MCB5472
Computer methods in molecular evolution

Lecture 3/22/2014

slides 1-46 and slide 104/105

Review of Alignments

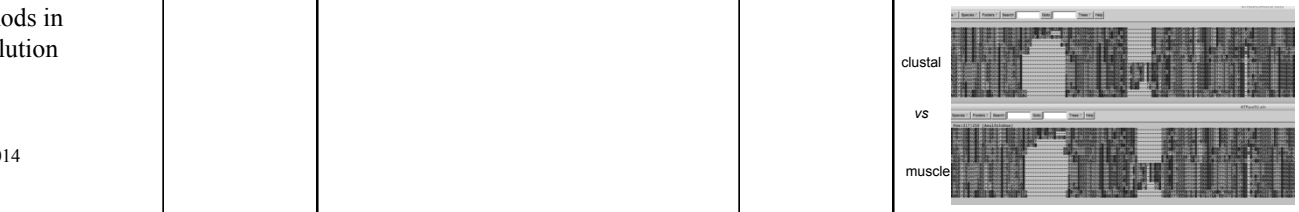
ATPase dataset

Alignment

clustal

vs

muscle

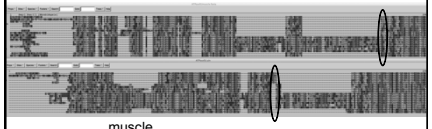


Conserved parts are aligned reproducibly

ATPase dataset

Alignment

clustal *vs*

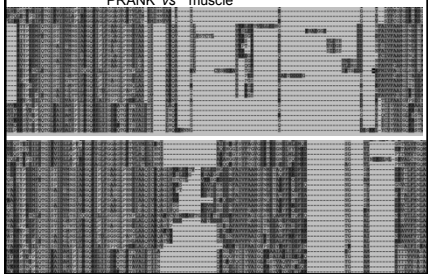


muscle

The alignment of the less conserved parts is questionable. Using the progressive alignment approach on these sequences can cause problems in downstream analyses.

ATPase dataset

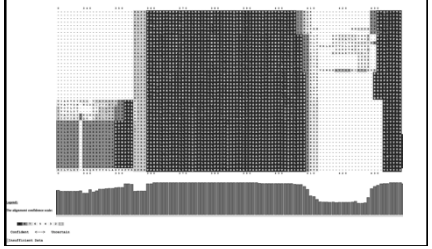
Alignment PRANK *vs* muscle



ATPase dataset

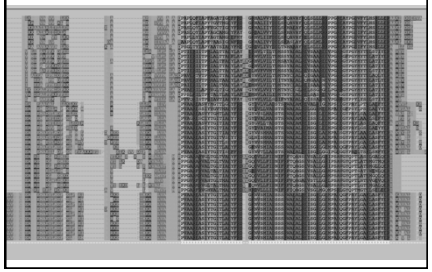
Reliably aligned positions determined with *guidance*

MSA color-coded by GUIDANCE scores



ATPase dataset

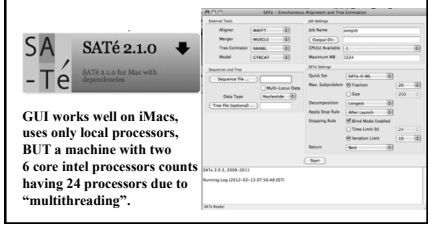
Reliably aligned columns determined with GBLOCKS



SATé

Simultaneous Alignment and Tree Estimation

http://phylo.bio.ku.edu/software/sate/sate.html



GUI works well on iMacs, uses only local processors, BUT a machine with two 6 core intel processors counts having 24 processors due to "multithreading".

Discussion

- What is the effect of throwing out sites that are not reliably aligned?
- What can one do to avoid phylogenetic reconstruction artifacts due to progressive alignment of non-homologous sites?
  (consider guide trees and gaps)

What's in a tree?

Trees form molecular data are usually calculated as unrooted trees (at least they should be - if they are not this is usually a mistake).
To root a tree you either can assume a
molecular clock (substitutions occur at a constant rate, again this assumption is usually not warranted and needs to be tested),
or you can use an **outgroup** (i.e. something that you know forms the deepest branch).
For example, to root a phylogeny of birds, you could use the homologous characters from a reptile as outgroup; to find the root in a tree depicting the relations between different human mitochondria, you could use the mitochondria from chimpanzees or from Neanderthals as an outgroup; to root a phylogeny of alpha hemoglobins you could use a beta hemoglobin sequence, or a myoglobin sequence as outgroup, to root the tree of life, you could use ancient paralogs, such as in the ATP synthases.
or you can use polarized characters.

Trees have a branching pattern (also called the **topology**), and **branch lengths**.

Often the branch lengths are ignored in depicting trees (these trees often are referred to as cladograms - note that cladograms should be considered rooted).
You can swap branches attached to a node, and in an unrooted you can depict the tree as rooted in any branch you like without changing the tree.

Terminology for Trees

- (Leaves, OTUs), (Branches, splits, bipartitions), Nodes
- In a rooted tree: clades (for unrooted trees sometimes the term clann is used)
- Mono-, Para-, polyphyletic groups, cladists and a natural taxonomy

The term cladogram refers to a strictly bifurcating diagram, where each clade is defined by a common ancestor that only gives rise to members of this clade. I.e., a clade is monophyletic (derived from one ancestor) as opposed to polyphyletic (derived from many ancestors). (Note: you do need to know where the root is!)

