EVE 161: Microbial Phylogenomics

Class #4: Phylogeny

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

Luryarchaea

EVE 161: Microbial Phylogenomics

Class #5: Phylogeny

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

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

isthokonts

Phylogeny

Desulfurococcus Aeropyrum Pyrobaculum rnermofimu Archaeoglobus Halophiles

Euryarchaea

Parts of a phylogenetic tree



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

Sister Taxa Relatedness

Sequence Alignment

| MSQNSL MSKLKEKREK MAEE MAEE | MAIDENK RLYEDKSVDK MSKLAEK MSKLAEK MAIDEDK MAINTDTSGK MAGTDR MAGTDR MANIDKDK AVVGIERASK MSVPDR 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 EEAI ELARVQ KRAL EAA I AV L KAL QAAMDK L QAL RMATEK KSYL EKAL KR SKAI ETAMSQ MKAL EYAL SS | I EKQFGKGS I I EKQFGKGSV I ERSFGKGSV I DKVFGKGAL I ERSFGKGAL I ERSFGKGAI I EKQFGKGSV I EKQFGKGSV I EKAFGKGSI I EKSFGKGSI I EKTFGKGAI I EKAFGKGSI I EKAFGKGSI I EKRFGKGSI | 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 - MIL 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 | |
|--|--|---|---|--|---|--|
| DVETISTGSL AVEVIPTGSL EIETISTGSL KVAVIPSGSV KIDAISTGSL RVETISTGAL PIEVIPTGST RTSTVPSGSL QMDAVSTGCL GIKSMSSGSI EISTTKTGAL QVEVIPTGSI DVSVIPSGSI | SLDIALGAGG MLDIALGIGG GLDIALGVGG GVDRALGVGG GLDLALGIGG TLDLALG-GG ALDVALGVGG ALDTALGIGG VLDEALGIGG SLDLALGVGG GLDLALGVGG ALNAALGVGG | L P MGRIVEIV L PKGRVVEIV V PKGRIIEIV VPRGRVVEVF VPKGRIIEIV I PRGRVVEVV VPRGRIIEIV VPKGRIIEIV VPRGRIIEIV VPRGRIIEIF VPKGRIIEIV VPRGRIIEIV | GPESSGKTT L GPESSGKTT L GPESSGKTT L GPESSGKTT L GPESSGKTT V GPESSGKTT V GPESSGKTT V GPESSGKTT L GPESSGKTT L GPESSGKTT L GPESSGKTT L GPESSGKTT L | TLQVIAAAQR TLQAIAECQK ALQTIAESQK TLHAIAQVQA SLHIIAECQK ALHAIAEVQK TLHAVAAAQK ALHVIAAAQQ ALHVVAAAQK TLQAIAEVQK ATHIVANAQK AIHAIAEAQK ALHAIAEAQK ALHAIAEAQK | 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 | |
| EVETIPTGSI | SLDIAEGVGG SLDIATGVGG | I PKGRITEI | GPESGKTTL GVESSGKTTL | ALH <mark>VIACAOK</mark> | Aquifex pyrophilus | |
| EGKTCAFIDA LGGTAAFIDA KGGICAFVDA AGGVAAFIDA NGGVCAFIDA EGGIAAFVDA | EHAL DP I YAR EHAL DP I YAA EHAL DP VYAR EHAL DVSYAR EHAL DVHYAK EQAL DP TYAS | KLGVDIDNLL KLGV <mark>NV</mark> DDLL KLGVDLQNLL KLGVRVEELL RLGVDTQNLL ALGVDIQNLL | CSQPDTGEQA LSQPDTGEQA ISQPDTGEQA VSQPDTGEQA VSQPDTGEQA VSQPDTGESA | LEI CDALARS LEI ADMLVRS LEI TDTLVRS LEI TEHLVRS LEI LET ITRS LEI VDQLV S | Escherichia coli Xanthomonas campestris Rhizobium phaseoli Myxococcus xanthus Helicobacter pylori Anabaena variabilis Stantomycos lividops | |
| Q- RTSAFIDA LGGAAAVIDA EGGIAAFIDA MGGVAAVIDA | EHAL DP VYAQ EHAL DP VYAK EHAL DP VYAK EHAL DP VYAK | KLGVDIDNLI KLGVNIEELL RLGVNIDDLV ALGVNVAELW LIGANINDLM | L SQPDTGEQ A VSQPDTGEQ A L SQPDTGEQ A I SQPDTGEQ A | LEI AEALVRS LEI TEALVRS LEI AEALI RS LSI AEALARS | Bacillus subtilis Clostridium perfringens Borrelia burgdorferi Chlamydia trachomatis | |
| AGGLAAFIDA AGGLAA <mark>I</mark> IDA MGGVAAFIDA AGG <mark>TC</mark> AFIDA RGGVA <mark>V</mark> FIDA | EHAFDRFYAA EHAFDRTYAE EHALDPVYAK EHALDPVYAR EHALDPKYAK | KLGVDVDNLF KLGV <mark>NV</mark> DNLW NLGVDLKSLL ALGVNADELL KLGVDVDNLV | I SQPDNGEQ A I SQPDNGEQ A I AQPDHGEQ A VSQPDNGEQ A I SQPDVGEQ A | LETAEQLIRS LETAEQLIRS LET <mark>VDE</mark> LVRS LET <mark>ME</mark> LLVRS LETAE <mark>SLTN</mark> S | 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

