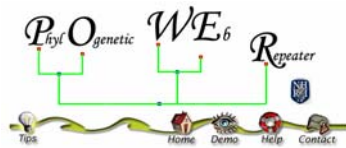


POWER: Phylogenetic WEb Repeater

- An integrated and user-optimized framework for biomolecular phylogenetic analysis



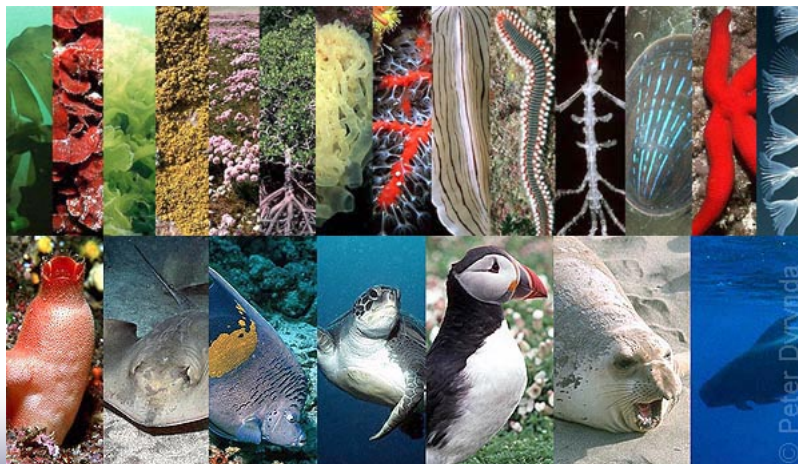
Speaker: Chieh -Hua Lin (林介華)

mammer@nhri.org.tw

GMBD Bioinformatics Core 推廣課程



What is phylogenetic ?



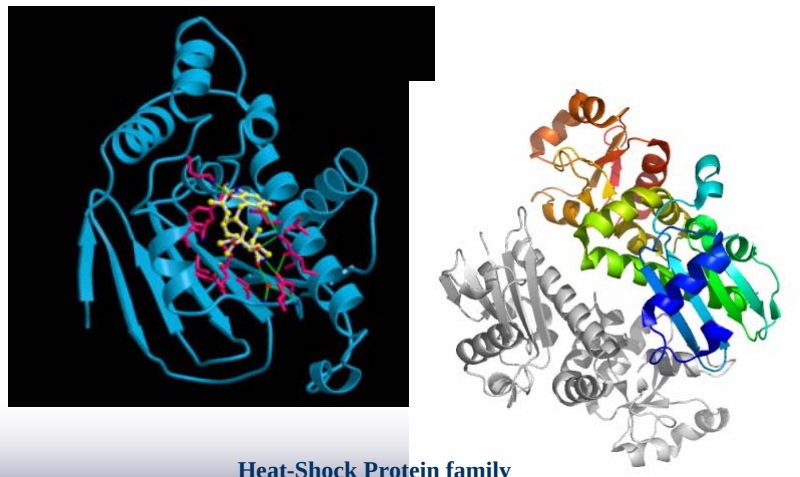


What is phylogenetic ?

- Phylogenetics is that field of biology which deals with **identifying and understanding the relationships between the many different kinds of life on earth**. This includes methods for collecting and analyzing data, as well as interpretations of those results as new biological information.



Molecular Phylogenetics



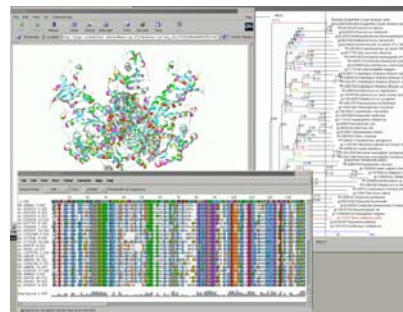
Heat-Shock Protein family



What is *Phylogenetic Analysis*



Classical phylogenetic analysis:
Morphology, Anatomy, Phenotype



Modern phylogenetic analysis:
DNA or Protein sequences



Phylogenetic Analysis

- Based on observable evidence to **estimate** or **infer** their evolution relationship
- Tree-thinking
 - Clustering orthologous genes
 - Inferring gene/protein functional diversification
 - Find the evolutionary history of genes/gene families or species



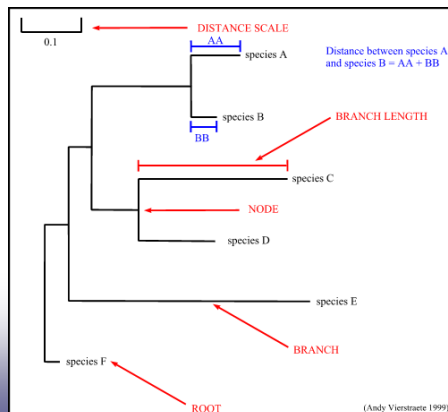
What is phylogenetic tree

- Phylogenetic tree can be used to represent the historical relationships of groups of organisms, not similarities
 - a graphical representation of the evolution relations among organisms or taxa. The taxa can be species, populations, individuals or genes.
- The challenge for biologist in phylogenetic analysis: programs
 - Why taxon? Why sequence?



