Introduction to Phylogenetics
Week 5

Distance methods

	Charater-based method	**Non-character based**
Explicit evolutionary model	Maximum likelihood	Pairwise distance
No explicit evolutionary model	Maximum parsimony	

I. Distance Methods

- Try to fit tree to genetic distance (*d*-distance) matrix
- *d*-distance determined based on observed distance (*p*-distance)
- Use with evolution/substitution models – attempt to determine how many sites have actually changed
- Determine topology based on generated *d*-distance matrix

I. Distance Methods

Align sequences

↓

Calculate observed distance (*p*-distance)

↓

Estimate evolutionary distance (*d*-distance)

↓

Calculate for every sequence vs sequence

I. Distance Methods

- Requires on accurate calculation *d*-distance (by default good alignment)
- Dependent on correct choice of evolutionary model
- Trees build using:
  - cluster analysis
  - minimum evolution
- Clustering sensitive to unequal evolutionary rates in different species

II. Cluster Analysis

- Based on taxonomic phenograms
- Applied to create ultrametric trees
  - Uses math based on triangle inequities
  - Ultrametric strengthened triangle inequity
  - Replace ‘sides’ with ‘taxa’

$$d_{AC} \pounds (d_{AB}, d_{BC})$$

  - Two of three distances are equal or larger than a third
  - Assumes molecular clock

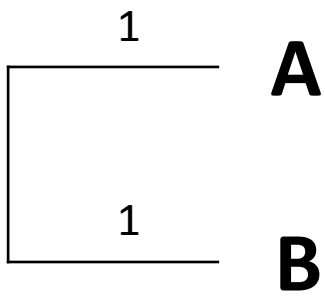
II. Cluster Analysis

- Two primary clustering methods:
  - Unweighted-pair group method with arithmetic mean (UPGMA)
  - Weighted-pair group method with arithmetic mean (WPGMA)
- Trees are built step-wise
- Grouping OTUs base on smallest genetic distance to create new *d* value
- Calculate distance from new node *u* to any other node *k*

WPGMA

$$d_{uk} = \frac{d_{(A,B)k} + d_{Ck}}{2}$$

	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8



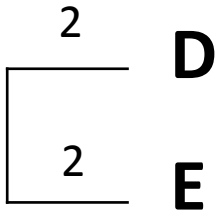
	AB	C	D	E
C	4			
D	6	6		
E	6	6	4	
F	8	8	8	8

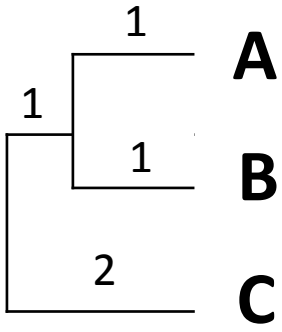
$$d_{(AB)C} = \frac{d_{(AC)} + d_{BC}}{2} = 4$$

$$d_{(AB)D} = \frac{d_{(AD)} + d_{BD}}{2} = 6$$

$$d_{(AB)E} = \frac{d_{(AE)} + d_{BE}}{2} = 6$$

$$d_{(AB)F} = \frac{d_{(AF)} + d_{BF}}{2} = 8$$





II. Cluster Analysis (UPGMA)

- Distances averaged on the number of OTUs
- Changes value of *u* and *k*

$$d_{uk} = \frac{(N_{AB} d_{(A,B)k} + N_C d_{Ck})}{(N_{AB} + N_C)}$$

N_{AB} = # OTUs in cluster AB

N_c = # OTUs in cluster C

- Mostly affects data that is not ultrametric

III. Minimum Evolution (ME)

- Clustering methods sensitive to variation
- NJ creates additive distance trees
- Uses four-point metric condition

$$d_{AB} + d_{CD} \pounds \max(d_{AC} + d_{BD}, d_{AD} + d_{BC})$$

- Are used to create an unrooted tree – sum of distance between pair OTUs is sum (rather than average) of length connecting them
- Tree algorithm cannot always satisfy four-point condition – tree distances different