A clade is recognized and defined by
shared derived characters (= synapomorphies).
Shared primitive characters (= symplesiomorphies, alternative spelling is symplesiomorphies) do not define a clade, but a paraphyletic group.
Characters that evolved twice independently are known as Homoplasies. They define polyphyletic groups (see in class example drawing ala Hennig).

To use these any terms ((aside from homoplasy) you need to have polarized characters; for most molecular characters you don't know which state is primitive and which is derived (exceptions:....).

See the wikipedia entries for Willi Hennig , cladistics, and monophyly – see holophyly and the link to Ashlock's paper for a divergent view.

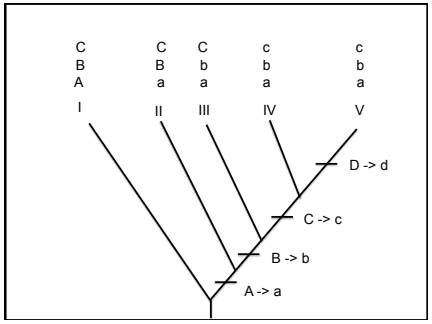
Terminology

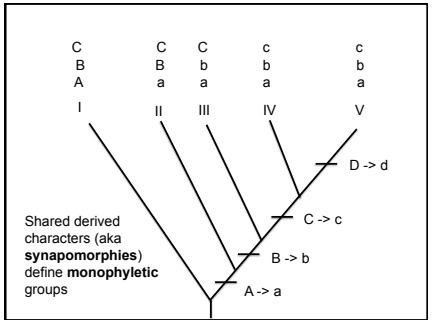
Related terms:
autapomorphy = a derived character that is only present in one group; an autapomorphic character does not tell us anything about the relationship of the group that has this character ot other groups.

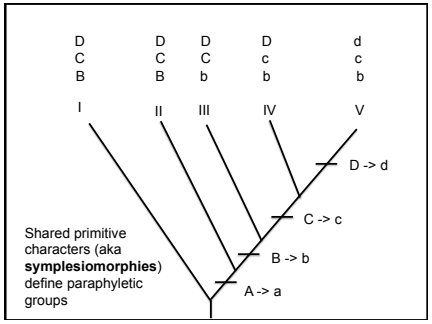
homoplasy = a derived character that was derived twice independently (convergent evolution). *Note that the characters in question might still be homologous (e.g. a position in a sequence alignment, frontlimbs turned into wings in birds and bats).*

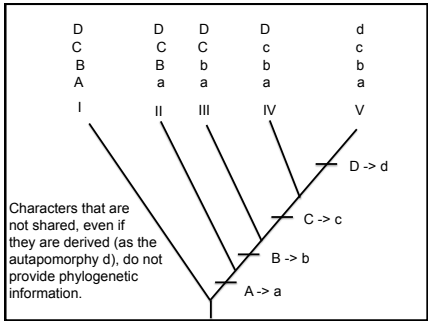
paraphyletic = a taxonomic group that is defined by a common ancestor, however, the common ancestor of this group also has decendants that do not belong to this taxonomic group. Many systematists despise paraphyletic groups (and consider them to be polyphyletic). Examples for paraphyletic groups are reptiles and protists. Many consider the archaea to be paraphyletic as well.

holophyletic = same as above, but the common ancestor gave rise only to members of the group.

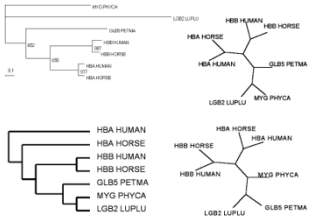








Test:Which of these trees is different?



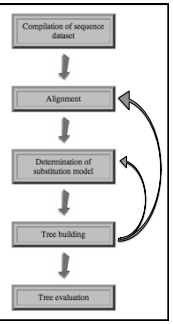
More tests here

Steps in phylogenetic analysis

Phylogenetic analysis is the inference of evolutionary relationships between genes or organisms.
Phylogenetics tries to answer the question "How did groups of organisms come into existence?"

Those relationships are usually represented by tree-like diagrams.

Note: the equation of biological evolution with a tree like process has limited validity at best.



Phylogenetic reconstruction - **How**

Distance analyses

calculate pairwise distances
(different distance measures, correction for multiple hits, correction for codon bias)

make distance matrix (table of pairwise corrected distances)

calculate tree from distance matrix

i) using optimality criterion
(e.g.: smallest error between distance matrix and distances in tree, or use
ii) algorithmic approaches (UPGMA or neighbor joining)
See wikipedia entry on Neighbor joining for a good illustration on how NJ works. See here for a worked example of UPGMA

Phylogenetic reconstruction - **How**

Parsimony analyses (see wikipedia entry for Parsimony)
find that tree that explains sequence data with minimum number of substitutions
(tree includes hypothesis of sequence at each of the nodes)

Maximum Likelihood analyses
given a model for sequence evolution, find the tree that has the highest probability under this model.
This approach can also be used to successively refine the model.

Bayesian statistics
use ML analyses to calculate posterior probabilities for trees, clades and evolutionary parameters. Especially MCMC approaches have become very popular in the last year, because they allow to estimate evolutionary parameters (e.g., which site in a virus protein is under positive selection), without assuming that one actually knows the "true" phylogeny.