Phylogenetic tree

- The tree is composed of nodes connected by branches.

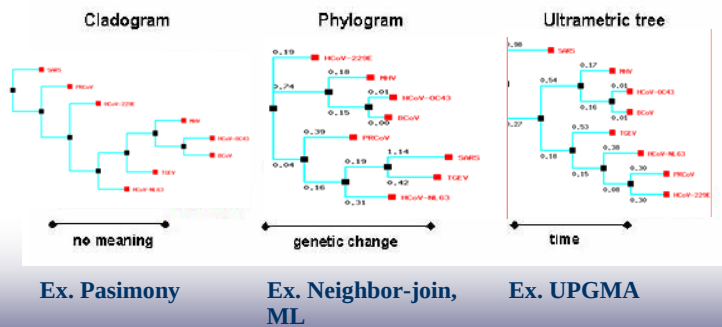


- **node** : a node represents a taxonomic unit.
 - Internal nodes
 - External nodes
- **branch (edge)**: defines the relationship between the taxa.
- **branch length** : often represents the number of changes that have occurred in that branch.
- **distance scale** : scale which represents the number of differences between sequences (e.g. 0.1 means 10 % differences between two sequences)
- **root** : is the common ancestor of all taxa.



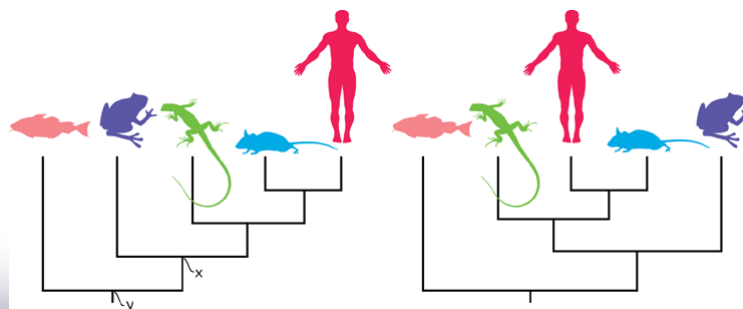
Phylogenetic tree cont.

- Branch: define relationship between nodes
 - Branch length: longer branch length, more sequence changes



Trees only represent the order of branching

- Same tree in a different style
 - Both trees have identical topologies, with some of the internal nodes rotated.



(David A. Baum et al., Science 11 November 2005:Vol. 310. no. 5750, pp. 979 – 980)



Phylogenetic Analysis Steps

- Sequences alignment
- Build substitution model
- Construct phylogenetic tree
- Evaluate phylogenetic tree



Phylogenetic Analysis Tool

~ <http://evolution.genetics.washington.edu/phylip/software.html>

Methods	By computer	Cross-referenced	Data types	New programs
Phylogeny Programs				
Changes	Waiting list	Other lists	Old programs	Not listed

Here are 265 of the phylogeny packages, and 31 free servers, that I know about. It is an attempt to be completely comprehensive. I have not made any attempt to exclude programs that do not meet some standard of quality or importance. Updates to these pages are made sporadically. (Note -- there were no updates of these pages from early 2001 until August, 2003. Updating has now resumed and I am slowly catching up. [Here](#) is a "waiting list" of new programs.) Many of the programs in these pages are available on the web, and some of them are also available from [ftp server machines](#).



Phylip documentation

<http://evolution.genetics.washington.edu/phylip/phylip.html>

v3.6

PHYLIP programs and documentation

PHYLIP, the PHYLogeny Inference Package, consists of 35 programs. There are documentation files for each program, in the form of web pages in HTML 3.2. There are also documentation web pages for each group of programs, and a main documentation file that is the basic introduction to the package. Before running any of the programs you should [read it](#).

Below you will find a list of the programs and the documentation files. The names of the documentation files are highlighted as links that will take you to those documentation files.

