Introduction to Phylogenies: Distance Methods

- Distance matrixes
- Mutational models
- Distance phylogeny methods

Distance Matrix

```
Human  aactc
Chimp  aagtc
Orang  tagtt
```

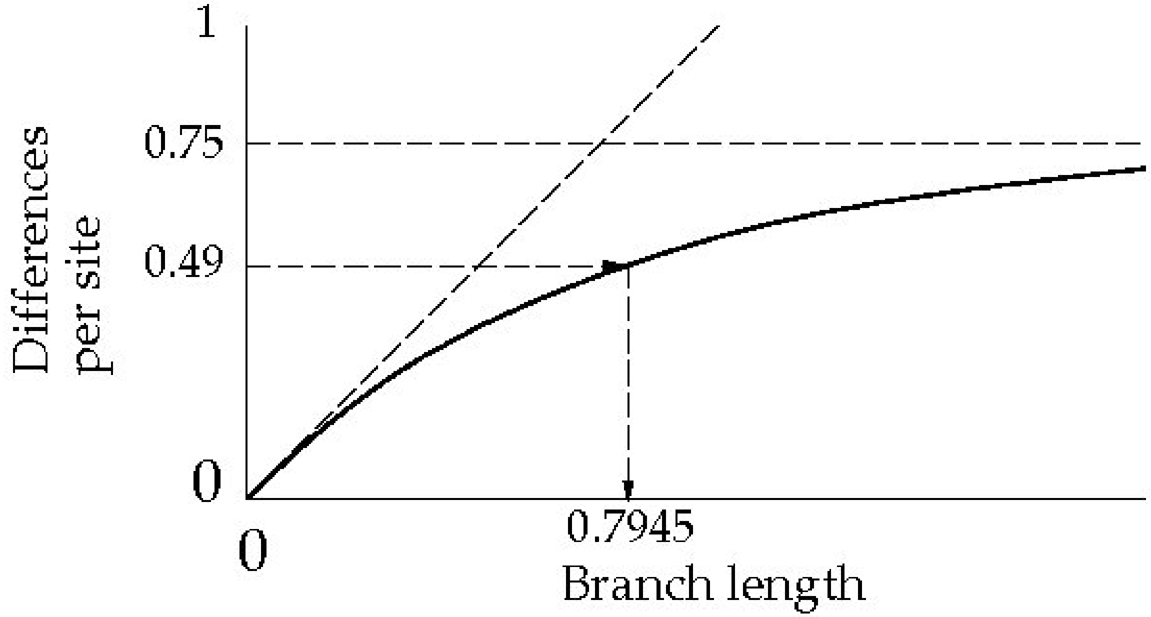
becomes

	H	C	O
H	-	1	3
C	1	-	2
O	3	2	-

Distance Methods

- Tree is built using distances rather than original data
- Only possible method if data were originally distances:
  - immunological cross-reactivity
  - DNA annealing temperature
- Can also be used on DNA, protein sequences, etc.

Large distances are underestimated by raw counts



A mutational model allows corrected distances

Jukes-Cantor model:

$$D = -\frac{3}{4} ln(1 - \frac{4}{3} D_s)$$

- D is the corrected distance (what we want)
- D_s is the raw count (what we have)
- ln is the natural log

Mutational models for DNA

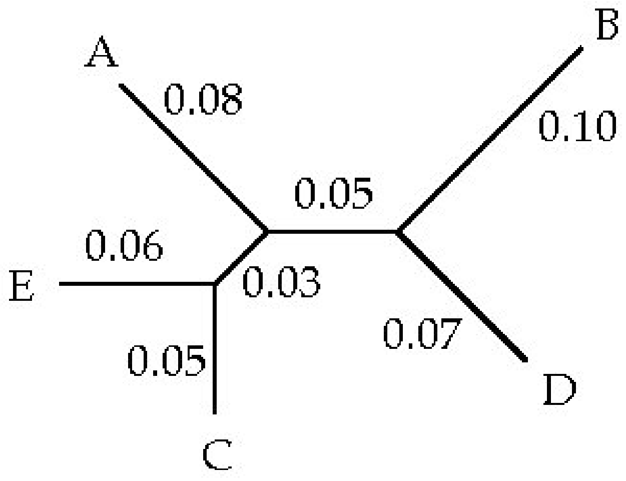
- Jukes-Cantor (JC): all mutations equally likely
- Kimura 2-parameter (K2P): transitions more likely than transversions
- Felsenstein 84 (F84): K2P plus unequal base frequencies
- Generalized Time Reversible (GTR): most general usable model

Models more complex than GTR would be useful but are very hard to work with.

