

Stat 536

Molecular Evolution and Phylogenetics

Misha Rajaram

December 3, 2008

Molecular Evolution

- The study of evolutionary biology can be broadly classified into the study of *mechanisms* that produce a change and the *pattern* that emerges due to these changes.
- The *pattern* refers primarily to the relatedness i.e phylogenetic relationships that can be established between resulting taxa.
- Molecular evolution is the study of mechanisms that shape the evolution of a group of organisms at the level of DNA, RNA and proteins.
- Phylogenetic analyses help in the study of the emergent pattern in these sequences.

Molecular Evolution cont'd.

- Active research in the last couple of decades has recorded remarkable advances in both areas, so much so, that today it is hard to clearly tell where one ends and the other begins.
- Evolution results primarily from two mechanisms: the genetic variability (different alleles) brought about by *Mutations* and with time, the change in allele frequency in the population.
- Mutant genes can propagate within a population by natural selection and/or genetic drift resulting in fixation and hence altered allele frequency.

Evolutionary hypotheses

- Recall that the main causes for change in allele frequency are *Mutation, Migration, Genetic Drift* and *Natural Selection*.
- Hypotheses of evolution vary depending on which of these forces they regard as most important in shaping evolution.
- The **Selectionist Hypothesis** regards Natural Selection as the main (or only) force shaping evolution i.e Darwinian. Difference between species were thought to consist of mainly mutations that had been fixed by positive selection.

- In 1980, Motoo Kimura put forth the **Neutral Theory** that claims that “the overwhelming majority of evolutionary changes at the molecular level are not caused by selection acting on advantageous mutants, but by random fixation of selectively neutral or very nearly neutral mutants through the cumulative effect of sampling drift (due to finite population number) under continued input of new mutations”.
- Note that the term *Selectively Neutral* does not mean the alternative alleles of a DNA/protein locus have no effect on fitness but that selection among different genotypes at that site is weak. This is the same as saying $N_e s < 1$ where s is the selection coefficient.
- This has major implications in theoretical models of DNA evolution and tests for selection.

Neutral Theory: The Null Hypothesis of evolution

- In observing a new allele at a locus in a population, the rates of two processes are involved:
 - Mutation rate: The rate at which a mutation occurs in a sequence or the probability that an offspring sequence has a different allele compared to its parent.
 - Substitution rate: An allele substitution occurs when the newly arisen allele becomes fixed in the population.
- item Under the neutral theory, assume that the mutation rate is μ for a site. In a population of N diploid individuals, $2N\mu$ mutations occur every generation at this site.
- In a diploid population, the probability of fixation of a new allele purely by genetic drift i.e. in the absence of selection is simply the proportion in which it exists in the population, $1/2N$ in this case.

Neutral Theory: The Null Hypothesis of evolution cont'd

- The substitution rate K can now be computed as

$$K = (2N\mu)(1/2N)$$

implying that $K = \mu$.

- Note that in the presence of natural selection, the rate of evolution K for an advantageous allele will be higher than the mutation rate $K > \mu$ and vice versa for a deleterious mutation.

Molecular Clock Hypothesis

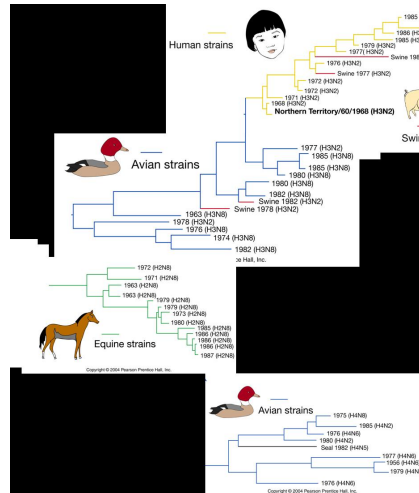
- In 1962, Zuckerkandl and Pauling noticed that the number of differences in amino acid hemoglobin sequences of various lineages corresponds roughly with their evolutionary divergence times as estimated from fossil records.
- The Hypothesis states that the rate of nucleotide or amino acid substitution is roughly constant over time and hence can be used to compute divergence times. (Note how it is linked to the Neutral Theory)
- It is hard to find a gene or protein that has a constant rate of evolution over a long period of time.
- Molecular clocks need not, however, be universal. As long as it works for a group of organisms, it is still useful.