Else:
spectral analyses, like evolutionary parsimony, look only at patterns of substitutions,

Another way to categorize methods of phylogenetic reconstruction is to ask if they are using

an optimality criterion (e.g.: smallest error between distance matrix and distances in tree, least number of steps, highest probability), or

algorithmic approaches (UPGMA or neighbor joining)

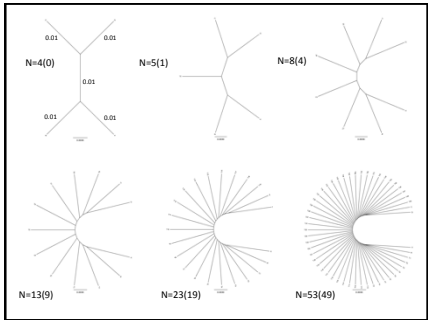
Packages and programs available: PHYLIP, phyml, MrBayes, Tree-Puzzle, PAUP*, clustalw, raxml, PhyloGenie, HyPhy

Bootstrap ?

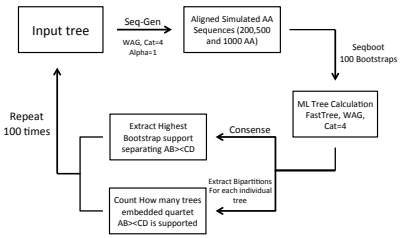
- See here and here
- **Non-parametric bootstrap**: bootstrap samples are generated through sampling with replacement
- **Parametric bootstrap** (not frequently used, should be used more often)
  samples are created through simulation of sequence evolution (with parameters estimated through ml). Good for testing hypotheses! (Examples?)

Boostrap Support Values for Embedded Quartets vs. Bipartitions:

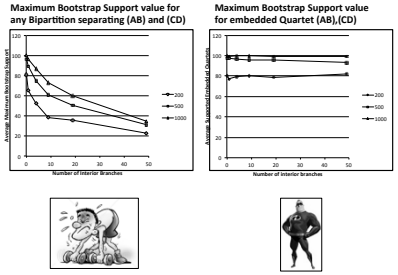
Performance evaluation using sequence simulations and phylogenetic reconstructions

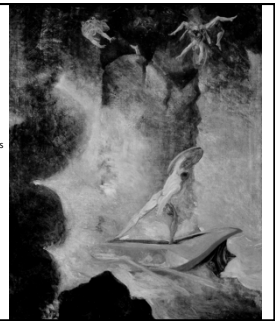


Methodology :



Results :

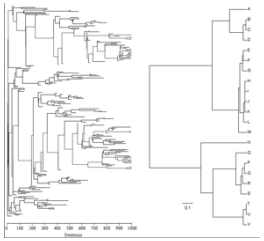




Johann Heinrich Füssli
Odysseus vor Scilla und Charybdis

From: http://en.wikipedia.org/wiki/File:Johann_Heinrich_F%C3%BCssli_054.jpg

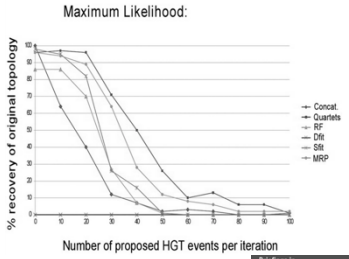
Genome simulation with EvolSimulator.



Lapierre P et al. Brief Bioinform 2012;bib.bbs050

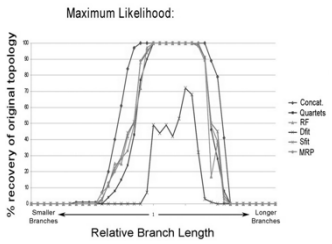
Briefings in Bioinformatics

The impact of HGT on phylogenomic reconstruction methods



Briefings in Bioinformatics

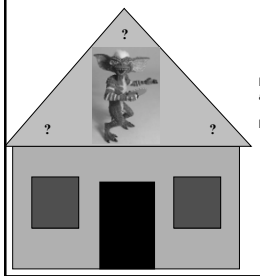
Percent recovery of original tree topology with increased and decreased substitution rates.



Lapierre P et al. Brief Bioinform 2012;bib.bbs050

Briefings in Bioinformatics

Elliot Sober's Gremlins



Observation: Loud noise in the attic

Hypothesis: *gremlins in the attic playing bowling*

Likelihood = *P(noise|gremlins in the attic)*

P(gremlins in the attic|noise)

Bayes' Theorem

Reverend Thomas Bayes (1702-1761)

Likelihood

describes how well the model predicts the data

$$P(model|data, I) = P(model, I)\,\frac{P(data|model, I)}{P(data,I)}$$

Posterior Probability

represents the degree to which we believe a given **model** accurately describes the situation given the available **data** and all of our prior information **I**

Prior Probability

describes the degree to which we believe the model accurately describes reality based on all of our prior information.

Normalizing constant

Likelihood estimates do not take prior information into consideration:

e.g., if the result of three coin tosses is 3 times head, then the likelihood estimate for the frequency of having a head is 1 (3 out of 3 events) and the estimate for the frequency of having a tail is zero.

$P(A,B) = P(A,B)$ The probability that both events (A and B) occur

$P(A|B) * P(B) = P(B|A) * P(A)$ Both sides expressed as conditional probability

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

If A is the model and B is the data, then
P(B|A) is the likelihood of model A
P(A|B) is the posterior probability of the model given the data.
P(A) is the considered the prior probability of the model.
P(B) often is treated as a normalizing constant.

