MCB 371/372

PHYLIP &
Exercises
4/13/05

Peter Gogarten
Office: *BSP 404*
phone: *860 486-4061,*
Email: *gogarten@uconn.edu*

input and output

Input and output files

For most of the PHYLIP programs, information comes from a series of input files, and ends up in a series of output files:

The programs interact with the user by presenting a menu. Aside from the user's choices from the menu, they read all other input from files. These files have default names. The program will try to find a file of that name - if it does not, it will ask the user to supply the name of that file. Input data such as DNA sequences comes from a file whose default name is `infile`. If the user supplies a tree, this is in a file whose default name is `intree`. Values of weights for the characters are in `weights`, and the tree plotting program need some digitized fonts which are supplied in `fontfile` (all these are default names).

Programs in PHYLIP are Modular

For example:

SEQBOOT take one set of aligned sequences and writes out a file containing bootstrap samples.

PROTDIST takes a aligned sequences (one or many sets) and calculates distance matices (one or many)

FITCH (or NEIGHBOR) calculate best fitting or neighbor joining trees from one or many distance matrices

CONSENSE takes many trees and returns a consensus tree

…. modules are available to draw trees as well, but often people use treeview or njplot

The Phylip Manual is an excellent source of information.

Brief one line descriptions of the programs are here

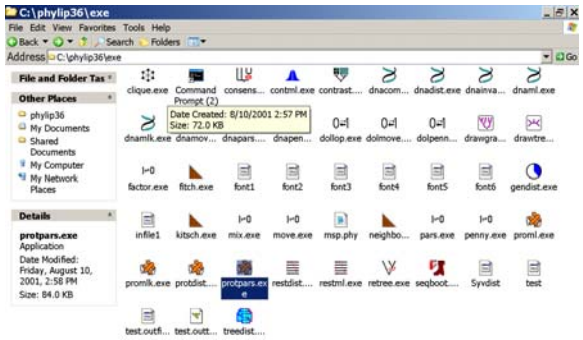
The easiest way to run PHYLIP programs is via a command line menu (similar to clustalw). The program is invoked through clicking on an icon, or by typing the program name at the command line.

```
> seqboot
> protpars
> fitch
```

If there is no file called infile the program responds with:

```
[gogarten@carrot gogarten]$ seqboot
seqboot: can't find input file "infile"
Please enter a new file name>
```

program folder



Example 1 Protpars

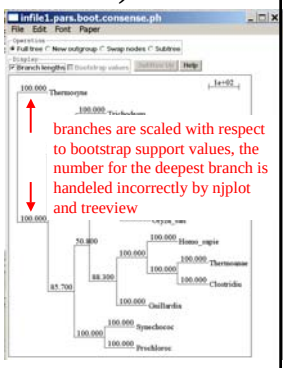
example: seqboot, protpars, consense on infile1

NOTE the bootstrap majority consensus tree does not necessarily have the same topology as the "best tree" from the original data!

threshold parsimony,
gap symbols - versus ?
(in vi you could use `:%s/-/?/g` to replace all – ?)
outfile
outtree compare to distance matrix analysis

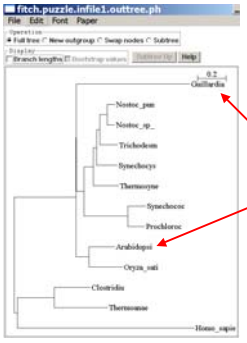
protpars (versus distance/FM)

```
Extended majority rule consensus tree

CONSENSUS TREE:
the numbers on the branches indicate the number
of times the partition of the species into the two sets
which are separated by that branch occurred
among the trees, out of 100.00 trees

                                        +------Prochloroc
                  +---------------------100.-|
                  |                     +------Synechococ
                  |
                  |                  +--------------------Guillardia
         +-85.7-|                   |
         |       |          +-88.3-|                  +------Clostridiu
         |       |          |      |           +-100.-|
         |       |          |      +-100.-|    +------Thermoanae
         |       |          +-50.8-|          |
         |       |                 |          +------------Homo sapie
+------|         |                 |
|      |         |                 +------Oryza sati
|      |         |              +-------------100.0-|
|      |         |              |                  +------Arabidopsi
|      |         |
|      |         |              +--------------------Synechocys
|      |         |              |
|      |      +--------------53.0-|                  +------Nostoc pun
|      |      |                   |           +-99.5-|
|      |      |                   +-38.5-|    +------Nostoc sp
|      |      |                          |
|      |      |                          +------------Trichodesm
|      |      |
+------+------+-------------------------------------------Thermosyne

 remember: this is an unrooted tree!
```



