Class 1:

# EVE 161: Microbial Phylogenomics

Class #4: Phylogeny

UC Davis, Winter 2018 Instructor: Jonathan Eisen Teaching Assistant: Cassie Ettinger

Euryarchaea

### Where we are going and where we have been

- Previous Class:
  - ■3. rRNA
- <u>Current Class:</u>
  4. Phylogeny
- <u>Next Class:</u>
  5. Tree of Life

## Phylogeny

## Eisen Make Up Office Hours Tomorrow

• <u>Storer 5331 10:30-11:30</u>

- <u>Cassie Ettinger Office Hours</u>
- Today Storer 5331 10:30-11:30

# Phylogeny

Phylogeny

Bacteria

### First Phylogenetic Tree?



**FIGURE 5.1.** Vertical inheritance as represented in the only figure in *On the Origin of Species*. The evolution of species over time is represented by the branching of the tree with a single species giving rise to one or more descendent species. The descendent species inherit traits from the parental species in the form of "vertical" inheritance. Modern species are represented by  $a^{10}$ ,  $f^{10}$ , and  $m^{10}$  with all other lineages having become extinct.  $a^{10}$  and  $f^{10}$  can trace their ancestry to a common ancestral species  $a^5$ . In turn,  $a^{10}$  and  $f^{10}$  can jointly trace their common ancestry with  $m^{10}$  to the species at the bottom of the figure.

5.1, reprinted from Darwin C., On the Origin of Species

Evolution © 2007 Cold Spring Harbor Laboratory Press

### Parts of a phylogenetic tree



Internal nodes represent hypothetical ancestral taxa

### **Branch Rotation**



**FIGURE 5.2.** Model phylogenetic trees showing tips, nodes, and branches. (*A*) All the key elements of a phylogenetic tree. This tree shows the evolutionary history of three operational taxonomic units (OTUs). OTUs can represent species, individuals, genomes, genes, or other entities having an evolutionary history. (*Thick blue lines*) The branches represent the evolution of the OTUs over time. In this tree, evolutionary time is shown progressing from bottom to top, and thus this is known as a vertical tree. When considering evolutionary time from the past to the present, nodes (*blue circles*) represent the points at which one lineage separated into two. When considering evolutionary time from the present to the past, nodes represent the common ancestor of the organisms above the node. In this case, Tip 2 and Tip 3 share a common ancestor at node B. All descendants from node B, including Tip 2 and Tip 3, can be considered a clade or monophyletic group (see Fig. 5.3 for more detail on clades). The separation of taxa on the horizontal axis and the angles of the branches have no real meaning; it is done to be visually pleasing. (*B*) The tree in *A* has been rotated 90°. Such horizontal trees contain the same information as vertical trees. In this case, evolutionary time progresses from left to right, and the separation on the *y*-axis has no meaning. (*C*) A T-branching tree. It too has the same information as the trees in *A* and *B*, but the branches are drawn with a T-shaped junction instead of a V-shaped junction. In each of these trees, the "root" of the tree is the branch leading up to the common ancestor of all taxa shown in the tree.

Evolution © 2007 Cold Spring Harbor Laboratory Press

## Types of Trees



estimated evolutionary difference).

### Phylogram (Tree with Lengths)



**FIGURE 5.5.** A phylogram (shown) is a phylogenetic tree that indicates the amount of evolution in addition to the branching order. The amount of evolution is represented by the branch lengths along the time axis (in this example, the vertical axis). In this tree, Tip 2 and Tip 3 share a common ancestor to the exclusion of Tip 1. However, during the time since they diverged from their common ancestor, Tip 2 has undergone more change. If Tip 2 and Tip 3 are modern organisms, this means that the rate of evolution in the lineage leading up to Tip 2 was greater than that leading up to Tip 3. Differences in rates of evolution are common and can be due to many factors such as different mutation rates, different population sizes, and different selective forces. Regardless of the cause, it is frequently very useful to incorporate such differences into evolutionary trees. In phylograms, a scale bar defines how much change is represented per unit length.

Evolution © 2007 Cold Spring Harbor Laboratory Press

### Phylogenetic Groups



Fig. 1. Trees are about groups: monophyletic (holophyletic), paraphyletic and 'polyphyletic'.

## Phylogenetic Groups



FIGURE 5.3. Different types of phylogenetic groups. In each panel, the phylogenetic group is indicated by a green shaded circle. (A) Monophyletic group. All species (C and D) in the group share a common ancestor (E) not shared by any of the other species. (B) Paraphyletic group. All species in the group share a common ancestor (F), but some species (D) have been excluded from the group. (C) Polyphyletic group. A grouping of lineages each more closely related to other species not in the group than they are to each other.

Evolution © 2007 Cold Spring Harbor Laboratory Press



# **Branch** rotation



**FIGURE 5.6.** Branch rotation does not change the information in a phylogenetic tree. The two trees are the same except the branch leading up to the ancestor of Tip 2 and Tip 3 has been rotated such that Tip 3 is on the left and Tip 2 is on the right. The tree drawings on *left* and *right* are simply different forms of the same tree. It is useful to consider a phylogenetic tree like a mobile: The branches are free to rotate, but the branching patterns never change.

Evolution © 2007 Cold Spring Harbor Laboratory Press

 A heritable feature of an organism is known as a character (also character trait or trait).