Visualizing Molecular Evolution

- Evolutionary processes shape DNA/Amino Acid sequences that we are able to sequence and observe.
- Our ultimate goal, however, is to study the difference we observe, in the context of relationships between these sequences.
- Suppose that you sequence gene A in species 1, 2, 3 and 4 and find that site j has the same amino acid for species 1 and 2 while species 3 and 4 have a different amino acid at that site. Could species 1 and 2 be closer to each other evolutionarily, than to 3 and 4?
- May be the change at that site modifies the resultant protein so that the animal is able to better digest fiber. Now for species 1 and 2 to be more closely related probably makes sense since they may be herbivorous while 3 and 4 may be carnivorous.
- If this hypothesis is true, was the common ancestor of all four species herbivorous or carnivorous?
- Phylogenetic Trees help up visualize these relationships.

Visualizing Molecular Evolution cont'd

- Phylogenetic trees also help resolve population structure.
- Consider the case of the H3N2 type Influenza outbreak in 1968 called the Hong Kong Flu.
- It was later termed “Bird Flu” because analysis of related sequences place it closest to the avian flu population.



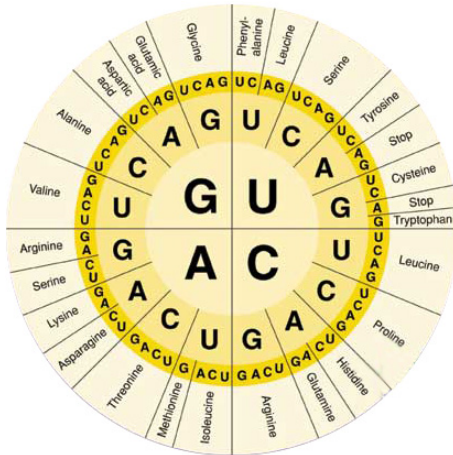
Types of Mutations

- Difference in sequences arise mainly in the form of mutations.
- Point Mutations : Change at a single nucleotide due to chemicals or malfunction of replication machinery. When they occur in a protein coding region of a gene they can be:
 - Silent/Synonymous mutations cause no change in resultant amino acid.
 - Missense/Non-synonymous mutations cause change in resultant amino acid.
 - Nonsense mutations code for premature stop codon.
- Insertions: Add one or more nucleotides to the DNA sequence. May cause *Frameshifts*.
- Deletions: Delete one or more nucleotides from the DNA sequence. May cause *Frameshifts*.

Codon Usage

- Three consecutive nucleotides in a DNA sequence that codes for a protein, form a *Codon*.
- Each codon codes for a specific amino acid that can then be added to a growing amino acid chain formed by the DNA translation machinery that scans the mRNA sequence and translates it into an amino acid chain.
- Every position in a codon can have one of 4 possible nucleotides, giving rise to $4^3 = 64$ possible codons.
- There are, however, only 20 possible amino acids.
- There is thus considerable redundancy in this coding system.

Genetic Code

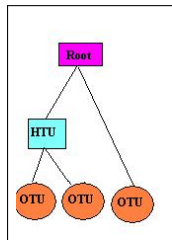


Types of Data

- Nucleotide sequences: Basic information. Ultimately, differences in any of the molecular markers we study (and of genetically-based morphological, behavioral, or physiological traits) are associated with some difference in the physical structure of DNA.
- Amino acid sequences: Many different nucleotide sequences can code for the same amino acid sequences due to redundancy in the genetic code.
- Secondary and higher-order structures: Different amino acid sequences may still end up forming similar structures upon folding.
- Sequence organization: The order of arrangement of genes within a genome or of introns and exons within a gene may hold important evolutionary information.
- Expression: Functional differences among individuals may arise due to different gene expression patterns.
- Copy number variation : Individuals of the same species may carry varying numbers of copies of the same gene. These are referred to as copy number polymorphisms.

