MCB 371/372

phyml &

treepuzzle

4/18/05

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

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.

create *.phy files

the easiest (probably) is to run clustalw with the phylip option:
For example (here):

```
#!/usr/bin/perl -w
print "# This program aligns all multiple sequence files with names *.fa \n
# found in its directory using clustalw,  and saves them in phyip format.\n“;
while(defined($file=glob("*.fa"))){
                @parts=split(/\./,$file);
                $file=$parts[0];
                system("clustalw -infile=$file.fa -align -output=PHYLIP");
                };
# cleanup:
system ("rm *.dnd");
exit;
```

Alternatively, you could use a web version of readseq – this one worked great for me ☺

run phylip programs from perl

An example on how to solve the homework assignment is here, the cmd files are here, here, and here:

```
#!/usr/bin/perl -w
print "# This program runs seqboot, protpars and consense on all multiple \n
# sequence files with names *.phy\n";

while(defined($file=glob("*.phy"))){

        @parts=split(/\./,$file);
        $file=$parts[0];

        system ("cp $file.phy infile");
        system ("seqboot < seqboot.cmd");
        system ("mv outfile infile");
        system ("protpars < protpars.cmd");
        system ("rm outfile");
        system ("mv outtree intree");
        system ("consense < consense.cmd");
        system ("mv outtree $file.outtree");
        system ("mv outfile $file.outfile");
};

# cleanup:
system ("rm infile");
system ("rm intree");
exit;
```

Alternative for entering the commands for the menu:

```
#!/usr/bin/perl -w
        system ("cp A.phy infile");
        system ("echo -e 'y\n9\n'|seqboot");
exit;
```

`echo` **returns the string in** `‘ ‘, i.e., y\n9\n.`
The `-e` **options allows the use of** `\n`
The `|` **symbol pipes the output from echo to seqboot**

phyml

PHYML - A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood

An online interface is here ;
there is a command line version that is described here (not as straight forward as in clustalw);
a phylip like interface is automatically invoked, if you type "phyml" – the manual is here.

Phyml is installed on bbcxsrv1.

Do example on atp_all.phy
Note data type, bootstrap option within program, models for ASRV (pinvar and gamma), by default the starting tree is calculated via neighbor joining.

phyml - comments

Under some circumstances the consensus tree calculated by phyml is wrong. It is recommended to save all the individual trees and to also evaluate them with *consense* from the phylip package.
Note: phyml allows longer names, but consense allows only 10 characters!

phyml is fast enough to analyze dataset with hundreds of sequences (in 1990, a maximum likelihood analyses with 12 sequences (no ASRV) took several days).

For moderately sized datasets you can estimate branch support through a bootstrap analysis (it still might run several hours, but compared to protml or PAUP, this is extremely fast).

The paper describing phyml is here,
a brief interview with the authors is here

TreePuzzle ne PUZZLE

TREE-PUZZLE is a very versatile maximum likelihood program that is particularly useful to analyze protein sequences. The program was developed by Korbian Strimmer and Arnd von Haseler (then at the Univ. of Munich) and is maintained by von Haseler, Heiko A. Schmidt, and Martin Vingron

(contacts see http://www.tree-puzzle.de/).

TREE-PUZZLE

- allows fast and accurate estimation of ASRV (through estimating the shape parameter alpha) for both nucleotide and amino acid sequences,
- It has a "fast" algorithm to calculate trees through quartet puzzling (calculating ml trees for quartets of species and building the multispecies tree from the quartets).
- The program provides confidence numbers (puzzle support values), which tend to be smaller than bootstrap values (i.e. provide a more conservative estimate),
- the program calculates branch lengths and likelihood for user defined trees, which is great if you want to compare different tree topologies, or different models using the **maximum likelihood ratio test**.
- Branches which are not significantly supported are collapsed.
- TREE-PUZZLE runs on "all" platforms
- TREE-PUZZLE reads PHYLIP format, and communicates with the user in a way similar to the PHYLIP programs.

Maximum likelihood ratio test

If you want to compare two models of evolution (this includes the tree) given a data set, you can utilize the so-called maximum likelihood ratio test.
If L_1 and L_2 are the likelihoods of the two models, d =2(logL_1-logL_2) approximately follows a Chi square distribution with n degrees of freedom. Usually n is the difference in model parameters. I.e., how many parameters are used to describe the substitution process and the tree. In particular n can be the difference in branches between two trees (one tree is more resolved than the other).
In principle, this test can only be applied if on model is a more refined version of the other. In the particular case, when you compare two trees, one calculated without assuming a clock, the other assuming a clock, the degrees of freedom are the number of OTUs – 2 (as all sequences end up in the present at the same level, their branches cannot be freely chosen) .

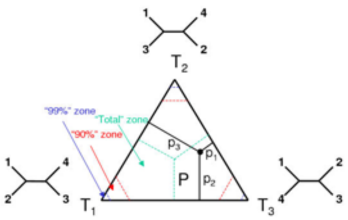
To calculate the probability you can use the CHISQUARE calculator for windows available from Paul Lewis.

