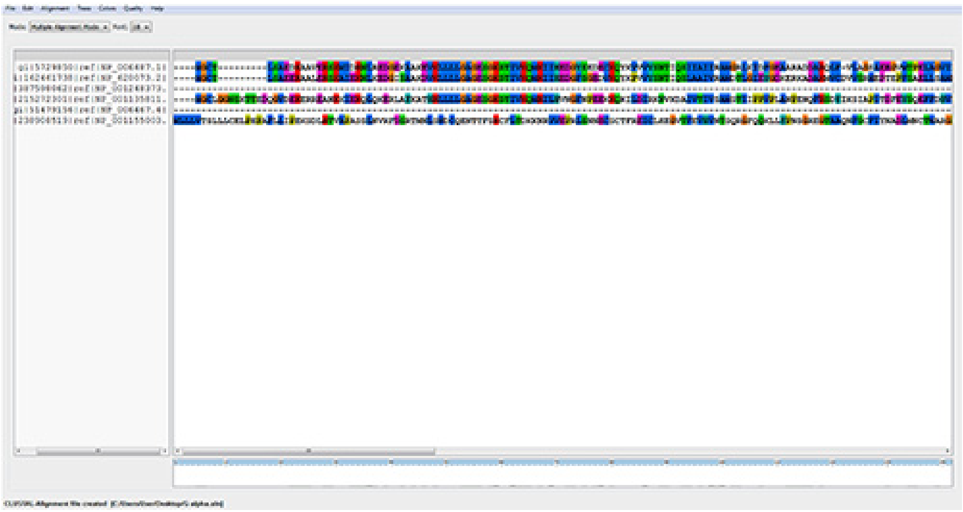
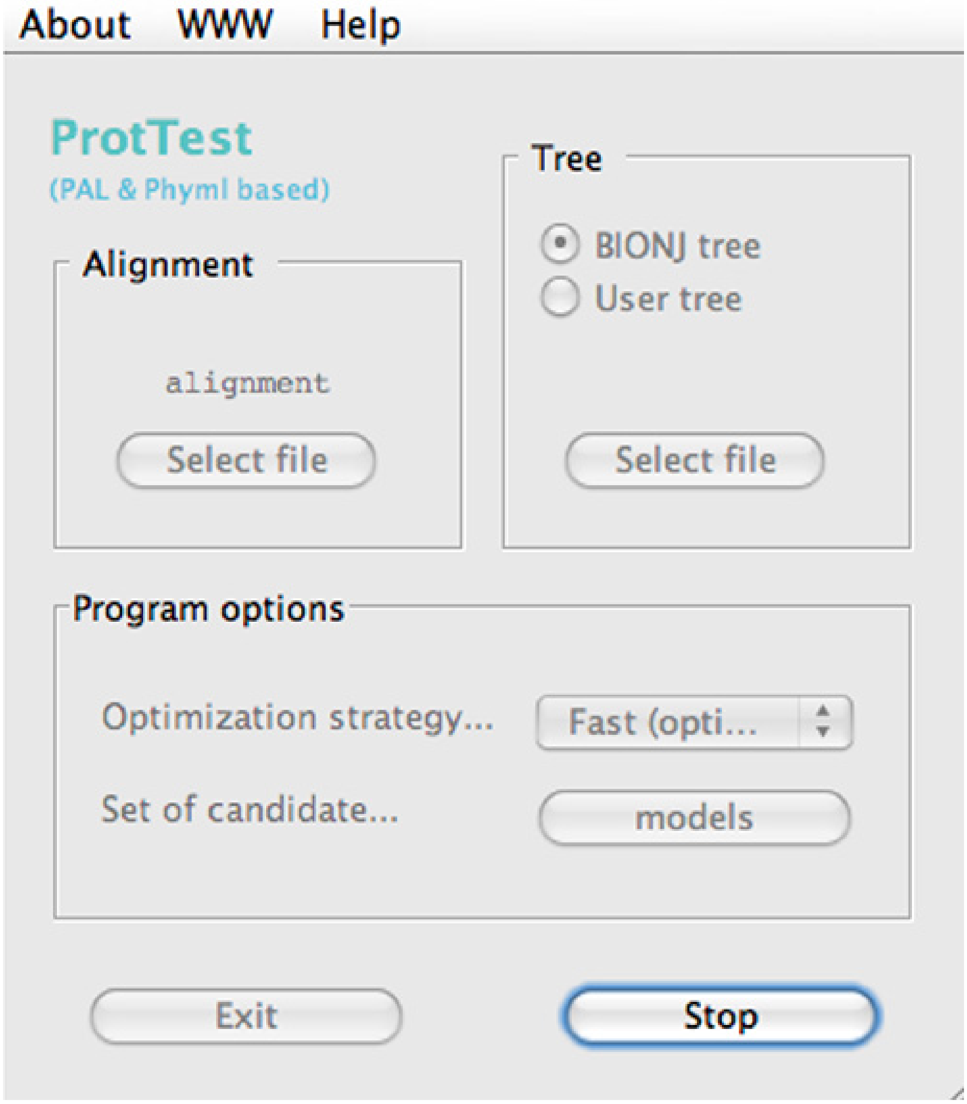