Alternative Approaches to Estimate Posterior Probabilities

Bayesian Posterior Probability Mapping with MrBayes
(Huelsenbeck and Ronquist, 2001)

Problem:

Strimmer's formula $p_i = \frac{L_i}{L_1+L_2+L_3}$ only considers 3 trees (those that maximize the likelihood for the three topologies)

Solution:

Exploration of the tree space by sampling trees using a biased random walk (implemented in MrBayes program)

Trees with higher likelihoods will be sampled more often

$p_i = \frac{N_i}{N_{total}}$,where N_i - number of sampled trees of topology *i*, *i*=1,2,3

N_{total} – total number of sampled trees (has to be large)

Illustration of a biased random walk

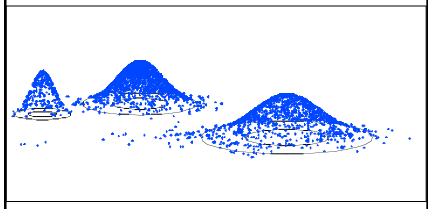


Figure generated using MCRobot program (Paul Lewis, 2001)

More reading on Phylogenetics is at

http://en.wikipedia.org/wiki/Computational_phylogenetics

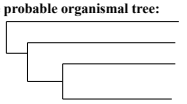
Why could a gene tree be different from the species tree?

- Lack of resolution
- Lineage sorting
- Gene duplications/gene loss (paralogs/ orthologs)
- Gene transfer
- Systematic artifacts (e.g., compositional bias and long branch attraction)

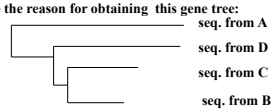
Trees – what might they mean?

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

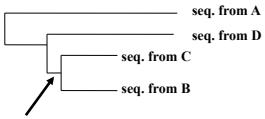
If this is the probable organismal tree:



what could be the reason for obtaining this gene tree:

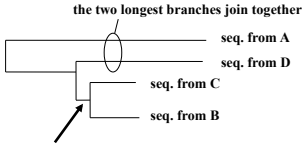


lack of resolution



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

long branch attraction artifact

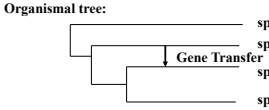


e.g., 100% bootstrap support for bipartition (AD)(CB)

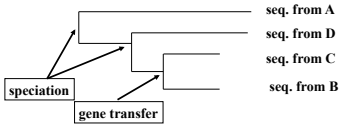
What could you do to investigate if this is a possible explanation?
use only slow positions,
use an algorithm that corrects for ASRV

Gene transfer

Organismal tree:

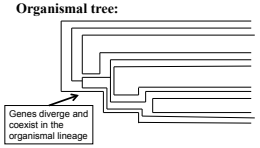


molecular tree:

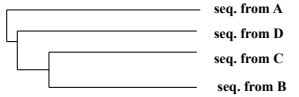


Lineage Sorting

Organismal tree:

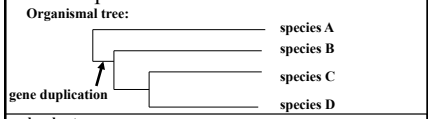


molecular tree:

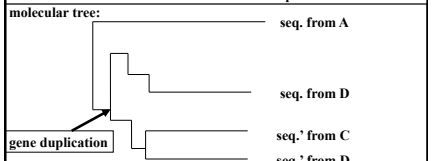


Gene duplication

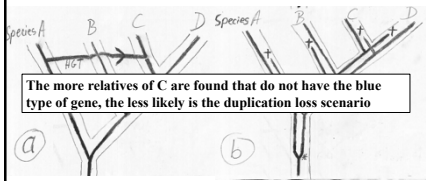
Organismal tree:



molecular tree:



Gene duplication and gene transfer are equivalent explanations.



Horizontal or lateral Gene

Ancient duplication followed by gene loss

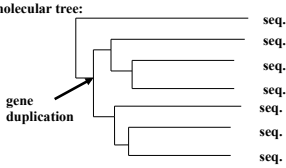
Note that scenario B involves many more individual events than A

1 HGT with orthologous replacement

1 gene duplication followed by 4 independent gene loss events

Function, ortho- and paralogy

molecular tree:



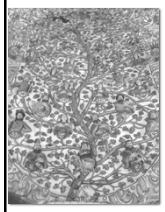
The presence of the duplication is a taxonomic character (shared derived character in species B C D).
The phylogeny suggests that seq' and seq have similar function, and that this function was important in the evolution of the clade BCD.
seq' in B and seq'in C and D are orthologs and probably have the 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)

Note:

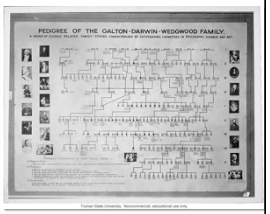
We likely will not cover the following slides on Phylip, phyml, etc.

Note the assignment for next week on the last slide.