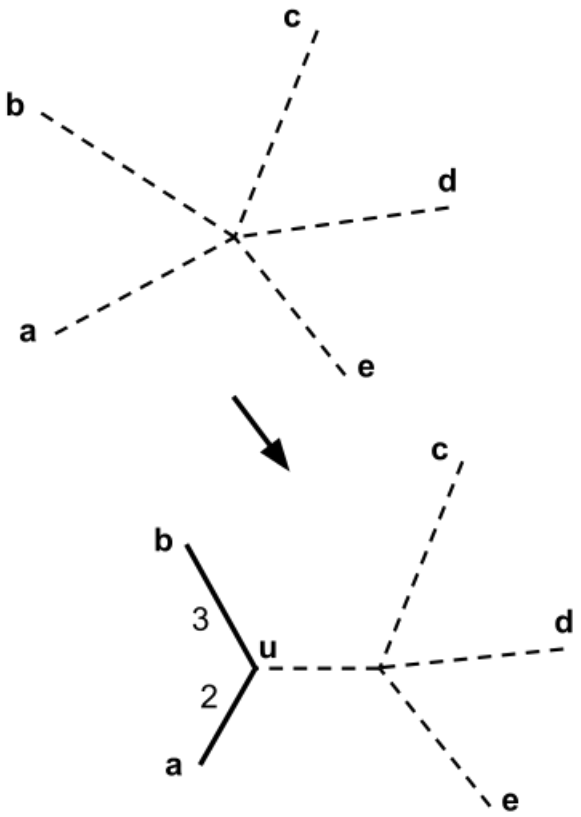
III. Minimum Evolution (ME)

- Clustering methods sensitive to variation
- NJ creates additive distance trees
- Uses four-point metric condition

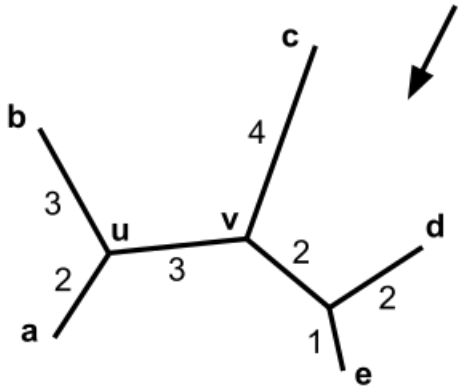
$$d_{AB} + d_{CD} \pounds \min(d_{AC} + d_{BD}, d_{AD} + d_{BC})$$

- Are used to create an unrooted tree – sum of distance between pair OTUs is sum (rather than average) of length connecting them
- Tree algorithm cannot always satisfy four-point condition – tree distances different

	a	b	c	d
b	5			
c	9	10		
d	9	10	8	
e	8	9	7	3



Branches = 2*n*-3



III. Minimum Evolution

- ME much better algorithm, but cannot take into account stochastic variation
- Minimum evolution method attempts to minimize branch lengths of tree
- Gives a best estimate of phylogeny

$$S = \sum_{i=1}^{2n-3} v_i$$

n = number of taxa in tree

v_i = ith branch

III. ME – Neighbor-Joining

- Can estimate branch length from distance matrix
- Optimizes so much that might not find best tree (which might be bigger)
- Have to explore many different topologies to find smallest tree - only works for smaller trees
- Can overcome some issues with NJ approach
- NJ does not assume an evolutionary clock

III. ME – Neighbor-Joining

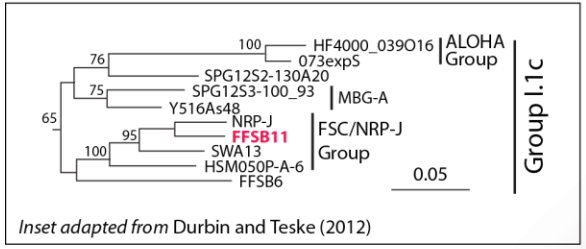
- Most common method to estimate distance trees
- Minimizes *S* value, but does for each branch added (*S* not globally minimized)
- Generally generates similar tree to ME method
- Can evaluate final trees using bootstrapping
- Can combine bootstrapping with maximum-likelihood (NJML)

