

Lecture 3: A Brief Overview of Molecular Phylogeny

There are approximately a zillion books out there on phylogeny. For a step-by-step walk-through of sequence-based phylogenetic analysis, mostly using the popular program PAUP, you may want: Hall, B.G., “Phylogenetic Trees Made Easy,” Sinauer, 2007, 3rd edition.

For a great intro to bioinformatics in general: Claverie, J.-M. and Notredame, C., “Bioinformatics for Dummies”, 2nd Ed., 2007.

Some time before Midterm Exam I you should complete the **Molecular Phylogeny Workshop**, available on the Class Website. There are two parts to the Workshop, one to introduce you to the process of sequence alignment and provide you with a rudimentary sequence editor; and the second to guide you through a set of phylogenetic analyses for class discussion and for your edification in that process. As discussed in the Workshop write-up, you should **save hard copy of the trees** you cast and turn these in at the time of the Midterm Exam. **They will be worth 20 points on the exam.**

1. The steps in a molecular phylogenetic analysis:

- Determine (or obtain) sequences [more later on this]
- “Align” sequences - identify homologous nt (or aa)
- Perform tree calculations
- Test tree(s)

2. The alignment process: the goal is to identify homologous nt in a collection of sequences:

position n+:	1	2	3	4	5	6	7	8	9	10
seq AA	A	A	C	U	U	G	U	U	U.....
seq BA	C	A	C	U	U	G	U	G	U.....
seq CA	G	A	U	U	U	-	U	C	U.....

A. Columns of nt constitute a specific hypothesis - those nt are homologs and changes reflect evolution.

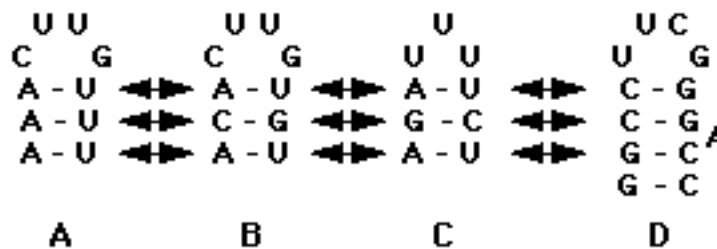
B. Note that seqs from different organisms may not be the same length - variation may be at the ends or internal. Hence use of “alignment gap,” or “indel” in seq. C above to bring that seq into continuity of alignment.

1. I.e., “Homologous” molecules are not necessarily “homologous” over their entire lengths, indeed length-variation of homologs in different critters is common.
2. There are automated protocols for sequence alignment (e.g. “clustal”), but they seldom do a perfect job unless seqs are highly similar. Generally, manual polishing is necessary, particularly if length variation is significant. Misalignment (or inclusion of non-homologous sequences) degrades the calculation – throws in random sequences.
3. In practice, don’t start at end of sequence and work forward: identify regions of clear homology and work out into regions of less clarity.
4. Note structural implications of the alignment process: homologous residues are expected to occur with the same spatial constraints - be in the same place/structure in the corresponding molecules in different organisms. You can predict structure from homology! (E.g. modeling tertiary structure by “threading” sequence onto a homolog with known structure.)
3. In the case of structured RNAs (or proteins), complementarities can often be used to establish register of the alignment.

For instance, how would you align these homologous blocks of sequence?

position n+:	1	2	3	4	5	6	7	8	9	10	11	12	13
seq AA	A	A	C	U	U	G	U	U	U		
seq BA	C	A	C	U	U	G	U	G	U		
seq CA	G	A	U	U	U	U	C	U			
seq DG	G	C	C	U	U	C	G	G	G	A	C	C

Not at all obvious, from sequence alone, but if you know something about the secondary structure of the RNA (paired regions), and look for paired elements:



alignment becomes straightforward:

position n+:	1	2	3	4	5	6	7	8	9	10	11	12	13	
seq A-	A	A	A	C	U	U	G	U	U	-	U	-..	
seq B-	A	C	A	C	U	U	G	U	G	-	U	-..	
seq C-	A	G	A	U	U	U	-	U	C	-	U	-..	
seq D	G	G	C	C	U	U	C	G	G	G	A	C	C

note convention:

- = alignment gap; no nt

• = nt present, but unknown

4. In the case of rRNAs, the high degree of conservation makes the secondary structures important alignment tools – (see secondary structures of rRNAs-available on Class Website).

