CAOS Documentation and Worked Examples

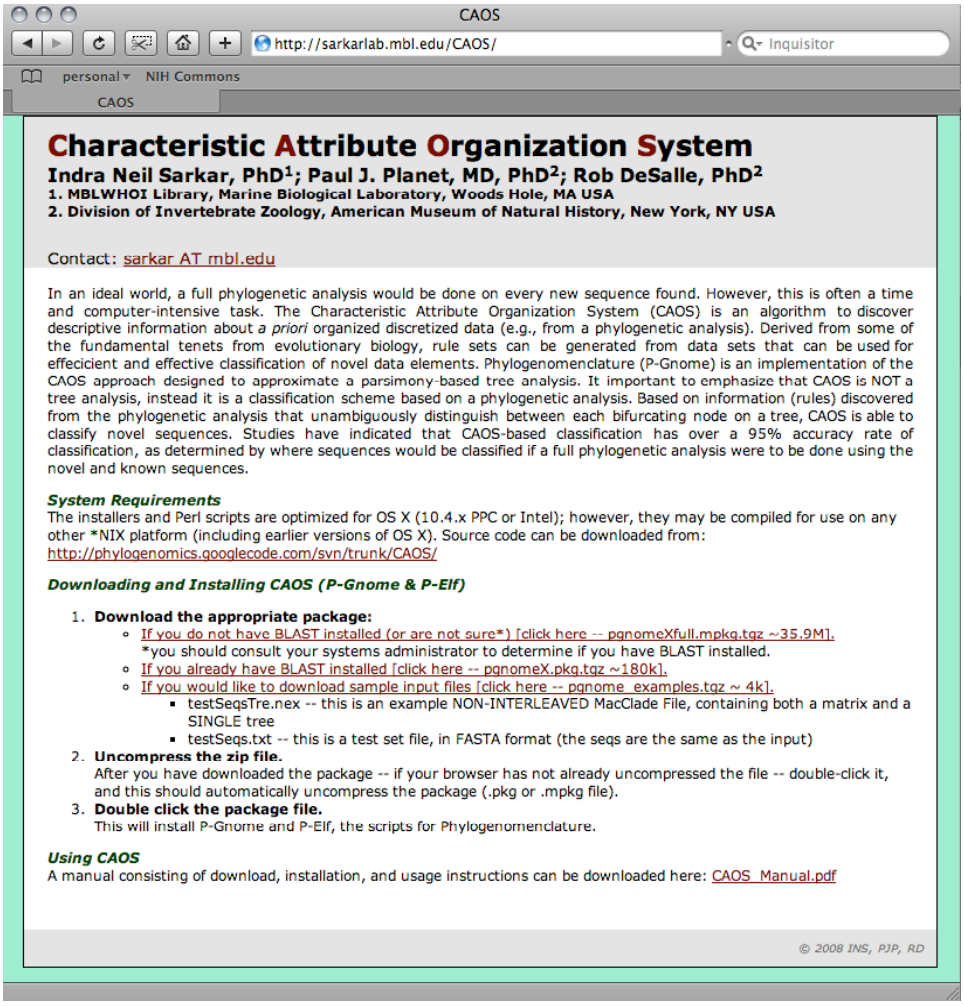
Neil Sarkar, Paul Planet and Rob DeSalle

Table of Contents

1. Downloading and Installing p-gnome

The program is downloadable from http://sarkarlab.mbl.edu/CAOS. The program is currently available in a Mac OS X installer (a UNIX version will also be made available from the authors). P-Gnome is based on a Columbia University Patent Pending Technology, CAOS. The program requires Perl. If the user does not have it installed (all OS X versions since 10.3 have it pre-installed), it is easiest to install the Apple Developer Tools, which comes with OS X or can be downloaded directly from Apple at http://developer.apple.com/tools.

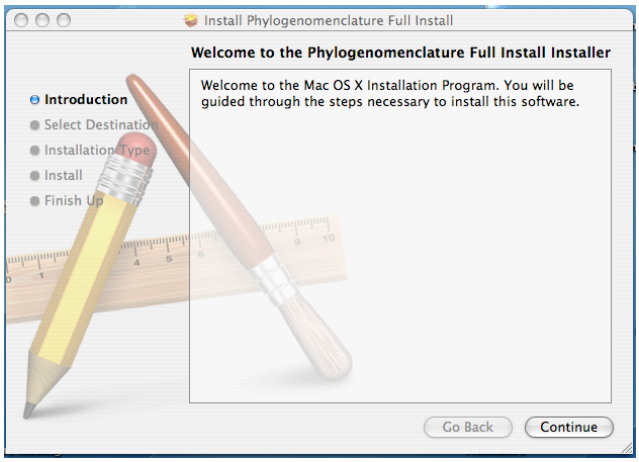
P-gnome runs from a command line interface. The program is installed in the /usr/bin directory. The following is a screen shot of the CAOS homepage.