- The form that a character takes is known as its state (also known as character state).
  - Note: Presence/absence can be a state
- Example:
  - Character = heart
  - Character state = present/absent
  - Character state = # of chambers

### Characters ancestry is critical to understand

- Characters that are inherited from a common ancestor are homologous.
- Species change over time
  - Known (generally) as divergence, or divergent evolution.
  - Species change over time due to the combined processes of mutation, recombination, drift, selection, etc

# Homology vs. Orthology vs. Paralogy



### Data matrices

TABLE 27.1. Character	matrix showing	character states	for nine traits	s (A–I) in eig	ht OTUs (1–8)
-----------------------	----------------	------------------	-----------------	----------------	---------------

				Cha	aracter Tr	aits			
OTUs	Α	В	С	D	E	F	G	Н	I.
							2		
1	α	20	1	α	+	+	2	1	—
2	β	10	1	α	+	-	2	ii	-
3	β	14	1	—	+	-	1	iii	-
4	α	23	1		_	-	1	ii	_
5	β	13	1		_	+	1	ii	_
6	α	23	1	α	_	+	2	i	_
7	β	14	1	β	_	+	2	ii	_
8	α	20	1	α	+	+	2	i	_

# Alignment

			Alignment Position							
		1	2	3	4	5	6	7	8	9
	A	A	Α	С	Т	Α	Т	G	G	С
	В	A	С	Т	A	Т	G	G	С	Α
s	С	A	Т	Α	С	Т	A	Т	G	G
15	D	A	Т	Α	С	Т	A	Т	G	G
0	E	A	Α	Т	G	G	С	С	С	Α
	F	A	A	С	Т	Т	Т	G	G	С
	G	A	С	Т	A	Т	G	G	С	Α
	Н	A	A	С	Т	A	Т	G	G	С

Euryarchaea

# Alignment with Gaps

			Alignment Position											
		1	2	3	4	5	6	7	8	9	10	11	12	13
	Α	Α	-	Α	C	Т	A	Т	G	G	C	I	-	-
	В	A	_	-	C	T	A	Т	G	G	C	Α	—	_
s	С	A	Т	A	C	T	A	Т	G	G	-	-	—	-
15	D	Α	Т	Α	C	Т	A	Т	G	G	-	I	-	_
Ó	E	A	-	-	-	-	A	Т	G	G	C	С	C	A
	F	A	—	A	C	Т	Т	Т	G	G	C	-	—	—
	G	Α	-	-	C	Т	A	Т	G	G	C	Α	-	_
	Н	Α	-	A	C	T	A	Т	G	G	C	_	—	_