A. Paired regions in the SSU rRNAs are not conjecture based on the occurrence of complements, but are “proven” by covariations in the sequence set that maintain the complementarity.

3. Calculation of trees: there are many ways to do this. Popular methods include trees based on:

- Evolutionary distance
- Maximum parsimony
- Maximum likelihood

A. In each case, computer-search is used to find “the tree” most consistent with the data set.

B. All the methods rely heavily on statistical analysis. One consideration in these calculations or any “discrete number” assessments is “Poisson statistics

--> Vignette: **Poisson Statistics** ←

4. Evolutionary distance:

- A. This method makes a “map” based on pairwise “**evolutionary distance**,” the number of sequence changes between all pairs of sequences (organisms) in the data sets.
- B. Recall that the number of differences you count between seqs is less than the number of changes that occurred - back mutations and multiple mutations.
- C. This can be estimated (from Poisson counting statistics) for any position as “Knuc,” the average extent of sequence change at any position in two homologous seqs:

$$K_{nuc} = -3/4[\ln(1-[4/3]D)]$$

where D=fractional difference in compared seqs.

- 1. For instance, between human and *E.coli* SSU rRNAs you *count* 50% difference=on average 0.5 changes/nt. The “real” extent is calculated from the expression as:

$$K_{nuc} = -3/4[\ln(1-[4/3]D)] = -3/4\ln(1-2/3) = -3/4\ln 0.33 = 0.825$$

So, more than half-again the changes you count “actually” occurred!

- 2. ”Evolutionary distance” (Knuc) is the calculated number of changes, not the number you count.
- 3. Note that this assumes that all positions in a molecule change at the same rate, which they do not. Some methods estimate the rate of change at each position (based on the data-set), and use that rate in the above calculation. In essence, however, change in sequences in the past are estimable, but fundamentally unknowable.
- 4. These “hidden” changes make treeing between the domains a chancy business.
- 5. Even between the bacterial phyla:

- a. Typical bacterial phylum-level differences (counts) are ca. 25% (75% identity), so:

$$D = 0.25$$

$K_{nuc}=0.3$: “only” about 15% ($0.05/0.3$) of your calculation-basis is inferred.

- b. Since the K_{nuc} calculation is non-linear (log function), the deeper you go in a tree, the more shaky your branching orders.

- D. An important concept in any tree construction is the amount of sequence that you use - more residues is better. The standard deviation as a sequence difference count can be estimated:

$$\text{Std. deviation} = \frac{3}{4} \sqrt{\frac{(D)(I)L}{1-D}}$$

D=fraction differences

I=fraction identities(1-D)

L=number of residues

e.g. for 50% differences and 1000 nt:

$$\text{std. deviation} = \frac{3}{4} \sqrt{\frac{(0.5)(0.5)/1000}{0.5-0.25}} = 0.024 \text{ position}$$

2xS.D. = ca. 50 positions, 5% of total positions counted

(Statistically, 95% of instances will fall within 2xS.D.)

but if you use only 100nt:

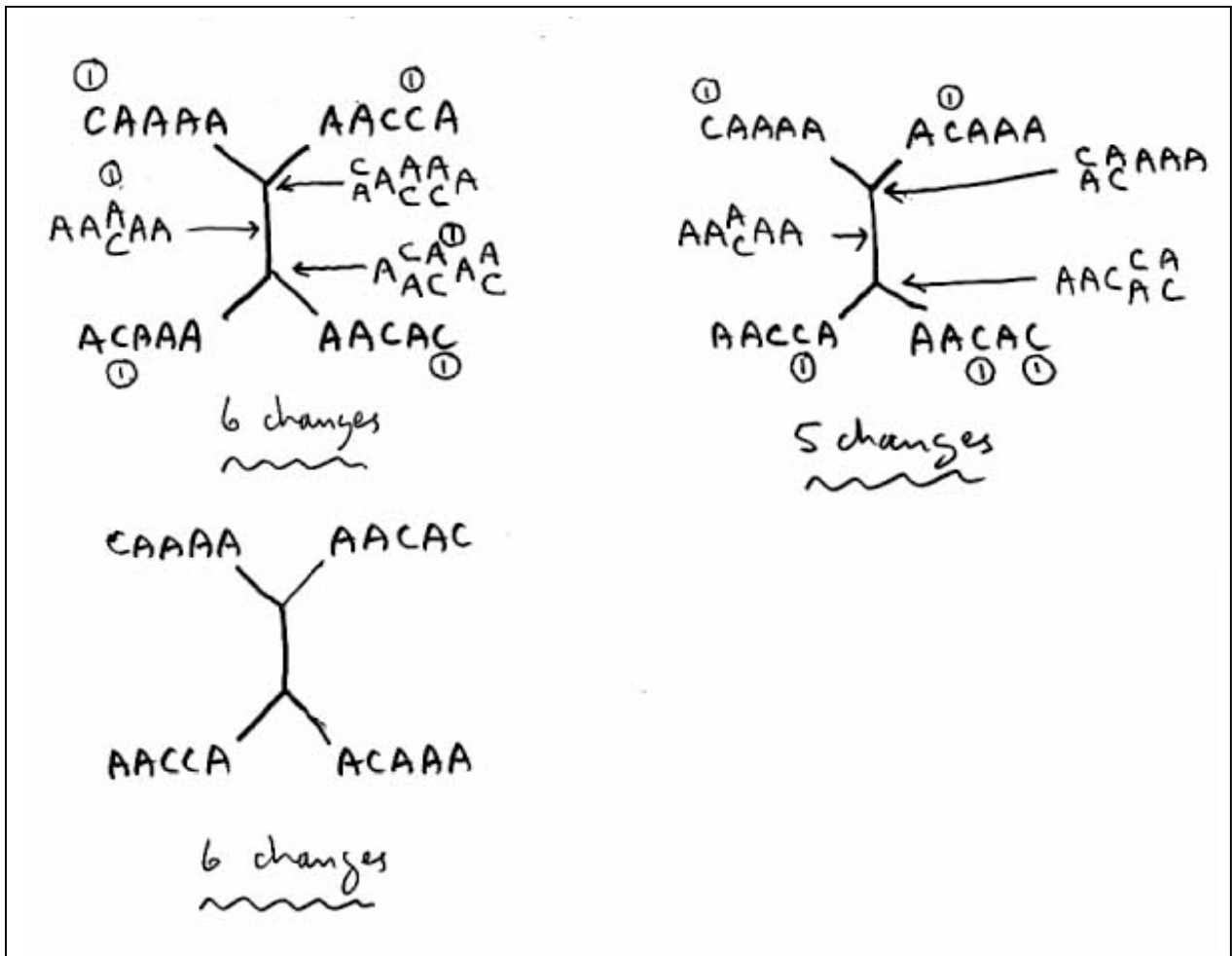
std. dev. = 0.075: 2xS.D = 15% of total is getting to be *not* very good - you can't rely on your counts to produce reliable trees!

- E. Computer programs use evolutionary distances to construct a tree most consistent with all the pairwise distances. Since biology is seldom regular, there is no single solution to all the pairwise distances. More later on what to do about that.

5. “**Parsimony**” methods of tree construction presume that the evolutionary path follows the fewest number of changes: the “correct tree” involves the fewest changes required to construct that particular topology.

A. Often called “ancestral sequence” methods, since inferred ancestral seqs are used to count changes.

B. E.g. there are 3 possible topologies of the following 4 seqs:



In this “heuristic” method (hunt best tree by testing alternatives, choosing “optimal” path in a succession of steps), all possible trees are examined and the “best” chosen by the least number of changes.