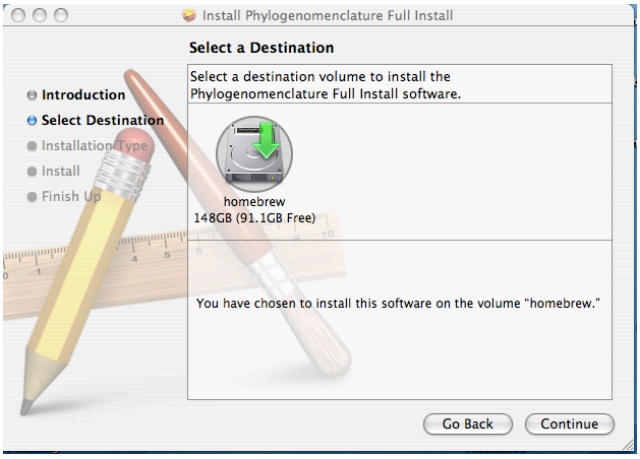
There are three download options. If you do not have BLAST installed on your computer or you are not sure, the option 1 should be clicked. If you do have BLAST installed then option 2 should be clicked. Option 3 is for downloading sample files.

The downloaded file should look like this:

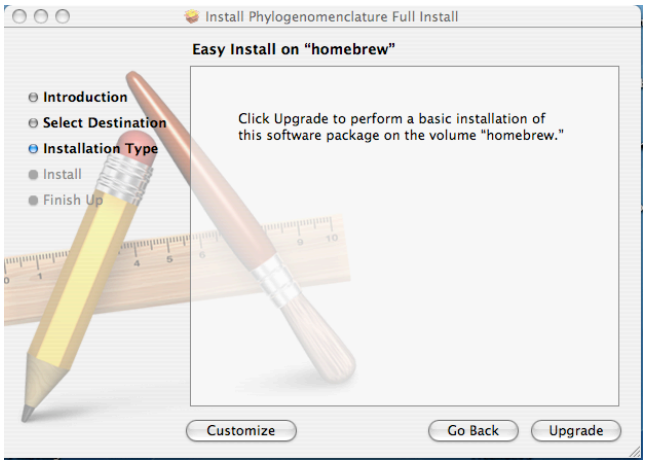
The top file is the .tar file that needs to be expanded. The bottom file is the installer file. Just double click on the bottom file. The following installer window should appear:



By clicking the "Continue" button the installation process will begin. The next window should be the following:



Click on the disc where the program will be installed and the click on “Continue”. The next screen will appear in the installer window:



Either an “Install” button or an “Upgrade” button will appear as in the above screenshot. Click on “Install” or “Upgrade” and the program will be completely installed.

2. Preparing your matrix for p-gnome

P-gnome uses the NEXUS format. This format can be manipulated easily in MacClade (Maddison and Maddison, 2002) or Mesquite (Maddison and Maddison, 2007). P-gnome uses a specific format for the input matrix. The following steps will assist the user in preparing a matrix for p-gnome.

1) First, perform an alignment of your DNA sequences with the output setting on NEXUS format. Alternatively if there is no NEXUS format setting in your alignment program then you can save the alignment in other formats and import them into MacClade and then export them in NEXUS format.

2) The next step is to perform a phylogenetic analysis on the matrix in PAUP, PHYLIP or any other program to generate a tree. Any approach (Neighbor Joining, Maximum Likelihood, Bayesian analysis or maximum parsimony) can be used, as a "rough" tree is all that is needed to proceed to the next step. Save the tree in NEXUS format.

3) The next step is to open or import the data file into MacClade or Mesquite (the screenshots below are all from MacClade; however the commands are equivocal to those in Mesquite). Within the application go to the tree window pulldown. The application will ask for a treefile so give the NEXUS treefile saved in the previous step (Figure 1).