Family Trees, Genealogies, Pedigrees



Genealogy (Church Ceiling, Santo Domingo, Oaxaca)



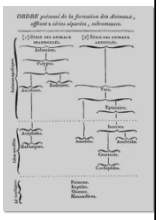
Charles Darwin's Family "Tree"

Russian naturalist **Peter Simon Pallas** (1766): "But the system of organic bodies is best of all represented by an **image of a tree** which immediately from the root would lead forth out of the most simple plants and animals a double, variously contiguous animal and vegetable trunk; the first of which **would proceed from** mollusks to fishes, with a large side branch of insects sent out between these, hence to amphibians and at the farthest tip it would sustain the quadrupeds, but below the quadrupeds it would put forth birds as an equally large side branch."
(translation from Latin by E. N. Genovese, from *J. David Archibald: "Edward Hitchcock's Pre-Darwinian (1840) 'Tree of Life'" Journal of the History of Biology (2009) 42:561–592*)

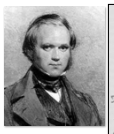
Tree of animal life, Depiction of the tree proposed by Peter Simon Pallas, from Carl Edward von Eichwald's ***Zoologia specialis* (1829)**. (From: *Mark A. Ragan: "Trees and networks before and after Darwin", Biol Direct. 2009 Nov 16;4:43.*)

Jean-Baptiste Lamarck

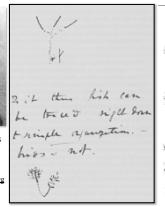
Tree from 1809



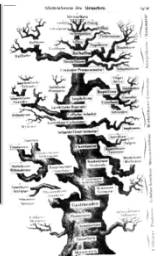
Tree-like classification from 1815



Charles Darwin, 1830s (by G. Richmond from Origins, Richard Leakey and Roger Lewin; source commons.wikimedia.org)

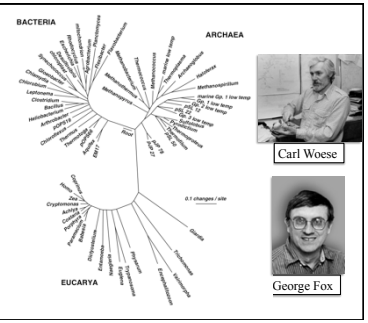


Sketch from Charles Darwin's (1809-1882) notebook (1837)



Lebensbaum from Ernst Haeckel, 1874

Small Subunit Ribosomal RNA based Tree of Life.



Carl Woese

George Fox

The Ribosomal "Tree of Life"

Rooting the Ribosomal Tree of Life using an Echo from the Early Expansion of the Genetic Code

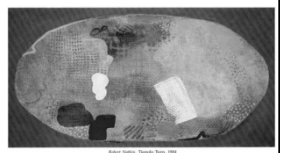
(Fournier and Gogarten, MBE 2010)

Fig. 3. Composite amino acid usage bias across universal ribosomal tree. Bias values reflect the geometric mean distances (normalized as SD) between observed and expected amino acid usages. Increased bias across the archaeal domain is largely due to widespread thermophily, halophily, and nucleotide composition bias. Aside from branches associated with haloarchaea, the branch leading to the bacteria (the bacterial root) contains the greatest bias.

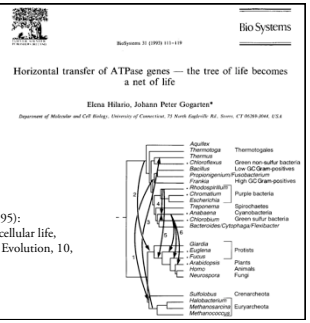
Ochiai, K., Yamanaka, T., Kimura, K., and Sawada, O. (1959)
Inheritance of drug resistance (and its transfer) between *Shigella* strains and Between *Shigella* and *E.coli* strains.
Hihon Iji Shimpor 1861: 34 (in Japanese)

Gray GS, Fitch WM (1983):
Evolution of antibiotic resistance genes: the DNA sequence of a kanamycin resistance gene from *Staphylococcus aureus*.
Mol Biol Evol 1983, 1(1):57-66.

Sorin Sonea (1988):
The global organism:
A new view of bacteria.
The Sciences, 28:38-45.



1993:


Bio Systems

Biosystems 31 (1993) 111–119

Horizontal transfer of ATPase genes — the tree of life becomes a net of life

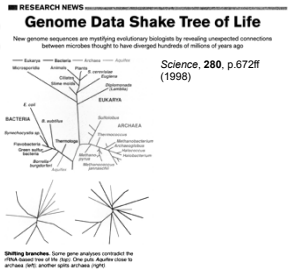
Elena Hilario, Johann Peter Gogarten*

Department of Molecular and Cell Biology, University of Connecticut, 75 North Eagleville Rd., Storrs, CT 06268-3044, USA

1995:

GOGARTEN, J.P. (1995):
The early evolution of cellular life,
Trends in Ecology and Evolution, 10,
147-151.

1998:

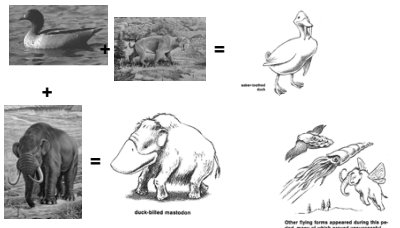

RESEARCH NEWS

Genome Data Shake Tree of Life

New genome sequences are mystifying evolutionary biologists by revealing unexpected connections between microbes thought to have diverged hundreds of millions of years ago

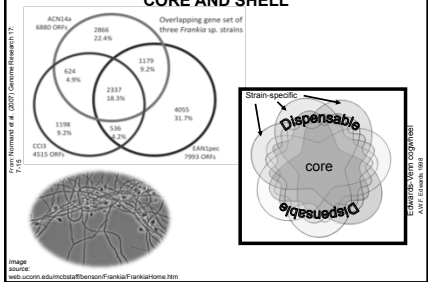
Science, **280**, p.672ff
(1998)

Evolutionary processes analogous to the ones proposed to occur in the microbial world