Introduction to PHYLIP

[main documentation file](#)

Molecular sequence methods

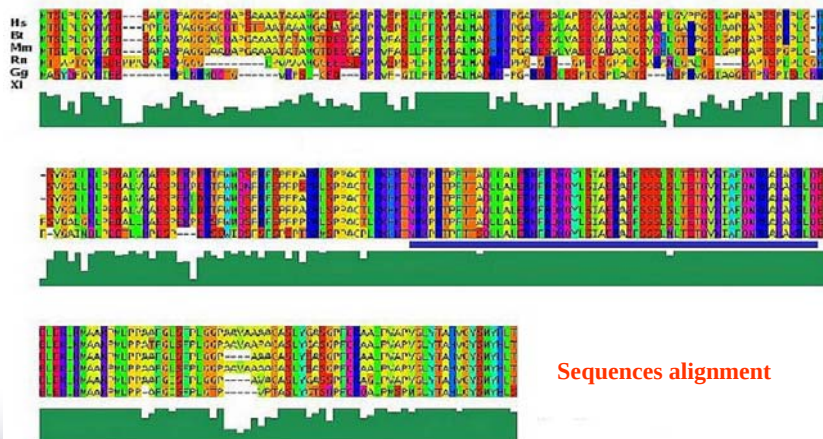
	molecular sequence programs documentation file
protpars	protein parsimony documentation file
dnapars	DNA sequence parsimony documentation file
dnapenny	DNA parsimony branch and bound documentation file
dnamove	interactive DNA parsimony documentation file
dnacomp	DNA compatibility documentation file
dnaml	DNA maximum likelihood documentation file
dnamlk	DNA maximum likelihood with clock documentation file
proml	Protein sequence maximum likelihood documentation file



Phylogenetic Analysis Tool

Parameters Collection
Pages (WEB)

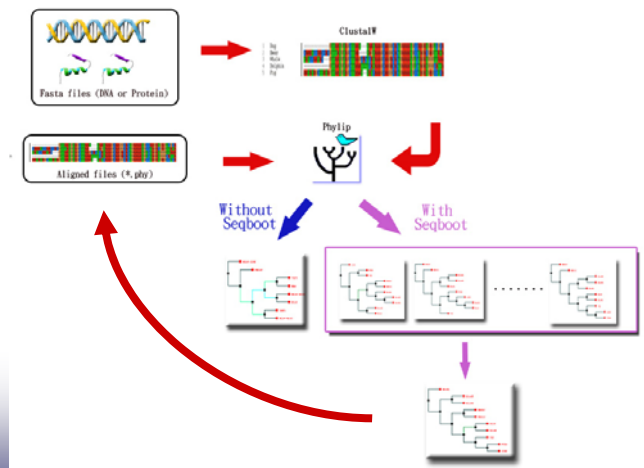
Result Display
Pages (WEB)



neighbor : web applications : Program : Database : Output file
 protis : web site : program : database : output file



Phylogenetic Web Repeater (POWER)



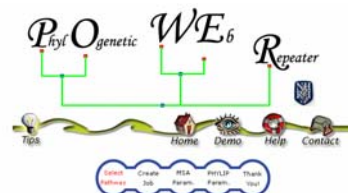
POWER: Phylogenetic Web Repeater - An integrated and user-optimized framework for biomolecular phylogenetic analysis

<http://power.nhri.org.tw>



POWER: Phylogenetic Web Repeater - an integrated and user-optimized framework for biomolecular phylogenetic analysis
Cheng-Yen Lin¹, Jue-Kai Lin, Chih-Hsin Lin, Li-Mei Lin, Wen-Yueh Hsu, Jue-Kuei Chen² and Chien-Hsien Shiao¹

Nucl. Acids Res. 2005 33: W553-W556



The Phylogenetic Web Repeater (POWER) allows users performing phylogenetic analysis with molecular data by most programs of PHYLIP package repeatedly. POWER provide two pipelines to process the analysis. One of them includes multiple sequence alignment (MSA) at the beginning of the pipeline whereas the other begin phylogenetic analysis with aligned sequence.

Please start your analysis by selecting the pipeline and the data type:

Pipeline	☒ MSA + Phylogenetic Analysis(Input the FASTA format)
	☐ Phylogenetic Analysis Only(Input the PHYLIP format)
Sequence Type	☒ DNA
	☐ Protein

- Sequences alignment
 - ClustalW, Jalview
- Build substitution model
 - Parameters determined
- Construct phylogenetic tree
 - Quantitatively determine the relationship between input sequences
- Evaluate phylogenetic tree
 - Get the same results with different methods

- The resource of sequence database: NCBI, EMBL, DDBJ
- Data format





The resource of phylogenetics

Tree of life (ToL) project : since 1995
~<http://tolweb.org/tree/>



NCBI Entrez Taxonomy
~<http://www.ncbi.nlm.nih.gov/>



Input file

- FASTA format: start with a symbol “>”
- The names of each sequence are not more than 10 characters.