Euryarchaea

### Sequence Alignment

MSQNSL MSKLKEKREK MAEE MSKDATKE	MAIDENK MDENK RLYEDKSVDK MSKLAEK MAIDEDK MAINTDTSGK MAGTDR MSDR MANIDKDK AVVGIERASK MSVPDR MSSEK KIPTVQDEKK MPEEKQK ISAPTDAKER ARVSENLSEK	Q AL AAAL GQ KRAL SAAL SQ SKAL EAAL SQ L KAVAAAVAS QKAI SL AI KQ QKAL TMVL NQ EKAL DAAL AQ QAAL DMAL KQ L KAI EMAMGQ EEAL ELARVQ KRAL EAA I AV L KAL QAAMDK L QAL RMATEK KSYL EKAL KR SKAI ET AMSQ MKAL EYAL SS	I EKQFGKGS I I EKQFGKGSV I ERSFGKGSV I DKVFGKGAL I ERSFGKGAI I ERQFGKGAV I EKQFGKGSV I EKQFGKGSV I EKAFGKGSI I EKSFGKGSI I EKSFGKGSI I EKRFGKGSI I EKRFGKGAV	MRL GEDRSM- MRMGDRYI E - MKL GSNENV I MTL GGEAREQ VRL GDKVQE - MRL GDAT RM- MRMGDRT NE - MKL GEKT DT - MKL GEKT DT - MKL GEQGAP - I KMGESPVGQ MSL GKHSSAH MKMGE - EVVE MNMGANT YE - MI L GDET QVQ MKL GAESKL - MPL KAYET V -	Escherichia coli Xanthomonas campestris Rhizobium phaseoli Myxococcus xanthus Helicobacter pylori Anabaena variabilis Steptomyces lividans Bacillus subtilis Clostridium perfringens Borrelia burgdorferi Chlamydia trachomatis Bacteroides fragilis Porphyromonas gingivalis Thermotoga maritima Deinococcus radiodurans Aquifex pyrophilus	
DVE <mark>T</mark> ISTGSL AVEVIPTGSL EIETISTGSL KVAVIPSGSV KIDAISTGSL BVETISTGAL	SLDIALG <mark>A</mark> GG MLDIALGIGG GLDIALGVGG GVDRALGVGG GLDLALGIGG TLDLALG-GG	L PMGRIVEIV L PKGRVVEIV L PKGRIIEIV V PRGRVVEVF V PKGRIIEIV L PRGRVIEIV	GPESSGKTT L GPESSGKTT L GPESSGKTT L GPESSGKTT L GPESSGKTT L GPESSGKTT V	TLQVIAAAQR TLQAIAECQK ALQTIAESQK TLHAIAQVQA SLHIIAECQK ALHAIAEVOK	Escherichia coli Xanthomonas campestris Rhizobium phaseoli Myxococcus xanthus Helicobacter pylori Anabaena variabilis	
P IEVIPTGST RISTVPSGSL QMDAVSTGCL GIKSMSSGSI EISTIKTGAL QVEVIPTGSI DVSVIPSGSI	ALDVALGVGG ALDTALGIGG DLDIALGIGG VLDEALGIGG SLDLALGIGG ALNAALGVGG GLDLALGVGG	IPRGRVVEV VPRGRIIEV VPKGRIIEIV VPRGRIIEIF VPKGRIVEIF VPRGRIIEIV	GPESSGKTTL GPESSGKTTV GPESSGKTTL GPESSGKTTL GPESSGKTTL GPESSGKTTL	TLHAVANAQK ALHVIAAAQQ ALHVVAAAQK TLQAIAEVQK ATHIVANAQK AIHAIAEAQK AIHAIAEAQK	Steptomyces lividans Bacillus subtilis Clostridium perfringens Borrelia burgdorferi Chlamydia trachomatis Bacteroides fragilis Porphyromonas gingivalis	
PVEVIPTGSL DVQV <b>VS</b> TGSL EVETIPTGSL	AIDIATGVGG SLDIALGVGG SLDIATGVGG	VPRGRIVEI <mark>F</mark> IP <mark>G</mark> GRITEIV IP <mark>K</mark> GRITEI <mark>F</mark>	G <mark>Q</mark> ESSGKTTL GPES <mark>G</mark> GKTTL G <mark>V</mark> ESSGKTTL	AL HA I AEAQK AL <mark>A I V</mark> AQAQK AL H <mark>V</mark> I AEAQK	Thermotoga maritima Deinococcus radiodurans Aquifex pyrophilus	
EG <mark>KTC</mark> AFIDA LGGTAAFIDA KGGICAF <mark>V</mark> DA AGGVAAFIDA NGGVCAFIDA EGGLAAFVDA	EHAL DPI YAR EHAL DPI YAA EHAL DPVYAR EHAL DVSYAR EHAL DVHYAK EQAL DPTYAS	KLGVD <mark>I</mark> DNLL KLGV <mark>NV</mark> DDLL KLGVDLQNLL KLGV <mark>RVEE</mark> LL RLGVDTQNLL	CSQPDTGEQA LSQPDTGEQA ISQPDTGEQA VSQPDTGEQA VSQPDTGEQA VSQPDTGESA	LEICDALARS LEIADMLVRS LEITDTLVRS LEITEHLVRS LEILETITRS	Escherichia coli Xanthomonas campestris Rhizobium phaseoli Myxococcus xanthus Helicobacter pylori Anabaena variabilis	
AGGQVAFVDA Q-RTSAFIDA LGGAAAVIDA EGGIAAFIDA MGGVAAVIDA	EHAL DPEYAK EHAL DPVYAQ EHAL DPVYAK EHAL DPVYAK EHAL DPNYAA	KLGVDIDNLI KLGVNIEELL RLGVNIDDLV ALGVNVAELW LIGANINDLM	L SQPDNGEQ A L SQPDTGEQ A VSQPDTGEQ A L SQPDTGEQ A I SQPDCGEDA	LEI VDMLVRS LEI AEALVRS LEI TEALVRS LEI AEALI RS LSI AEALARS	Steptomyces lividans Bacillus subtilis Clostridium perfringens Borrelia burgdorferi Chlamydia trachomatis	
AGGIAAFIDA AGGLAAIIDA MGGVAAFIDA AGGTCAFIDA RGGVAVFIDA	EHAFDRFYAA EHAFDRTYAE EHALDPVYAK EHALDPVYAR EHALDPKYAK	KLGVDVDNLF KLGVNVDNLW NLGVDLKSLL ALGVNADELL KLGVDVDNLV	I SQPDNGEQA I SQPDNGEQA I AQPDHGEQA VSQPDNGEQA I SQPDVGEQA	LEIAE <mark>QLI</mark> RS LEIAEQLIRS LEIVDELVRS LEIMELLVRS LEIAESLINS	Bacteroides fragilis Porphyromonas gingivalis Thermotoga maritima Deinococcus radiodurans Aquifex pyrophilus	

**FIGURE 5.16.** Example of a multiple sequence alignment, showing alignment of a portion of the RecA proteins from different bacterial species. Amino acids conserved across most or all of the species are highlighted in *red. Letters* are the abbreviations of different amino acids (see Fig. 2.23).

Evolution © 2007 Cold Spring Harbor Laboratory Press

### Match 1

Query	AATTAATTAACC		
Database	AATTAATTAACC		
Score	111111111111 =	=	Total 12

Match 2

Query Database Score

AATTAATTAACC AACCAATTAACC 110011111111 = Total 10

# **TABLE 27.3.** DNA substitution matrix and some word matches

Query		Databas	e Seque	nce
Sequence	Α	С	G	Τ
А	1	0	0	0
С	0	1	0	0
G	0	0	1	0
Т	0	0	0	1

