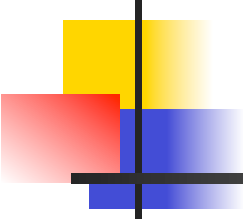




Maximum Likelihood Tree Construction

Dennis Pearl, April 6, 2010



Idea: Find the tree which maximizes the likelihood of the observed sequences under a particular evolutionary model.

This lecture:

- Quick Example
- Probability models
- Calculating the likelihood

April 20th:

- Algorithms



Quick Example

Leitner's Swedish Social Network

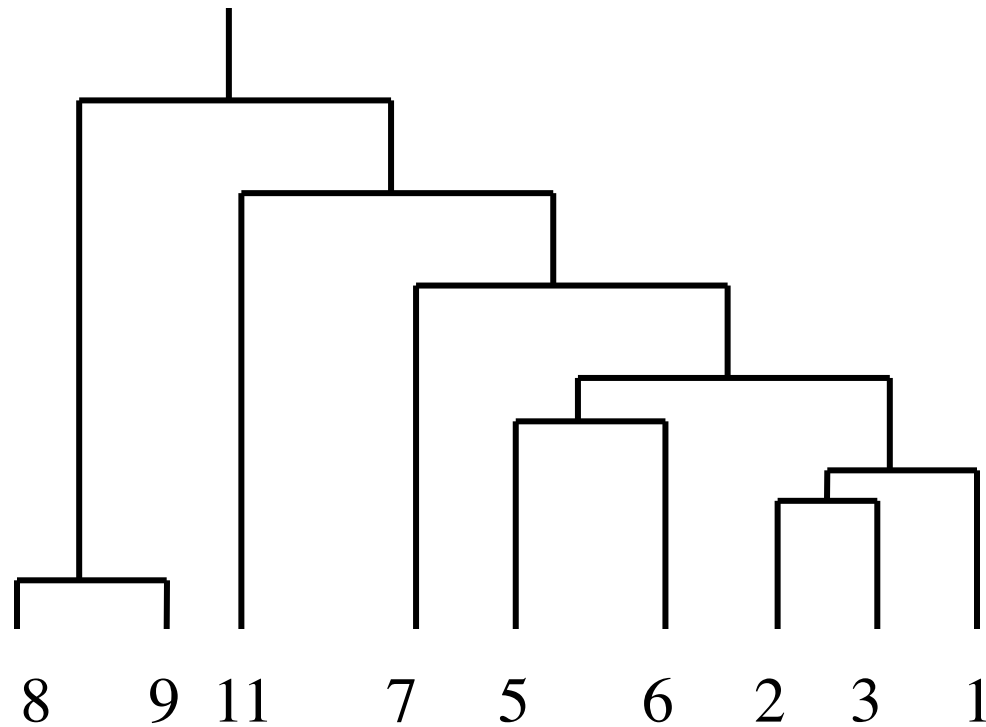
- Index case patient 1 transmitted HIV to female patients 8, 11, 7, 5, and 2.
- Patient 5 transmitted the virus to male patient 6.
- Patient 2 transmitted the virus to child patient 3
- Patient 8 transmitted the virus to child patient 9
- Times for each transmission are known within a few months.

Subset of the data:

| | |
|------------|--|
| P1 | GTAGTAATTAGATCTGAAAAC TTCTCGAACCAATGCTAAAACCATAA |
| P2 | -----A----- |
| P3 | A-----A----- |
| P5 | -----A-----G----- |
| P6 | -----A-----G----- |
| P7 | A-----A----- |
| P8 | A-----A--G----- |
| P9 | A-----A--G----- |
| P11 | -----C-----A--G----- |