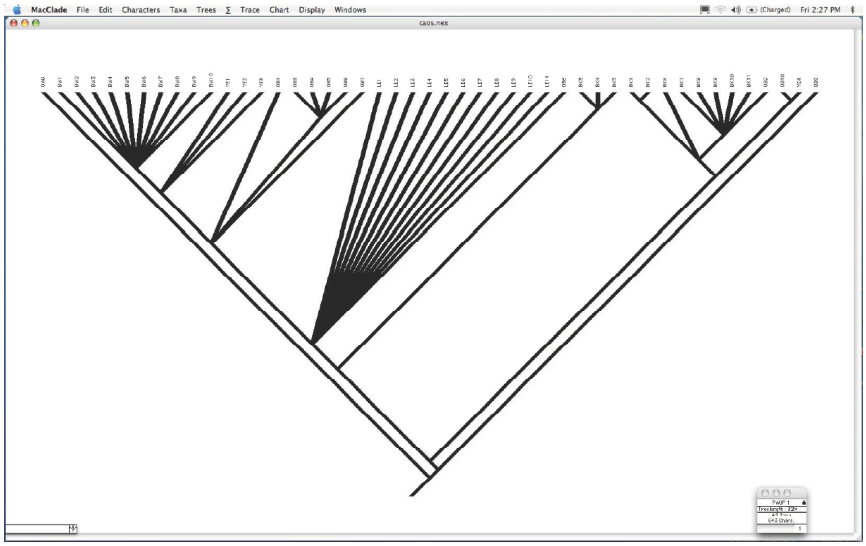


Figure _1_

4) The tree saved in Step 2 will appear. With the move branch button in the tool palette, group your individual sequences into clades according to pre-described species boundaries or hypothetical species groupings. This can be accomplished by moving terminals in the tree around to form monophyletic groups representing your pre described species.

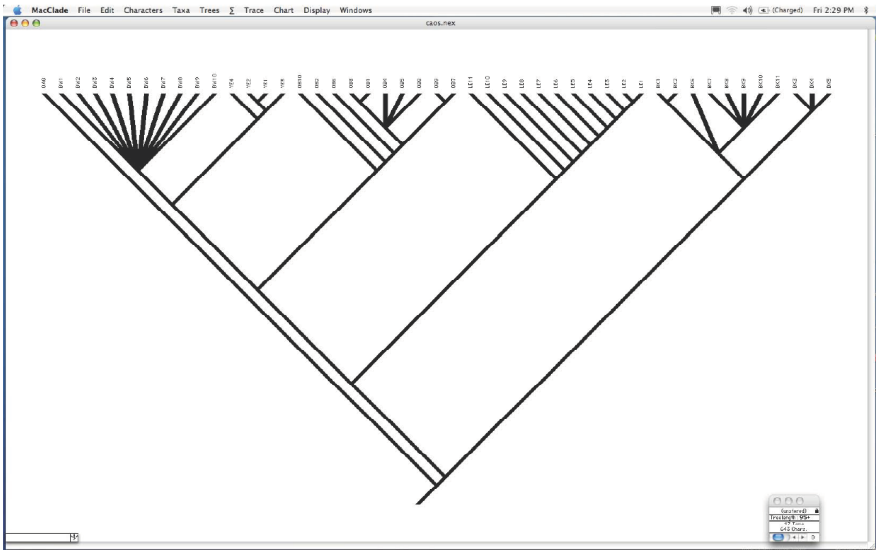


Figure_2_

5) Before proceeding to generate diagnostics, the nodes within species can be collapsed using the "collapse clade" palette tool.

Figure _3_

The collapse clade operation leaves a tree with unresolved monophyletic groups that represent the individuals in pre-described species boundaries. The branches that are collapsed in this fashion eliminate some structure in the tree but structure within each species is irrelevant to discovering diagnostics.