### **Blosum Matrix**

S Η R v Υ Ρ Α G Ν Ε Κ М L W т Q F  $\mathbf{C}$  $\mathbf{D}$ Ι 0 -11 2 2 -22 1 1 2 0 5 1 2 5 1 0 0 4 1 С 2 -20 0 1 1 -1S 0 -10 1 0 1 -1 0 0 0 0 1 С 9 2 -1-1 0 0 0 0  $^{-1}$ 1 0 1 3 Т -1 0 0 0 1 -1  $^{-1}$ 2  $^{-1}$ S -2-1 0 -1 -1 -1 0 0 2 1 P 4 1 0 -1 0 1 2  $^{-1}$ -11 -2-2 -1 0 1 1 1 2 Α Т 5 0 1 0 0 0 1 -3 -1  $^{-1}$ 2 0 -1 -1 -1 2 Ρ  $^{-2}$ 1 0 4 G 7 0 0 0 1 1 -1 0  $^{-1}$ 3 -1 1  $^{-1}$ -1 0 -10 0 Α 1 0 0 0 N 4 0 0 -1 2 -1 0 -1 0 2 1 3 -3-2-2 $^{-1}$ 0 D G 0 0 0 0 6 1 0 2 2 2 0  $^{-1}$ 2 -31 -2-21 0 4 Ε  $\mathbf{N}$ 0 0 6 0 3 0 -20 -1 1 3 -3-1 -1 -2 -11 6 1 1 0 0 0 D 0 2 2  $^{-1}$ 2 2 Е -4 -1 $^{-1}$ -1-20 5 0 1 0 -1 1 Η 0 0 0 -3 0 -1 $^{-1}$  $^{-1}$ -2 0 0 2 5  $^{-1}$ -1 0  $^{-1}$ 1 0 3 -4 R 1 -22 -1 -2-2 -2 -21 -1 1 3 1 Κ Η -3 1 -10 0 8 1 5 R -2 $^{-1}$ -20 -20 1 -2 -1-10 1 2 4 Μ -10 -3-1-21 1 -1 2 -1 1 0 1 3 -30  $^{-1}$  $^{-1}$ 0  $^{-1}$ 0 Τ ĸ -15 -2-2-1 0 -11 2  $\mathbf{L}$ -2 $^{-1}$ -3-2 -3  $^{-1}$ Μ 0 -1 -15 -1-1 0 2 -3 -31 -1 -2-1 -3  $^{-1}$ -4 -3-3 -3 -3-3 1 4 v Т 4 -1 -2-3 -3 -2 -3 -22 2 F -3 -1 -4 -21 -1 -2 $^{-1}$ -4 4 ь -2 1 3 -12 -2-2-3-3-3 -2 -3-3-21 4 Υ v -1 0 0 -2-2 -3 -3 -3 -3 -3 -3 0 0 0  $^{-1}$ 6 -1 W F -4 -1 -2-2-3 -2-2-3 -2 -3 -2-3 -2-1 2 -2-2 -1 Y -2 $^{-1}$ -1 -1 7 3 -2-3-2-4 -3 -2 -4 -3 -2 -2-3 -3 -1-3 -2-3 2 11 W -4 1 C S т P Α G Ν D Ε Η R Κ Μ Τ  $\mathbf{L}$ Υ Q v F W

### Guide Tree for Alignment

(a) Guide tree D G Н Κ Α В С Е F J (b) Sequence addition order E + F Step 1 A + B I + J EF + G Step 2 IJ + K AB + С EFG + H ABC Step 3 D + ABCD EFGH Step 4 + ABCDEFGH Step 5 IJK + TRENDS in Genetics

**Fig. 4**. Steps in progressive sequence alignment. (a) The first step is to calculate the guide tree. (b) This determines the order in which sequences are added to the growing alignment.

### **Refining Alignment**

(a)							
taxon			10				
Fu	Nosema.40928	QFGLFSP	EEIRASS	VALIRYPE	TLENGVE	KESGLVCAG	HFGHIELVK
Fu	Aspergillus.	QFGLFSP	EEIKRMS	VVHVEYPE	TMDEQRQRE	RTKGLECPGI	HFGHIELAT
Ap	Plasmodium.3	ELGVLDP	EIIKKIS	VCEIVNVD	IYK <mark>DG</mark> FF	REGGLYCPG	HFGHIELAK
An	Cricetulus.2	QFGVLSP	DELKRMS	<b>VTEGGIKYPE</b>	TTEGGRI	KLGGLECPG	HFGHIELAK
An	Homo.7434727	QFGVLSP	DELKRMS	<b>VTEGGIKYPE</b>	TTEGGRI	KLGGLECPGI	HFGHIELAK
An	Drosophila.9	QFGILSP	DEIRRMS	VTEGGVQFAE	TMEGGRE	KLGGLECPGI	HFGHIDLAK
An	Celegans.133	QFGILGP	EEIKRMS	VAHVEFPE	VYENGKE	KLGGLDCPGI	HFGHLELAK
Fu	Spombe.54881	QFGILSP	EEIRSMS	VAKIEFPE	TMDESGORE	RVGGLDCPG	HFGHIELAK
Pl	Athaliana.40	QFGILSP	DEIROMS	VIHVEH	ISETTEKGKE	KVGGLECPG	HFGYL <mark>E</mark> LAK
My	Ddiscoideum.					ECPGI	HFGHIELAK
Rh	Porphyra.316					ECPGI	HFGFIELAK
Kt	Tbrucei.1021	QFEIFKE	RQIKSYA	VCLVEHAKSY	ANAAI	Q <b>SGEAECPG</b>	HFGYIELAE
Kt	Leishmania.7	QFEVFKE	AQIKAYA	K <b>¢IIEHAKS</b> Y	EHGQI	VRGGIECPG	HFGYVELAE
I							

