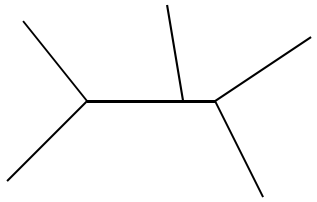
Creating Phylogenetic Trees with MEGA

Prat Thiru



Outline

- MEGA Features
- Background on Phylogenetic Trees
- Brief Overview of Tree Building Methods
- MEGA Demo

MEGA

- Easy-to-use software with multiple features
- Features:
  - Aligning sequences
  - Estimating evolutionary distances
  - Building trees using several methods
  - Testing tree reliability
  - Marking Genes/Domains
  - Testing for selection
  - Computing sequence statistics

Phylogenetics

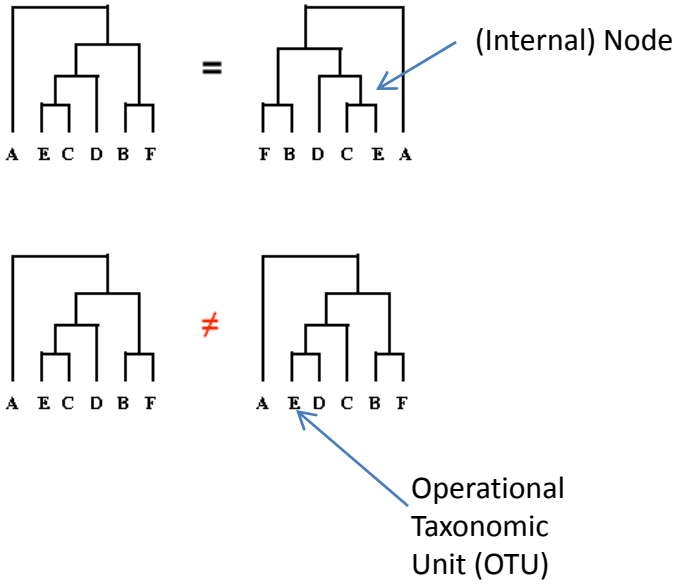
- Study of evolutionary relationship
- Phylogenetic tree is a graphical representation of the evolutionary relationships
- Phylogeny of…
  - Species
  - Strains
  - Genes
  - Metabolic Pathways
- Trees can be inferred by morphology or molecular information

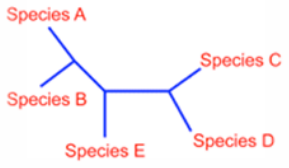
Why Create Phylogenetic Trees?

- Reconstruct evolutionary history
- Draw conclusions of biological functions which might not be apparent
- Pre-computed trees (eg. Ensembl, Pfam) might not include genes or species of interest

Parts of a Tree

- Nodes: taxonomic units (eg. genes, species, etc.)
  - Internal: ancestral state
    - Bifurcating
    - Multifurcating
  - External: Operational Taxonomic Units (OTUs)
- Branches: relationships among the taxonomic units (ie. ancestor-descendent relationship).
  - Clade
- Branch Length: number of changes that have occurred
- Topology: branching pattern
- Rooted vs Unrooted tree





Steps to Create Phylogenetic Trees

- Identify and acquire the sequences that are to be included on the tree
- Align the sequences (MSA using ClustalW, T-Coffee, MUSCLE, etc.)
- Estimate the tree by one of several methods
- Draw the tree and present it

From Hall, B.G. (p.3 - see *Further Reading* Slide)

Tree Building Methods

Distance-based

Character-based
-Parsimony
-Probabilistic

Example:

Distance Matrix

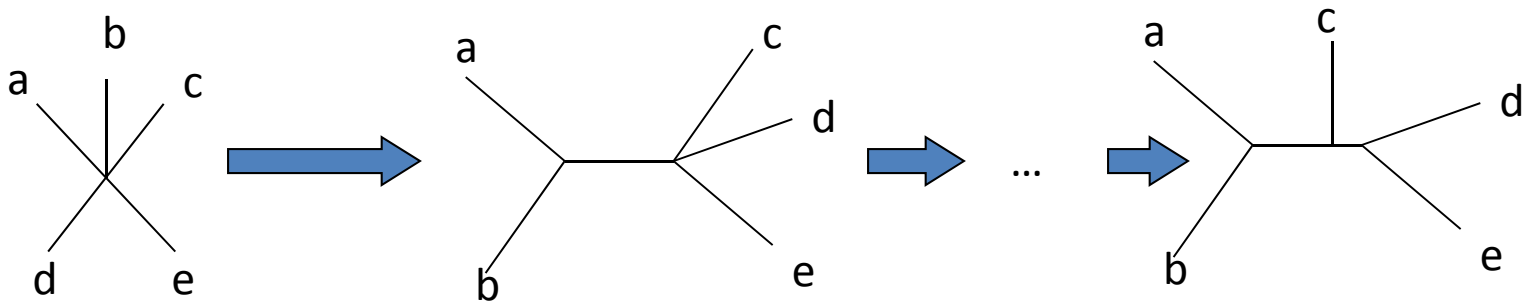
	A	B	C	D
A	1	0.8	0.4	0.6
B	0.8	1	0.6	0.4
C	0.4	0.6	1	0.8
D	0.6	0.4	0.8	1

Character State Matrix

Species A	ACTTC
Species B	AGTTC
Species C	CGTAC
Species D	CCTAC

Distance-based

- Unweighted Pair Group Method with Arithmetic mean (UPGMA)
- Neighbor-Joining (NJ)
  - Start with all taxa in a single node and *decompose* with each iteration
  - Pair of nodes pulled out (grouped) at each iteration are chosen so that the total length of the branches on the tree is minimized
  - Mutation rates are not constant