Structure Guided 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

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

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



Many possible trees

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

| Таха | Rooted Trees ^a | Unrooted Trees ^b |
|------|---------------------------|-----------------------------|
| | | |
| 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 |

^a $N_{\rm r} = (2n-3) \times (2n-5) \times (2n-7) \times \cdots \times 3 \times 1 = (2n-3)!/[2_{n-2} \times (n-2)!].$ ^b $N_{\rm u} = (2n-3) \times (2n-5) \times (2n-7) \times \cdots \times 3 \times 1 = (2n-5)!/[2_{n-3} \times (n-3)!].$

Other Tree Related Issues



Bootstrapping



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

| Species | 1 | ATGTT | GGA | TG | GT | ' <mark>GA</mark> T |
|---------|---|-------|-----|----|----|---------------------|
| Species | 2 | ATGTT | GGA | AG | GA | .G <mark>A</mark> A |
| Species | 3 | ATGTT | AGG | AG | AA | .G <mark>A</mark> A |
| Species | 4 | ATGTC | AGC | AC | CC | G <mark>C</mark> C |

Bootstrapping alignment #1

Species 1 Species 2 Species 3 Species 4 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

A

| Protein | 1 | DEMGLGKKRVESTR |
|---------|---|--------------------|
| Protein | 2 | DEMGVGKRGHLESRK |
| Protein | 3 | DDMGLGRWKRVESTE |
| Protein | 4 | DDMGLGHKKKVDSTK |
| Mask | | 111111100000011111 |
| | | |

В

| Protein | 1 | DEMGLGKKVESTR |
|---------|---|---------------|
| Protein | 2 | DEMGVGKRLESRK |
| Protein | 3 | DDMGLGRKVESTE |
| Protein | 4 | DDMGLGHKVDSTK |