6. **“Maximum Likelihood”** methods are not intuitively interpreted; they calculate the probability that each node in a proposed tree (the heuristic search) is consistent with the particular data set. Statisticians consider this method to be the most “robust” (least sensitive to idiosyncrasies injected by a particular sequence) of any of the treeing methods; certainly it is the most statistically valid, since it is statistics-based.

A. Several other probabilistic methods are around, e.g. “Bayesian methods” that calculate the probability of a tree topology based on the data (<http://mrbayes.csit.fsu.edu/>).

B. Phylogeneticists often argue about the ‘best’ method for phylogenetic analysis, but they all work about equally well given the constraints of the particular method and with appropriate ‘corrections’ for variable rates, base compositions, etc. The best approach, as usual in science, is all of the above.

7. How to validate a particular tree (topology)?

A. Construct tree with different taxa; since the tree is dependent on which seqs you include, use of a different suite of seqs. will test whether associations observed in one tree are consistent in the context of different taxa.

B. Use all methods available to test specific associations of particular seqs of interest.

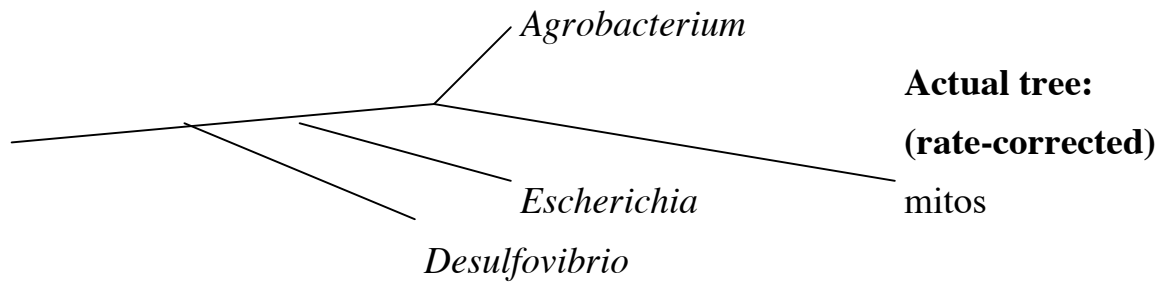
C. ‘Bootstrap’ analysis: resolve tree many times using random subsets of data set. E.g., compile data set for each analysis by drawing alignment columns from the data set at random, with replacement, to compile data set for each tree calculation.

1. This tests anomalies in tree calculations and to some extent whether particular sequence blocks in an alignment cause weird behavior.

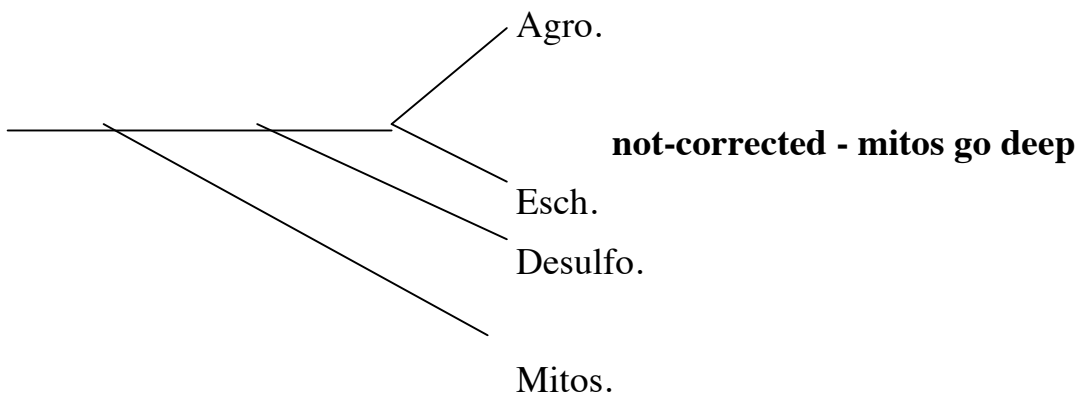
2. E.g. of ‘Bootstrap support’ for BigTree nodes: Support for nodes are marked: ML/parsimony; note different results with different methods. that is, a particular node is observed in ##% of bootstrap trees.