Figure 6. A sample alignment using ClustalX. Similar matches are color coded and gaps are inserted as a "-" sign. The menu bar is seen in the top-left. Click here to view larger image.

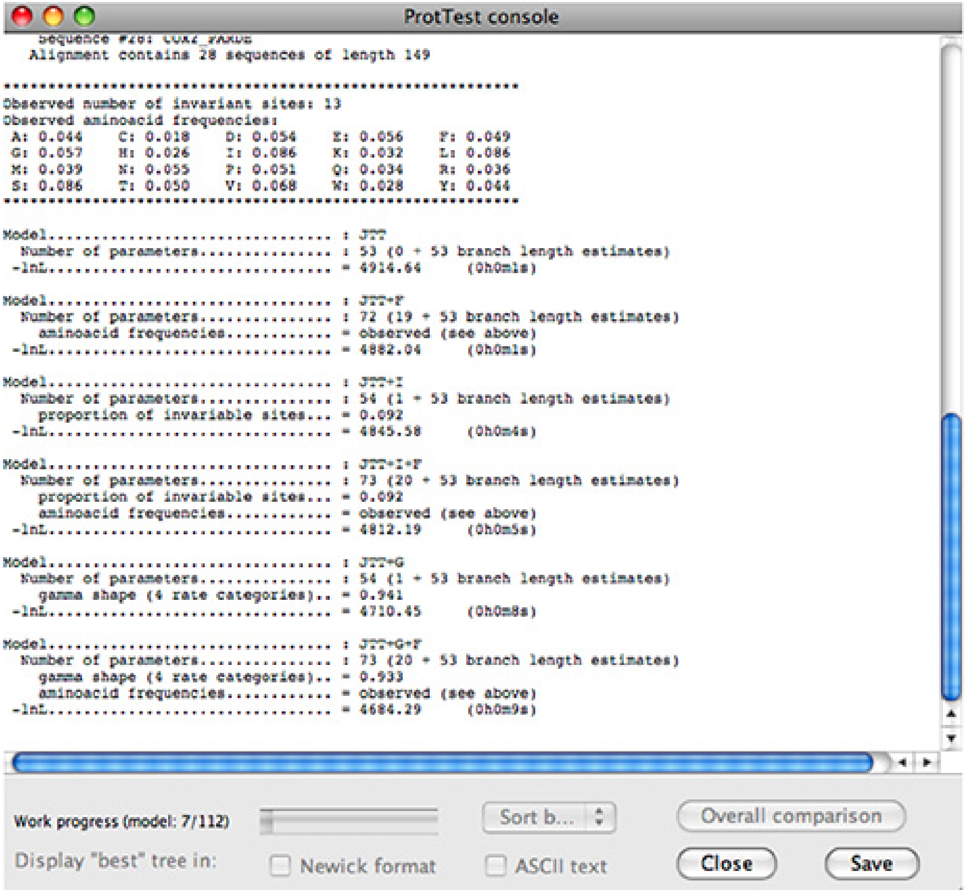


● BIONJ tree
○ User tree

Optimization strategy... Fast (opti...

