Simple Analysis with the Graphical User Interface of POY

Andrés Varón

July 25, 2008

1 Introduction

This tutorial concentrates in the use of the Graphical User Interface (GUI) of POY 4.0. The GUI provides an easy access to the basic program functions and simplifies the learning process for first time users. For advanced users the GUI constitutes a fast interface to execute complex scripts that have already been personalized.

The GUI is available for Windows XP, Windows Vista, Mac OS X Tiger, Mac OS X Leopard (both intel and ppc processors), and Linux (with intel processors). It can be obtained from the American Museum of Natural History's website at `http://research.amnh.org/scicomp/projects/poy.php`.

This tutorial will guide you through the process of analyzing a few genes and various morphological features of unnamed species. The data files can be downloaded from `ftp://ftp.amnh.org/pub/group/molecular/poy/POY4/docs/data/Analyses.zip`. Before starting download them, and decompress the file in your Desktop.

Upon completing this tutorial, you should be able to perform a basic phylogenetic analysis using POY.

2 Starting POY

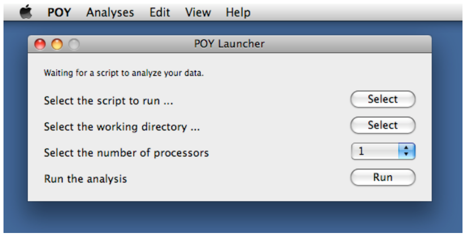
Mac OS X Open Finder, go to the applications folder and double click on POY. The application should then start.

Windows Click in the Start menus, select programs, click on POY, and then click in the POY application. The program should then start.

3 Graphical User Interface Overview

1. Upon starting, POY welcomes you with a small window known as the *POY Launcher* (Figure 1). This window provides a quick access for advanced

users to run custom made scripts. We are, however, not interested in learning the advanced features of the application yet, but instead running some simple analyses.


POY
Analyses
Edit
View
Help
POY Launcher
Select the script to run ...
Select
Select the working directory ...
Select
Select the number of processors
1
Run the analysis
Run

Waiting for a script to analyze your data.

Figure 1: The POY launcher Window. This is the look of the program when you first start it

2. To do so, select in the Analyses menu the Timed Search item (Figure 2). A new window with a number of options should appear. This is the basic window where your analyses will be performed (Figure 3).

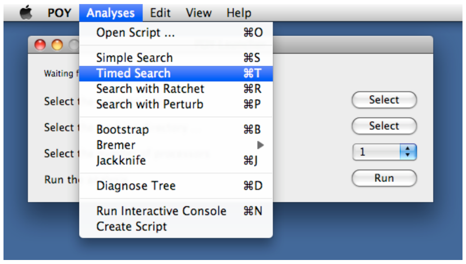


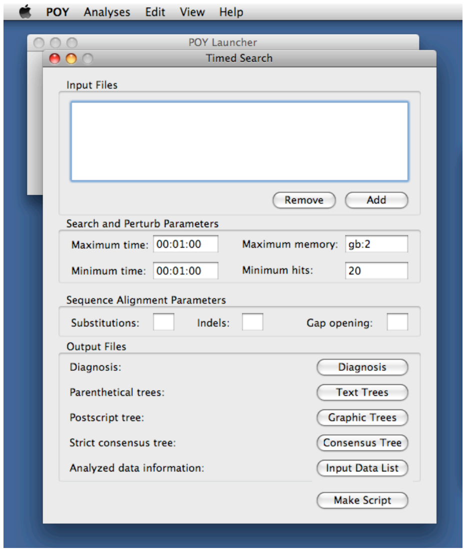
Figure 2: Selecting the menu item for the timed search.

3. The Timed Search window is divided in 4 sections (Figure 4):

   Input Files where the files containing the genes that will be analyzed should be added.