| u) | | |
|--|--|--|
| axon | | |
| Fu | Nosema 40928 | OFGLESPEETRASSVALTEYPETLENGVEKESGLVCAGHEGHTELV |
| Fu | Aspergillus | OFGLESPEEIKRMSVUHVEVPETMDEORORPRTKGLECPGHEGHTELA |
| An | Plasmodium. 3 | ELGVLDPETIKKISVCEUVNVDIVKDGFREGGLVCPGHEGHIELA |
| An | Cricetulus.2 | OFGVI.SPDELKRMSVTEGGIKYPETTEGGREKLGGLECPGHEGHIELA |
| An | Homo, 7434727 | OFGVI.SPDELKRMSVTEGGIKYPETTEGGREKLGGLECPGHEGHIELA |
| An | Drosophila.9 | OFGILSPDEIRRMSVTEGGVOFAETMEGGREKLGGLECPGHEGHIDLA |
| An | Celegans, 133 | OFGILGPEEIKRMSVAHVEFPEVYENGKEKLGGLDCPGHEGHLELA |
| Fu | Spombe, 54881 | OFGILSPEEIRSMSVAKIEFPETMDESGORERVGGLDCPGHEGHIELA |
| P1 | Athaliana.40 | OFGILSPDEIROMSVIHVEHSETTEKGKEKVGGLECPGHEGYLELA |
| Mv | Ddiscoideum. | |
| Rh | Porphyra.316 | ECPGHFGFTELA |
| Kt. | Tbrucei.1021 | OFETEKEROTKSYAVCLVEHAKSYANAADOSGEAECPGHEGYTELA |
| K+ | Leishmania.7 | OFEVEKEAOIKAYAKOITEHAKSYEHGOFVRGGIECPGHEGYVELA |
| NJ | | |
| ayon | | Parabasalids |
| axon | leariids | ···· ····10··· ···20··· ···30··· ···40··· ···· |
| axon Fu | Nosema.40928 | QFGLFSPEEIRASSVALIRYPETLE-NGVPKESGLVCAGHFGHIEL |
| Fu Fu | Nosema.40928 Aspergillus. | QFGLFSPEEIRASSVALIRYPETLE-NGVPKESGLVCAGHFGHIELY QFGLFSPEEIKRMSVVH-VEYPETMDEQRQRPRTKGLECPGHFGHIELY |
| Fu Fu Fu Fu | Nosema.40928 Aspergillus. Spombe.54881 | 10 20 30 40 QFGLFSPEEIRASSVALIRYPETLENGVPKESGLVCAGHFGHIEL QFGLFSPEEIKRMSVVHVEYPETMDEQRQRPRTKGLECPGHFGHIEL QFGILSPEEIRSMSVAKIEFPETMDESGQRPRVGGLDCPGHFGHIEL |
| Fu Fu Fu Fu Ap | Nosema.40928 Aspergillus. Spombe.54881 Plasmodium.3 | 10 20 30 40 QFGLFSPEEIRASSVALIRYPETLENGVPKESGLVCAGHFGHIELM QFGLFSPEEIKRMSVVHVEYPETMDEQRQRPRTKGLECPGHFGHIELM QFGILSPEEIRSMSVAKIEFPETMDESGQRPRVGGLDCPGHFGHIELM ELGVLDPEIIKKISVCEIVNVDIYKDGFPREGGLYCPGHFGHIELM |
| Fu Fu Fu Ap An | Nosema.40928 Aspergillus. Spombe.54881 Plasmodium.3 Cricetulus.2 | 10 20 30 40 QFGLFSPEEIRASSVAI IRYPETLENGVPKESGLVCAGHFGHIELZQFGLFSPEEIKRMSVVHVEYPETMDEQRQRPRTKGLECPGHFGHIELZQFGILSPEEIRSMSVAKIEFPETMDESGQRPRVGGLDCPGHFGHIELZELGVLDPEIIKKISVCEIVNVDIYKDGFPREGGLYCPGHFGHIELZQFGVLSPDELKRMSVTEGGIKYPETTEGGRPKLGGLECPGHFGHIELZ |
| Fu Fu Fu Ap An An | Nosema.40928 Aspergillus. Spombe.54881 Plasmodium.3 Cricetulus.2 Homo.7434727 | 10 20 30 40 2QFGLFSPEEIRASSVAL IRYPETLENGVPKESGLVCAGHFGHIELMQFGLFSPEEIRRMSVVHVEYPETMDEQRQRPRTKGLECPGHFGHIELMQFGILSPEEIRSMSVAKIEFPETMDESGQRPRVGGLDCPGHFGHIELMELGVLDPEIIKKISVCEIVNVDIYKDGFPREGGLYCPGHFGHIELMQFGVLSPDELKRMSVTEGGIKYPETTEGGRPKLGGLECPGHFGHIELMQFGVLSPDELKRMSVTEGGIKYPETTEGGRPKLGGLECPGHFGHIELM |