Set of candidate... models

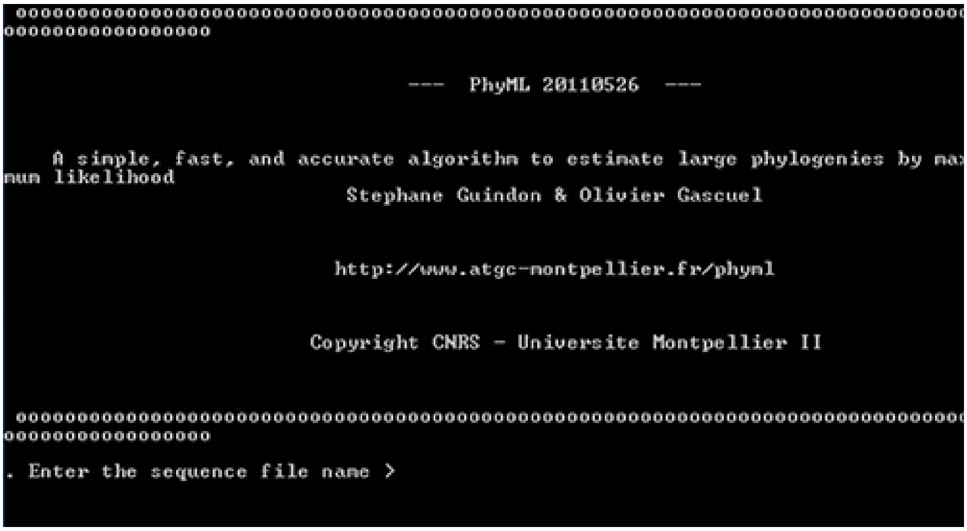
Figure 7. The ProtTest program interface. Click here to view larger image.



```
    Sequence #28: COX2_PARDE
   Alignment contains 28 sequences of length 149

************************************************************
Observed number of invariant sites: 13
Observed aminoacid frequencies:
 A: 0.044    C: 0.018    D: 0.054    E: 0.056    F: 0.049
 G: 0.057    H: 0.026    I: 0.086    K: 0.032    L: 0.086
 M: 0.039    N: 0.055    P: 0.051    Q: 0.034    R: 0.036
 S: 0.086    T: 0.050    V: 0.068    W: 0.028    Y: 0.044
************************************************************

Model................................ : JTT
  Number of parameters............... : 53 (0 + 53 branch length estimates)
 -lnL................................ = 4914.64     (0h0m1s)

Model................................ : JTT+F
  Number of parameters............... : 72 (19 + 53 branch length estimates)
    aminoacid frequencies............ = observed (see above)
 -lnL................................ = 4882.04     (0h0m1s)

Model................................ : JTT+I
  Number of parameters............... : 54 (1 + 53 branch length estimates)
    proportion of invariable sites... = 0.092
 -lnL................................ = 4845.58     (0h0m4s)

Model................................ : JTT+I+F
  Number of parameters............... : 73 (20 + 53 branch length estimates)
    proportion of invariable sites... = 0.092
    aminoacid frequencies............ = observed (see above)
 -lnL................................ = 4812.19     (0h0m5s)

Model................................ : JTT+G
  Number of parameters............... : 54 (1 + 53 branch length estimates)
    gamma shape (4 rate categories).. = 0.941
 -lnL................................ = 4710.45     (0h0m8s)

Model................................ : JTT+G+F
  Number of parameters............... : 73 (20 + 53 branch length estimates)
    gamma shape (4 rate categories).. = 0.933
    aminoacid frequencies............ = observed (see above)
 -lnL................................ = 4684.29     (0h0m9s)
```