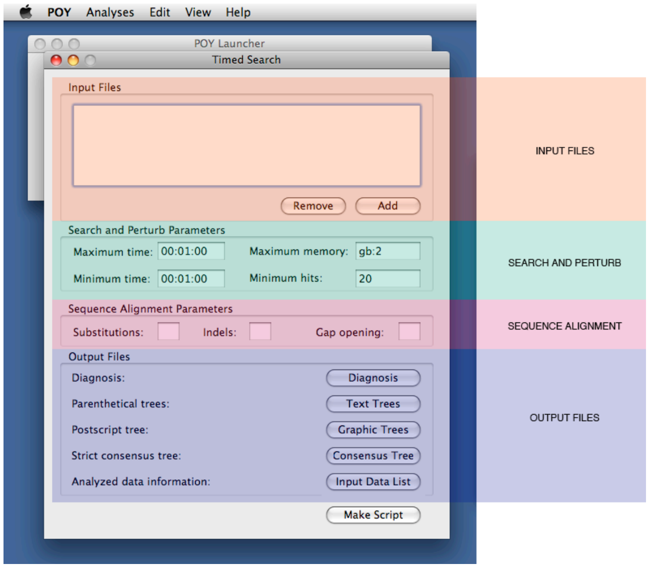
   Search and Perturb Parameters specify some details of how the search will be performed. We will learn a little bit more about this in Section 5.

   Sequence Alignment Parameters specify what the cost of the pairwise sequence alignments in the genes will be. There are three parameters to select: substitutions, indels, and gap opening. Each one of those parameters is explained in Figure 5.



Maximum time: 00:01:00
Maximum memory: gb:2
Minimum time: 00:01:00
Minimum hits: 20

Figure 3: The Timed Search window.



Maximum time: 00:01:00
Maximum memory: gb:2
Minimum time: 00:01:00
Minimum hits: 20

Figure 4: The four sections of the Timed Search window.

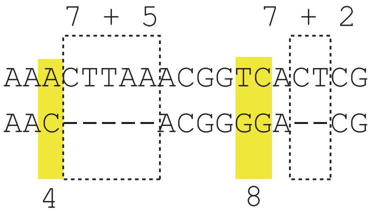


Figure 5: The different sequence alignment parameters. During evolution, DNA sequences suffer modifications that modify their lengths. In order to compare them, we need to find locations where those insertions and deletions that have changed their length may have happened. This is also an optimization problem (like the Traveling Salesman Problem). Therefore we have to assign a cost to the three basic events we are looking for: substitutions (for example when an A becomes a C), insertions (when the sequence AC becomes AGC), or deletions (when a sequence AGC becomes AG). In this example the substitution cost is 4, the gap opening cost is 7, and indels have cost 1. Insertions (and deletions) are represented with a $-$. Observe that there are 3 substitutions (the filled boxes), for a total cost of 12. And there are 7 individual indels, divided in two blocks, one with 5 (left dashed box), and another one with 2 (right dashed box). The cost of each block is the gap opening cost (7 for each box), and the cost of each individual insertion and deletion is 1 what the indels cost represents, in this case 1. Therefore we have $7 + 5$ in the left box and $7 + 2$ in the right box. The total cost of this alignment is $4 + 8 + (7 + 5) + (7 + 2) = 33$.

Output Files define what files should contain the results of your analyses. Probably the most important files that you want to produce are the parenthetical trees, and the postscript trees.

Lets start using those windows to be able to perform our first analysis.

4 File Selection

In order to perform the analysis we should first give the program a number of files containing all the information about the species we are interested in. For this initial stage of the tutorial we will first use a set of files that have been prepared for this course. However, we will learn later how to directly collect raw sequences from genbank so that you can analyze sequences of your species of interest.

1. From the Timed Search window (to remember how to open it go to Section 3), in the Input Files frame, click in the Add button. A file selection window should appear in your screen (Figure 6).

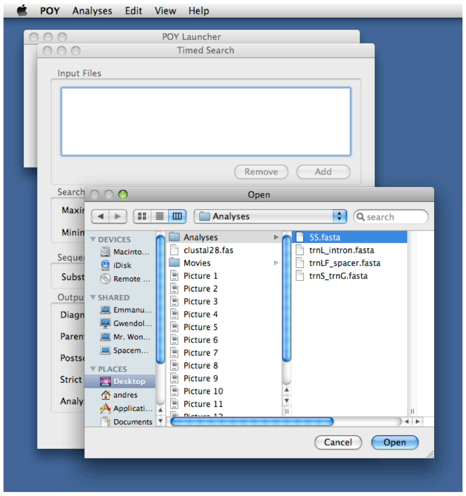
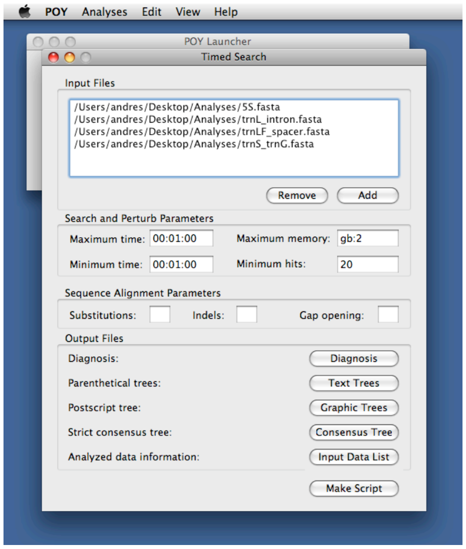


Figure 6: Adding the 5S.fasta file to the analysis.