❑ I: [AF029248](#). Reports Mouse hepatitis v...[gi:2641127]

```
>gi|2641127|gb|AF029248.1| Mouse hepatitis virus strain MHV-A59 C12 mutant, complete genome
TATAAGAGTGATTGGCGTCCGTACGTACCTCTCAACTCTAAACTCTTGTAGTTTAAATCTAATCTAAA
CTTTATAAACCGGCACTTCCCTGCGGTGCCATGCCCGCGGGCTGGTCTTGTCTAGTGTGCTGACATTTGTAG
TTCCCTTGACTTTTCGTTCTCTGCCAGTGACGTGTCCATTCCGGCGCCAGCAGCCACCCATAGGTTGCATAA
TGGCAAGATGGGCAAAATACGGTCTCGGCTTCAAATGGGCCCCAGAATTTCCATGGATGCTTCCGAACGC
ATCGGAGAAGTTGGGTAACCTGAGAGGTGAGAGGAGGATGGGTTTGGCCCTCTGCTGCCGAAGAACCG
AAAATTTAAAGGAAAAAATTTGGTTAATCAGCTGAGGGTGAATTTGTAGCCGGCTTCCAGCTTTGGAATGCT
GTGTTTCAGTCTGCCATAATCCGTGATATTTTGTAGATGAGGATCCCCAGAAGTGGAGGGCTCAACTAT
GATGGCATTGCAAGTTCCGTTAGTCCGCTCTTGGTTAAGCCATCCAAAGCGCTTGTCTATTACAGGCTGGAAT
AATTTGGGTGTGCTTCCCAAAACAGCTGCCATGGGTTGTTCAAGCGCGTCTGCCTGTGTAAACACAGGG
AGTGTCTTGTGACGCCCAAGTGGCTTTTCACTTTTACGGTCCAAACCGATGGTGTATGCTGGGTAA
TGGCCGTTTATAGGCTGGTTCGTTCCAGTCAACGCCATACCGGAGTATGCGAAGCAGTGGTTGCAACCC
TGGTCCATCCTTCTCGTAAGGGTGGTAACAAAGGGTCTGTGACATCCGGCCACTTCCGCCCGCGCTGTTA
CCATGCCCTGTGTATGACTTTAATGTAGAGGATGCTTGTGAGGAGGTTCACTTTAACCAGAAAGGTAAGTA
CTCCTGCAAGGGTATGCTCTTCTTAAGGGCTATCGCGGTGTTAAGCCCATCCTGTTTGTGGACCAAGTAT
```

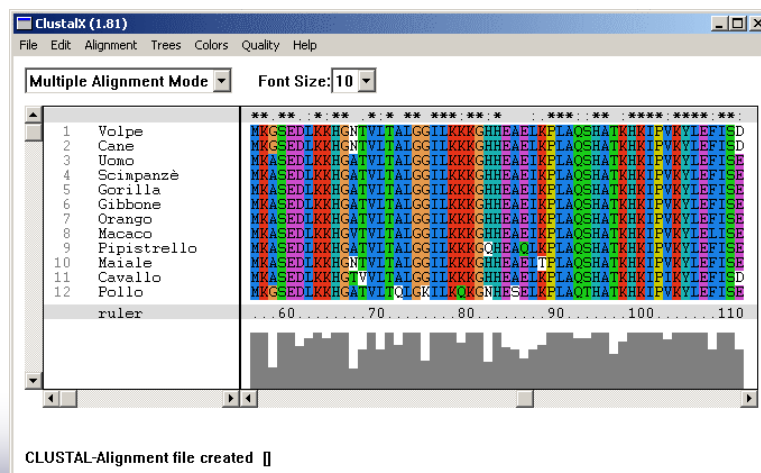


How to choose the sequences

- Homologs sequences: two taxa/sequences are descended from a common ancestor
 - Using BLAST program
 - Identical or Mixed sequences set
- Sample can answer the questions you are interested
- The representative samples
- The informative samples
- DNA or protein sequence: far related can use protein sequence



Alignment program-ClustalX (for Windows)



<http://bips.u-strasbg.fr/fr/Documentation/ClustalX/>



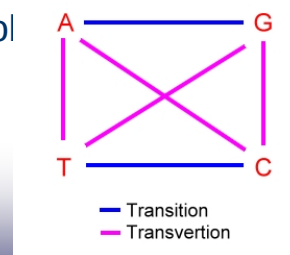
Phylogenetic Analysis Step 2

- Sequences alignment
 - ClustalW, Jalview
- Build substitution model
 - Parameters determined
- Construct phylogenetic tree
 - Quantitatively determine the relationship between input sequences
- Evaluate phylogenetic tree
 - Get the same results with different methods