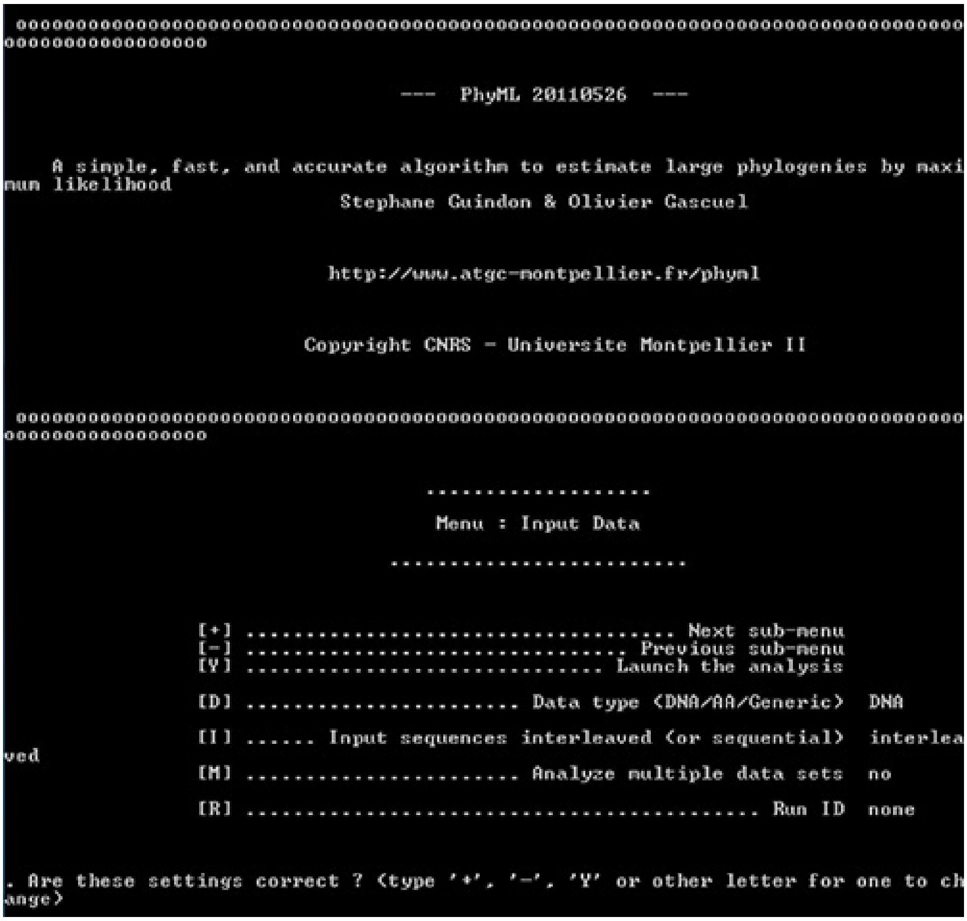
Work progress (model: 7/112)

Figure 8. The ProtTest console. ProtTest console while running an analysis. The progress bar indicates how many models have been completed, and the main window displays the log likelihood score for each model. Click here to view larger image.



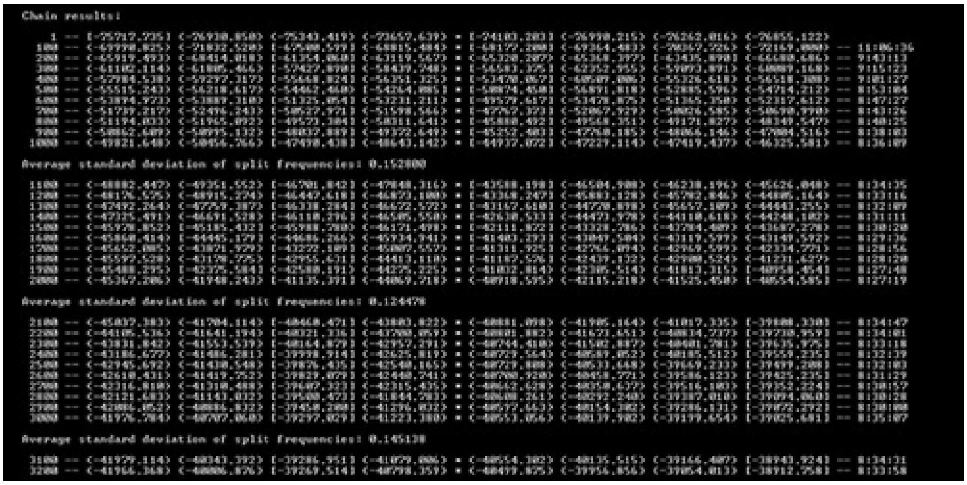
```
 oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
ooooooooooooooooo


                              ---  PhyML 20110526  ---


    A simple, fast, and accurate algorithm to estimate large phylogenies by ma
num likelihood
                       Stephane Guindon & Olivier Gascuel


                      http://www.atgc-montpellier.fr/phyml


                    Copyright CNRS - Universite Montpellier II


 oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
ooooooooooooooooo

. Enter the sequence file name >
```