2. The following instructions are for Mac OS X, but the overall location of the files is the same in Windows. On the left panel of the Open window select Desktop, then select the Analyses folder, and within it select the file 5S.fasta. It contains one gene for all the species that we will be analyzing in this tutorial (Figure 6).

3. Click in the Open button to add the file to our list of input files.



```
/Users/andres/Desktop/Analyses/5S.fasta
/Users/andres/Desktop/Analyses/trnL_intron.fasta
/Users/andres/Desktop/Analyses/trnLF_spacer.fasta
/Users/andres/Desktop/Analyses/trnS_trnG.fasta
```

Maximum time: 00:01:00
Maximum memory: gb:2
Minimum time: 00:01:00
Minimum hits: 20

Figure 7: View of the Timed Search window once all of the files that will be included in this analysis have been added.

4. Repeat this procedure to add the rest of the files contained in the Analyses folder. In total, 4 files should be listed at the end, as shown in Figure 7.

5 Search and Perturb Parameters

The search and perturb parameters frame is used to specify values that define how careful the tree search will be. There are 4 parameters that can be selected. Let's fill them one by one:

Maximum time defines for how long the program will be executing. As we have learned already, an analysis is a *heuristic search* for a good solution. Given that the computational problem is so hard to solve, the program provides a function to say for how long will the user be willing to run an analysis. That is what this maximum time box sets.

For example, when preparing the analyses for your project, you might want to leave the program running all night long, searching for the best phylogenetic hypothesis. To make sure that things will be ready in the morning, you can set this parameter to 12 hours. In this way, if the program starts a 6 PM today, tomorrow at 6 AM it will finish and you will have a hypothesis to start working on.

The format of the box is:

DAYS:HOURS:MINUTES

(a typical error of many users is to write hours:minutes:seconds, be careful!). For example, if we input the value 01:12:46, we will have set the maximum time to 1 day, 12 hours, and 46 minutes.

For this tutorial, set the value to 0 days, 0 hours, and 3 minutes (00:00:03). As this is not a complete analysis, we just want to get some results fast!

Minimum time is located below Maximum time. This parameter is better specified after the remaining two are addressed. So let's first do Maximum memory, then Minimum hits, and only then check this parameter.

Maximum memory specifies, as the name says, how much memory can be used to *store trees*. Imagine that your computer has only 512 MB of RAM. The RAM is the memory where the programs that are currently executing are being stored. If a program runs out of memory – meaning that there is not enough memory available to fill its needs –, then the program will crash, fail and leave you with nothing. We don't want an analysis to crash, so we can tell POY how much memory are we willing to use.

By default the program specifies 2 GB of RAM. We will take a more conservative value and use 512 MB of RAM (1 GB = 1024 MB). To do this write in the Maximum memory field `mb:512`.

Minimum hits defines how many times, at least, should the shortest tree be found. The number of hits is a good way to guess wether or not the search is good enough. For example, if in a small data set, the shortest tree has been found 20 times, then it could be reasonable to stop the search there.

POY will stop before the Maximum time is reached if the minimum number of hits is found. We will set this value to 20 in this exercise.

Minimum time which we skipped before, can now be explained better. A user may want to have at least 20 hits, but not stop the analysis before the first 12 hours, to ensure that good trees where found. The minimum time sets a minimum amount of time that the program should spend in the analyses before the minimum hits is taken into consideration to stop the search. We will leave this value to the default 00:00:00.

6 Sequence Alignment Parameters

1. For this initial test we will use a cost regime that is known as *affine gap cost*, where insertions and deletions occur in blocks. To do this, we assign a cost to each block of indels (that is the gap opening parameter), a cost to each individual insertion and deletion (the indel parameter), and a cost for each substitution (the substitution parameter) (see Figure 5 for a graphical explanation).