Mutational models for protein sequence

- We have already seen these in alignment (BLOSUM etc.)
- Protein models are usually built from empirical data

Distances into trees



	A	B	C	D	E
A	0	0.23	0.16	0.20	0.17
B	0.23	0	0.23	0.17	0.24
C	0.16	0.23	0	0.15	0.11
D	0.20	0.17	0.15	0	0.21
E	0.17	0.24	0.11	0.21	0

Distances into trees

- Not all sets of distances fit a tree perfectly
- For those that do, finding the tree is simple
- If no tree fits perfectly, which one is best?

Least squares

- Least squares rule: prefer the tree for which the sum of

  $$(observed - expected)^2$$

  is minimized.

- This means that getting a long branch wrong is penalized much more heavily than getting a short branch wrong

- Some least-squares methods add weights to this calculation to allow for long distances being less accurately measured than short ones

Minimum evolution and neighbor-joining

- Minimum evolution rule: for each topology, find the best branch lengths by least-squares

- Then, choose the topology with the lowest total branch lengths

- The popular neighbor-joining algorithm is a very fast approximation to ME

- Neighbor-joining gains its speed by considering very few trees

- It uses a clustering approach rather than a tree search

- Surprisingly, it works quite well

The molecular clock

- The molecular clock is the hypothesis that the rate of evolution is constant over time and across species

- This is almost never true

- It is most nearly true:
  - among closely related species
  - among species with similar generation time and life history
  - for genetic regions with the same function in all species, or no function

The molecular clock

- Even when the clock is doubtful, it is often assumed in order to:
  - put a root on the tree
  - infer the times at which species arose
  - estimate the rate of mutation
- When the data are not really clocklike, assuming a clock will often result in inferring the wrong tree
  - Branch lengths will certainly be wrong
  - Topology will often be wrong
- Statistical tests for clock violation are available and should be used

Practical example: UPGMA

- UPGMA is a clock-requiring algorithm similar to neighbor-joining
- Algorithm:
  - Connect the two most similar sequences
  - Assign the distance between them evenly to the two branches
  - Rewrite the distance matrix replacing those two sequences with their average
  - Break ties at random
  - Continue until all sequences are connected
- This is too vulnerable to unequal rates to be reliable
- However, it is easy to learn and understand, so used in teaching

UPGMA example

	A	B	C	D	E
A	-	5	1	8	9
B	5	-	4	10	11
C	1	4	-	9	9
D	8	10	9	-	2
E	9	11	9	2	-

UPGMA example

	A	B	C	D	E
A	-	5	1	8	9
B	5	-	4	10	11
C	1	4	-	9	9
D	8	10	9	-	2
E	9	11	9	2	-

Group A and C to form AC, with branches of length 0.5

	AC	B	D	E
AC	-	4.5	8.5	9
B	4.5	-	10	11
D	8.5	10	-	2
E	9	11	2	-

UPGMA example

	AC	B	D	E
AC	-	4.5	8.5	9
B	4.5	-	10	11
D	8.5	10	-	2
E	9	11	2	-

Group D and E to form DE, with branches of length 1.0

	AC	B	DE
AC	-	4.5	8.75
B	4.5	-	10.5
DE	8.75	10.5	-

UPGMA example

	AC	B	DE
AC	-	4.5	8.75
B	4.5	-	10.5
DE	8.75	10.5	-

Group B with AC to form ABC, with branches of length 2.25

	ABC	DE
ABC	-	9.625
DE	9.625	-

UPGMA example

	ABC	DE
ABC	-	9.625
DE	9.625	-

Group ABC with DE, with branches of length 4.80

Distance methods summary

- All distance methods lose some information in making the distances
- Which algorithm you use is much less important than a good distance correction
- The more you know about the evolutionary process, the better you can correct the distances
- Distance methods are popular because they are fast and can be used with a variety of models

Judging tree-inference methods

Points to consider:

- Consistency: would it get the right answer with infinite data and a correct model?
  - Parsimony is not consistent
  - Distance methods with properly corrected distances are
- Robustness: how much is it hurt by a wrong model?
  - Distance methods can be highly vulnerable
  - Parsimony is more robust
- Power: how well can it do with limited data?
- Speed: can I stand to run it?
  - Methods that are consistent, robust and powerful tend to be slow

Judging tree-inference methods

Points to consider:

- Availability: can I find a program to do this?
  - The PHYLIP package is a good free source of phylogeny programs
  - http://evolution.gs.washington.edu/phylip.html
  - Links to huge list of other available programs
- Intended use: what do I need from my phylogenies?