Figure 9. The PhyML interface. Click here to view larger image.



```
 oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
ooooooooooooooooo

                                 ---  PhyML 20110526  ---

    A simple, fast, and accurate algorithm to estimate large phylogenies by maxi
mum likelihood
                         Stephane Guindon & Olivier Gascuel

                         http://www.atgc-montpellier.fr/phyml

                      Copyright CNRS - Universite Montpellier II

 oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
ooooooooooooooooo

                                    ..................

                                     Menu : Input Data

                                 .........................

                [+] .................................... Next sub-menu
                [-] ................................ Previous sub-menu
                [Y] .............................. Launch the analysis

                [D] ........................ Data type (DNA/AA/Generic)  DNA

                [I] ...... Input sequences interleaved (or sequential)  interlea
ved
                [M] ........................ Analyze multiple data sets  no

                [R] ............................................ Run ID  none

. Are these settings correct ? (type '+', '-', 'Y' or other letter for one to ch
ange)
```

Figure 10. The PhyML interface menu. Once sequences are loaded into PhyML the first menu appears, which can be navigated by typing the letter or symbol in the square bracket. Submenus can be reached by typing the "+" sign. Click here to view larger image.



```
  Chain results:

      1 -- [-25717.235] (-26930.850) (-25343.419) (-23657.639) * [-24103.283] (-26990.215) (-26262.016) (-26855.122)
    100 -- (-67770.025) (-71032.520) [-67500.577] (-68815.404) * [-68172.200] (-67364.403) (-70367.726) (-72169.000) -- 11:06:36
    200 -- (-65917.493) (-68414.018) [-61354.060] (-63119.567) * (-65320.207) (-65368.397) [-63435.890] (-66600.686) -- 9:43:13
    300 -- (-61102.114) (-61805.464) [-57427.890] (-58439.748) * [-56583.375] (-62352.955) (-59073.891) (-60087.168) -- 9:15:23
    400 -- (-57984.538) (-59297.317) [-55668.824] (-56351.325) * [-53470.067] (-60509.006) (-55326.618) (-56518.388) -- 9:01:27
    500 -- (-55515.243) (-56218.617) (-54462.460) [-54264.085] * [-50874.458] (-56891.818) (-52885.596) (-54714.212) -- 8:53:04
    600 -- (-53874.973) (-53889.310) [-51325.054] (-53231.211) * [-49579.617] (-53478.875) (-51365.350) (-52317.612) -- 8:47:27
    700 -- (-51739.217) (-52476.243) [-50527.971] (-51598.566) * [-47767.373] (-52067.929) (-50026.585) (-50690.990) -- 8:43:26
    800 -- (-51194.033) (-51965.092) [-49523.384] (-50311.641) * [-45888.492] (-49568.351) (-49171.827) (-48349.547) -- 8:40:25
    900 -- (-50862.609) (-50975.132) [-48037.889] (-49372.649) * [-45252.480] (-47760.185) (-48066.146) (-47004.516) -- 8:38:03
   1000 -- (-49821.648) (-50456.766) [-47490.438] (-48643.142) * [-44937.072] (-47229.114) (-47419.437) (-46325.581) -- 8:36:09

  Average standard deviation of split frequencies: 0.152800

   1100 -- (-48882.447) (-49351.552) [-46701.042] (-47848.316) * [-43588.198] (-46504.900) (-46238.196) (-45626.048) -- 8:34:35
   1200 -- (-48176.575) (-48915.274) [-46447.618] (-46873.100) * [-43368.247] (-45883.138) (-45702.046) (-44805.164) -- 8:33:16
   1300 -- (-47492.264) (-47759.387) [-46338.284] (-46622.572) * [-43167.610] (-44720.898) (-45657.109) (-44443.255) -- 8:32:09
   1400 -- (-47325.491) (-46691.528) [-46110.396] (-46505.558) * [-42630.533] (-44473.978) (-44110.618) (-44248.182) -- 8:31:11
   1500 -- (-45978.852) [-45185.432] (-45988.780) (-46171.498) * [-42111.872] (-43328.786) (-43704.409) (-43687.278) -- 8:30:20
   1600 -- (-45860.414) [-44445.177] (-44686.266) (-45934.741) * [-41403.273] (-43047.504) (-43119.597) (-43149.592) -- 8:29:36
   1700 -- (-45652.085) (-43871.979) [-43272.189] (-45007.557) * [-41311.925] (-42766.094) (-42969.599) (-42334.771) -- 8:28:56
   1800 -- (-45597.528) (-43178.775) [-42955.631] (-44413.110) * [-41187.576] (-42439.132) (-42900.524) (-41231.627) -- 8:28:20
   1900 -- (-45488.295) [-42375.584] (-42588.191) (-44275.225) * (-41032.814) (-42305.514) (-41813.315) [-40958.454] -- 8:27:48
   2000 -- (-45367.206) (-41948.243) [-41135.391] (-44069.718) * (-40918.595) (-42115.218) (-41525.458) [-40554.585] -- 8:27:19

  Average standard deviation of split frequencies: 0.124478

   2100 -- (-45037.383) (-41704.114) [-40460.471] (-43803.822) * (-40881.098) (-41905.164) (-41017.335) [-39808.338] -- 8:34:47
   2200 -- (-44106.536) (-41641.194) [-40321.336] (-43700.059) * (-40801.882) (-41673.651) (-40814.727) [-39738.959] -- 8:34:01
   2300 -- (-43831.842) (-41553.539) [-40164.879] (-42957.791) * (-40744.410) (-41502.887) (-40401.781) [-39635.975] -- 8:33:18
   2400 -- (-43186.677) (-41486.281) [-39998.914] (-42625.819) * (-40729.564) (-40589.052) (-40185.512) [-39559.235] -- 8:32:39
   2500 -- (-42745.692) (-41438.548) [-39876.435] (-42540.165) * (-40729.888) (-40533.668) (-39649.233) [-39499.288] -- 8:32:03
   2600 -- (-42618.431) (-41419.752) [-39829.079] (-42448.741) * (-40700.920) (-40458.771) (-39586.123) [-39425.235] -- 8:31:29
   2700 -- (-42316.810) (-41318.488) [-39607.323] (-42315.435) * (-40662.628) (-40358.637) (-39516.103) [-39352.224] -- 8:30:57
   2800 -- (-42121.683) (-41141.032) [-39500.473] (-41844.583) * (-40608.361) (-40292.248) (-39387.018) [-39094.060] -- 8:30:28
   2900 -- (-42086.052) (-40886.832) [-39458.288] (-41276.832) * (-40577.663) (-40154.382) (-39286.131) [-39072.292] -- 8:30:00
   3000 -- (-41976.784) (-40707.060) [-39297.029] (-41223.380) * (-40553.056) (-40139.982) (-39199.654) [-39025.681] -- 8:35:07

  Average standard deviation of split frequencies: 0.145138

   3100 -- (-41979.114) (-40343.392) [-39286.951] (-41079.006) * (-40554.382) (-40135.515) (-39166.487) [-38943.924] -- 8:34:31
   3200 -- (-41966.368) (-40006.876) [-39269.514] (-40798.359) * (-40499.875) (-39956.856) (-39054.013) [-38912.758] -- 8:33:58
```