#### (b)

taxon			
Fu	Nosema.40928	QFGLFSPEEIRASSVALIRYP	ETLENGVPKESGLVCAGHFGHIELVK
Fu	Aspergillus.	QFGLFSPEEIKRMSVVHVEYF	ETMDEQRQRPRTKGLECPGHFGHIELAT
Fu	Spombe.54881	QFGILSPEEIRSMSVAKIEFF	<b>ETMDES</b> GQRPRVGGLDCPGHFGHIELAK
Ap	Plasmodium.3	ELGVLDPEIIKKISVCEIVNV	DIYKDGFPREGGLYCPGHFGHIELAK
An	Cricetulus.2	QFGVLSPDELKRMSVTEGGIKYF	ETTE GGRPKLGGLECPGHFGHIELAK
An	Homo.7434727	QFGVLSPDELKRMSVTEGGIKYF	ETTE GGRPKLGGLECPGHFGHIELAK
An	Drosophila.9	QFGILSPDEIRRMSVTEGGVQFA	ETME GGRPKLGGLECPGHFGHIDLAK
An	Celegans.133	QFGILGPEEIKRMSVAHVEFF	EVYE NGKPKLGGLDCPGHFGHLELAK
Pl	Athaliana.40	QFGILSPDEIRQMSVIHVEHS	ETTE KGKPKVGGLECPGHFGYLELAK
My	Ddiscoideum.		ECPGHFGHIELAK
Rh	Porphyra.316		ECPGHFGFIELAK
Kt	Tbrucei.1021	QFEIFKERQIKSYAVCLVEHA	KSYANAADQSGEAECPGHFGYIELAE
Kt	Leishmania.7	QFEVFKEAQIKAYAKCIIEHA	KSY-EHGQPVRGGIECPGHFGYVELAE
1			

TRENDS in Genetics

**Fig. 5**. Refining an alignment. (a) The raw output from a ClustalX alignment of rpb1 sequences, which predicts six insertion/deletion events (boxed), some of which are blatantly inconsistent with known taxonomy. (b) The refined alignment makes much better evolutionary sense, because it shows only two insertion events in well-defined taxonomic groups (animals and higher fungi). Taxon labels are Fu (fungi), An (animals), PI (green plant), Ap (apicomplexan), Rh (rhodophyte), My (mycetozoan), Kt (kinetoplastids). In (b), the sequence from *Saccharomyces pombe* has been placed adjacent to the other fungi to make these relationships more obvious.

### Structure Guided Alignment



### Structure Guided Alignment



# HMM Alignment



### Tree reconstruction methods

### TABLE 27.4. Molecular phylogenetic methods

Method	Description
Parsimony	Possible trees are compared and each is given a score that is a reflection of the minimum number of character state changes (e.g., amino acid substitutions) that would be required over evolutionary time to fit the se- quences into that tree. The optimal tree is considered to be the one re- quiring the fewest changes (the most parsimonious tree).
Distance	The optimal tree is generated by first calculating the estimated evolu- tionary distance between all pairs of sequences. These distances are then used to generate a tree in which the branch patterns and lengths best represent the distance matrix.
Maximum likelihood	This method is similar to parsimony methods in that possible trees are compared and given a score. The score is based on how likely the given sequences are to have evolved in a particular tree given a model of amino acid or nucleotide substitution probabilities. The optimal tree is considered to be the one that has the highest probability.
Bayesian	A variant of maximum likelihood in which the likelihood of a tree itself is calculated.

Based on Table 3 in Eisen J.A. 1998. Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.* 8:163–167.

### **Distance Phylogenetics**

- Calculate distances between taxa
- Use an algorithm to infer a tree from those distances
- Based on principle that divergence occurs after organisms separate

# **Parimony Phylogenetics**

- Based on Principle of Parsimony / Occam's Razor
- Possible trees are given a score of the fewest number of sequence changes required to fit data into tree
- Score for all possible trees

![](_page_29_Picture_4.jpeg)

## Likelihood Phylogenetics

- Based on Bayes' Theorem
- Trying to calculate the Probability of the Data, Given a Hypothesis
- For each tree, calculate probability that the data (e.g., sequence alignment) could come from that tree
- Score for all possible trees

- $Prob(H|D) = Prob(H \text{ and } D) \div Prob(D)$
- $Prob(H|D) = Prob(D|H) \times Prob(H) \div Prob(D)$

### **Bayesian Phylogenetics**

- Based on Bayes' Theorem
- Trying to calculate the Probability of the Hypothesis, Given the Data
- For each tree, calculate probability that the tree could come from that data ((e.g., sequence alignment)
- Score for all possible trees

- $Prob(H|D) = Prob(H \text{ and } D) \div Prob(D)$
- **Prob**(*H*|*D*) =  $\underline{Prob}(D|H) \times Prob(H) \div Prob(D)$

### Bayesian

- Based on Bayes' Theorem
- $Prob(H|D) = Prob(H \text{ and } D) \div Prob(D)$

•  $\underline{\operatorname{Prob}(H|D)} = \operatorname{Prob}(D|H) \times \operatorname{Prob}(H) \div \operatorname{Prob}(D)$ 

![](_page_32_Picture_4.jpeg)

### Many possible trees

### TABLE 27.5. Number of possible branching patterns versus number of OTUs

Таха	Rooted Trees <sup>a</sup>	Unrooted Trees <sup>b</sup>
3	3	1
4	15	3
5	105	15
6	945	105
7	10,395	945
8	135,135	10,395
9	2,027,025	135,135
10	34,459,425	2,027,025

<sup>a</sup> $N_r = (2n - 3) \times (2n - 5) \times (2n - 7) \times \cdots \times 3 \times 1 = (2n - 3)!/[2_{n-2} \times (n - 2)!].$ <sup>b</sup> $N_u = (2n - 3) \times (2n - 5) \times (2n - 7) \times \cdots \times 3 \times 1 = (2n - 5)!/[2_{n-3} \times (n - 3)!].$ 