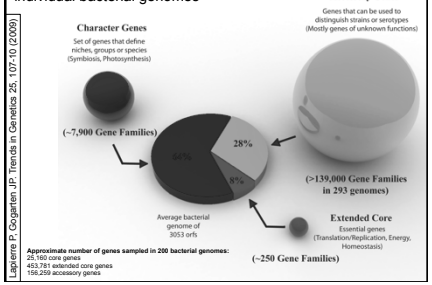


Cartoons from *Science Made Stupid*, T. Weller, 1986.
from http://www.besse.at/sms/ *Slide kindly provided by Kenneth Noll.*

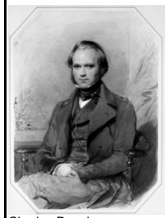
GENOMES OF CLOSELY RELATED ORGANISMS: CORE AND SHELL



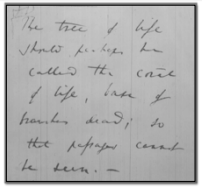
Frequencies of gene types in individual bacterial genomes



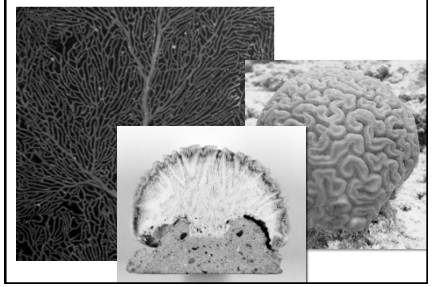
Tree, Web, or Coral of Life?


Charles Darwin
painted by George Richmond in the late 1830

"The tree of life should perhaps be called the coral of life, base of branches dead"

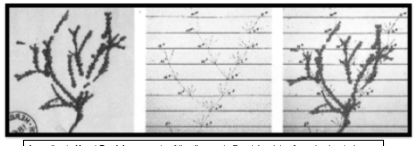

Page B26 from Charles Darwin's (1809-1882) notebook (1837/38)

Which Type of Coral ?



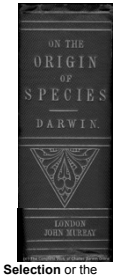
Darwin's Coral a Red Algae?

(Bossea orbignyana)


According to **Horst Bredekamp**, parts of the diagram in Darwin's origin of species (center) may reflect the branching properties of a specimen Darwin collected himself.

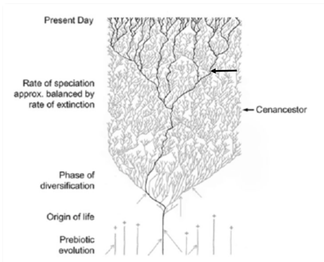
From Florian Maderspacher:
"The captivating coral--the origins of early evolutionary imagery."
Current Biology 16: R476-8 2006

"As buds give rise by growth to fresh buds, and these, if vigorous, branch out and overtop on all sides many a feebler branch, so by generation I believe it has been with the great Tree of Life, which fills with its dead and broken branches the crust of the earth, and covers the surface with its ever-branching and beautiful ramifications."



Charles Darwin in
"On the **Origin of Species** by **Means of Natural Selection** or the **Preservation of Favoured Races in the Struggle for Life**",
p 162 ff, London, John Murray, 1859

The Coral of Life (Darwin)



ZHAXYBAYEVA and GOGARTEN (2004): Cladogenesis, Coalescence and the Evolution of the Three Domains of Life. Trends in Genetics 20 (4): 182- 187

THE POTATO OF LIFE

Successive independent crystallization of the three domains of life from a population of pre-cells

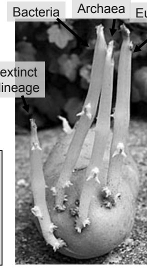


photo from www.sonnentaler.net

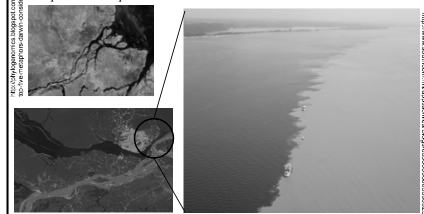
Glansdorff N, Xu Y, Labedan B: *The Last Universal Common Ancestor: emergence, constitution and genetic legacy of an elusive forerunner.* *Biol Direct* 2008, **3**:29.

Kandler O (1994) *The early diversification of life.* In *Early Life on Earth, Nobel Symposium*, vol. 84, pp. 152-509: Columbia Univ. Press.

Woese CR (2002) *On the evolution of cells. PNAS* 99: 8742-7

The Watershed of Life

Jonathan Eisen in his Tree of Life Blog on
"Top Five Metaphors Charles Darwin Considered"



Meeting of the meeting of Solimões and Negro rivers at Origin of the Ama

Can one reconstruct microbial phylogeny?

PHYLOGENY:
from Greek phylon, race or class, and -geneia, born.
"*the origin and evolution of a set of organisms, usually of a species*" (Wikipedia);

Phylogeny does not necessarily occur in a tree-like process!

Rope as a metaphor to describe an organismal lineage

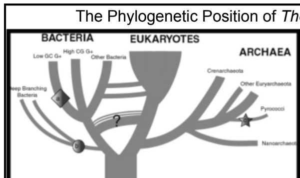
(Graham Cairns-Smith, Gary Olsen)

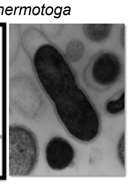
Individual fibers = genes that travel for some time in a lineage.

While no individual fiber present at the beginning might be present at the end, the rope (or the organismal lineage) nevertheless has continuity.