Character-based: Parsimony

- Preferred phylogenetic tree is the one with the fewest evolutionary steps
- Identify informative sites
- For each possible tree, calculate the number of changes at each informative site
- Sum total number of changes for each possible tree, the tree with the smallest number of changes is selected as the most likely tree

	Site								
	1	2	3	4	5	6	7	8	9
Sequence									
1	A	A	G	A	G	T	G	C	A
2	A	G	C	C	G	T	G	C	G
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	G
					*		*		*

*informative site

Character-based: Probabilistic

- Maximum Likelihood
  - At each site, the likelihood is determined by evaluating the probability that a certain evolutionary model (eg. BLOSSUM or PAM matrices) has generated the observed data.
  - The likelihood's for each site are then multiplied to provide likelihood for each tree
  - Choose the tree with maximum likelihood
- Bayesian Inference
  - Recent variant of ML
  - Finds a set of trees with the greatest likelihood given the data

Comparison of Methods

- Distance-based
  - Results in a single tree
  - UPGMA: reliable only for closely related species; replaced by NJ
  - NJ: fast, suitable for large dataset
- Character-based
  - Multiple trees will be found
  - MP: long branch attraction problem
  - ML: statistically well founded but slow for large dataset
  - BI: faster to assess support for tree, prior distribution parameters must be specified

Hall, B.G. (p.60 - see *Further Reading* slide)

Which Method to Use?

- Accuracy
  - looking at how *close* the estimated tree is to the *true* tree resulted (in order of decreasing accuracy):

Bayesian Inference → Maximum Likelihood → Maximum Parsimony → Neighbor Joining

- Ease of Interpretation
  - single vs multiple trees
- Time and Convenience*

Data Set	Neighbor Joining	Maximum Parsimony	Maximum Likelihood	Bayesian Inference
Small Data	1 sec	3 sec	6 sec	-
Small Data**	9 sec	10 min	1 hr 34 min	29 min 40 sec
Large Data	1 sec	22 sec	3 min 29 sec	-
Large Data**	86 sec	10 hr 2 min	58 hr	6 hr 33 min

* MacPro dual processor (2.6 GHz); NJ and MP was using MEGA 4.0, ML by PHYML, and BI using MrBayes
** with reliability estimate

Hall, B.G. (p.160 - see *Further Reading* slide)

Tree Reliability: Bootstrapping

- A statistical re-sampling procedure commonly used for providing confidence to branches in phylogenetic trees
- A measure of repeatability, the probability that the branch would be recovered if the taxa were sampled again
- Bootstrapping values are typically presented from 1000 repeated calculations
- Bootstrap values of >70% is recommended

Some Tips

- Use more than one method
- Use more than one software package
- Examine more than one tree if multiple trees are generated
- Bootstrap your data
- Homology vs Homoplasy
- Consider another intermediate taxa to resolve relationship, if needed

Phylogenetic Tree Software

- PHYLIP (the *PHYL*ogeny *I*nference *P*ackage):
  http://evolution.genetics.washington.edu/phylip.html
- PAUP*:
  http://paup.csit.fsu.edu
- MrBayes:
  http://mrbayes.csit.fsu.edu
- MEGA(Molecular Evolutionary Genetic Analysis)
  http://www.megasoftware.net

A More Comprehensive Listing of Phylogeny Programs:
http://evolution.genetics.washington.edu/phylip/software.html

*Commercial software

Summary

Table 1 | **Comparison of methods**

Method	Advantages	Disadvantages	Software
Neighbour joining	Fast	Information is lost in compressing sequences into distances; reliable estimates of pairwise distances can be hard to obtain for divergent sequences	PAUP* MEGA PHYLIP
Parsimony	Fast enough for the analysis of hundreds of sequences; robust if branches are short (closely related sequences or dense sampling)	Can perform poorly if there is substantial variation in branch lengths	PAUP* NONA MEGA PHYLIP
Minimum evolution	Uses models to correct for unseen changes	Distance corrections can break down when distances are large	PAUP* MEGA PHYLIP
Maximum likelihood	The likelihood fully captures what the data tell us about the phylogeny under a given model	Can be prohibitively slow (depending on the thoroughness of the search and access to computational resources)	PAUP* PAML PHYLIP
Bayesian	Has a strong connection to the maximum likelihood method; might be a faster way to assess support for treesthan maximum likelihood bootstrapping	The prior distributions for parameters must be specified; it can be difficult to determine whether the Markov chain Monte Carlo (MCMC) approximation has run for long enough	MrBayes BAMBE

Nature Reviews Genetics 4, 275-284 (2003)

Further Reading

- Kumar, S., Dudley, J., Nei, M., and Tamura, K., **MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences.** *Briefings in Bioinformatics* 9: 299-306 (2008)
- Hulsenbeck, J.P., and Ronquist, F., **MyBayes: Bayesian inferences of phylogeny.** *Bioinformatics* 17:754-755 (2001)
- Holder, M., and Lewis, P.O., **Phylogeny estimation: traditional and Bayesian approaches.** *Nature Reviews Genetics* 4: 275-284 (2003)

Books:

- Phylogenetic Trees Made Easy: A How-to Manual 3rd Ed. Hall, B.G. (2008)
- Inferring Phylogenies. Felsenstein, J. (2003)