Figure 11. MrBayes Interface. When MrBayes is launched the progress can be viewed in specific intervals set using the "printfreq=X" command. Although the program cannot be stopped during a run, after the specified number of generations are computed the user will be asked if they want to run more generations. Click here to view larger image.

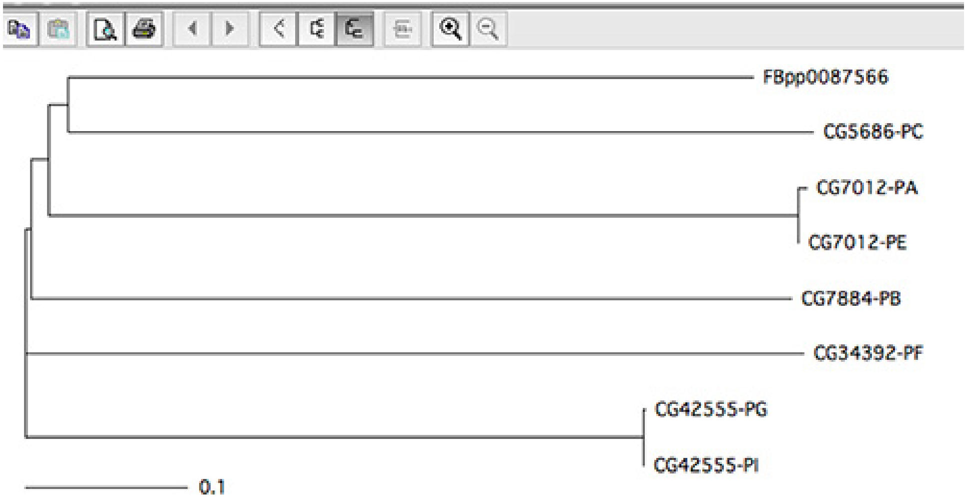


Figure 12. The TreeView interface. In this figure the TreeView window displays a sample tree of proteins from *Flybase* (http://flybase.org/). Files are imported by clicking the "open" option, and selecting an appropriate file type (*e.g*. Newick format). Click here to view larger image.

