Introduction to Phylogenetics
Week 4

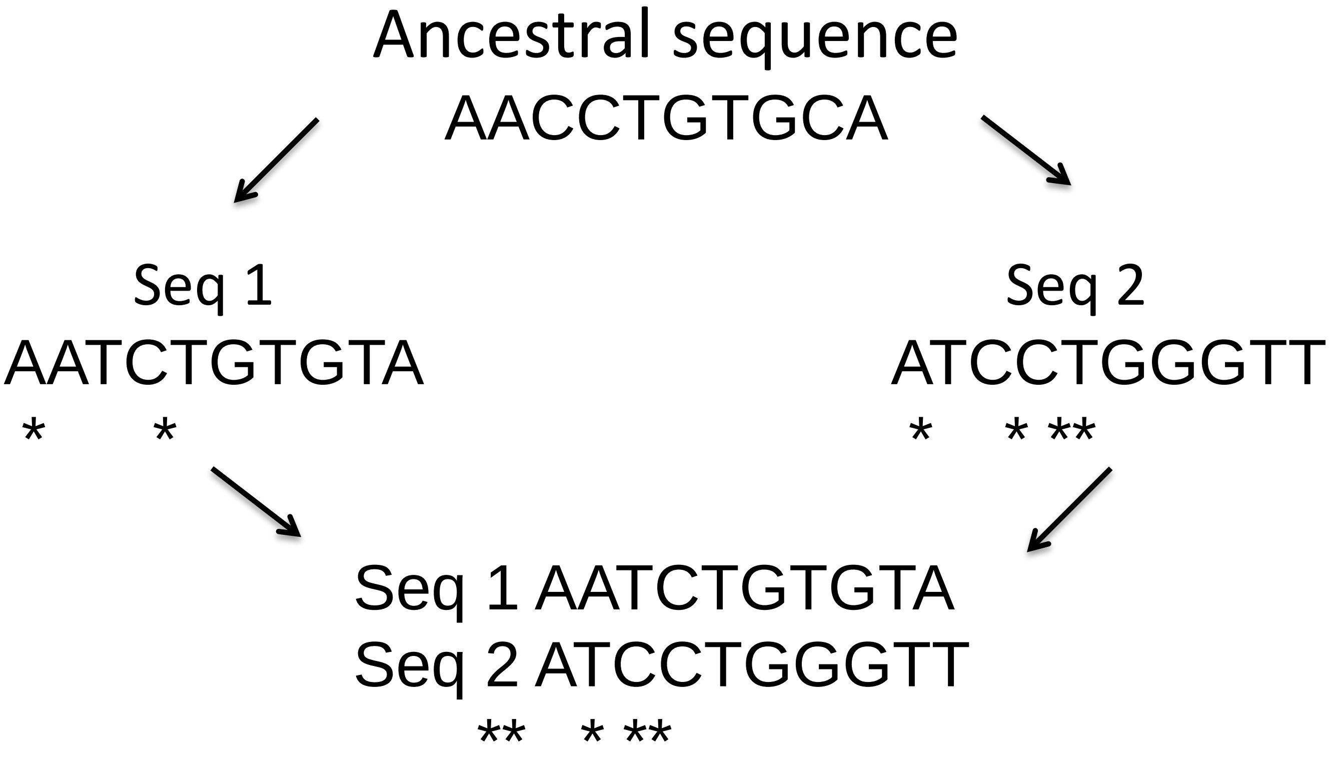
Phylogenetic Models

I. Models

- Genetic distance
  - Used to determine divergence between sequences
  - Two identical sequences will diverge based on standard evolutionary rates
  - Rate depends a lot on how you model evolution
  - The evolutionary model you use is critical to obtaining a robust phylogenetic structure