However, the genome as a whole will acquire the character of the incoming genes (the rope turns solidly red over time).

The Phylogenetic Position of *Thermotoga*





(a) Most genome based approaches
(b) according to 16S & other ribosomal markers
(c) according to phylogenetically discordant genes

Zhaxybayeva O, Fournier G, Lapierre P, Swithers K, Nesbø C, Gogarten JP, Doolittle WF, Noll KM: *Phylogenetic position of Thermotogales revisited: conflicting signals and evidence for thermophilic ancestral proteome*, PNAS 3/24/2009

Gophna U, Doolittle WF & Charlebois RL: *Weighted genome trees: refinements and applications. J. Bacteriol.* (2005)

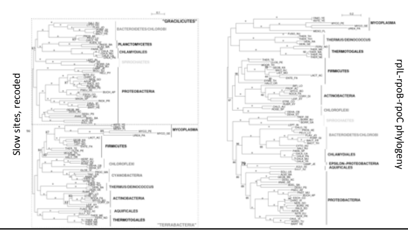
Gogarten JP & Townsend JP: Horizontal gene transfer, genome innovation, and evolution Nature Reviews in Microbiology 3(9) 679-687(2005)

Research article Highly accessed Open Access

Accounting for horizontal gene transfers explains conflicting hypotheses regarding the position of aquificales in the phylogeny of Bacteria

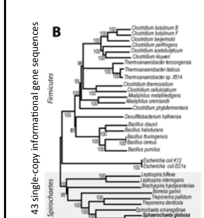
Bastien Boussau*, **Laurent Guéguen** and **Manolo Gouy**

BMC Evolutionary Biology 2008, **8**:272 doi:10.1186/1471-2148-8-272

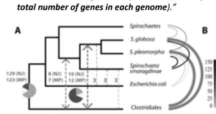


The Chimeric Genome of *Sphaerochaeta*: Nonspiral Spirochetes That Break with the Prevalent Dogma in Spirochete Biology

A. Caro-Quintero,[a] K. M. Ritalahti,[b,c] K. D. Cusick,[b] F. E. Löffler,[b,c,d] and K. T. Konstantinidis[a,e]



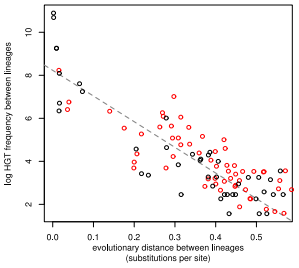
*"The Sphaerochaeta genomes are highly enriched in fermentation and carbohydrate metabolism genes relative to other spirochetes, indicating a fermentative lifestyle. Remarkably, most of the enriched genes appear to have been acquired from non spirochetes, particularly clostridia, in several massive horizontal gene transfer events (**>40% of the total number of genes in each genome**)."*



the values next to the arrows indicate the numbers of genes exchanged (out of a total of 178 genes examined).

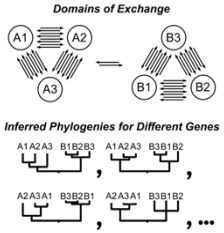
mBio

Gene transfer and recombination in haloarchaea



Analyses by David Williams in collaboration with Thane Papke- GBE in press

HGT can generate reproducible patterns in molecular phylogenies.



Gogarten, J.P., Doolittle, W.F., Lawrence, J.G., Mol. Biol. Evol. 2002 19:2226-2238.

MOLECULAR BIOLOGY and EVOLUTION

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

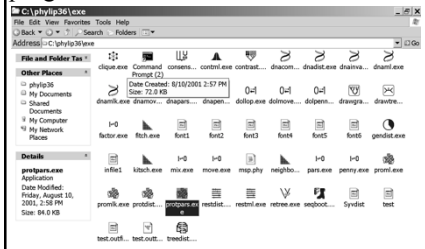
The easiest way to run PHYLIP programs is via a command line menu (similar to clustalw). The program is invoked through clicking on an icon, or by typing the program name at the command line.

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

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

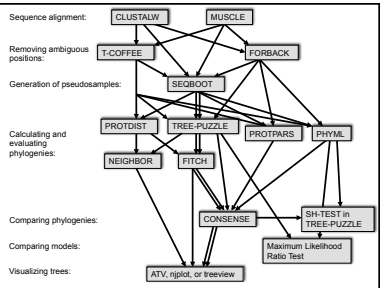
```
[gogarten@carrot gogarten]$ seqboot
seqboot: can't find input file "infile"
Please enter a new file name>
```

program folder



menu interface

example: seqboot and protpars on infile1



Phylip programs can be combined in many different ways with one another and with programs that use the same file formats.

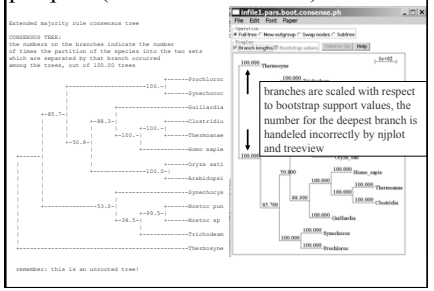
Example 1 Protpars

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!