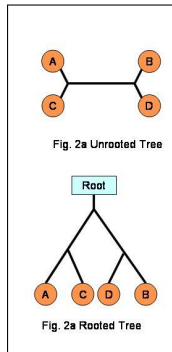
Phylogenetic trees

- Phylogenetic trees help illustrate the relationship between biological entities.
- Each node may have a set of descendants and is their Most Recent Common Ancestor (MRCA).
- Each node is called a Taxonomic Unit (TU). Leaf nodes are known as Observed Taxonomic Units (OTUs) and internal nodes are called Hypothetical Taxonomic Units (HTUs) since their actual value can not be observed.



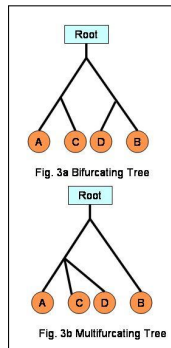
Types of Trees :Rooted and Unrooted Trees

- A *rooted* tree is a directed tree and has a unique node (root) corresponding to the MRCA of the OTUs in the tree.
- In practice, trees are rooted by including a distant related sequence called the *outgroup* or by introducing additional assumptions about the relative rates of evolution on each branch, such as an application of the molecular clock hypothesis.
- An *unrooted* tree illustrates the relationship of the OTUs without making assumptions of common ancestry.



Types of Trees: Bifurcating and Multifurcating

- Both rooted and unrooted tree could be *bifurcating* or *multifurcating*.
- A *bifurcating* tree is one in which each internal node gives rise to a maximum of two children.
- A *multifurcating* tree allows internal nodes to have any number of descendants.



How many Trees?

- The number of possible trees for a set of n OTUs varies depending on whether the trees are rooted or unrooted, bifurcating or multifurcating and no so.
- For n OTUs, there are

$$\frac{(2n - 3)!}{2^{(n-2)}(n - 2)!}$$

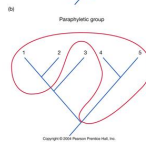
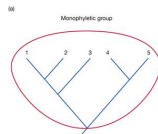
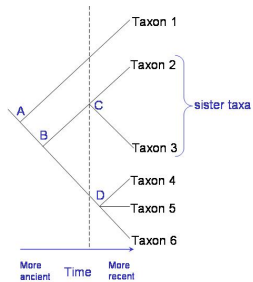
possible rooted bifurcating trees and,

$$\frac{(2n - 5)!}{2^{(n - 3)}(n - 3)!}$$

possible unrooted bifurcating trees.

- Note that the number of unrooted trees for n OTUs is equal to the number of rooted trees for $n - 1$ OTUs.

Reading phylogenetic trees



Maximum Parsimony Methods

- Originally developed for morphological characters.
- Preference to the evolutionary tree that involves “the minimum net amount of evolution” (Edwards and Cavalli-Sforza 1963).
- It involves examining all possible trees for a set of sequences and assigning, for each tree a “parsimony score” .
- The tree with the lowest score is said to be maximally parsimonious as it reflects a path from the ancestral nodes to the observed taxa, that requires minimal changes.

Maximum Parsimony trees- An illustration

- Consider an alignment of four DNA sequences for which you want to infer a phylogeny.

Alpha	A	C	G	T
Beta	G	C	G	C
Gamma	C	C	A	A
Delta	C	C	G	T

- Suppose we want to build a maximum parsimony unrooted tree based on the first column in the alignment.
- For 4 sequences, there are 15 possible rooted phylogenies.

Maximum Parsimony trees- An illustration

- Consider an alignment of four DNA sequences for which you want to infer a phylogeny.

Alpha	A	C	G	T
Beta	G	C	G	C
Gamma	C	C	A	A
Delta	C	C	G	T

- Suppose we want to build a maximum parsimony unrooted tree based on the first column in the alignment.
- For 4 sequences, there are 15 possible rooted phylogenies.

Maximum Parsimony trees- An illustration

- Consider an alignment of four DNA sequences for which you want to infer a phylogeny.

