

Lab 8: Using POY from your desktop and through CIPRES

In this lab we're going to try to look at the effects of different methods of DNA alignment. We'll try different settings in ClustalW. We'll go through how to convert the output from ClustalW to a Nexus file, and how to manipulate that Nexus file. We'll also use POY, which does alignment and tree searching together. Trying different alignment strategies and seeing how they affect your phylogenetic hypothesis is a great way to add some interest to your project, since it helps explore the question of just how much changes at this step affect the trees you build.

1. CIPRES Basics



CIPRES is a computational server that hosts popular phylogenetic research tools. There are four necessary steps to use these tools: create an account, upload your data, submit your tasks, and analyze your output.

How easy! How useful! Of course, this means CIPRES attracts oodles of task submissions each day. Depending on many factors (the type of task, whether you submit during peak hours, etc.), it may take anywhere from seconds to hours for your task to be processed. This makes CIPRES wonderful for overnight jobs, but less reliable for urgent matters.

Usage statistics: http://www.phylo.org/sub_sections/portal/statistics.php

Since we only have an hour or so for lab, we'll begin by submitting a task, which may or may not complete by the end of the lab.

Account creation

Register your CIPRES account here: <http://www.phylo.org/portal2/register.action>. You can run analyses as a guest, but how could you forgo these benefits!

Not registered yet? Benefits of registering:

1. Ability to save and retrieve data;
2. Ability to organize jobs into different folders;
3. Ability to see history of your job submissions

[Register](#) | [Proceed without Registering](#)

First, you'll want to create a folder. Click on My Workbench, click Create New Folder, and name it "IB200A". Each folder contains your data and tasks, and helps keep your multiple projects organized.

Upload data

Next, we'll upload our data file. We'll use a FASTA-formatted file since it is a simple format and unlikely to confuse whichever tool we decide to use. Get some FASTA DNA sequences – either from your own groups, or you can use the FASTA file of COI sequences from some cephalopod species that you can download from the lab website.

1. Click on the My Workbench tab.
2. Click on Data under your IB200A folder.
3. Click on the Upload button.
4. Give your data a creative label, fill out the upload form appropriately, and click Save.
5. Verify your data was saved into your folder.

Submit task

Submitting a task to CIPRES is very similar to saving data.

1. Click on the My Workbench tab.
2. Click on Tasks under your IB200A folder.
3. Click on the Create New Task button.

4. Select your FASTA file as your input. Select ClustsalW as your tool. Give your task a creative description. The “Set Parameters” allows you to run your tool under non-default settings (we’ll be doing this with ClustalW later in the lab).
5. Under your Tasks list, find your task and click “Run Task”. The “Run Task” button should change to a “View Status” button. When your task is complete. CIPRES will send you an email telling you

Analyze output

Once you receive your confirmation email, you can see your task output. These steps should feel familiar by now.

1. Click on the My Workbench tab.
2. Click on Tasks under your IB200A folder.
3. Find your completed task and click on the “View Output” button.
4. Different tools produce different outputs. For ClustalW, outfile.aln is probably what you’re most interested in. If you’re new to using a tool, it’s good to have a manual in hand and to be familiar with all of the output files that tool produces.
5. Two of these files are “meta” task files that tell you whether your task completed successfully or not: STDERR and STDOUT. These files are especially important to check, since they may indicate that your data was not formatted correctly, or that your parameters could not be applied as you intended.
6. If you see a massive STDERR file and none of the expected output files, something probably went wrong with your task, which you’ll need to troubleshoot. Try running the task again under default parameters or under a known working dataset, then start increasing the complexity of your task gradually until you can identify the issue.

You should have an idea of how to use CIPRES. The general same procedure should apply to any of the tools they have available. Now, let’s look at some of the details for ClustalW and POY.

2. POY (Desktop)

- 1) Get some FASTA DNA sequences – either from your own groups, or you can use the FASTA file of COI sequences from some cephalopod species that you can download from the lab website.
- 2) Google up POY and download/install it.

Note: The lab instructions were written assuming you were using POY on the windows machines in the computer lab. So you will have to mentally adjust a little, depending on where/how you install it on your laptops.

Notes from 2010: We had trouble (i.e. no success) getting the POY command line to load/see the *moll.txt* file on POY installed on peoples' Windows laptops, despite trying many different ways of doing it. On Mac laptops it worked fine, so we had the Windows users look over the shoulders of Mac users.

* also, apparently the Mac version of POY deposits the *tutorial1_trees.txt* and *tutorial1_stats.txt* files in the home directory (e.g. */user/Nick*), without telling you.

Turn in: just email me short answers to the questions as they come up.

ClustalW

ClustalW is the most commonly used program for multiple sequence alignments. It works by first doing pairwise alignments of each pair of sequences to calculate distances between them. It uses those distances to derive a tree. That tree then guides the production of the multiple sequence alignment.

For this lab, we're going to access *ClustalW* directly through web page of the European Bioinformatics Institute, where you can control each of the search parameters. Since ClustalW is the industry standard, we're going to use it for comparison to *POY*, and to see the effects that parameters can have on the alignment.