branches are scaled with respect to bootstrap support values, the number for the deepest branch is handeled incorrectly by njplot and treeview

(protpars versus) distance/FM



Tree is scaled with respect to the estimated number of substitutions.

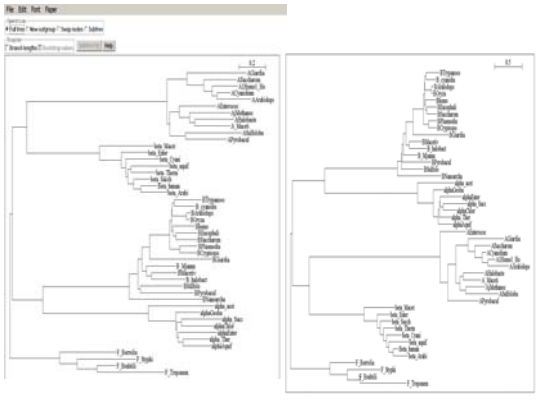
what might be the explanation for the red algae not grouping with the plants?

If time: demo of njplot

protdist

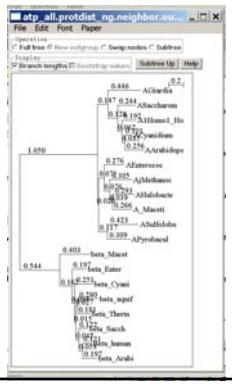
PROTdist
Settings for this run:
 P Use JTT, PMB, PAM, Kimura, categories model? Jones-Taylor-Thornton matrix
 G Gamma distribution of rates among positions? No
 C One category of substitution rates? Yes
 W Use weights for positions? No
 M Analyze multiple data sets? No
 I Input sequences interleaved? Yes
 0 Terminal type (IBM PC, ANSI)? ANSI
 1 Print out the data at start of run No
 2 Print indications of progress of run Yes

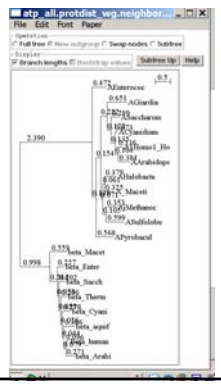
without and with correction for ASRV



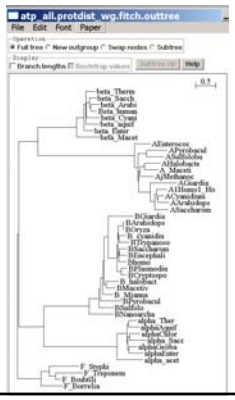
subtree with branch lengths

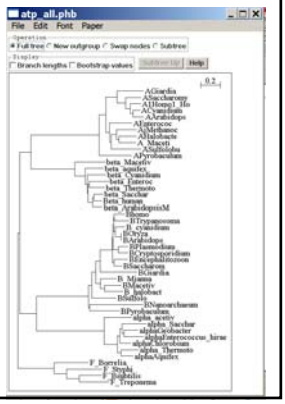
without and with correction for ASRV



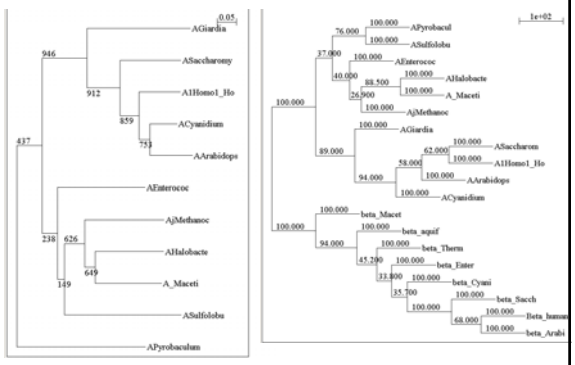


compare to trees with FITCH and clustalw – same dataset





bootstrap support ala clustal protpars (gaps as ?)



