MCB 371/372

PHYLIP

how to make sense out of a tree

4/11/05

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

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

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.

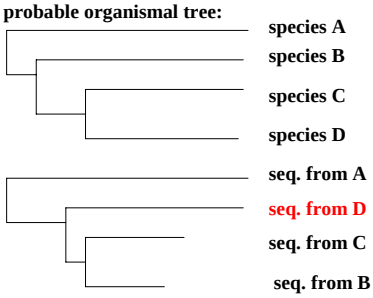
Write a perl script that executes muscle or clustalw and
1) aligns the sequences within each file
2) successively calculates profile alignments between all aligned sequences.

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

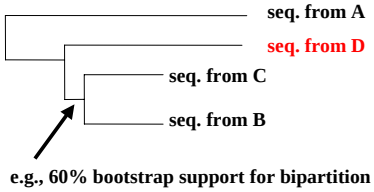
Trees – what might they mean?

Calculating a tree is comparatively easy, figuring out what it might mean is much more difficult.

If this is the probable organismal tree:

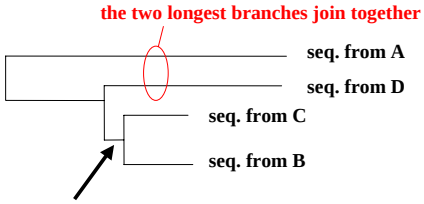


lack of resolution



e.g., 60% bootstrap support for bipartition (AD)(CB)

long branch attraction artifact

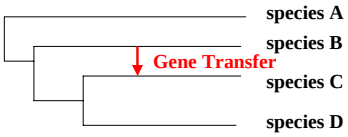


e.g., 100% bootstrap support for bipartition (AD)(CB)

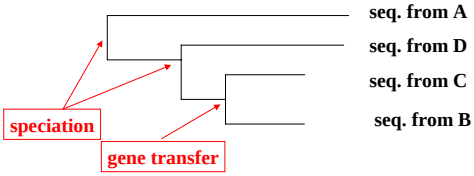
What could you do to investigate if this is a possible explanation?
use only slow positions,
use an algorithm that corrects for ASRV

Gene transfer

Organismal tree:

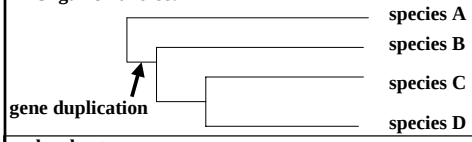


molecular tree:

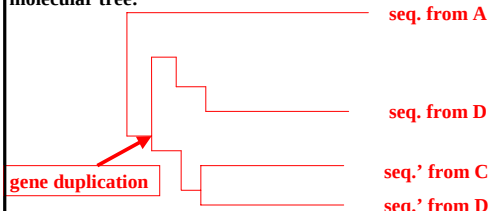


Gene duplication

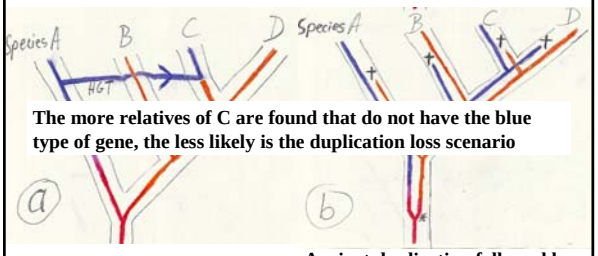
Organismal tree:



molecular tree:



Gene duplication and gene transfer are equivalent explanations.



The more relatives of C are found that do not have the blue type of gene, the less likely is the duplication loss scenario

Horizontal or lateral Gene

Ancient duplication followed by gene loss

Note that scenario B involves many more individual events than A

1 HGT with orthologous replacement

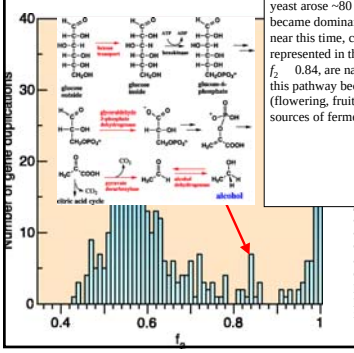
1 gene duplication followed by 4 independent gene loss events

What is it good for?

Gene duplication events can provide an outgroup that allows rooting a molecular phylogeny.
Most famously this principle was applied in case of the tree of life – the only outgroup available in this case are ancient paralogs (see http://gogarten.uconn.edu/cvs/Publ_Pres.htm for more info).
However, the same principle also is applicable to any group of organisms, where a duplication preceded the radiation (example).
Lineage specific duplications also provide insights into which traits were important during evolution of a lineage.

e.g. gene duplications in yeast

from Benner et al., 2002



The chemical pathway that converts glucose to alcohol in yeast arose ~80 Ma, near the time that fermentable fruits became dominant. Gene families that suffered duplication near this time, captured in the episode of gene duplication represented in the histogram in Fig. 1 by bars at f_2 0.84, are named in red. According to the hypothesis, this pathway became useful to yeast when angiosperms (flowering, fruiting plants) began to provide abundant sources of fermentable sugar in their fruits.

at ~80 Ma, whereby yeast gained its ability to ferment sugars found in fruits created by angiosperms. Also noticeable are recent duplications of genes that enable yeast to speed DNA synthesis, protein synthesis, and malt degradation, presumably representing yeast's recent interaction with humans.

e.g. gene duplications in yeast

from Benner et al., 2002

Also noticeable are recent duplications of genes that enable yeast to speed DNA synthesis, protein synthesis, and malt degradation, presumably representing yeast's recent interaction with humans.