### Other Tree Related Issues

![](_page_34_Picture_1.jpeg)

![](_page_35_Figure_1.jpeg)

**Fig. 6.** Bootstrap analysis proceeds in three steps. The dataset is randomly sampled with replacement to create multiple pseudo-datasets of the same size as the original ((a), three are shown in this example). (b) Individual trees are constructed from each of the pseudo-datasets. (c) Each of the pseudo-dataset trees are scored for which nodes (groupings) appear and how often. In this case, a node uniting seqA plus seqB is found in two of the three replicate trees. This gives a bootstrap support for this grouping of 2/3 or 67%.

http://tigs.trends.com
#### Bootstrapping

Euk

#### A Bootstrapping

Alignment of sequences

ГG	GΤ	'G <mark>A</mark>	T
AG	GA	.G <mark>A</mark>	A
AG	AA	.G <mark>A</mark>	A
AG	CC	GC	C:
		F <mark>G</mark> GT AGGA AGAA AGCC	FGGTGA AGGAGA AGAAGA AGCCGC

Bootstrapping alignment #1

Species 1 Species 2 Species 3 Species 4

S S S

> ATGTTGGAGGGTAAT ATGTTGGAGGGAAAA ATGTTAGGGGAAAAA ATGTCAGCGGCCCCC

Bootstrapping alignment #2

Species 1 Species 2 Species 3 Species 4 ATGGGGGGATGGTGAT ATGGGGGGAAGGAGAA ATGGGAGGAGAAGAA ATGGGAGCAGCCGCC

#### Jacknifing

Plant

#### B Jackknifing

#### Alignment of sequences

Species	1	ATGTTGGATGGTGAT
Species	2	ATGTTGGAAGGAGAA
Species	3	ATGTTAGGAGAAGAA
Species	4	ATGTCAGCAGCCGCC

#### Jackknifed alignment 1 (10 columns kept)

Species	1
Species	2
Species	3
Species	4

AT-T-GGA-GG-GA-AT-T-GGA-GG-GA-AT-T-AGG-GA-GA-AT-T-AGC-GC-GC-

Jackknifed alignment 2 (10 columns kept)

Species	1	GTTGGGGT-AT
Species	2	GTTGGGGA-AA
Species	3	GTTAGGAA-AA
Species	4	GTCAGGCC-CC

#### Congruence



#### Masking

#### Α

- Protein 1 Protein 2 Protein 3 Protein 4 Mask
- DEMGLGKK---R--VESTR DEMGVGKRGH---LESRK DDMGLGRWK--R-VESTE DDMGLGHKKK---VDSTK 111111100000011111

#### B

Protein 1 DEMGLGKKVESTR Protein 2 DEMGVGKRLESRK Protein 3 DDMGLGRKVESTE Protein 4 DDMGLGHKVDSTK A Gene alignments **B** Concatenated alignment 3 

#### Homoplasy

#### A Sequence 1 GACGAG Sequence 2 GCCGAC

#### В

Sequence 1Sequence 2GACGAGGCCGAC $\uparrow G \rightarrow Aat$  $\uparrow C \rightarrow Aat$  $\uparrow position 5$  $\bigcirc CGGCC$ GACGGGGCCGCC $\uparrow G \rightarrow Aat$  $\bigcirc G \rightarrow Cat$  $\uparrow position 2$  $\bigcirc G \rightarrow Cat$ GGCGGGGCCGGG $G \rightarrow Cat$ position 6GCCGGGGCCGGGGCCGGG

#### Clustering vs. Distance Trees



# UPGMA

# Unweighted Pair Group Method with Arithmetic mean (UPGMA) algorithm

#### The True Tree



### Distance Matrix For True Tree

#### TABLE 27.6. Distance matrix

OTUs	Α	В	С	D	Ε	F
А	0	2	4	6	6	8
В	2	0	4	6	6	8
С	4	4	0	6	6	8
D	6	6	6	0	4	8
E	6	6	6	4	0	8
F	8	8	8	8	8	0

# Collapse Diagonal

#### TABLE 27.7. Diagonal matrix—Step 1

OTUs	Α	B	С	D	Ε
В	2				
С	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

# Identify Lowest D

OTUs	A	B	С	D	E
B	<u>2</u>				
С	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

#### Join Those Two Taxa

• Create branch with length = D

• 2 • A-----B

### Make D from Node Equal



#### Create New Distance Matrix

• Merge Two OTUs joined in previous step (AB)