(Note that many peripheral nodes have funky bootstrap values because too many ‘outgroups’ included - below ‘Troubles’.)

e.g. with mitochondria:



if you don't correct for "fast clock" of mitos, they try to get away from otherwise close relatives, to jump deeper into tree.



1. Such rate-effects can be compensated to some extent by adding more sequences to trees, and by using systematic rate-correction calculations, but variable rates is a big problem.
2. This phenomenon has been called "long branch attraction," but it is really "short branch rejection of long branches."
3. Misalignment creates long branches, probably the most common mistake for neophytes and too many pros. Note that machine alignments (e.g. BLAST) commonly MISalign non-homologous stretches – any final alignment process is best manually guided if the sequence representation is broad in diversity. With rRNA seqs the alignment is helped LOTS by structural elements that serve as landmarks.

(For LOTS on alignment issues: Korf et al. "BLAST" (O'Reilly, 2003, 339 pp.)

B. Base-composition biases.

1. Variation in genomic G+C composition is reflected in all genes, and can make seqs. behave spuriously as a function of the organism compositions of trees. To test the problem, do "transversion analysis:"

Count only changes that are "transversions."

Transversion= $R \Leftrightarrow Y$

Transition= $R \Leftrightarrow R$ ($A \Leftrightarrow G$)

$Y \Leftrightarrow Y$ ($T \Leftrightarrow C$)

R=purine

Y=pyrimidine

Hence, you don't see G+C vs. A+T difference. (and you lose a lot of data, about half)

2. What causes variation in genome/rRNA base composition (genomes are more radically different in base comp than rRNA seqs?)

C. Taxon selection:

1. Taxa included in a tree have influence on topology of any particular association because of random similarities/ dissimilarities - to test any association or branching order, solve tree with different suites of taxa.
 2. In general, minimize the number of "outgroup" sequences used to "root" a cluster of interest.
9. "Signature Analysis" - use of simple features to test a tree or assign seqs. to "clades" (relatedness groups)

A. This is just like use of morphological or other qualities to identify taxa: the properties are e.g.:

1. Occurrence of specific nt or sequences at particular positions.

- 2, Occurrence (or lack) of structural elements (e.g. helices) at particular positions. Note that LACK of a property is not a specific property. It is only useful in comparison to presence of the property.
3. Or any other diagnostic feature.

B. Woese used oligonucleotide signatures to first define the clades Eucarya/Archaea/Bacteria.

- 1."oligonucleotide" = a short seq, usually <10-20 nt.
- 2.However, domain clades are even evident with single nt distribution (below).
Similar sorts of compendia could be produced for distinguishing (assignment to) any real clade, and can be very useful for testing controversial tree solutions.
- 3.Anyone interested in assembling a phylum-level signature table? It needs to be done with the bacterial phylogenetic divisions. It can't be done with unique signatures, probably, but perhaps with a progressive key

Nucleotide signatures were used early-on to bolster the concept of 3 domains (Woese)

position in se- quence*	All Archaea	All Bacteria	All Eucarya	position in se- quence	Archaea	Bacteria	Eucarya
113	C	G	C	962	G	C	U
314	G	C	G	966	U	G	U
338	G	A	A	973	C	G	G
339	G	C	C	1016	G	A	A
358	G	U	G	1060	C	U	C
377	C	G	C	1087	C	G	U
386	G	C	G	1098	G	C	G
399	C	G	-	1110	G	A	G
403	A	C	-	1197	G	A	G
507	G	C	G	1211	G	U	U
585	C	G	U	1212	A	U	A
675	U	A	U	1229	G	A	G
716	C	A	C	1381	C	U	C
756	G	C	A	1393	C	U	U
923	G	A	A	1415	C	G	C
952	C	U	C	1485	G	U	G

*Eco numbering - position in Eco SSU rRNA sequence in the alignment; not necessarily the same in all rRNAs.

10. Discussion of Workshop results.