Lecture 20, Phylogeny & RNA: Pfold
12/4; D. Langworthy

Modeling Sequence Evolution
Make simplifying assumptions
Most mutations are neutral – even in protein coding regions
Mutations are markov order 1 – not true in general – in plants some mutations have been seen to revert – controversial

Simple Example: Jukes-Cantor
Rule matrix is a simple model
Not talking about an individual mutation – talking about a mutation that has grown to pervade the species.

A rate matrix can be scaled to arbitrary durations. T=0 is the identity matrix. T=Infinity the observed distribution.

Other Models
Other models correct for more features, e.g., unequal frequencies of A C G T

General Reversible Model
Evolutionary Models—calendar time is not molecular time. E.g. bacteria divide very quickly in relation to humans.

Uses Example 4
Don't need to assume molecular clock because it can be inferred
Only ever use product of rate and time
The model is reversible so the "root" can be located at any point between x and y.
X could have evolved to Y, Y could have evolved to X, or the both could have evolved from a common ancestor.

Uses Example 5
Actually need another data point to root tree, an outlier.
The statistical approach was very controversial in the beginning. Parsimony was king. Parsimony can lead to different results with enough data.
Can "bootstrap" statistical approach for confidence intervals (repeat analysis with random subsequences)
Naive calculation is exponential in the number of samples.

Felsenstein Recurrence
Makes cost of computing a probabilities for a fixed tree topology manageable. (Searching over all alternative topologies is more difficult, but people do approximate it, e.g. via MCMC.)
Tree assumption not true for prokaryotes

Rate Matrix (Paired)
Count GC and CG together even though there is a bias in the training data.
Gaps could be treated as another character but this is not good. Treating them as "unknown", i.e. a background-frequency mixture of all characters, is also not good, but more realistic alternatives are still too expensive.