Progress in Unix

wget is a command that is implemented in most modern UNIX flavors (linux, darwin ..) You execute it from the command line.
For example:
`>wget http://web.uconn.edu/gogarten/MCB372/Laboratories/test1.fa`
will copy test1.fa from the web address
http://web.uconn.edu/gogarten/MCB372/Laboratories/test1.fa into your current working directory.

Using this in conjunction with
<ctrl> <copy> (PC), <apple> <copy> (MAC), or
<shift> <insert> (linux/darwin) this saves a lot of typing.
(control c in UNIX terminates whatever program you are running.
Therefore the copy shortcut usually is <ctrl><ins>)

Hasan and Lina volunteered to do a short UNIX intro this afternoon in the lab.

perl assignment #1:

On Monday we used the perl script extract_lines.pl .

Modify the script so that it prints out an additional column that for each blasthit it alse gives the bitscore of the alignment divided by the alignment lengths.

Hints:
`chomp ($line);` removes the \n character at the end of $line
`$parts[i]` contains the (i+1)s entry of the line.
If $a[11] and $a[3] are variables in an array that contain numbers you can assign a new variable to the ratio using the command
`$ratio_of_values=$a[11]/$a[3];`
To print things in a line in the output separated by tabs and a new line symbol at the end you could say for example:
`print OUT "$line\t$parts[12]\t$ratio\n";`

Go over new*.pl in from_peter/temp/ on carrot
One working program is here.

perl assignment #2:

Assume that you have the following non-aligned multiple sequence files in a directory:

A.fa : vacuolar/archaeal ATPase catalytic subunits ;
B.fa : vacuolar/archaeal ATPase non-catalytic subunits;
alpha.fa : F-ATPases non-catalytic subunits,
beta.fa : F-ATPases catalytic subunits,
F.fa : ATPase involved in the assembly of the bacterial flagella.

Write a perl script that executes muscle or clustalw and
1) aligns the sequences within each file
2) successively calculates profile alignments between all aligned sequences.

Hints:
`system (command)` executes "command"as if you had typed `command` in the command line

A working solution is here

questions and comments

```
@parts=split(/\./,$file);  #why is the \ in \. necessary?

if ($counter==0);  #why would one "=" fail?

system("clustalw -profile1=result.out
-profile2=$filename.aln -outfile=result.out -profile
-align");
```
#Notes: the profile alignment is written back into the file from which profile1 originated. The file does not need to end on .aln -- but it is nice to keep the convention (see below).
Perl replaces the variables (`$filename` with their content e.g. `beta`).

Note the use of indentations to clarify the control structures.

`$counter=$counter+1;` and `$counter += 1` are equivalent
= is an assignment operator.
`$counter=$counter+1` is not a mathematical equation

Perl assignment #3

Write a script that takes all phylip formated aligned multiple sequence files present in a directory, and perfomes a bootstrap analyses using maximum parsimony.
I.e., the script should go through the same steps as we did in the exercises #4 tasks 1a and 1c

Files you might want to use are A.fa, B.fa, alpha.fa, beta.fa, and atp_all.phy. BUT you first have to convert them to phylip format AND you should replace some or all gaps with ?
(In the end you would be able to answer the question "does the resolution increase if a more related subgroup is analyzed independent from an outgroup?)

hints

Rather than typing commands at the menu, you can write the responses that you would need to give via the keyboard into a file (e.g. your_input.txt)

You could start and execute the program protpars by typing

protpars < your_input.txt

your input.txt might contain the following lines:

```
infile1.txt
r
t
10
y
r
r
```

in the script you could use the line

```
system ("protpars < your_input.txt");
```

The main problem are the owerwrite commands if the oufile and outtree files are already existing. You can either create these beforehand, or erase them by moving (mv) their contents somewhere else.