IV. Evaluation of trees: Bootstrapping

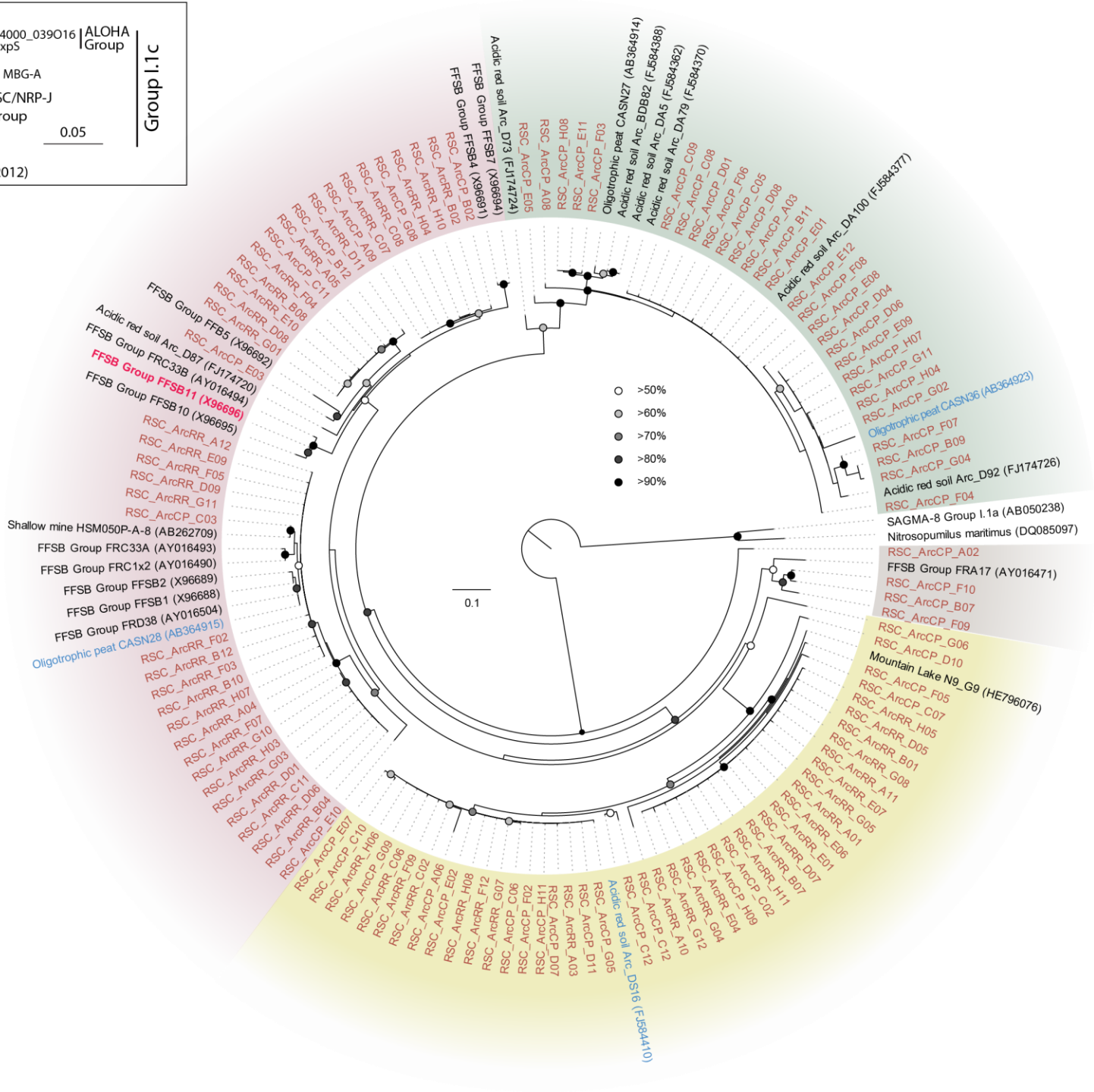
- Calculates statistical error when using an unknown sampling distribution
- Allows approximation underlying distribution
- Resamples original **sequence data** with replacement
- Samples the alignment – creates a tree
- Proportion of clades from bootstrap tree agree with calculated tree
- Provides confidence of grouping/clades calculated