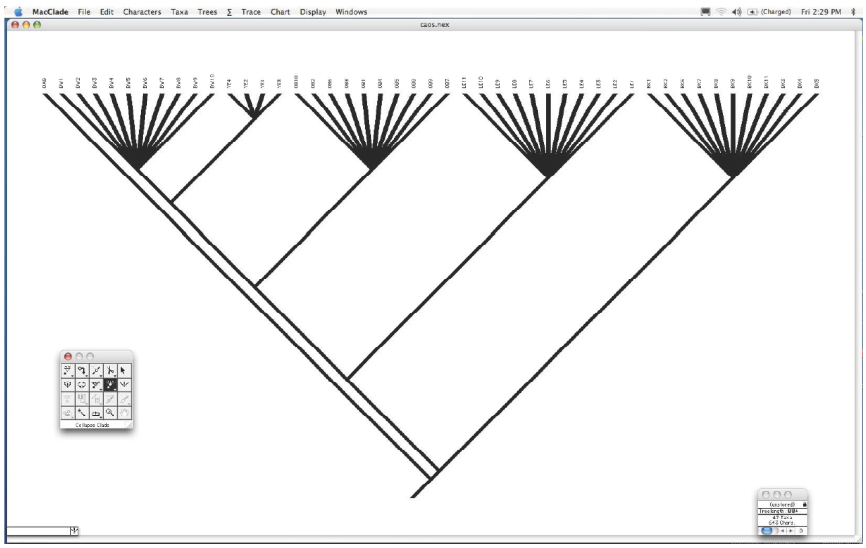


Figure _4_

6) Save the NEXUS file by going to the File pulldown and click on “Options for Saving > NEXUS format”.

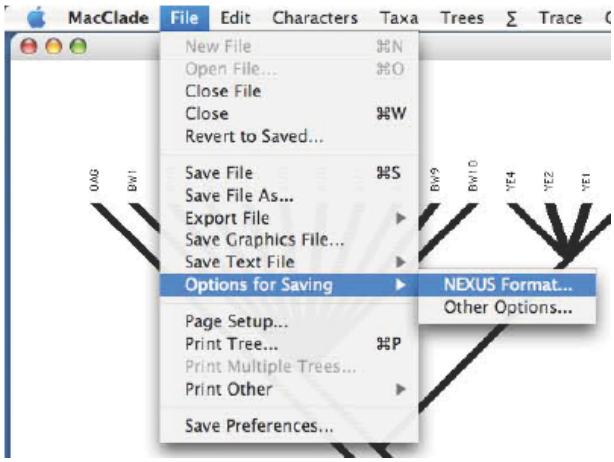


Figure_5_

Turn off the “interleave” option (p-gnome will not operate on interleaved matrices). Also make sure that the file is saved with a translation table. Figure 6 shows the MacClade menu options for saving that are necessary for p-gnome.

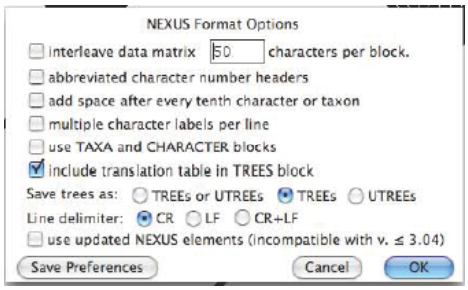


Figure _6_

Your matrix should now be ready for processing in p-gnome.

3. Running p-gnome

1) Open a terminal on an Apple OS-X machine. Simply type in "p-gnome" and the program will be accessed.

2) P-gnome will give three prompts. First the program will prompt for the input file. When a valid input NEXUS file is typed in and the return key is hit, the second prompt will appear.

3) This second prompt asks for a name for the folder that will contain the p-gnome output files. When this name is typed and the return key is hit CAOS searches for diagnostics and generates diagnostic rules.

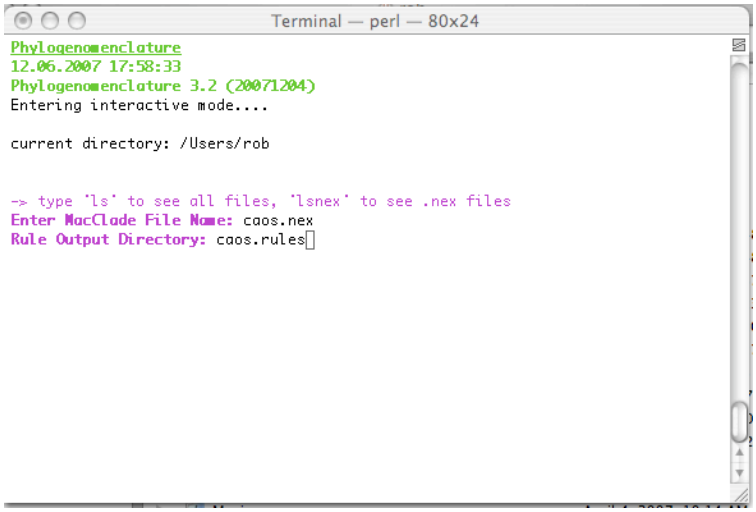
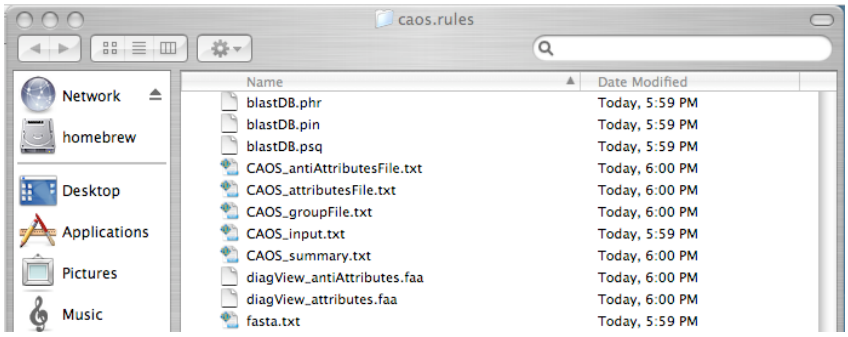