2. Set the substitution parameter to 2

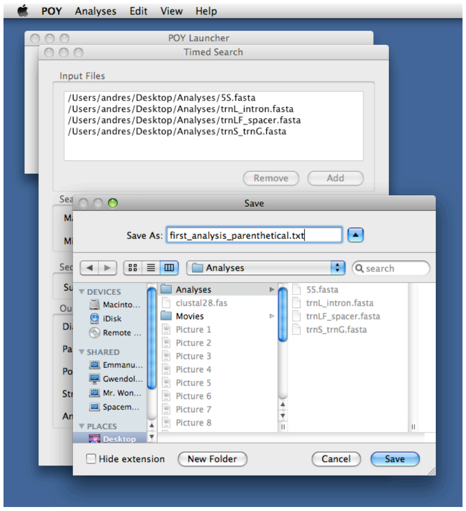
3. Set the indel parameter to 1

4. Set the gap opening parameter to 4

7 Output Files

Now that we have filled all the information required for the analysis itself, we are only left with the task of selecting where the results should be stored. That is what the buttons in the Output Files frame let's do. For now we will only care about storing three files: the Parenthetical trees, the Postscript tree, and the Analyzed data information. All of the output from this tutorial will be stored in the Analyses folder located in your Desktop.

1. Click in the text trees button. That opens a window that allows you to select the file where the phylogenetic trees should be stored. Navigate to the Destop/Analyses folder and give it as name “first_analysis_parenthetical.txt” (Figure 8).



Input Files

/Users/andres/Desktop/Analyses/5S.fasta
/Users/andres/Desktop/Analyses/trnL_intron.fasta
/Users/andres/Desktop/Analyses/trnLF_spacer.fasta
/Users/andres/Desktop/Analyses/trnS_trnG.fasta

Save As: first_analysis_parenthetical.txt

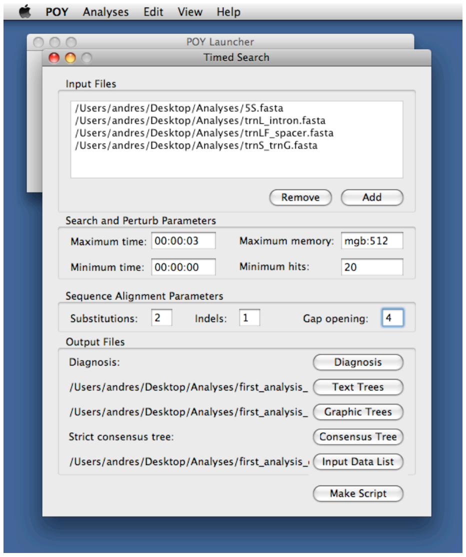
Figure 8: The Save window for parenthetical tree output

2. Repeat the procedure with the Graphic Trees button. The file you select there will contain a graphical representation of the tree that you can use for your presentation. Name this file “first_analysis_graphical.ps”.

3. Repeat the procedure with the Input Data List button. The file you select there will contain the report of what data was analyzed. This is useful for

future reference when you have forgotten what you did ... we will name this file “first_analysis_data.txt”.

After selecting those files, your Timed Search window should look as in Figure 9.



Input Files

/Users/andres/Desktop/Analyses/5S.fasta
/Users/andres/Desktop/Analyses/trnL_intron.fasta
/Users/andres/Desktop/Analyses/trnLF_spacer.fasta
/Users/andres/Desktop/Analyses/trnS_trnG.fasta

Search and Perturb Parameters

Maximum time: 00:00:03
Maximum memory: mgb:512
Minimum time: 00:00:00
Minimum hits: 20

Sequence Alignment Parameters

Substitutions: 2
Indels: 1
Gap opening: 4

Output Files

Diagnosis:
/Users/andres/Desktop/Analyses/first_analysis_
/Users/andres/Desktop/Analyses/first_analysis_
Strict consensus tree:
/Users/andres/Desktop/Analyses/first_analysis_

Figure 9: The four sections of the Timed Search window.

8 Run the Analysis

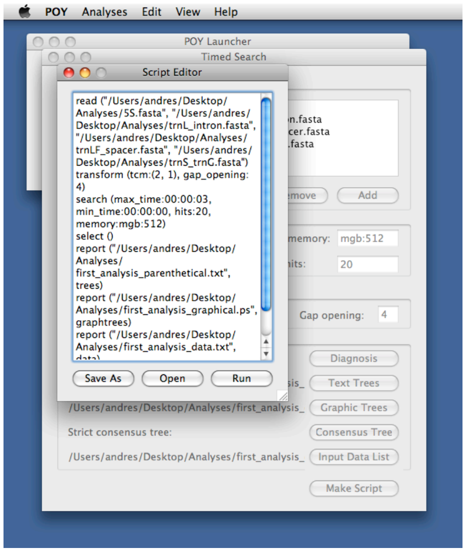
1. Click in the Make Script button on the lower right corner of the Timed Search window. A new window with the title “Script Editor” should appear as in Figure 10.