| Fu Fu Fu Ap An An An | Nosema.40928 Aspergillus. Spombe.54881 Plasmodium.3 Cricetulus.2 Homo.7434727 Drosophila.9 | 10 20 30 40 QFGLFSPEEIRASSVAI IRYPETLENGVPKESGLVCAGHFGHIELM QFGLFSPEEIKRMSVVHVEYPETMDEQRQRPRTKGLECPGHFGHIELM QFGILSPEEIRSMSVAKIEFPETMDESGQRPRVGGLDCPGHFGHIELM ELGVLDPEIIKKISVCEIVNVDIYKDGFPREGGLYCPGHFGHIELM QFGVLSPDELKRMSVTEGGIKYPETTEGGRPKLGGLECPGHFGHIELM QFGVLSPDELKRMSVTEGGIKYPETTEGGRPKLGGLECPGHFGHIELM QFGILSPDEIRRMSVTEGGVQFAETMEGGRPKLGGLECPGHFGHIELM |
| Fu Fu Fu Ap An An An An | Nosema.40928 Aspergillus. Spombe.54881 Plasmodium.3 Cricetulus.2 Homo.7434727 Drosophila.9 Celegans.133 | 10 20 30 40 QFGLFSPEEIRASSVAL IRYPETLENGVPKESGLVCAGHFGHIELM QFGLFSPEEIKRMSVVHVEYPETMDEQRQRPRTKGLECPGHFGHIELM QFGILSPEEIRSMSVAKIEFPETMDESGQRPRVGGLDCPGHFGHIELM ELGVLDPEIIKKISVCEIVNVDIYKDGFPREGGLYCPGHFGHIELM QFGVLSPDELKRMSVTEGGIKYPETTEGGRPKLGGLECPGHFGHIELM QFGILSPDEIRRMSVTEGGVQFAETMEGGRPKLGGLECPGHFGHIELM QFGILSPDEIRRMSVTEGGVQFAETMEGGRPKLGGLECPGHFGHIDLM QFGILGPEEIKRMSVAHVEFPEVYENGKPKLGGLDCPGHFGHLELM |
| Fu Fu Fu Ap An An An An Pl | Nosema.40928 Aspergillus. Spombe.54881 Plasmodium.3 Cricetulus.2 Homo.7434727 Drosophila.9 Celegans.133 Athaliana.40 | 10 20304040 QFGLFSPEEIRASSVAI IRYPETLENGVPKESGLVCAGHFGHIELA QFGLFSPEEIKRMSVVHVEYPETMDEQRQRPRTKGLECPGHFGHIELA QFGILSPEEIRSMSVAKIEFPETMDESGQRPRVGGLDCPGHFGHIELA ELGVLDPEIIKKISVCEIVNVDIYKDGFPREGGLYCPGHFGHIELA QFGVLSPDELKRMSVTEGGIKYPETTEGGRPKLGGLECPGHFGHIELA QFGULSPDELKRMSVTEGGIKYPETTEGGRPKLGGLECPGHFGHIELA QFGILSPDEIRRMSVTEGGVQFAETMEGGRPKLGGLECPGHFGHIELA QFGILSPDEIRRMSVTEGGVQFAETMEGGRPKLGGLECPGHFGHIELA QFGILSPDEIRRMSVTEGGVQFAETMEKGKPKVGGLECPGHFGHLELA |
| Fu Fu Fu Ap An An An An Pl My | Nosema.40928 Aspergillus. Spombe.54881 Plasmodium.3 Cricetulus.2 Homo.7434727 Drosophila.9 Celegans.133 Athaliana.40 Ddiscoideum. | 10 203040 QFGLFSPEEIRASSVAL IRYPETLE NGVPKESGLVCAGHFGHIELM QFGLFSPEEIRRMSVVH VEYPETMDEQRQRPRTKGLECPGHFGHIELM QFGILSPEEIRSMSVAK IEFPETMDESGQRPRVGGLDCPGHFGHIELM ELGVLDPEIIKKISVCE IVNVDIYK DGFPREGGLYCPGHFGHIELM QFGVLSPDELKRMSVTEGGIKYPETTE GGRPKLGGLECPGHFGHIELM QFGVLSPDELKRMSVTEGGIKYPETTE GGRPKLGGLECPGHFGHIELM QFGILSPDEIRRMSVTEGGVQFAETME