Figure __

5) P-gnome is finished generating rules when the prompt reads "P-Gnome Rule Generation Complete." p-gnome will generate diagnostic rules and place the results in a folder (named as above) in the same folder as the home of the program.

4. Interpreting p-gnome results

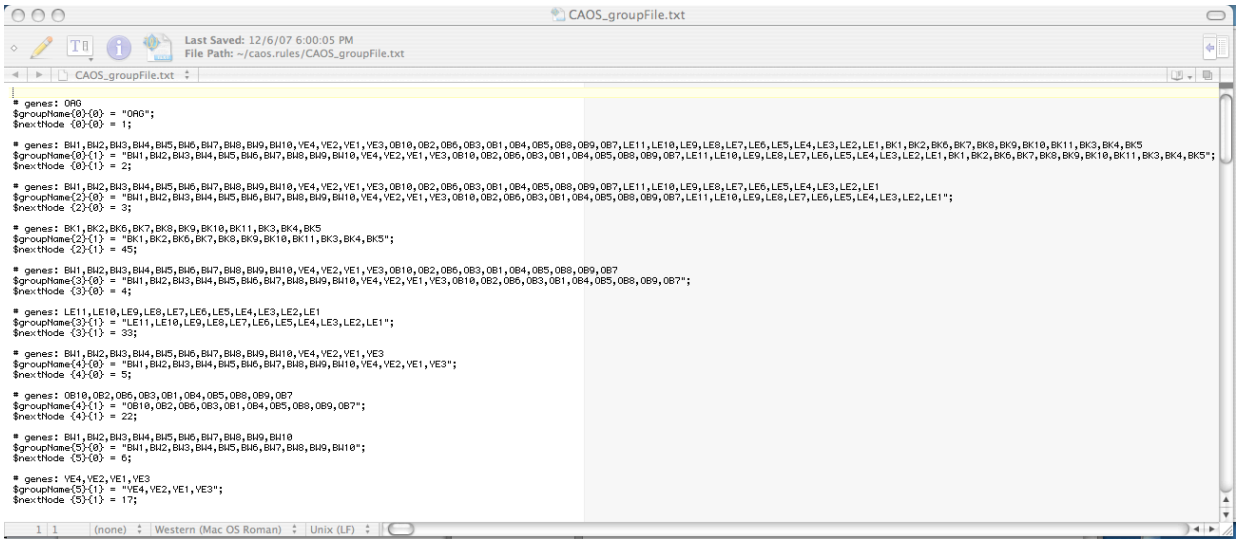
p-gnome generates the following files:



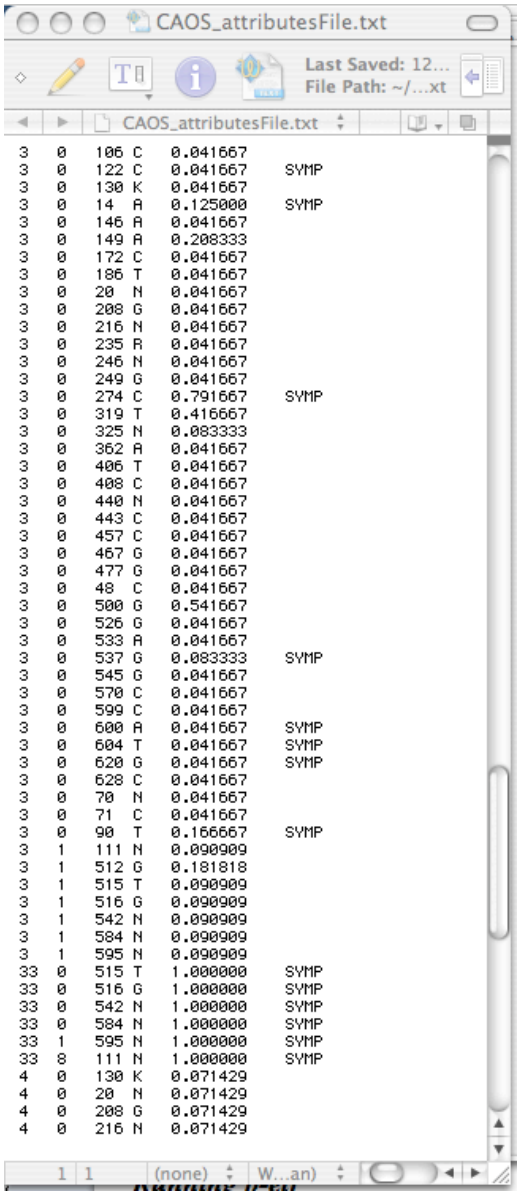
Here is a brief description of each of the files generated by P-Gnome:

- CAOS_input.txt - data matrix and tree data in CAOS format
- CAOS_attributesFile.txt - the attributes found through performed CAOS analysis
- CAOS_groupFile.txt - grouping file for names of groups, and keeping track of nodes during classification
- fasta.txt - data sequences in FastA format
- blastDB.phr - used by BLAST for classification
- blastDB.psq - used by BLAST for classification
- blastDB.pin - used by BLAST for classification

By opening the CAOS_groupFile.txt the user can view the groups that were created in MacClade and the numbering system for the groups that will appear in the CAOS_attributesFile.txt file. For each node in the tree a rules group will be generated. Only the groups that hold the specimen names that represent the pre-described species are relevant to the diagnosis search. The first few lines of the groups file will look like the following:



By opening CAOS_attributesFile.txt the diagnostics can be viewed. The screenshot below shows how this file should look.



There are six columns in the attributes output file. The first and second columns indicate the group coordinates that are established in the group file (see above). The third column indicates the base pair position of the residue that is a potential diagnostic and the fourth column indicates the level of confidence in the diagnostic.

Only diagnostics with values of 1.00 are considered pure diagnostics. For the table below we have extracted the group coordinates that correspond to prescribed potential species units and have placed the attributes output for each of these groups below them.