II. Observed and expected

- Simplest approach – count differences



Ancestral sequence
AACCTGTGCA

Seq 1
AATCTGTGTA
*   *

Seq 2
ATCCTGGGTT
*   * **

Seq 1 AATCTGTGTA
Seq 2 ATCCTGGGTT
**   * **

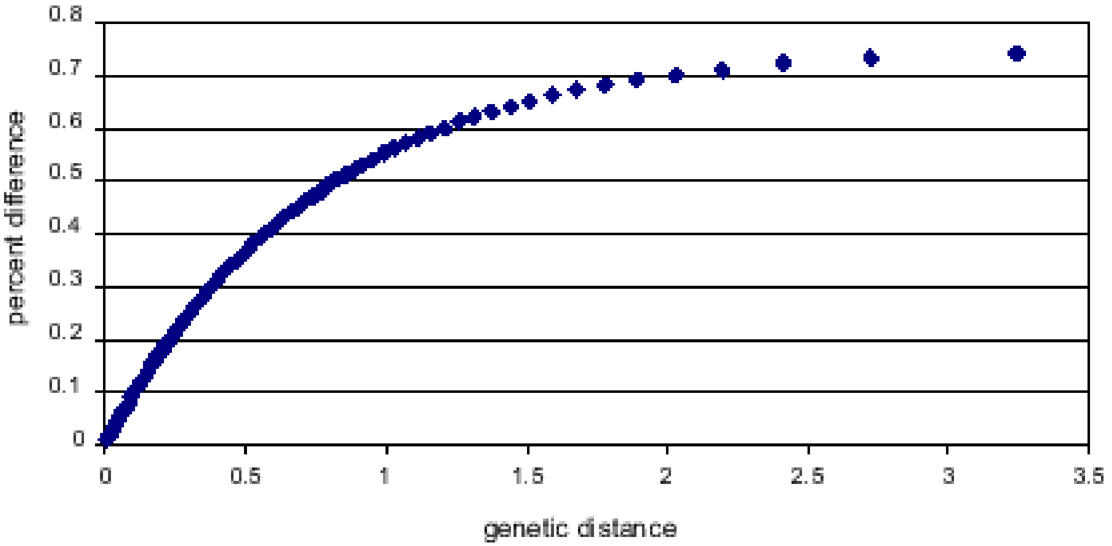
- Proportion is observed distance or *p*-distance

II. Observed and expected

- Relatedness based on *p*-distance (Hamming distance) easy to imagine
- Only based on identity between sequences
- Cannot account for type of change/back-mutations
- No account of evolutionary processes (i.e. transition vs. transversion)
- Observed distance (*p*) underestimates true genetic distance (*d*)

II. Observed and expected

- Overtime substitutions at each site accumulate and sequences are saturated



- Can use this to determine genetic distance for tree – not very robust.

III. Mutations and time

- Substitutions assumed to be random event
- Substitution model provides statistical description of stochastic process
- Number of mutations X(*t*) over time *t*.
- Use Poisson (P) distribution
  - Discrete probability distribution
  - Most basic model assumes mutation equally likely at each site
  - Occurs at rate μ
  - P_n(t) probability n mutations in *t*

III. Mutations and time

$$P_n(t) = [(\mu t)^n \exp(- \mu t)] / n!$$

- Number of substitutions up to time *t* is distributed with factor μt with variance μt
- Nucleotide substitution rate therefore is tied to *t*

IV. Calculating Nucleotide Subs.

- Rates of substitution a Markov process
  - Stochastic model
  - Random system changes state according to transition rule
  - Can make predictions on future based on present state of system
- Basically – you know what the nucleotide is now, you can determine likelihood of assuming future state

IV. Calculating Nucleotide Subs.

- Q matrix specifies rate of change for each nucleotide
- Way of describing all possible changes between states
- Probability departing from state *i*, arriving at state *j*
- Assumes state prior to *i* has no impact on probability of *j*
- Rate of change modeled differently by different evolutionary models

$$Q=\begin{array}{cccc} \mathrm{A} & \mathrm{C} & \mathrm{G} & \mathrm{T} \\ -\mu(a\pi_C+b\pi_G+c\pi_T) & a\mu\pi_C & b\mu\pi_G & c\mu\pi_T \\ g\mu\pi_A & -\mu(g\pi_A+d\pi_G+e\pi_T) & d\mu\pi_G & e\mu\pi_T \\ h\mu\pi_A & i\mu\pi_C & -\mu(h\pi_A+j\pi_C+f\pi_T) & f\mu\pi_T \\ j\mu\pi_A & k\mu\pi_C & l\mu\pi_G & -\mu(i\pi_A+k\pi_C+l\pi_T) \end{array}$$