Alpha	A	C	G	T
Beta	G	C	G	C
Gamma	C	C	A	A
Delta	C	C	G	T

- Suppose we want to build a maximum parsimony unrooted tree based on the first column in the alignment.
- For 4 sequences, there are 15 possible rooted phylogenies.

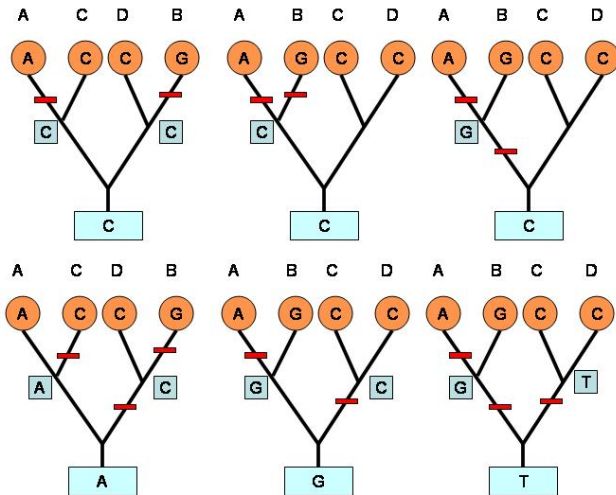
Maximum Parsimony trees- An illustration

- Consider an alignment of four DNA sequences for which you want to infer a phylogeny.

Alpha	A	C	G	T
Beta	G	C	G	C
Gamma	C	C	A	A
Delta	C	C	G	T

- Suppose we want to build a maximum parsimony unrooted tree based on the first column in the alignment.
- For 4 sequences, there are 15 possible rooted phylogenies.

Computing Parsimony Scores




Estimating branch lengths

- Branch lengths can be computed by considering all evolutionary pathways at each variable site and computing the average number of substitutions for each internal or external branch.
- When there is only one substitution, this can be assigned to the external branch leading to the OTU.
- For two or more substitution, there are several pathways of assigning these (as we saw before).
- The branch length is then simply the average of these.
- This method is called the *Average pathway method*

Problems with the Maximum Parsimony approach

- The most parsimonious tree may not be unique.
- Lineages that undergo rapid rates of evolution have long branches in a tree and these tend to “attract” each other, often clustering in the resulting tree. This phenomenon was identified by Felsenstein and termed *long branch attraction* and parsimony methods are particularly sensitive to it often leading to erroneous phylogenies.
- Parsimony is biased when the base composition of the DNA sequence is skewed. Even with highly conserved sequences, it overestimates the number of changes, once again leading to erroneous phylogenies.

Markov Chains

- Markov Chains are stochastic processes with Markov property which states that conditional on the present state, future states are independent of the past states.
- Consider a simple binary system that can, at each time point, be in one of two states.
- The transition observed in this system can be represented using a finite state machine: 
- or a *State Transition Matrix*

$$P = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$$

- Markov property states that the conditional probability of all future states given the current state and all past states depends only on the current state and is independent of all past states.
- A Markov Chain can then be described as a sequence of random variables $X_1, X_2, X_3, X_4 \dots X_n$ such that

$$Pr(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1} \dots, X_0 = i_0) = Pr(X_{n+1} = j | X_n = i)$$

where i and j are states in the Markov Chain.

Transition Probability Matrix

- Probabilities of the nature of $Pr(X_{n+1} = j | X_n = i)$, notated as $p(i, j)$ for convenience, are referred to as *Transition Probabilities*
- Every Markov Chain is associated with a Transition Probability Matrix that contains probabilities of transitioning from a starting state i to any other state j in the state space in one time step.
- The probability of transitioning from state i to state j in n time steps can be given by

$$p_{ij}^{(n)} = Pr(X_n = j | X_0 = i)$$

and be computed as

$$p_{ij}^{(n)} = \sum_{r \in S, r \neq j} p_{ir}^{(k)} p_{rj}^{(n-k)}$$

Properties of Markov Chains

- Reducibility : A state j is said to be **accessible** from another state i if the the transition probability p_{ij} over $n \geq 0$ time steps.
- Communicating states: States i and j are said to communicate if i is accessible from j and vice versa.
- Closed Communicating class : A set of states C in which every pair of states are communicating.
- A Markov chain is *irreducible* if all its states belong to single communicating class.