TREE-PUZZLE allows (cont)

- TREEPUZZLE calculates distance matrices using the ml specified model. These can be used in FITCH or Neighbor.
PUZZLEBOOT automates this approach to do bootstrap analyses – WARNING: this is a distance matrix analyses!
The official script for PUZZLEBOOT is here – you need to create a command file (puzzle.cmds), and puzzle needs to be envocable through the command puzzle.
Your input file needs to be the renamed outfile from **seqboot**
A slightly modified working version of puzzleboot_mod.sh is here, and here is an example for puzzle.cmds . Read the instructions before you run this!
- Maximum likelihood mapping is an excellent way to assess the phylogenetic information contained in a dataset.
- ML mapping can be used to calculate the support around one branch.

@@@ Puzzle is cool, don't leave home without it! @@@

ml mapping



From: Olga Zhaxybayeva and J Peter Gogarten ***BMC Genomics*** **2002, 3:4**

ml mapping

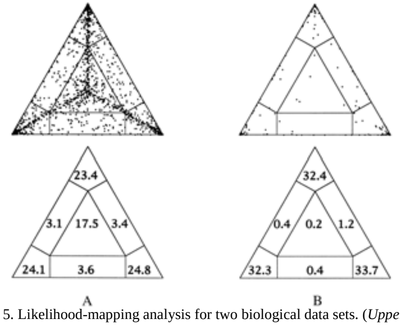


Figure 5. Likelihood-mapping analysis for two biological data sets. (*Upper*) The distribution patterns. (*Lower*) The occupancies (in percent) for the seven areas of attraction.

(*A*) Cytochrome-*b* data from ref. 14. (*B*) Ribosomal DNA of major arthropod groups (15).

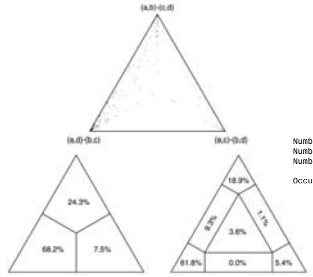
From: Korbinian Strimmer and Arndt von Haeseler Proc. Natl. Acad. Sci. USA Vol. 94, pp. 6815-6819, June 1997

ml mapping (cont)

If we want to know if *Giardia lamblia* forms the deepest branch within the known eukaryotes, we can use ML mapping to address this problem.
To apply ml mapping we choose the "higher" eukaryotes as cluster a, another deep branching eukaryote (the one that competes against Giardia) as cluster b, Giardia as cluster c, and the outgroup as cluster d. For an example output see this sample ml-map.

An analysis of the carbamoyl phosphate synthetase domains with respect to the root of the tree of life is here.

Application of ML mapping to comparative Genome analyses
see here for a comparison of different probabil;ity measures
see here for an approach that solves the problem of poor taxon sampling that is usually considered inherent with quartet analyses is.



```
Cluster a: 14 sequences
outgroup (prokaryotes)

Cluster b: 20 sequences
other Eukaryotes

Cluster c: 1 sequences
Plasmodium

Cluster d: 1 sequences
Giardia
```

```
              (a,b)-(c,d)
                   /\
                  /  \
                 /    \
                /   1  \
               / \    / \
              /   \  /   \
             /     \/     \
            /  3    :   2  \
           /        :       \
          /__________________\
(a,d)-(b,c)                  (a,c)-(b,d)

Number of quartets in region 1: 68 (= 24.3%)
Number of quartets in region 2: 21 (= 7.5%)
Number of quartets in region 3: 191 (= 68.2%)

Occupancies of the seven areas 1, 2, 3, 4, 5, 6, 7:

              (a,b)-(c,d)
                   /\
                  /  \
                 /  1 \
                / \  / \
               /   /\   \
              / 6 /  \ 4 \
             /   /  7 \   \
            / \ /______\ / \
           / 3  :   5  :  2 \
          /__________________\
(a,d)-(b,c)                  (a,c)-(b,d)

Number of quartets in region 1: 53 (= 18.9%)
Number of quartets in region 2: 15 (= 5.4%)
Number of quartets in region 3: 173 (= 61.8%)
Number of quartets in region 4: 3 (= 1.1%)
Number of quartets in region 5: 0 (= 0.0%)
Number of quartets in region 6: 26 (= 9.3%)
Number of quartets in region 7: 10 (= 3.6%)
```

TREE-PUZZLE – PROBLEMS/DRAWBACKS

- The more species you add the lower the support for individual branches. While this is true for all algorithms, in TREE-PUZZLE this can lead to completely unresolved trees with only a few handful of sequences.

- Trees calculated via quartet puzzling are usually not completely resolved, and they do not correspond to the ML-tree: The determined multi-species tree is not the tree with the highest likelihood, rather it is the tree whose topology is supported through ml-quartets, and the lengths of the resolved branches is determined through maximum likelihood.

puzzle example

archaea_euk.phy in puzzle_temp

usertree

check outfile