Substitution model

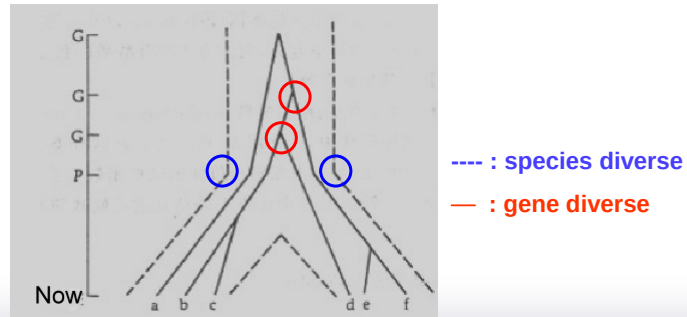
- Transition: purine to purine, pyrimidine to pyrimidine
- Transversion: purine to pyrimidine (and pyrimidine to purine)
- Base frequencies
- Among-site rate variation: Wol
 - Invariable, Variable sites
- Outgroup: to root a tree





Using Substitution model to infer evolution progress

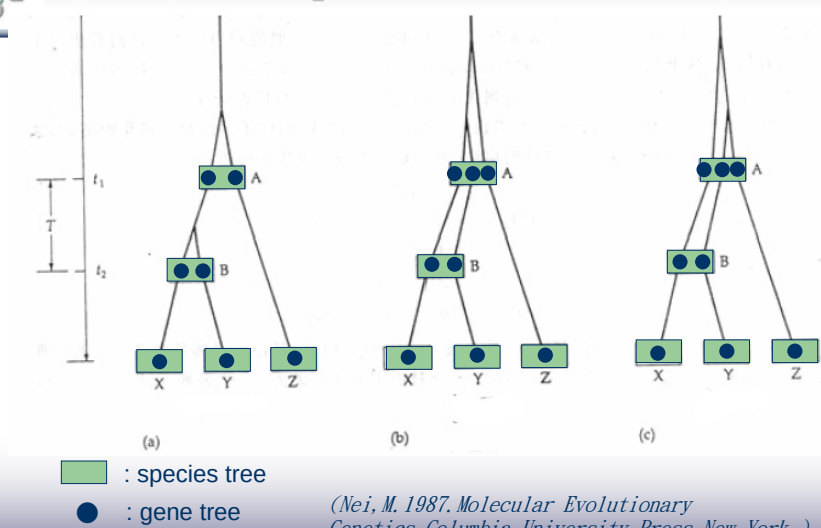
- Gene tree v.s Species tree : Gene tree is different from species tree



(Nei, M. 1987. *Molecular Evolutionary Genetics*. Columbia University Press, New York.)



Gene tree vs Species tree



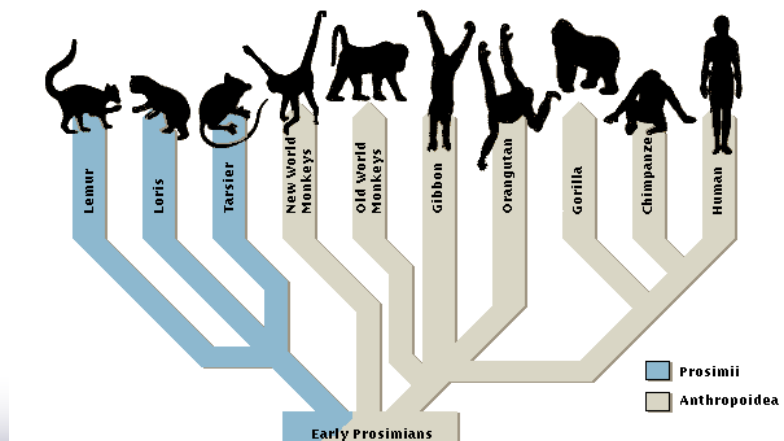


Chromosome and Genome evolution

- Molecular evolution
 - Mutation: change at level of genes and proteins
 - Genetic recombination
 - Natural selection
 - Reproductive isolation
- Accumulated genetic change: polymorphisms
 - Quasispecies: for example, virus genome
 - Speciation



Speciation event





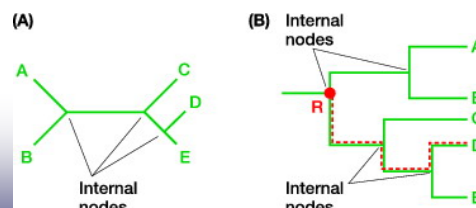
Substitution model

- Transition: purine to purine, pyrimidine to pyrimidine
- Transversion: purine to pyrimidine (and pyrimidine to purine)
- Base frequencies
- Among-site rate variation
 - Invariable, Variable sites
- Outgroup: to root a tree