GGRPKLGGLECPGHFGHIELM QFGILSPDEIRRMSVAH VEFPEVYE NGKPKLGGLECPGHFGHIELM QFGILSPDEIRQMSVIH VEHSETTE KGKPKVGGLECPGHFGHIELM QFGILSPDEIRQMSVIH VEHSETTE KGKPKVGGLECPGHFGHIELM QFGILSPDEIRQMSVIH VEHSETTE KGKPKVGGLECPGHFGHIELM QFGILSPDEIRQMSVIH VEHSETTE KGKPKVGGLECPGHFGHIELM QFGILSPDEIRQMSVIH VEHSETTE KGKPKVGGLECPGHFGHIELM QFGILSPDEIRQMSVIH VEHSETTE KGKPKVGGLECPGHFGHIELM QFGILSPDEIRQMSVIH VEHSETTE KGKPKVGGLECPGHFGHIELM |
| Fu Fu Fu Ap An An An An Pl My Rh | Nosema.40928 Aspergillus. Spombe.54881 Plasmodium.3 Cricetulus.2 Homo.7434727 Drosophila.9 Celegans.133 Athaliana.40 Ddiscoideum. Porphyra.316 | 10 20 30 40 QFGLFSPEEIRASSVAI IRYPETLE NGVPKESGLVCAGHFGHIELMQFGLFSPEEIRRMSVVH VEYPETMDEQRQRPRTKGLECPGHFGHIELMQFGILSPEEIRSMSVAK IEFPETMDESGQRPRVGGLDCPGHFGHIELMELGVLDPEIIKKISVCE IVNVDIYK DGFPREGGLYCPGHFGHIELMQFGVLSPDELKRMSVTEGGIKYPETTE GGRPKLGGLECPGHFGHIELMQFGVLSPDELKRMSVTEGGIKYPETTE GGRPKLGGLECPGHFGHIELMQFGILSPDEIRRMSVTEGGVQFAETME GGRPKLGGLECPGHFGHIELMQFGILSPDEIRRMSVTEGGVQFAETME GGRPKLGGLECPGHFGHIELMQFGILSPDEIRRMSVTEGGVQFAETME GGRPKLGGLECPGHFGHIELMQFGILSPDEIRRMSVTEGGVQFAETME GGRPKLGGLECPGHFGHIELMQFGILSPDEIRRMSVTEGGVQFAETME GGRPKLGGLECPGHFGHIELMQFGILSPDEIRRMSVTEGGVQFAETME GGRPKLGGLECPGHFGHIELMQFGILSPDEIRRMSVTEGGVQFAETME GGRPKLGGLECPGHFGHIELMQFGILSPDEIRRMSVTEGGVQFAETME GGRPKLGGLECPGHFGHIELMQFGILSPDEIRQMSVIH VEFPEVYE NGKPKLGGLECPGHFGHIELMQFGILSPDEIRQMSVIH VEHSETTE KGKPKVGGLECPGHFGYLELM |
| Fu Fu Fu Fu Ap An An An An Pl My Rh Kt | Nosema.40928 Aspergillus. Spombe.54881 Plasmodium.3 Cricetulus.2 Homo.7434727 Drosophila.9 Celegans.133 Athaliana.40 Ddiscoideum. Porphyra.316 Tbrucei.1021 | 10 20 30 40 2QFGLFSPEEIRASSVAI IRYPETLE NGVPKESGLVCAGHFGHIELMQFGLFSPEEIRASSVAI VEYPETMDEQRQRPRTKGLECPGHFGHIELMQFGILSPEEIRSMSVAK IEFPETMDESGQRPRVGGLDCPGHFGHIELMELGVLDPEIIKKISVCE IVNVDIYK DGFPREGGLYCPGHFGHIELMQFGVLSPDELKRMSVTEGGIKYPETTE GGRPKLGGLECPGHFGHIELMQFGILSPDEIRRMSVTEGGIVQFAETME GGRPKLGGLECPGHFGHIELMQFGILSPDEIRRMSVTEGGVQFAETME GGRPKLGGLECPGHFGHIELMQFGILSPDEIRRMSVTEGGVQFAETME GGRPKLGGLECPGHFGHIELMQFGILSPDEIRRMSVTEGGVQFAETME GGRPKLGGLECPGHFGHIELMQFGILSPDEIRRMSVTEGGVQFAETME CGRPKLGGLECPGHFGHIELMQFGILSPDEIRRMSVTEGGVQFAETME CGRPKLGGLECPGHFGHIELMQFGILSPDEIRQMSVIH VEFPEVYE NGKPKLGGLECPGHFGHIELMQFGILSPDEIRQMSVIH VEHSETTE CGRPKLGGLECPGHFGYLELMQFGILSPDEIRQMSVIH VEHAKSYA NAADQSGEAECPGHFGYIELM |