-Go to <http://www.ebi.ac.uk/Tools/clustalw2/index.html>.

For the first pass let's just use the defaults.

-From the output **format menu** select '**aln wo/numbers**'. Hit the **Browse** button and select your FASTA file. Then hit **run**.

A web page will come up, letting you know that your job is in progress. It will automatically refresh, so that when your job is finished, a new web page will come up.

-Pay attention to how long it takes, and when the job is finished, make a note of the alignment score.

There will be four files available for you to download. The output file is a description of the alignment process. The *.aln* file is your actual sequence alignment in Clustal format.

The .dnd file is the guide tree used to make the alignment. The input file is the file that you uploaded to *ClustalW*.

-Download the alignment and guide tree files and give them unique names.

-Repeat the same alignment only this time run a **'fast'** rather than a **'full'** alignment. Don't forget to select **'aln wo/numbers'**.

It probably won't make a difference for this alignment, because you don't have a very big matrix, but for more species or a longer sequence a fast alignment can lead to a big savings in time. The 'fast' alignment works by only considering a chunk of the sequence at a time. The options on the second line refer only to the 'fast' alignment and have to do with how big a chunk of the sequence you consider.

How long did it take? Did it seem faster to you?

Is your alignment score as good as it was last time?

-Save the alignment file again, but don't bother with the guide tree.

Let's do this one more time, but this time let's give it really unrealistic parameters to see how the parameters do effect the alignment.

-Set the **gap open penalty** to 1 and the **gap extension penalty** to 5. This will make it easier to start a gap then to extend it, a highly unrealistic situation.

-Run the same sequences and save the alignment file. Don't forget to select **'aln wo/numbers'**.

Comparing the Alignments

You can use Mesquite to open up the alignments you made with different input parameters and see how much the results changed. You may also wish to export the resulting alignments for use with either Phylip or Nona/TNT and see whether the different alignments produce changes in the phylogenetic reconstruction of these clades.

POY

POY is a program by Ward Wheeler, David Gladstein, and Jan De Laet that is freely available on the web (<http://research.amnh.org/scicomp/projects/poy.php>). It views the alignment as a problem, in which gaps and deletions are events that happen along the tree. In other words, POY implements a way to build the alignment and the tree at the same time. As its creators describe it: "An essential feature of POY4 is that it implements the concept of dynamic homology allowing optimization of unaligned sequences.... POY4 provides a unified approach to co-optimizing different types of data, such as morphological and molecular sequence data. In addition, POY4 can analyze entire chromosomes and genomes and take into account large-scale genomic events (translocations, inversions, and duplications)." As a result, the output from POY is a tree, not an alignment per se.

- Create a folder on your desktop called 'POY folder'. Drag the unaligned sequence file into this folder.
- Find poy4.exe (**Start>Programs>poy4>poy4.exe**) and mol1.txt (**Start>Programs>poy4>docs>mol1.txt**). Copy and paste both of these into your new folder. The docs folder also contains a file called "tutorial commands" or something like that. You can open this folder if you want and copy and paste all of the commands below.
- Go to Analyses->Interactive Console to open the command prompt.
- Open mol1.txt using a text editor. You'll see that it's just a series of input sequences, not aligned yet. Close it.
- Open poy4 by double-clicking on it. An interface will appear that has three separate boxes. The box on the lower right is the one where you can type commands. You can move the cursor to the box above it (which contains responses from the program) by pressing the up arrow, and you can move the cursor back into the command line box by pressing return.
- Type: `read ("`, then drag your mol.txt file from its folder and drop it on the poy window. Poy will enter the path name to your file. Once it has been entered, type `)`, and hit enter. This will: Import the DNA sequence datafile mol1.txt.
- Type: `build (100)`, hit enter. This will: Generate 100 random addition sequence Wagner trees.
- Type: `select (unique)`, hit enter. This will: Discard duplicate trees.
- Type: `swap (threshold:10)`, hit enter. This will: Alternate SPR and TBR branch swapping of all current trees as well as all new trees found that are up to 10% longer than the current tree being swapped.
- Type: `select ()`, hit enter. This will: Discard suboptimal and duplicate trees (i.e., retain only optimal trees).
- Type: `report (asciitrees)`, hit enter. This will: Draw optimal trees in POY4 Output window. You'll see a little tree appear in the top poy window. Press the up arrow to scroll through it, then press return when you are done to move back to the command window.
- Type: `report ("tutorial1_trees.txt", trees)`, hit enter. This will: Output all current trees to file tutorial1 trees.txt
- Type: `report ("tutorial1_stats.txt", treestats)`, hit enter. This will: Output basic tree statistics to file tutorial1 stats.txt
- View cladogram(s) in parenthetical format by opening tutorial1 trees.txt in notepad.
- View basic tree statistics by opening tutorial1 stats.txt in notepad.

There is much more to be learned about POY, and the manual includes several tutorials that you could use were you interested in learning more about POY's advanced features.