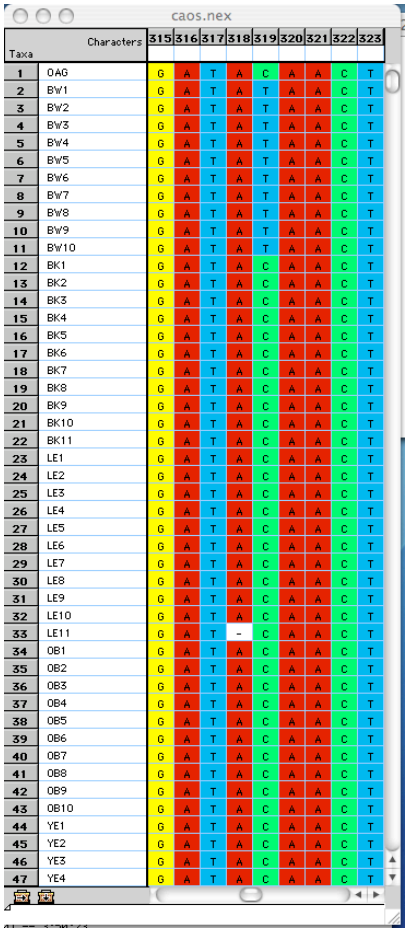
```
# genes: BK1,BK2,BK6,BK7,BK8,BK9,BK10,BK11,BK3,BK4,BK5
```

```
$groupName{2}{1} = "BK1,BK2,BK6,BK7,BK8,BK9,BK10,BK11,BK3,BK4,BK5";
$nextNode {2}{1} = 45;
2       1       552     N       0.090909        SYMP
2       1       595     G       0.181818        SYMP


# genes: LE11,LE10,LE9,LE8,LE7,LE6,LE5,LE4,LE3,LE2,LE1
$groupName{3}{1} = "LE11,LE10,LE9,LE8,LE7,LE6,LE5,LE4,LE3,LE2,LE1";
$nextNode {3}{1} = 33;
3       1       111     N       0.090909
3       1       512     G       0.181818
3       1       515     T       0.090909
3       1       516     G       0.090909
3       1       542     N       0.090909
3       1       584     N       0.090909
3       1       595     N       0.090909


# genes: OB10,OB2,OB6,OB3,OB1,OB4,OB5,OB8,OB9,OB7
$groupName{4}{1} = "OB10,OB2,OB6,OB3,OB1,OB4,OB5,OB8,OB9,OB7";
$nextNode {4}{1} = 22;
4       1       106     C       0.100000
4       1       122     C       0.100000
4       1       146     A       0.100000
4       1       149     A       0.500000
4       1       172     C       0.100000
4       1       186     T       0.100000
4       1       235     R       0.100000
4       1       246     N       0.100000
4       1       249     G       0.100000
4       1       362     A       0.100000
4       1       406     T       0.100000
4       1       443     C       0.100000
4       1       457     C       0.100000
4       1       467     G       0.100000
4       1       477     G       0.100000
4       1       48      C       0.100000
4       1       526     G       0.100000
4       1       533     A       0.100000
4       1       537     G       0.200000
4       1       545     G       0.100000
4       1       570     C       0.100000
4       1       600     A       0.100000
4       1       604     T       0.100000
4       1       620     G       0.100000
4       1       628     C       0.100000
4       1       71      C       0.100000


# genes: BW1,BW2,BW3,BW4,BW5,BW6,BW7,BW8,BW9,BW10
$groupName{5}{0} = "BW1,BW2,BW3,BW4,BW5,BW6,BW7,BW8,BW9,BW10";
$nextNode {5}{0} = 6;
5       0       20      N       0.100000
5       0       208     G       0.100000
5       0       216     N       0.100000
5       0       319     T       1.000000
5       0       325     N       0.200000
5       0       70      N       0.100000


# genes: YE4,YE2,YE1,YE3
$groupName{5}{1} = "YE4,YE2,YE1,YE3";
$nextNode {5}{1} = 17;
5       1       130     K       0.250000
5       1       14      A       0.250000        SYMP
```

5	1	274	T	0.250000	SYMP
5	**1**	**319**	**C**	**1.000000**	**SYMP**
5	1	408	C	0.250000	
5	1	440	N	0.250000	
5	1	500	A	0.250000	SYMP
5	1	599	C	0.250000	
5	1	90	T	0.250000	SYMP

In the above table there are two such positions – one in the BW entity (in red) and one in the YE entity (in blue). Note however that the YE diagnostic at position 319 has a SYMP in the sixth column. This last column indicates whether or not the diagnostic is symplesiomorphic (ie present in other more basal entities in the analysis) and in this case indicates that the C in position 319 in the YE entitiy is not diagnostic for that entity.

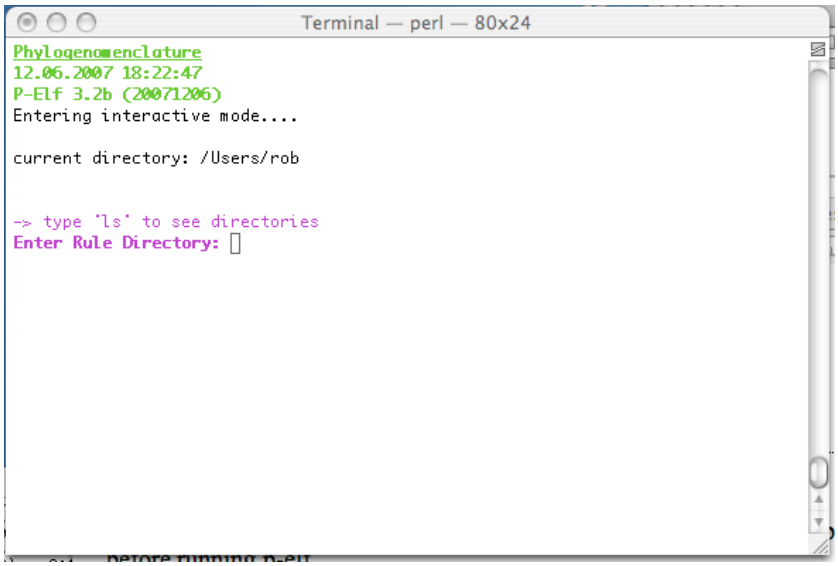
To validate this diagnostic we show the MacClade text editor screen shot for this position in the caos.nex matrix. Note that position 319 is fixed for T in the BW entity and different in all others.