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.

Luryarchaea

A Gene alignments **B** Concatenated alignment 3

Homoplasy

Α

Sequence 1 GACGAG Sequence 2 GCCGAC

В

Sequence 1Sequence 2GACGAGGCCGAC $\uparrow G \rightarrow Aat$ $\uparrow C \rightarrow Aat$ $\mid position 5$ $\uparrow C \rightarrow Aat$ GACGGG $\downarrow C \rightarrow Cat$ $\uparrow G \rightarrow Aat$ $\downarrow G \rightarrow Cat$ $\mid position 2$ $G \rightarrow Cat$ GGCGGG $G \rightarrow Cat$ $G \rightarrow Cat$ $\downarrow position 5$ $G \rightarrow Cat$ $\downarrow position 6$ $G \subseteq CGGG$ $G \subset CGGG$ $\downarrow constrained and constrained$

Orthologs and Paralogs



Fig. 3. The problem with paralogues. (a) Paralogous genes are created by gene duplication events. Gene X is duplicated in a common ancestor to species A and B resulting in two paralogous genes, X and X'. All subsequent species inherit both copies of the gene (unless one or the other is lost somewhere along the way). (b) Phylogenetic analysis of the X/X' gene family gives two parallel phylogenies. All sequences of gene X are orthologues of each other, and all the sequences of gene X' are orthologues of each other. However, X and X' are paralogues. Both the X and X' subtrees show the true relationships among the three species. The subtrees are also each other's natural outgroup, and as a result each subtree is rooted with the other (reciprocally rooting). (c) A tree of the X/X' gene family can be misleading if not all the sequences are included (because of incomplete sampling or gene loss). If the broken branches are missing, then the true species relationships are misrepresented.

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 | E | |
|------|----|---|---|---|--|
| C | 1 | | | | |
| | 4 | 6 | | | |
| D | 6 | 6 | | | |
| E | 6 | 6 | 4 | | |
| F | 8 | 8 | 8 | 8 | |
| | | | | | |