- μ is mean instantaneous substitution rate
- a, b, c... relative substitution rate (i.e. A to C)
- π_G, $\pi_{A,}$ π_T... nucleotide frequencies
- Diagonal values so each row = 0 (no change)
- How you parameterize matrix determines how you model evolution

V. Time Reversible Models

- Basic substitution models are probably not biologically relevant
- Do allow us to model stochastic events
- Time-reversible models assume rate of change *i* to *j* is the same as *j* to *i* (a = g, etc)
- Probability of nucleotide change at any site during evolutionary time (*t*)

$$P(t) = \exp(\mathrm{Q}t)$$

V. Jukes Cantor Model (JC69)

- When the probabilities of change $P(t)$ are known, can determine evolutionary distance between two sequences
- JC69 equilibrium frequency nucleotide = 25%

$$\pi_G = \pi_A = \pi_T = \pi_C = 1/4$$

- JC69 any nucleotide replaced by any other

$$a = b = c = d = e = f = g\ldots = 1$$

- Probability of nucleotide not changing $P_{ii}(t)$
- Probability of nucleotide replacement $P_{ij}(t)$

V. Jukes Cantor Model (JC69)

Using JC69 Q matrix and given *P*(*t*) = exp(Q*t*)

$P_{ii}(t) = \frac{1}{4} + \frac{3}{4}\exp(-mt)$ and $P_{ij}(t) = \frac{1}{4} - \frac{1}{4}\exp(-mt)$

Comparing two sequences: $p = \frac{3}{4}[1 - \exp(-2mt)]$

Solving for μt:

$$mt = -\frac{1}{2}\log(1 - \frac{4}{3}p)$$

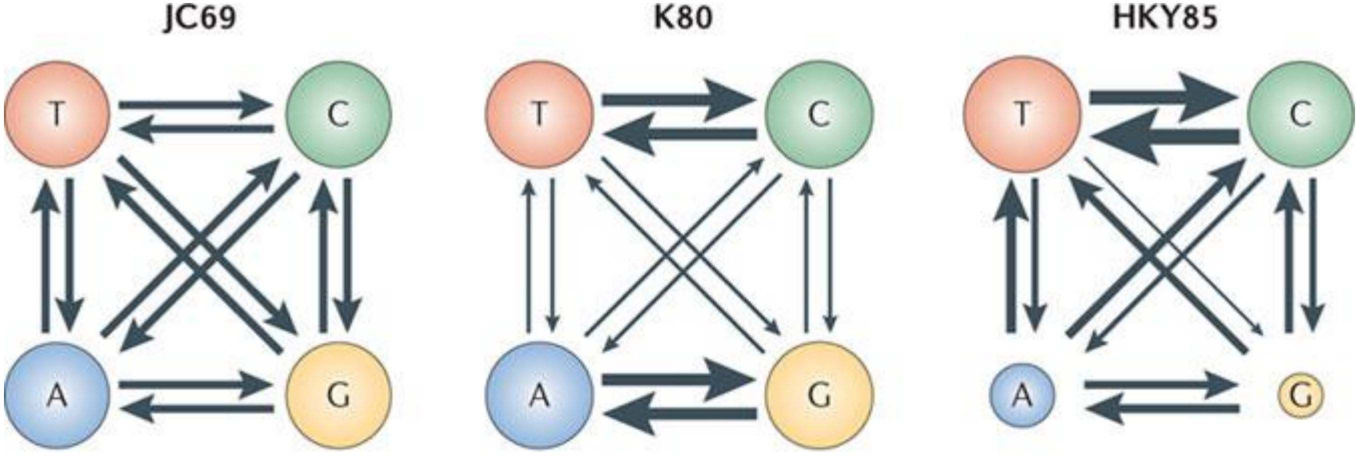
V. Jukes Cantor Model (JC69)