• 
$$D_{x,AB} = 0.5 * (D_{x,A} + D_{x,B})$$

### New Matrix

#### TABLE 27.8. Diagonal matrix—Step 2

OTUs	AB	С	D	Ε	
C	4				
	4	6			
F	6	6	1		
L	0	0	4	0	
Г	0	0	0	0	

# UPGMA

TABLE 27.9. Example of UPGMA tree construction							
Step	Cycle 1	Cycle 2	Cycle 3	Cycle 4	Cycle 5	Cycle 6	
Distance matrix	OTUs     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	OTUs AB C D E C 4 D 6 6 E 6 6 4 F 8 8 8 8	OTUS AB C DE C 4 DE 6 6 DE 8 8 8	OTUS ABC DE DE <mark>6</mark> F 8 8	OTUs <u>ABCDE</u> F <mark>8</mark>	No new matrix	
ldentify smallest D	$A \leftrightarrow B = 2$	$AB \leftrightarrow C = 4$ $D \leftrightarrow E= 4$	$AB \leftrightarrow DE = 6$ $C \leftrightarrow DE = 6$	ABC↔DE	ABCDE↔F		
Taxa joined	A and B	D and E	AB and C	ABC and DE	ABCDE and F		
Subtree	1 A 1 B	2 D 2 E	1 1 A 2 C	1 1 A 1 2 C 1 2 D E	1 1 A 1 2 C 1 2 D 4 F	Root 1 1 A 1 1 B 2 C 1 2 D 4 F	
Comments on tree drawing	The distance between A and B is 2 units. A sub- tree is drawn with the branch point halfway between the two. Thus, each branch is 1 unit in length.	Branching done as in Step 1. Because the distance from AB to C is also 4, that pair could have been selected as well.	First a subtree is drawn with AB and C: 2 AB C The the AB subtree is attached to the AB branch at a point equal to the length of the A and B branches.	The tree is first done as in Step 3 with the ABC and DE subtrees replacing the branches.	The tree is now complete but unrooted.	The tree can then be rooted using midpoint rooting which tries to balance all the tips to reach the same end point. Note this is the tree that we started with to build the distance matrix.	

### Compare to True Tree



#### But ...

#### • What is evolutionary rates not equal



# UPGMA with Unequal rates

TABLE 27.10. UPGMA tree construction errors							
	Cycle 1	Cycle 2	Cycle 3	Cycle 4	Cycle 5	Cycle 6	
Distance matrix	A       B       C       D       E         B       5       -       -       -       -         C       4       7       -       -       -         D       7       10       7       -       -         E       6       9       6       5       -         F       8       11       8       9       8	ACBDEB4D710-E695-F81189	ACBDEB4-DE6.59.5F811	ACB DE DE <mark>8</mark> F 9.5 9.5	ABCDE F 9	No new matrix	
ldentify smallest D	$A \leftrightarrow C = 4$	$D \leftrightarrow E = 5$	$AC \leftrightarrow B = 4$	$ACB \leftrightarrow DF = 8$	$ABCDE \leftrightarrow F = 9$		
Taxa joined Subtree	A and C $ \begin{array}{c} 2 \\ 2 \\ C \end{array} $	D and E 2.5 2.5 E	A and C with B 1  2  A 2  C 3  B	ABC with DE	ABCDE with F	$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & B \\ 1 & 2 & C \\ 1 & 2 & C \\ 1 & 2 & E \\ 4 & F \end{bmatrix}$	
Comments					F Note how this is not the same as the starting tree		

#### Compare to True Tree



# Neighbor Joining

## Start with Star Tree



# Calculate S

- For each OTU, calculate a measure (S) as follows.
- S is the sum of the distances (D) between that OTU and every other OTU, divided by N-2 where N is the total number of OTUs.
- This is a measure of the distance an OTU is from all other OTUs

# Calculate Pairwise Distances

• Calculate the distance D<sub>ij</sub> between each OTU pair (e.g., I and J).

# Identify Closest Pair

• Identify the pair of OTUs with the minimum value of  $D_{ij} - S_i - S_j$ 

# Joining Taxa

• As in UPGMA, join these two taxa at a node in a subtree.

# Calculate Branches

- Calculate branch lengths. Unlike UPGMA, neighbor joining does not force the branch lengths from node X to I (D<sub>xi</sub>) and to J (D<sub>xj</sub>) to be equal, i.e., does not force the rate of change in those branches to be equal. Instead, these distances are calculated according to the following formulas
- $D_{xi} = 1/2 D_{ij} + 1/2 (S_i S_j)$
- $D_{xj} = 1/2 D_{ij} + 1/2 (S_j S_i)$

# Calculate New Matrix

• Calculate a new distance matrix with I and J merged and replaced by the node (X) that joins them. Calculate the distances from this node to the other tips (K) by:

• 
$$D_{xk} = (D_{ik} + D_{jk} - D_{ij})/2$$

### Distance Matrix



- $S_A = (5+4+7+6+8) / 4 = 7.5$
- $S_B = (5+7+10+9+11) / 4 = 10.5$
- $S_C = (4+7+7+6+8) / 4 = 8$
- $S_D = (7+10+7+5+9) / 4 = 9.5$
- $S_E = (6+9+6+5+8) / 4 = 8.5$
- $S_F = (8+11+8+9+8) / 4 = 11$

# M

•  $M_{ij} = D_{ij} - S_i - S_j$ 

- Smallest are
- $M_{AB} = 5 7.5 10.5 = -13$
- $M_{DE} = 5 9.5 8.5 = -13$
- Choose one of these (AB here)

# New Tree

- Create a node (U) that joins pair with lowest  $M_{ij}$  such that
- $S_{IU} = D_{IJ}/2 + (S_I S_J)/2$
- Merge U with other taxa
- Join I and J according to S above and make all other taxa in form of a star