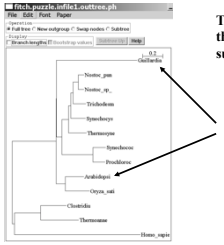
threshold parsimony,
gap symbols - versus ?
(in vi you could use `:%s/-/?/g` to replace all – ?)
outfile
outtree compare to distance matrix analysis

protpars (versus distance/FM)



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



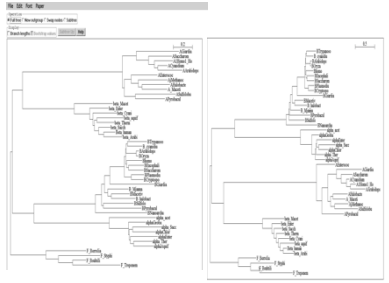
Tree is scaled with respect to the estimated number of substitutions.

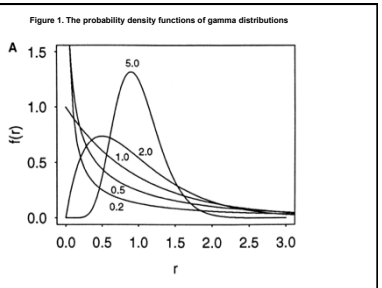
what might be the explanation for the red algae not grouping with the plants?

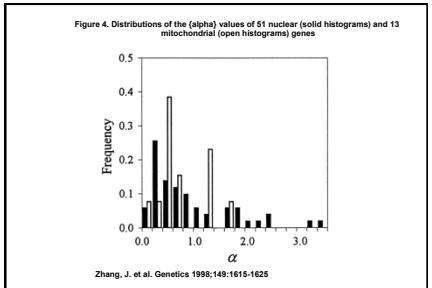
protdist

```
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
```

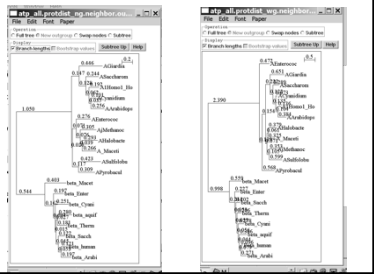
without and with correction for ASRV



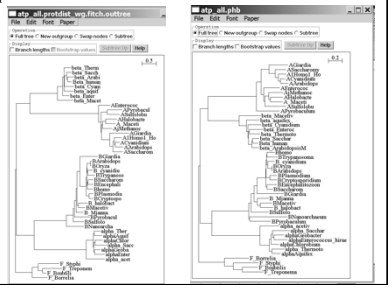




subtree with branch lengths
without and with correction for ASRV

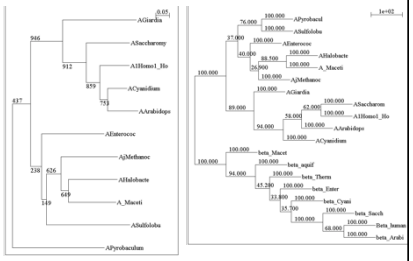


compare to trees with FITCH and clustalw – same dataset



bootstrap support ala clustal

protpars (gaps as ?)



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.

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, $d = 2(\log L_1 - \log L_2)$ 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.

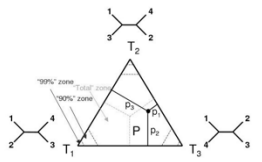
@@@ Puzzle is cool, don't leave home without it! **@@@**

TREE-PUZZLE – PROBLEMS/DRAWBACKS

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

ml mapping



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

ml mapping

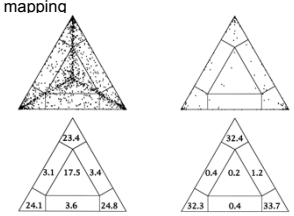
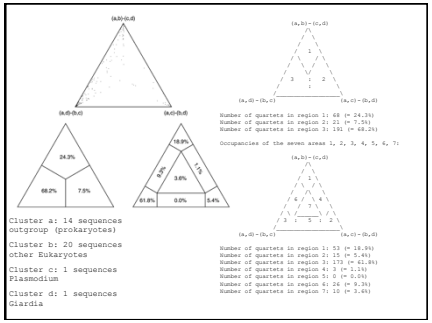
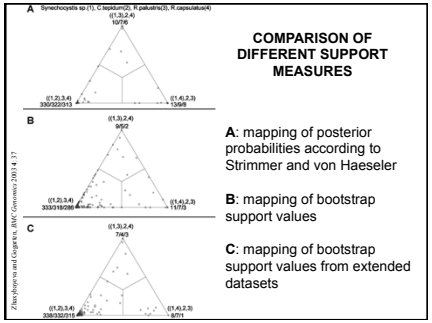


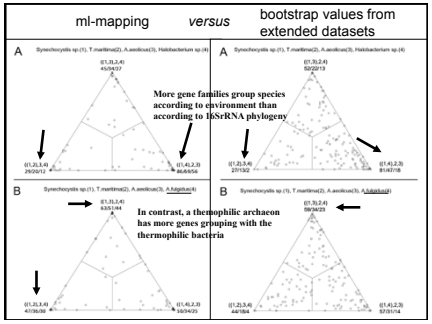
Figure 5. Likelihood-mapping analysis for two biological data sets. (*Upper*) The distribution patterns. (*Lower*) The occupancies (in percent) for the seven areas of attraction.

(*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







Assignments for next week Monday:

1. **Write a script that prints out a hash sorted on the keys in alphabetical order.**
2. **How can you remove an entry in a hash (key and value)?**
3. **Write a program that it uses hashes to calculates mono-, di-, tri-, and quartet-nucleotide frequencies in a genome.**
4. **Turn your project outline into a step by step to-do list / pseudocode**

assignments continued

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 clustalw2 and

- aligns the sequences within each file
- successively calculates profile alignments between all aligned sequences.