Tree topology

- Rooted: common ancestor
- Unrooted tree: the phylogenetic relationships only (Fig. A)
- Rooted tree: provide the direction of evolution (Fig. B)
 - More correct, but less informative



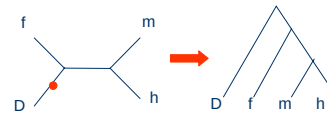
External nodes: A~E
Root: R



To root a tree

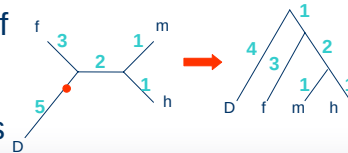
- Outgroup rooting

- One or more taxon that are distantly related to all of the taxa being considered



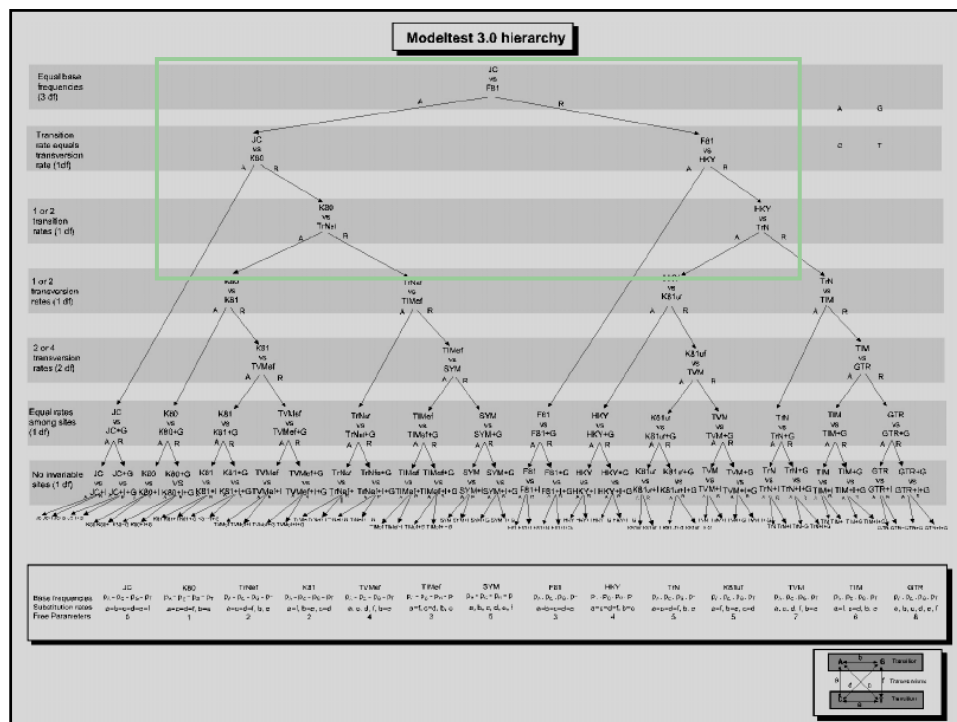
- Midpoint rooting

- Put the root at the middle of the longest path (sum of branches between any two operational taxonomic units (OTUs))



How to choose the suitable model : Model selection

- Which method would be better for our data matrix?
- Using a variety of method to analysis the data matrix
- Software: Modeltest 3.7 
 - <http://darwin.uvigo.es/software/modeltest.html>



Phylogenetic Analysis Step 3

- Sequences alignment
 - ClustalW, Jalview
- Build substitution model
 - Parameters determined
- Construct phylogenetic tree
 - Quantitatively determine the relationship between input sequences
- Evaluate phylogenetic tree
 - Get the same results with different methods



Construct tree methods

- Character state method:
 - Maximum parsimony
- Distance method
 - Neighbor-joining (variable evolution rate) and UPGMA method :
 - Fitch-Margoliash method
- Maximum likelihood methods
 - determinate evolution model first, then construct system trees



Distance matrix

30

Sequence 1	GTATAGGGGATATACTGAGAGCTATTTACA
Sequence 2	GTATTGGCGATATTCGAGACCTATTTACT
Sequence 3	GTATTGGCCATATTCCGAGACCTATTTACT
Sequence 4	GTATAGGCGATATACCGAGACCTATTTACT

Distance matrix	1	2	3	4
1	—	0.20	0.233	0.133
2	—	—	0.20	0.067
3	—	—	—	0.10
4	—	—	—	—



Phylogenetic Analysis Step 4

- Sequences alignment
 - ClustalW, Jalview
- Build substitution model
 - Parameters determined
- Construct phylogenetic tree
 - Quantitatively determine the relationship between input sequences
- Evaluate phylogenetic tree