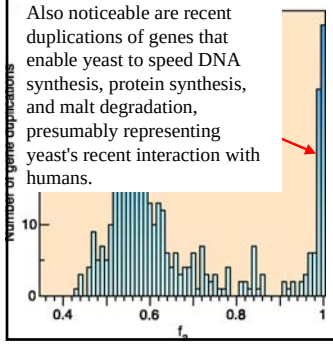
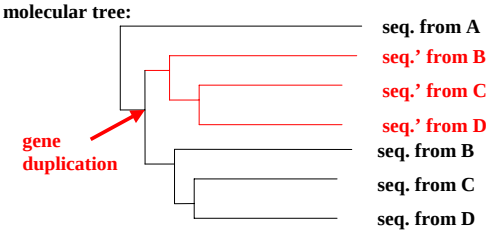


Figure 1. The number of duplicated gene pairs (vertical axis) in the genome of the yeast *Saccharomyces cerevisiae* versus f_2, a metric that models divergence of silent positions in twofold redundant codon systems via an approach-to-equilibrium kinetic process and therefore acts as a logarithmic scale of the time since the duplications occurred. Recent duplications are represented by bars at the right. Duplications that diverged so long ago that equilibrium at the silent sites has been reached are represented by bars where f_2 0.55. Noticeable are episodes of gene duplication between the two extremes, including a duplication at f_2 0.84. This represents the duplication, at ~80 Ma, whereby yeast gained its ability to ferment sugars found in fruits created by angiosperms. Also noticeable are recent duplications of genes that enable yeast to speed DNA synthesis, protein synthesis, and malt degradation, presumably representing yeast's recent interaction with humans.

Function, ortho- and paralogy

molecular tree:



The presence of the duplication is a taxonomic character (shared derived character in species B C D).
The phylogeny suggests that seq' and seq have similar function, and that this function was important in the evolution of the clade BCD.
seq' in B and seq'in C and D are orthologs and probably have the same function, whereas seq and seq' in BCD probably have different function (the difference might be in subfunctionalization of functions that seq had in A. – e.g. organ specific expression)

Phylip

written and distributed by Joe Felsenstein and collaborators (some of the following is copied from the PHYLIP homepage)

PHYLIP (the *PHYL*ogeny *I*nference *P*ackage) is a package of programs for inferring phylogenies (evolutionary trees).

PHYLIP is the most widely-distributed phylogeny package, and competes with PAUP* to be the one responsible for the largest number of published trees. PHYLIP has been in distribution since 1980, and has over 15,000 registered users.

Output is written onto special files with names like "outfile" and "outtree". Trees written onto "outtree" are in the Newick format, an informal standard agreed to in 1986 by authors of a number of major phylogeny packages.

Input is either provided via a file called "infile" or in response to a prompt.

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).

What's in PHYLIP

Programs in PHYLIP allow to do parsimony, distance matrix, and likelihood methods, including bootstrapping and consensus trees. Data types that can be handled include molecular sequences, gene frequencies, restriction sites and fragments, distance matrices, and discrete characters.

Phylip works well with protein and nucleotide sequences
Many other programs mimic the style of PHYLIP programs. (e.g. TREEPUZZLE, phyml, protml)

Many other packages use PHYIP programs in their inner workings (e.g., PHYLO_WIN)

PHYLIP runs under all operating systems

Web interfaces are available

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

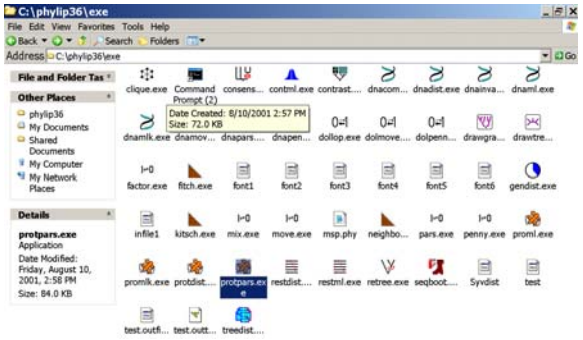
The easiest way to run PHYLIP programs is via a command line menu (similar to clustalw). The program is invoked through clicking on an icon, or by typing the program name at the command line.

```
> 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



menu interface

Example 1 Protpars

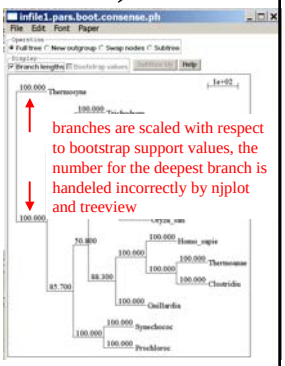
example: seqboot, protpars, consense on infile1

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

threshold parsimony,
gap symbols - versus ?
outfile
outtree compare to distance matrix analysis

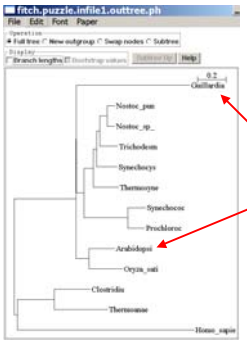
protpars (versus distance/FM)

```
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