# New Matrix

•  $D_{XU} = D_{IX} + D_{JX} - D_{IJ}$  where I and J are those selected from above.
TABLE 27.11. Neighbor-joining example					
	Cycle 1	Cycle 2	Cycle 3	Cycle 4	Cycle 5
Distance matrix	A     B     C     D     E       B     5     -     -     -       C     4     7     -     -       D     7     10     7     -       E     6     9     6     5       F     8     11     8     9     8	U <sub>1</sub> C D E C 3 D 6 7 E 5 6 5 F 7 8 9 8	$\begin{array}{c cccc} U_1 & C & U_2 \\ C & 3 & & \\ U_2 & 3 & 4 & \\ F & 7 & 8 & 6 \end{array}$	$ \begin{array}{cccccccc} U_2 & U_3 \\ U_3 & 2 \\ F & 6 & 6 \end{array} $	U <sub>4</sub> F 5
Step 1					
S calculations $S_x = (\text{sum all } D_x)/(N-2),$ where N is the # of OTUs in the set.	$\begin{split} S_{\rm A} &= (5{+}4{+}7{+}6{+}8)/4 = 7.5\\ S_{\rm B} &= (5{+}7{+}10{+}9{+}11)/4 = 10.5\\ S_{\rm C} &= (4{+}7{+}7{+}6{+}8)/4 = 8\\ S_{\rm D} &= (7{+}10{+}7{+}5{+}9)/4 = 9.5\\ S_{\rm E} &= (6{+}9{+}6{+}5{+}8)/4 = 8.5\\ S_{\rm F} &= (8{+}11{+}8{+}9{+}8)/4 = 11 \end{split}$	$S_{U_1} = (3+6+5+7)/3 = 7$ $S_C = (3+7+6=8)/3 = 8$ $S_D = (6+7+5+9)/3 = 9$ $S_E = (5+6+5+8)/3 = 8$ $S_F = (7+8+9+8)/3 = 10.6$	$S_{U_1} = (3+3+7)/2 = 6.5$ $S_C = (3+4+8)/2 = 7.5$ $S_{U_2} = (3+4+6)/2 = 6.5$ $S_F = (7+8+6)/2 = 10.5$	$S_{\cup 2} = (2+6)/1 = 8$ $S_{\cup 3} = (2+6)/1 = 8$ $S_{\rm F} = (6+6)/1 = 12$	Because N – 2 = 0, we cannot do this calculation.
Step 2					
Calculate pair with smallest ( <i>M</i> ), where $M_{ij} = D_{ij} - S_i - S_j$ .	Smallest are $M_{AB} = 5 - 7.5 - 10.5 = -13$ $M_{DE} = 5 - 9.5 - 8.5 = -13$ Choose one of these (AB here).	Smallest is $M_{CU_1} = 3 - 7 - 8 = -12$ $M_{DE} = 5 - 9 - 8 = -12$ Choose one of these (DE here).	Smallest is M <sub>CU1</sub> = 3 – 6.5 – 7.5 = –11	Smallest is $M_{U_2F} = 6 - 8 - 12 = -14$ $M_{U_3F} = 6 - 8 - 12 = -14$ $M_{U_2U_3} = 2 - 8 - 8 = -14$ Choose one of these ( $M_{U_2U_3}$ here).	
Step 3				2 9	
Create a node (U) that joins pair with lowest $M_{ij}$ such that $S_{I\cup} = D_{ij}/2 + (S_i - S_j)/2$ .	U <sub>1</sub> joins A and B: $S_{AU1} = D_{AB}/2 + (S_A - S_B)/2 = 1$ $S_{BU1} = D_{AB}/2 + (S_B - S_A)/2 = 4$	U <sub>2</sub> joins D and E: $S_{DU2} = D_{DE}/2 + (S_D - S_E)2 = 3$ $S_{EU2} = D_{DE}/2 + (S_E - S_D)/2 = 2$	U <sub>3</sub> joins C and U <sub>1</sub> : $S_{CU_3} = D_{CU_1}/2 + (S_C - S_{U_1})/2 = 2$ $S_{U_1U_3} = D_{CU_1}/2 + (S_{U_1} - S_C)/2 = 1$	U <sub>4</sub> joins U <sub>2</sub> and U <sub>3</sub> : $S_{U_2U_4} = D_{U_2U_3}/2 + (S_{U_2} - S_{U_3})/2 = \frac{1}{2}$ $S_{U_3U_4} = D_{U_2U_3}/2 + (S_{U_3} - S_{U_2})/2 = \frac{1}{2}$	For last pair connect 1 with branch = D. 1 Here $D_{\cup 4\mathrm{F}}$ = 5.
Step 4					
Join <i>i</i> and <i>j</i> according to <i>S</i> above and make all other taxa in form of a star. Branches in black are of unknown length. Branches in red are of known length.	C = C = B C = A $C = BC = BC = BC = BC = BC = AB = A$	$ \begin{array}{c} C \\ D \\ 3 \\ U_2 \\ U_1 \\ 1 \\ A \end{array} $	$ \begin{array}{cccc} D & & C & B \\  & & U_2 & 2 & U_1 & 4 \\  & & U_2 & 2 & U_1 & 1 \\  & & E & 2 & U_3 & 1 & A \\  & & & F & F \end{array} $	$ \begin{array}{ccccc} D & 3 & C & 4 \\ & & 2 & 2 & 0 \\ & & & 2 & 0 \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & $	$ \begin{array}{cccc} D & 3 & C & B \\ & & & 2 & U_1 & 4 \\ & & & U_2 & U_3 & U_1 & 1 \\ & & & E & 2 & U_4 & A \\ & & & 5 & F \\ \end{array} $
Step 5	F	F			Comments
Calculate new distance					Note this is the same

tree we started with

(drawn in unrooted

form here).

Calculate new distance matrix of all other taxa to U with  $D_{x\cup} = D_{ix} + D_{jx} - D_{ij}$ , where *i* and *j* are those selected from above.

## Compare to True Tree



## Long branch attraction



## Rooting



## Rooting TOL Review