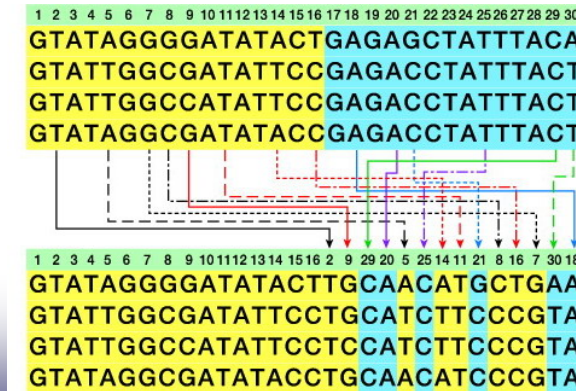
Evaluate reliability of tree

- Seqboot
 - Bootstrap
 - Jackknife
 - Permutate
- Get the same results with different methods



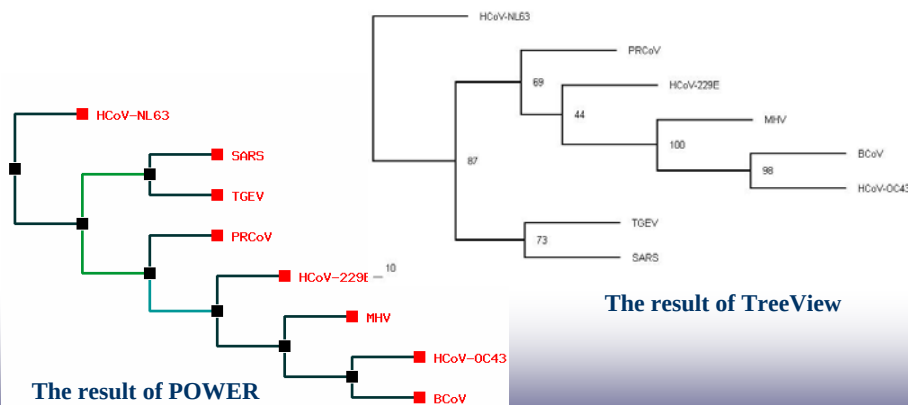
Bootstrapping

- Generate pseudo-sequence alignment randomly
- Compare with original result to get bootstrap value



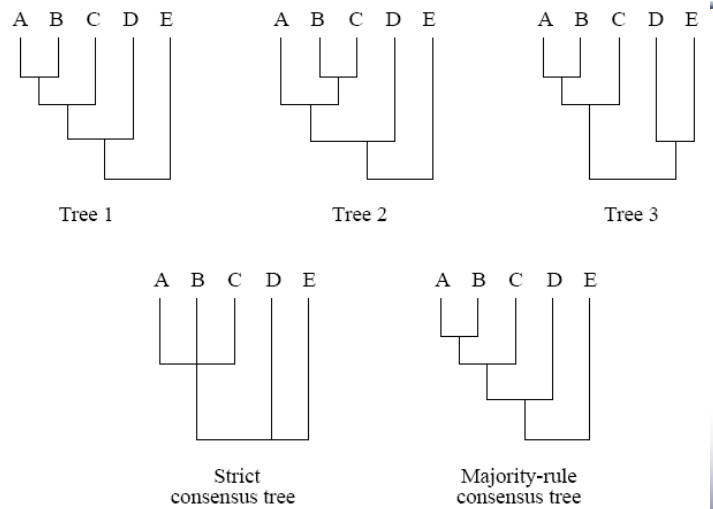
Construct trees with Bootstrapping

Reliability measurement: The probability that the nodes of a given clade are always nodes of that clade





Consensus trees



Reconstruct tree completely??

- Species evolution is a history, which can't be reconstructed fully
- All models are questionable but some are useful
- Narrow down difference between gene tree and species tree



POWER



- ClustalW
 - Step of multiple sequence alignment (**MSA**)
- Phylip
 - Character state method
 - Maximum parsimony
 - Distance method
 - Neighbor-joining and UPGMA method
 - Fitch-Margoliash method
 - Maximum likelihood methods
 - with molecular clock



Pylogenetic analysis for DNA

Character state methods

- [Maximum parsimony\(heuristic search\) method](#)
- [Maximum parsimony\(branch and bound search\) method](#)
- [Compatibility method](#)

Distance methods

- [Neighbor-joining and UPGMA method](#)
- [Fitch-Margoliash and least squares method](#)
- [Fitch-Margoliash and least squares method with molecular clock](#)

Maximum likelihood methods

- [Maximum likelihood method](#)
- [Maximum likelihood method with molecular clock](#)

(Warning: these methods cost much memory. So, if the number of bases of your sequences data exceeds 40000, the job may fail.)