Discussion

Our hope for this article is that it will serve as a starting point to guide researchers or students that are new to phylogenetics. Genome sequencing projects have become less expensive over the last few years and as a consequence the user demand for this technology is increasing, and now the production of large sequence datasets is commonplace in small labs. These datasets often provide researchers with sets of genes that require a phylogenetic framework to begin to understand their function. Furthermore, because phylogenetics is finding a home in an ever increasing number of research labs, we also intend for this article to serve as an educational device for students interested broadly in biological research. By providing user information on the "why", "how", and "where" for commonly used tree-building tools, we provide a framework for the reader to begin to familiarize themselves with these applications and how they work. However, we advise the reader to play around with all the settings within each tool in an attempt to understand how the various parameters can influence their sequence data, and to ensure compatibility between platform and software in each case. The analysis outlined above was computed using a Dell Optiplex 990 with Intel core i7 processor and a MacBook laptop with an Intel Core 2 Duo processor, however, the speed of analysis and also the specific binaries (*e.g*. 32 bit or 64 bit) will depend on the user's platform.

A challenge when compiling a user guide like this one for phylogenetics, is that the field of phylogenetics, and bioinformatics as a whole, is a rapidly expanding area of research that constantly releases new software aimed at providing better alignments, similarity predictions, or phylogenetic trees. To mitigate this problem, we tried to focus on programs that have been around for a number of years and are still popular on account of how well they work. That said, we want to point out that there are many other tools available to tackle the problems we have outlined in this article, and so encourage the reader to exploit this and incorporate multiple applications into their analyses.

Disclosures

We have nothing to disclose.

Acknowledgements

We thank members of the O'Halloran lab for comments on the manuscript. We thank The George Washington University Department of Biological Sciences and Columbian College of Arts and Sciences for Funding to D. O'Halloran.

References

1. Altschul SF, Carroll RJ, Lipman DJ. Weights for data related by a tree. *J. Mol. Biol*. **207**(4), 647-653 (1989).
2. Akaike H. A new look at the statistical model identification. *IEEE Trans. Automat. Contr*. **19**(6), 706-723 (1974).
3. Schwarz G. Estimating the dimension of a model. *Ann. Stat*. **6**(2), 461-464 (1978).
4. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol*. **52**(5), 696-704 (2003).
5. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*. **17**(8), 754-755 (2001).
6. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. **22**(22), 4673-4680 (1994).
7. Lassmann T, Sonnhammer EL. Kalign--an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*. **6**, 298. doi: 10.1186/1471-2105-6-298 (2005).

8. Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*. **33**(2), 511-518. doi: 10.1093/nar/gki198 (2005).
9. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: A novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res*. **30**(14), 3059-3066 (2002).
10. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. **32**(5), 1792-1797. doi: 10.1093/nar/gkh340 (2004).
11. Notredame C, Higgins DG, Heringa J. T-coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol*. **302**(1), 205-217, doi: 10.1006/jmbi.2000.4042 (2000).
12. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res*. **15**(2), 330-340, doi: 10.1101/gr.2821705 (2005).
13. Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: Fast selection of best-fit models of protein evolution. *Bioinformatics*. **27**(8), 1164-1165, doi: 10.1093/bioinformatics/btr088; 10.1093/bioinformatics/btr088 (2011).
14. Page RD. TreeView: An application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci*. **12**(4), 357-358 (1996).
15. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: More models, new heuristics and parallel computing. *Nat. Methods*. **9**(8), 772, doi: 10.1038/nmeth.2109; 10.1038/nmeth.2109 (2012).
16. Chevenet F, Brun C, Banuls AL, Jacq B, Christen R. TreeDyn: Towards dynamic graphics and annotations for analyses of trees. *BMC Bioinformatics*. **7**, 439, doi: 10.1186/1471-2105-7-439 (2006).