$$Q = \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix} \; P = \begin{pmatrix} \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} \end{pmatrix}$$

$$d = -\frac{3}{4}\ln(1 - \frac{4}{3}p)$$

VI. Nucleotide substitution models

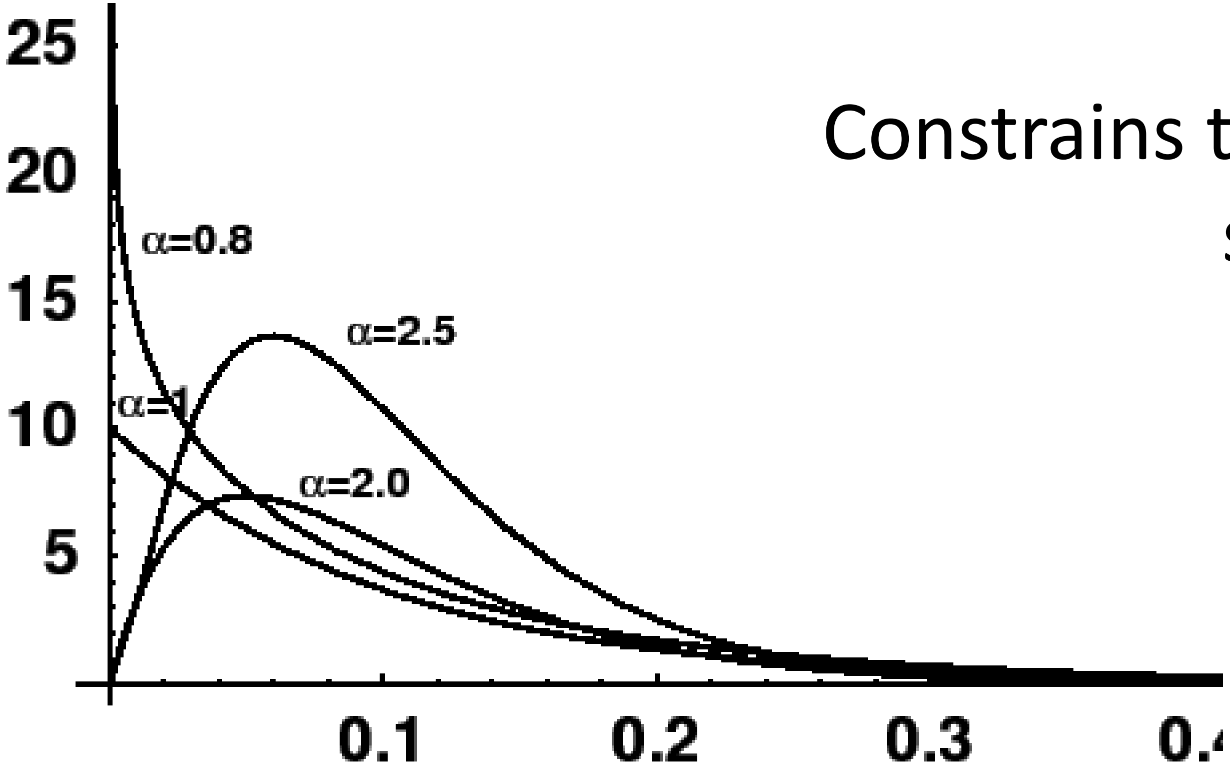
- If all parameters of Q matrix determined – considered a general time reversible model (**GTR**)
- Parameterization reflects more of biological processes – rate heterogeneity between sites



Nature Reviews | Genetics

VI. Nucleotide substitution models

- Consider distribution of nucleotide changes
- Standard probability distributions - Gamma



Constrains the amount of site variability