Pylogenetic analysis for Protein

Character state methods

- [Maximum parsimony\(heuristic search\) method](#)

Distance methods

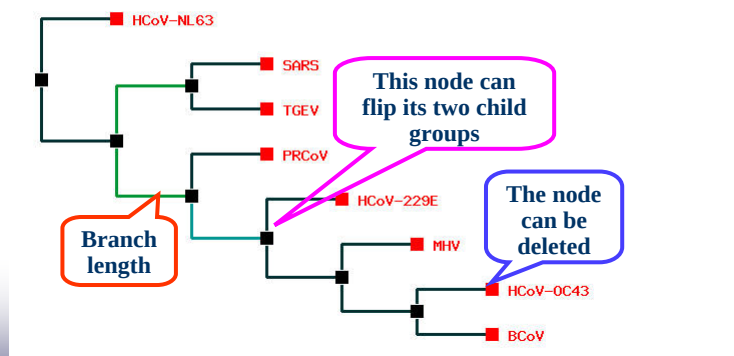
- [Neighbor-joining and UPGMA method](#)
- [Fitch-Margoliash and least squares method](#)
- [Fitch-Margoliash and least squares method with molecular clock](#)



The Result of POWER

Color of line indicates how many times the group which consists of the species to the right of the fork occurred.

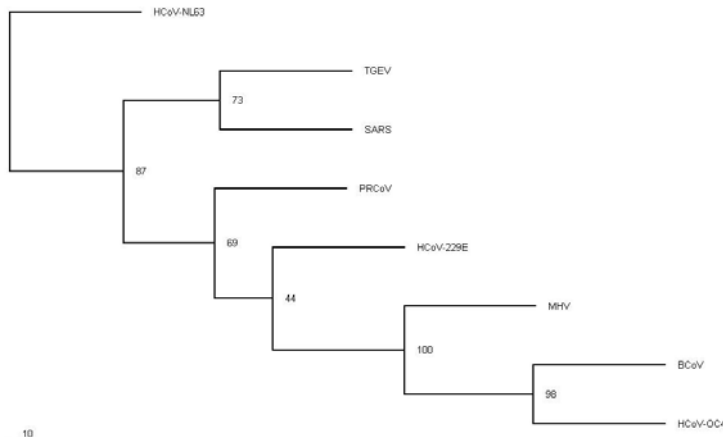
—: <=20 —: >20 & <=40 —: >40 & <=60 —: >60 & <=80 —: >80





View the result by TreeView

(((((BCoV:100.0,HCoV-OC43:100.0):98.0,MHV:100.0):100.0,HCoV-229E:100.0):44.0,PRCoV:100.0):69.0,(TGEV:100.0,SARS:100.0):73.0,HCoV-NL63:100.0);



How to use the downloaded files

- Original fasta file
- .phy(generated by clustalw)
- Tree Image
- outfile(generated by phylip programs) :
consensus
- treefile



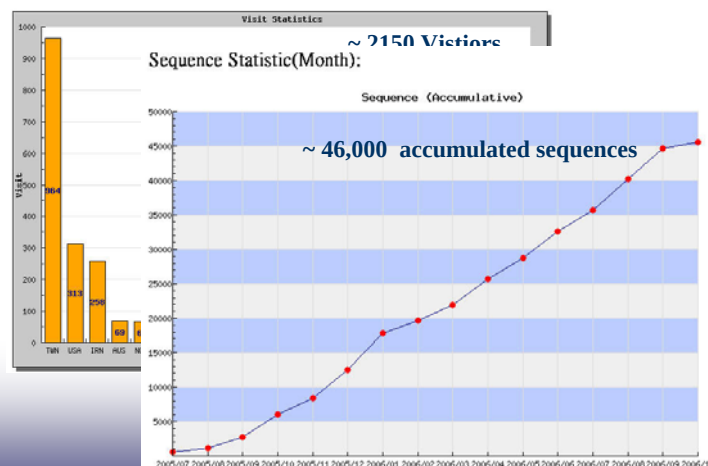
The problem would be asked

- What kind of datasets should be chosen
- How to do model selection
- Which tree topology should be used
- What kind of result is more reliable



Service Usage of POWER from 2005 July.

Visit Statistic(Country):





Acknowledgement

Chia-Ling Chen
Chia-Ming Chang
Chi-Shiang Cho
Chen-Zen Lo
Shu-Jun Hsu
Li-Ming Wang
Wei-Chen Chen
Fan-kai Lin
Chieh-Hua Lin
Li-Wei Lai
Pao-Yang Chen



Thanks for Your Attention!



National Health Research Institutes
Division

of
Biostatistics &
Bioinformatics