UPGMA

| TABLE 27.9. Example of UPGMA tree construction | | | | | | | |
|--|---|--|--|---|--|--|--|
| Step | Cycle 1 | Cycle 2 | Cycle 3 | Cycle 4 | Cycle 5 | Cycle 6 | |
| Distance matrix | OTUsABCDEB2C44D666E6664-F88888 | 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 | $ \begin{array}{c} 1 \\ 1 \\ 1 \\ 2 \\ C \end{array} $ | 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 A 1 2 C 1 2 D E 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 2 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_{ij} 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_{xi}) and to J (D_{xj}) 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 ₁ 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}{ccccccc} U_2 & U_3 \\ U_3 & 2 \\ F & 6 & 6 \end{array} $ | U ₄ F 5 | | | | |
| Step 1 | | | | | | | | | |
| S calculations $S_x = (\text{sum all } D_x)/(N-2),$ where N is the # of OTUs in the set. | $\begin{split} S_{\rm A} &= (5{+}4{+}7{+}6{+}8)/4 = 7.5\\ S_{\rm B} &= (5{+}7{+}10{+}9{+}11)/4 = 10.5\\ S_{\rm C} &= (4{+}7{+}7{+}6{+}8)/4 = 8\\ S_{\rm D} &= (7{+}10{+}7{+}5{+}9)/4 = 9.5\\ S_{\rm E} &= (6{+}9{+}6{+}5{+}8)/4 = 8.5\\ S_{\rm F} &= (8{+}11{+}8{+}9{+}8)/4 = 11 \end{split}$ | $S_{U_1} = (3+6+5+7)/3 = 7$ $S_C = (3+7+6=8)/3 = 8$ $S_D = (6+7+5+9)/3 = 9$ $S_E = (5+6+5+8)/3 = 8$ $S_F = (7+8+9+8)/3 = 10.6$ | $S_{U_1} = (3+3+7)/2 = 6.5$ $S_C = (3+4+8)/2 = 7.5$ $S_{U_2} = (3+4+6)/2 = 6.5$ $S_F = (7+8+6)/2 = 10.5$ | $S_{\cup 2} = (2+6)/1 = 8$ $S_{\cup 3} = (2+6)/1 = 8$ $S_{F} = (6+6)/1 = 12$ | 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 $M_{CU_1} = 3 - 6.5 - 7.5 = -11$ | 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 | | | | | | | | | |
| 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 ₁ joins A and B: $S_{AU_1} = D_{AB}/2 + (S_A - S_B)/2 = 1$ $S_{BU_1} = D_{AB}/2 + (S_B - S_A)/2 = 4$ | U ₂ joins D and E: $S_{DU_2} = D_{DE}/2 + (S_D - S_E)2 = 3$ $S_{EU_2} = D_{DE}/2 + (S_E - S_D)/2 = 2$ | U ₃ joins C and U ₁ : $S_{CU_3} = D_{CU_1}/2 + (S_C - S_{U_1})/2 = 2$ $S_{U_1U_3} = D_{CU_1}/2 + (S_{U_1} - S_C)/2 = 1$ | $U_4 \text{ joins } U_2 \text{ and } U_3:$ $S_{\cup_2 \cup_4} = D_{\cup_2 \cup_3}/2 + (S_{\cup_2} - S_{\cup_3})/2 =$ $S_{\cup_3 \cup_4} = D_{\cup_2 \cup_3}/2 + (S_{\cup_3} - S_{\cup_2})/2 =$ | 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 3 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. | $K \xrightarrow{D} \xrightarrow{C} \xrightarrow{A} \xrightarrow{B} \xrightarrow{B} \xrightarrow{A}$ | $ \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 \\ \end{array} $ | $ \begin{array}{cccccc} D & 3 & C & B \\ & & U_2 & U_3 & U_1 & 4 \\ & & E & 2 & U_4 & 1 \\ & & F & F \end{array} $ | $ \begin{array}{cccccc} D & 3 & C & B \\ & & & 2 & U_1 & 4 \\ & & & U_2 & U_3 & 1 \\ & & E & 2 & U_4 & A \\ & & & 5 & F \\ \end{array} $ | | | | |
| Step 5 | F | F | | | Comments | | | | |
| Calculate new distance matrix of all other taxa | | | | | Note this is the same tree we started with | | | | |

(drawn in unrooted

form here).

matrix of all other taxa to U with $D_{x\cup} = D_{ix} + D_{jx} - D_{ij}$,

 $D_{x\cup} = D_{ix} + D_{jx} - D_{ij}$, where *i* and *j* are those selected from above.

Compare to True Tree



Long branch attraction



Rooting



Rooting TOL Review



Other Tree Issues

- More characters vs. more taxa
- Lateral gene transfer
- Polytomies
- Paralogs and Orthologs
- Species Tree vs. Gene Tree
- Networks vs. Trees
- Clusters vs. Trees

Unrooted Tree of Life from Woese









Can You Tell Which Groups are Sister Taxa from this?



What is Missing?











Gene Duplications Prior to MRCA of All Life


Gene Duplications Prior to MRCA of All Life



Semi-resolved Phylogeny of Elongation Factors



If AB Together Should Be For Both EF2 and EFG



If AE Together Should Be For Both EF2 and EFG



WInner







Slides by Jonathan Eisen for BIS2C at UC Davis Spring 2016





Alternative Position of Eukaryote Branch





Slides by Jonathan Eisen for BIS2C at UC Davis Spring 2016