2. Click in the Run button of the Script Editor. POY will ask you to save the script in a file. As with every other file we have produced, save it in the Desktop/Analyses folder, under the name “first_analysis_script.poy”.

3. Once its saved, the analysis should start. It is time now to relax! you have to wait for three minutes for the results to come out.

9 Looking at the results

1. Upon completion, POY will open an Output Window. This window can always be accessed from the Window menu.



```
read ("/Users/andres/Desktop/
Analyses/5S.fasta", "/Users/andres/
Desktop/Analyses/trnL_intron.fasta",
"/Users/andres/Desktop/Analyses/
trnLF_spacer.fasta", "/Users/andres/
Desktop/Analyses/trnS_trnG.fasta")
transform (tcm:(2, 1), gap_opening:
4)
search (max_time:00:00:03,
min_time:00:00:00, hits:20,
memory:mgb:512)
select ()
report ("/Users/andres/Desktop/
Analyses/
first_analysis_parenthetical.txt",
trees)
report ("/Users/andres/Desktop/
Analyses/first_analysis_graphical.ps",
graphtrees)
report ("/Users/andres/Desktop/
Analyses/first_analysis_data.txt",
data)
```

n.fasta
cer.fasta
.fasta

memory: mgb:512

hits: 20

Gap opening: 4

/Users/andres/Desktop/Analyses/first_analysis_

Strict consensus tree:

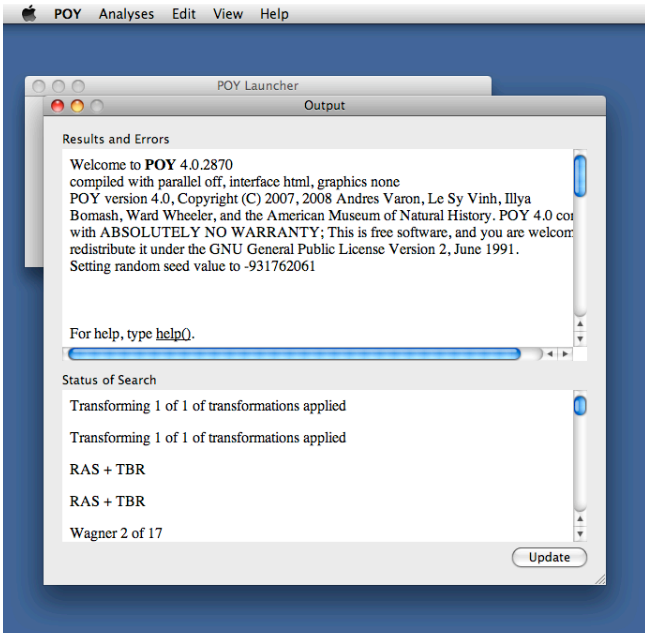
/Users/andres/Desktop/Analyses/first_analysis_

Figure 10: The script that will be executed, as shown by the script editor window.

2. The Output window (Figure 11) is divided in two halves, the upper containing the results of the program (if the user requests the results on screen, we did not do that), and the lower, showing how the analysis was made.

3. We can now look at the resulting tree by opening the Analyses folder in your Desktop, and opening the "first_analysis_graphical.ps". This is a postscript file that Mac OS X will convert automatically into a PDF that can be placed in a presentation or a paper. In windows, this file format is easily read by any vectorial image editing application, such as Adobe Illustrator, or Ghostscript.

4. Excellent, we just finished a very simple phylogenetic analysis of DNA sequences using POY 4.0!



```
Welcome to POY 4.0.2870
compiled with parallel off, interface html, graphics none
POY version 4.0, Copyright (C) 2007, 2008 Andres Varon, Le Sy Vinh, Illya
Bomash, Ward Wheeler, and the American Museum of Natural History. POY 4.0 co
with ABSOLUTELY NO WARRANTY; This is free software, and you are welcom
redistribute it under the GNU General Public License Version 2, June 1991.
Setting random seed value to -931762061



For help, type help().
```

```
Transforming 1 of 1 of transformations applied

Transforming 1 of 1 of transformations applied

RAS + TBR

RAS + TBR

Wagner 2 of 17
```

Figure 11: Look of the Output Window when the analysis has been completed.