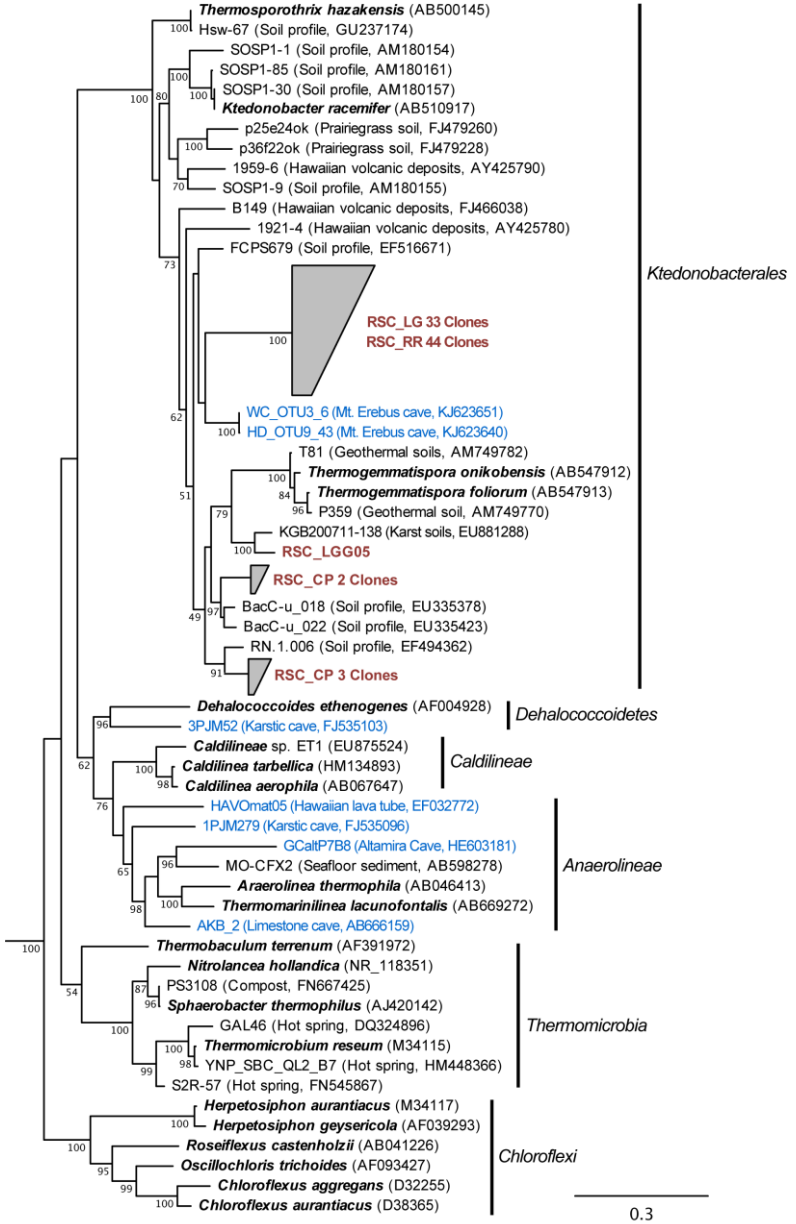
IV. Evaluation of trees: Bootstrapping

- Can be shown as a majority rule consensus
- Can be shown as values on tree
- Bootstrapping demonstrates statistical noise in data
- If your tree is bad, your bootstrap data will not support the topology
  - Poor alignment
  - Wrong substitution model
  - Incorrect assumption distribution



Inset adapted from Durbin and Teske (2012)