$$\begin{pmatrix} 0.2 & 0.3 & 0 & 0.5 \\ 0.3 & 0.1 & 0.1 & 0.5 \\ 0.1 & 0.2 & 0.6 & 0.1 \\ 0.2 & 0.2 & 0.2 & 0.4 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Limiting Distributions

- A Stationary distribution of a Markov chain with transition matrix P is a vector that satisfies

$$\pi_j = \sum_i p_{i,j} \pi_i$$

and

$$\sum_j \pi_j = 1$$

- A finite state irreducible Markov chain has a limiting distribution that is also unique when other conditions are satisfied.

Continuous Time Markov Chains

- A CTMC is characterized by state changes that can occur at any arbitrary time. Thus, state space is still discrete, index is now continuous.
- A CTMC is thus a discrete-state continuous-time stochastic process in which for any time $0 \leq s_0 \leq s_1 \leq s_2 \dots \leq s$ the conditional *pmf* satisfies the Markov property:

$$\Pr(X(t) = j | X(t_n) = i, X(t_{n-1}) = i_{n-1} \dots, X(0) = i_0) = \Pr(X(t) = j | X(t_n) = i)$$

- Transition Probability functions (over an interval) are given by:

$$p_{i,j}(s, t) = \Pr(X(t) = j | X(s) = i)$$

for $0 \leq s \leq t$.

$$\sum_{j \in S} p_{i,j}(s, t) = 1$$

Time homogeneous CTMC

- A time homogeneous CTMC has a conditional *pmf* that satisfies:

$$p_{i,j}(t) = \Pr(X(t+s) = j | X(s) = i)$$

in other words,

$$p_{i,j}(s, t) = p_{i,j}(t - s)$$

i.e. The rate of change is fixed.

- We will restrict the rest of this discussion to time homogeneous CTMCs.

- Like in a DTMC

$$p_{i,j}(s, t) = \sum_{k \in S} p_{i,k}(s, u) p_{k,j}(u, t)$$

- Unlike DTMC though, the transition probabilities are now functions of elapsed time as opposed to the number of elapsed steps.
- We thus introduce the notion of *rates of transitions*
- Define rates (probabilities per unit time)
 - q_j is the rate at which the chain leaves state j
 - $q_{i,j}$ is the rate from state i to state j

Rate Matrix Q

- For a small instance of time h , transition probabilities can now be defined as:

$$p_{i,j}(t, t + h) = p_{i,j}(h) = q_{i,j}.h + o(h)$$

for $i \neq j$

$$p_{j,j}(h) = 1 - q_j.h + o(h)$$

- The matrix Q of all $q_{i,j}$ is called the infinitesimal rate matrix.

Properties of the rate Matrix



$$\sum_j q_{i,j} = 0$$

$$q_{i,i} = -q_i = -\sum_{j \neq i} q_{i,j}$$



$$\frac{dP(t)}{dt} = P(t)Q$$

$$\frac{d\pi(t)}{dt} = \pi(t)Q$$



$$\pi Q = 0$$

- For an irreducible homogeneous CTMC the following limits exist.

$$\pi_i = \lim_{t \rightarrow \infty} \pi_i(t) = \lim_{t \rightarrow \infty} P_{ji}(t)$$

- Given the rate matrix Q , the corresponding transition probability matrix can be computed from the relation

$$P(t) = e^{Qt}$$

Reversibility

- Consider a Markov chain with a unique limiting distribution π and realizations $\dots, X_{n-1}, X_n, X_{n+1} \dots$
- Now retrace your steps to get a sequence $\dots, X_{n+1}, X_n, X_{n-1} \dots$. This also turns out to be a Markov Chain.
- A time reversible Markov chain satisfies

$$p(i, j)\pi_i = p(j, i)\pi_j$$

DNA evolution as a CTMC

Consider a DNA sequence of length m evolving in time by base substitution. The state space for a site is $S = \{A, C, G, T\}$, therefore
Assume that the m sites are iid and constant in time with probability of being one of the four bases given by

$$\pi(t) = (\pi_A(t), \pi_C(t), \pi_G(t), \pi_T(t))$$

Also for each site, let μ_{xy} represent the probability of replacement of base x by base y where $x, y \in S$. We can now use Q , the matrix containing probabilities of a nucleotide being replaced by another at an instant of time as the rate matrix