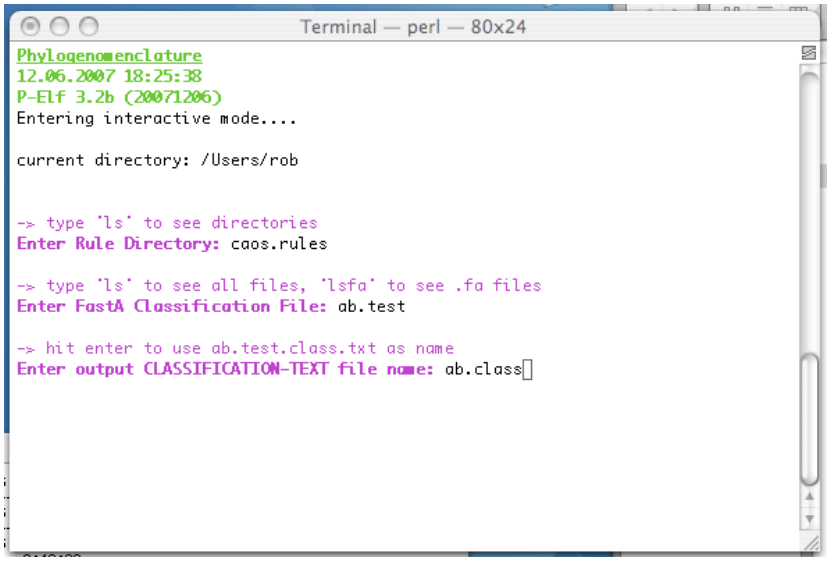
5. *Running p-elf*

1. Input files for p-elf are FASTA files. The files can have gaps or gaps can be removed before running p-elf.

2. Open a terminal window on a MacIntosh-OSX and type in "p-elf". The program should start to run. The first prompt is for the rules folder (generated by p-gnome) that will be used as the rules to do the identifications. The screen should look like the following.

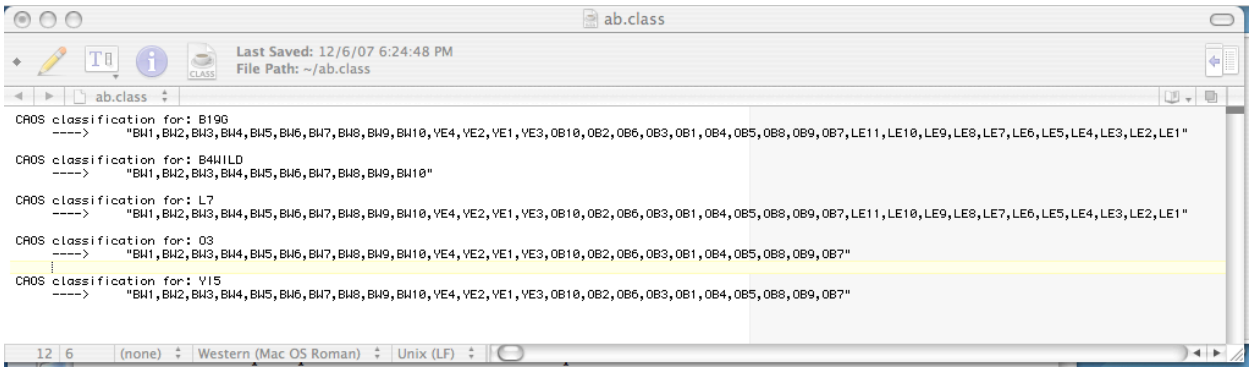


3. The next two prompts ask for the file with the sequences that need to be classified or identified and the name that will be given to the file with the output from the classification step. The finished file input should look like the following screen.



6. *Interpreting p-elf output*

The p-elf program uses the rule set to identify the query sequences and returns the named file with the classifications in it. The file should look like this:



Note that in this file five query sequences were classified using a rule set with only one of the five entities in the study having a pure diagnostic. p-elf returns classifications for these five queries that show the potential groups that the query can be classified to. In this case only one query can be classified down to a single entity – BW. All other queries are not classifiable to the pre-described entities for their origin. These results are consistent with the presence of a single pure diagnostic in the BW entity.