Thermosporothrix hazakensis (AB500145)
Hsw-67 (Soil profile, GU237174)
SOSP1-1 (Soil profile, AM180154)
SOSP1-85 (Soil profile, AM180161)
SOSP1-30 (Soil profile, AM180157)
Ktedonobacter racemifer (AB510917)
p25e24ok (Prairiegrass soil, FJ479260)
p36f22ok (Prairiegrass soil, FJ479228)
1959-6 (Hawaiian volcanic deposits, AY425790)
SOSP1-9 (Soil profile, AM180155)
B149 (Hawaiian volcanic deposits, FJ466038)
1921-4 (Hawaiian volcanic deposits, AY425780)
FCPS679 (Soil profile, EF516671)
RSC_LG 33 Clones
RSC_RR 44 Clones
WC_OTU3_6 (Mt. Erebus cave, KJ623651)
HD_OTU9_43 (Mt. Erebus cave, KJ623640)
T81 (Geothermal soils, AM749782)
Thermogemmatispora onikobensis (AB547912)
Thermogemmatispora foliorum (AB547913)
P359 (Geothermal soil, AM749770)
KGB200711-138 (Karst soils, EU881288)
RSC_LGG05
RSC_CP 2 Clones
BacC-u_018 (Soil profile, EU335378)
BacC-u_022 (Soil profile, EU335423)
RN.1.006 (Soil profile, EF494362)
RSC_CP 3 Clones
Ktedonobacterales
Dehalococcoides ethenogenes (AF004928)
3PJM52 (Karstic cave, FJ535103)
Dehalococcoidetes
Caldilineae sp. ET1 (EU875524)
Caldilinea tarbellica (HM134893)
Caldilinea aerophila (AB067647)
Caldilineae
HAVOmat05 (Hawaiian lava tube, EF032772)
1PJM279 (Karstic cave, FJ535096)
GCaltP7B8 (Altamira Cave, HE603181)
MO-CFX2 (Seafloor sediment, AB598278)
Araerolinea thermophila (AB046413)
Thermomarinilinea lacunofontalis (AB669272)
AKB_2 (Limestone cave, AB666159)
Anaerolineae
Thermobaculum terrenum (AF391972)
Nitrolancea hollandica (NR_118351)
PS3108 (Compost, FN667425)
Sphaerobacter thermophilus (AJ420142)
GAL46 (Hot spring, DQ324896)
Thermomicrobium roseum (M34115)
YNP_SBC_QL2_B7 (Hot spring, HM448366)
S2R-57 (Hot spring, FN545867)
Thermomicrobia
Herpetosiphon aurantiacus (M34117)
Herpetosiphon geysericola (AF039293)
Roseiflexus castenholzii (AB041226)
Oscillochloris trichoides (AF093427)
Chloroflexus aggregans (D32255)
Chloroflexus aurantiacus (D38365)
Chloroflexi
0.3

IV. Evaluation of trees: Bootstrapping

- Bootstrapping can be fooled by sequences
  - Artificial grouping of data – sequences too similar compared with other sequence data (high G+C)
- Long-branch attraction
  - Incorrect sequence data or rapidly evolving sequences – distant sequences will group together
  - Divergent sequences drawn toward the root

IV. Evaluation of trees: Jackknifing

- Also a subsampling technique
- Deletes random sequences from alignment – recreates tree
- Looks at number of trees match best-tree topology
- Gives a jackknife value that represents % of trees that match clade topology

V. PAUP* and PHYLIP

- Traditionally used for distance/NJ methods
- PAUP* is command-line only – costs $$
- PHYLIP free, but very difficult to use due to disjointed program development

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

- PHYLIP gateway now on Pasteur Server

- http://mobyle.pasteur.fr/cgi-bin/portal.py#welcome

V. PHYLIP – Programs

DISTMAT Calculates evolutionary distance (d) values between sequences. Generates distance matrix for tree-building distance methods.

DNADIST Computes four different distances between species from nucleic acid sequences. The distances can then be used in the distance matrix programs. The distances are the Jukes-Cantor formula, one based on Kimura's 2- parameter method, the F84 model used in DNAML, and the LogDet distance.

PROTDIST Computes a distance measure for protein sequences, using maximum likelihood estimates based on the Dayhoff PAM matrix, the JTT matrix model, the PBM model, Kimura's 1983 approximation to these, or a model based on the genetic code plus a constraint on changing to a different category of amino acid

FITCH Estimates phylogenies from distance matrix data under the additive tree model.

KITSCH Estimates phylogenies from distance matrix data under the ultrametric model.

NEIGHBOR An implementation of the UPGMA (Average Linkage clustering) method. Methods are very fast and thus can handle much larger data sets.

V. PHYLIP Distance Methods

Align sequences

↓

Calculate model/substitution rates

↓

Calculate distance matrix (*d*-distance)

↓

Calculate evolutionary tree (FITCH/KITSCH)

↓

Use TreeView to view tree with bootstrap values

V. PHYLIP Distance Methods

- PHYLIP will provide trees – image quality not great
- If you generate NEWICK format trees, can use FigTree

http://tree.bio.ed.ac.uk/software/figtree/

- NEWICK standard tree format in nested parentheses

((A,B),(C,D)) ≡