- Using notation from above,

$$Q = \begin{pmatrix} -\mu_A & \mu_{AC} & \mu_{AG} & \mu_{AT} \\ \mu_{AC} & -\mu_C & \mu_{CG} & \mu_{CT} \\ \mu_{GA} & \mu_{GC} & -\mu_G & \mu_{GT} \\ \mu_{TA} & \mu_{TC} & \mu_{TG} & -\mu_T \end{pmatrix}$$

- In the context of DNA substitution/evolution we will refer to the stationary distribution π as base frequencies and the μ s as substitution/mutation rates.
- We can now compute the probability that a sequence transitioned from having base x at a position to now having base y , in time t as

$$P(t) = e^{Qt}$$

Models of Nucleotide Substitution

These are CTMCs that make varying levels of assumptions in specifying the base frequencies π_i and the substitution rates μ .
JC69 model (Jukes and Cantor 1969)

- Simplest model that assumes equal base frequencies so $\pi_A = \pi_C = \pi_G = \pi_T = 1/4$ and equal mutation rates.
- These assumptions reduce the parameter space to a single parameter μ , the mutation rate and the rate matrix is given as:

$$Q = \begin{pmatrix} - & \mu & \mu & \mu \\ \mu & - & \mu & \mu \\ \mu & \mu & - & \mu \\ \mu & \mu & \mu & - \end{pmatrix}$$

Kimura Two-Parameter Model- K80 (Kimura 1980)

- Extends the JC69 model .
- Assumes a different rate for *transition* (A \rightarrow G or C \rightarrow T) , say α and a different rate for *transversion* (purine \rightarrow pyrimidine) , say β .
- Base frequencies are still assumed to be equal ($=1/4$).
- The parameter space now grows to include two parameters α and β , expressed in the Q matrix as a ratio $\kappa = \frac{\alpha}{\beta}$.

$$Q = \begin{pmatrix} - & \beta & \alpha & \beta \\ \beta & - & \beta & \alpha \\ \alpha & \beta & - & \beta \\ 1 & \alpha & \beta & \beta \end{pmatrix}$$

$$Q = \begin{pmatrix} - & 1 & \kappa & 1 \\ 1 & - & 1 & \kappa \\ \kappa & 1 & - & 1 \\ 1 & \kappa & 1 & 1 \end{pmatrix}$$

- F81 model (Felsenstein 1981)
 - Assumes equal mutation rates but unequal base frequencies.

$$Q = \begin{pmatrix} - & \pi_C & \pi_G & \pi_T \\ \pi_A & - & \pi_G & \pi_T \\ \pi_A & \pi_C & - & \pi_T \\ \pi_A & \pi_C & \pi_G & - \end{pmatrix}$$

- HKY85 (Hasegawa, Kishino and Yang 1985)
 - Combines ideas from K80 and F81. Assumes unequal base frequencies and different rates for transition and transversion.

$$Q = \begin{pmatrix} - & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & - & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & - & \pi_T \\ \pi_A & \pi_C & \kappa\pi_G & - \end{pmatrix}$$

TN93 model (Tamura and Nei 1993)

Assumes unequal base frequencies and different transition rates α_1 and α_2 , for purines (A \leftrightarrow G) and pyrimidines (C \leftrightarrow T) respectively. A single transversion ratio β is used. This leads to two transition- transversion ratios, κ_1 and κ_2 .

$$Q = \begin{pmatrix} - & \pi_C & \kappa_1 \pi_G & \pi_T \\ \pi_A & - & \pi_G & \kappa_2 \pi_T \\ \kappa_1 \pi_A & \pi_C & - & \pi_T \\ \pi_A & \pi_C & \kappa_2 \pi_G & - \end{pmatrix}$$

Generalized Time Reversible Model

- So far all models have been *reversible*:

$$p(i, j)\pi_i = p(j, i)\pi_j$$

implying that if we see base i at the end of a branch and base j on the other, there is no way to decide which belongs to the ancestor and which to the descendant.

- This is the basic reason that rooting trees is often not possible forcing us to infer unrooted trees.
- The most general time reversible model assumes different base frequencies and a different substitution rate for each pair of bases.

$$Q = \begin{pmatrix} - & \alpha\pi_C & \beta\pi_G & \gamma\pi_T \\ \alpha\pi_A & - & \delta\pi_G & \epsilon\pi_T \\ \beta\pi_A & \delta\pi_C & - & \eta\pi_T \\ \gamma\pi_A & \epsilon\pi_C & \eta\pi_G & - \end{pmatrix}$$

Rate variation between sites

- So far we have assumed that rates of change at all sites are equal.
- We may know that rates differ across sites but may not know *a priori* which sites have high rates and which have low rates.
- Rate variation can be modelled using a gamma distribution.
- It is of utility in computing distances between a pair of sequences.
- Additionally, there may be reason to believe that a site is invariant i.e. never changes. Such sites can be assigned a probability f_0 of being invariant.

Codon models and beyond

- Substitutions in the third position of the codon or sometimes the second, if they are synonymous, could be under relaxed selective pressure when compared to substitutions that will be non-synonymous and hence potentially deleterious to the protein.
- The data can thus be partitioned into two sets under different selective pressures.
- Also, these account for the fact that sites are non-independent and treat codons as independent units.
- Another scenario of non-independent sites arises from *compensating mutations* that occur in rRNA molecules.
- Two sites, when paired in a secondary structure are favored by natural selection to continually be paired, so as to maintain the structure.
- Substitution in one may lead to a corresponding substitution in the other.

Distance matrix methods

- The general idea is to calculate a measure of distance between each pair of sequences and then find a tree that predicts the observed set of distances as closely as possible
- Consider each distance to be an estimated of the branch length separating the two species.
- Each distance then refers to the best unrooted tree for that pair of sequences.
- In effect, we then have a large number of estimated two-taxa trees using which we are trying to infer the n -taxa tree that is implied by these.
- Branch lengths reflect expected amounts of evolution and are not simply a function of time.

Computing evolutionary distances

- The evolutionary distance between a pair of sequences can be computed under a model of substitution.
- Consider the JC69 model. To calculate the distances we need to compute transition probabilities under this model.
- Let μ denote the substitution rate as defined earlier and recall that $P(t) = e^{Qt}$.
- Using the Q matrix for the JC69 model, the transition probabilities over time t can be computed as

$$P_{i,j}(t) = 1/4(1 + 3e^{-\mu t})$$

when $i = j$ and

$$P_{i,j}(t) = 1/4(1 - e^{-\mu t})$$

when $i \neq j$.

Computing evolutionary distances

cont'd

- Let us now define d as the expected number of nucleotide substitutions separating the two sequences diverging for a time t at any one position.
- For the JC69 model $d = \frac{3\mu t}{2}$ since the total rate of change is $3\mu/4$
- The probability, therefore, that there is change at a site for the pair of sequences is

$$1 - P_{i,i}(2t) = 3P_{i,j}(2t) = p$$

where $i \neq j$.

$$= \frac{3}{4} - \frac{3}{4}e^{-3\mu t/4}$$

$$d = -\frac{3}{4}\ln\left[1 - \frac{4}{3}p\right]$$

- An estimate \hat{d} can be obtained by using an estimate \hat{p} in the equation.
- \hat{p} can be the MOM estimator of p i.e. observed proportion of changed sites which has a Binomial distribution.
- The large sample variance of \hat{d} can then be derived using the delta function, since d is a function of p

$$V(\hat{d}) = \frac{9p(1-p)}{n(3-4p)}$$

Computing evolutionary distances

cont'd

- For the K80 model which proposes a rate α for transitions and rate 2β for transversions, the rate of substitution per site per time unit is $\alpha + 2\beta$.
- Counting all transition substituted pairs as P and all transversion substituted pairs as Q , we can express these as

$$P = (1/4)(1 - 2e^{-4(\alpha+\beta)t} + e^{-8t\beta})$$

$$Q = (1/2)(1 - e^{-8t\beta})$$

where t is the time since divergence.

Computing evolutionary distances cont'd

- The expected number of nucleotide substitutions is once again $d = 2\mu t$ except here $\mu = \alpha + 2\beta$ i.e. $d = 2t(\alpha + 2\beta)$ or

$$d = -(1/2)\ln(1 - 2P - Q) - (1/4)\ln(1 - 2Q)$$

- Estimate \hat{d} can be obtained by replacing P and Q with observed values. Variance of \hat{d} is given by

$$V(\hat{d}) = \frac{1}{n} [c_1^2 P + c_2^2 Q - (c_1 P + c_2 Q)^2]$$

where $c_1 = 1/(1 - 2P - Q)$ and $c_2 = 1/(1 - 2Q)$

Gamma distances

- When the substitution rate across sites varies with a gamma distribution, when the nucleotide substitution follows the JC69 model, the gamma distance can be computed as

$$d = \frac{3}{4}\alpha\left[\left(1 - \frac{4}{3}(1 - q)\right)^{-1/\alpha} - 1\right]$$

where α is the shape parameter of the gamma distribution.

Building a distance based tree using the Neighbor Joining method

	Cow	Deer	Whale	Hippo	Pig	Peccary	Camel
Deer	0.073						
Whale	0.150	0.197					
Hippo	0.148	0.197	0.053				
Pig	0.264	0.270	0.197	0.217			
Peccary	0.340	0.412	0.266	0.287	0.129		
Camel	0.284	0.347	0.216	0.236	0.291	0.340	
Outgroup	0.306	0.340	0.241	0.261	0.311	0.306	0.210

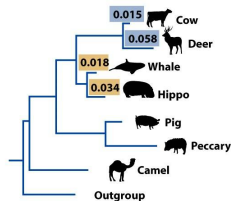


Figure 4-10 Evolutionary Analysis, 4/e
© 2007 Pearson Prentice Hall, Inc.

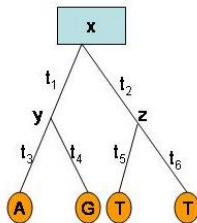
Likelihood-Based Phylogeny

- Consider an alignment of n sequences, m sites long.
- We are given a phylogeny with branch lengths and an evolutionary model that allows us to compute transition probabilities along the tree.
- In particular, we use the model to compute specific values for $P_{i,j}(t)$, the probability that nucleotide j exists at the end of branch of length t given that start of the branch was at nucleotide i .
- We now make two assumptions:
 - Assume that all sites are iid.
 - Assume that different lineages are independent.

- The first assumption allows us to decompose the likelihood of the tree as a product of the likelihood at each site

$$L = P(D|T) = \prod_i P(D^{(i)}|T)$$

where $D^{(i)}$ is data at the i th site.



- Consider the tree from site i . The likelihood for this tree is the sum, over all possible nucleotides that may have existed at the interior nodes of the tree, of the probabilities of each scenario of events

$$P(D^{(i)}|T) = \sum_x \sum_y \sum_z P(A, G, T, T, x, y, z|T)$$

- The second assumption allows us to decompose this probability into a product of terms.

$$P(A, G, T, T, x, y, z | T) = P(x)P(y|x, t_1)P(z|x, t_2)P(A|y, t_3)P(G|y, t_4)P(T|z, t_5)P(T|z, t_6)$$

- There are methods that economize this computation to recursively compute the likelihood of subtrees at a node and move up the tree towards the root.
- Note that as long as the model of substitution used to compute the underlying transition probabilities is reversible, the inferred tree is *unrooted*

Finding the maximum likelihood tree

- Essentially, we are searching in a space of trees with branch lengths.
- We need to find the optimum branch lengths for each given tree topology.
- We also need to search all tree topologies for one that has a set of branch lengths that gives the highest likelihood.
- Various algorithms have been proposed to efficiently search the tree space to accomplish these tasks.