Note “-” used to indicate same nucleotide as index patient 1

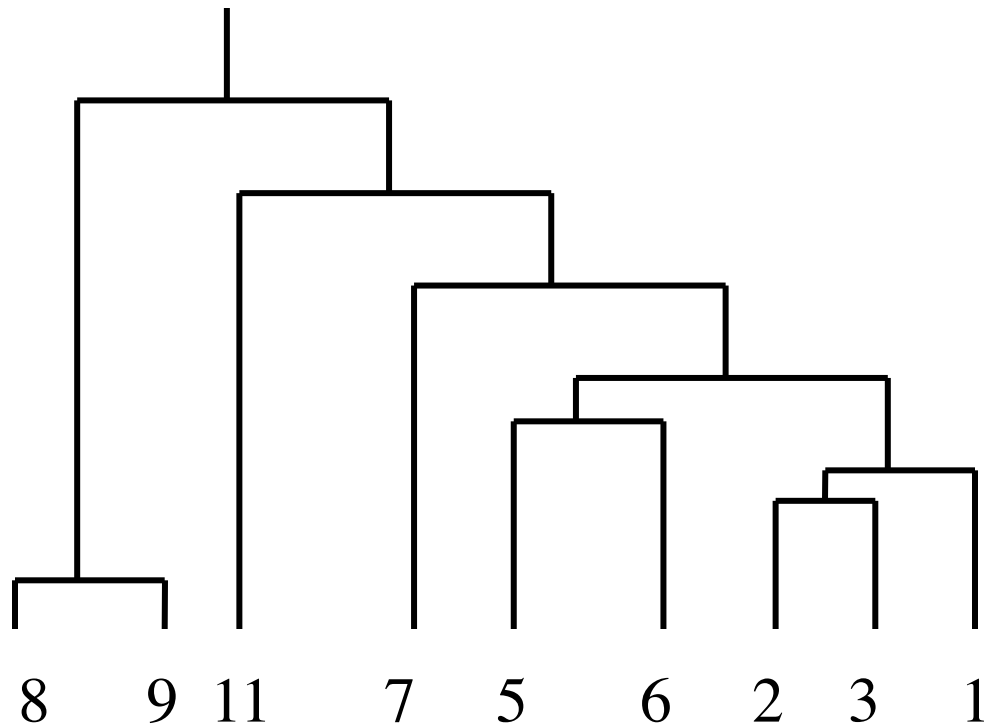
The Swedish Social Network - (Tree of highest likelihood)



$P(\text{Tree}) =$

$P(\text{Topology})P(\text{Split Times}|\text{Topology})P(\text{Sequences}|\text{Topology, Times})$

The Swedish Social Network - (Tree of highest likelihood)



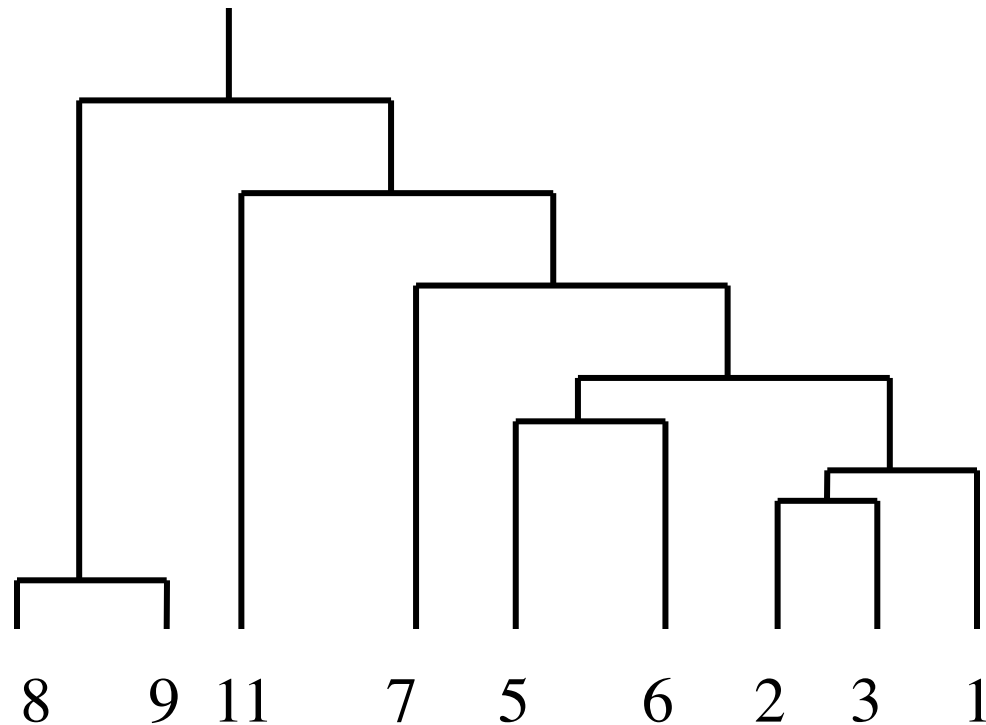
Data | Parameter



$P(\text{Tree}) =$

$P(\text{Topology})P(\text{Split Times}|\text{Topology})P(\text{Sequences}|\text{Topology, Times})$

The Swedish Social Network - (Tree of highest likelihood)

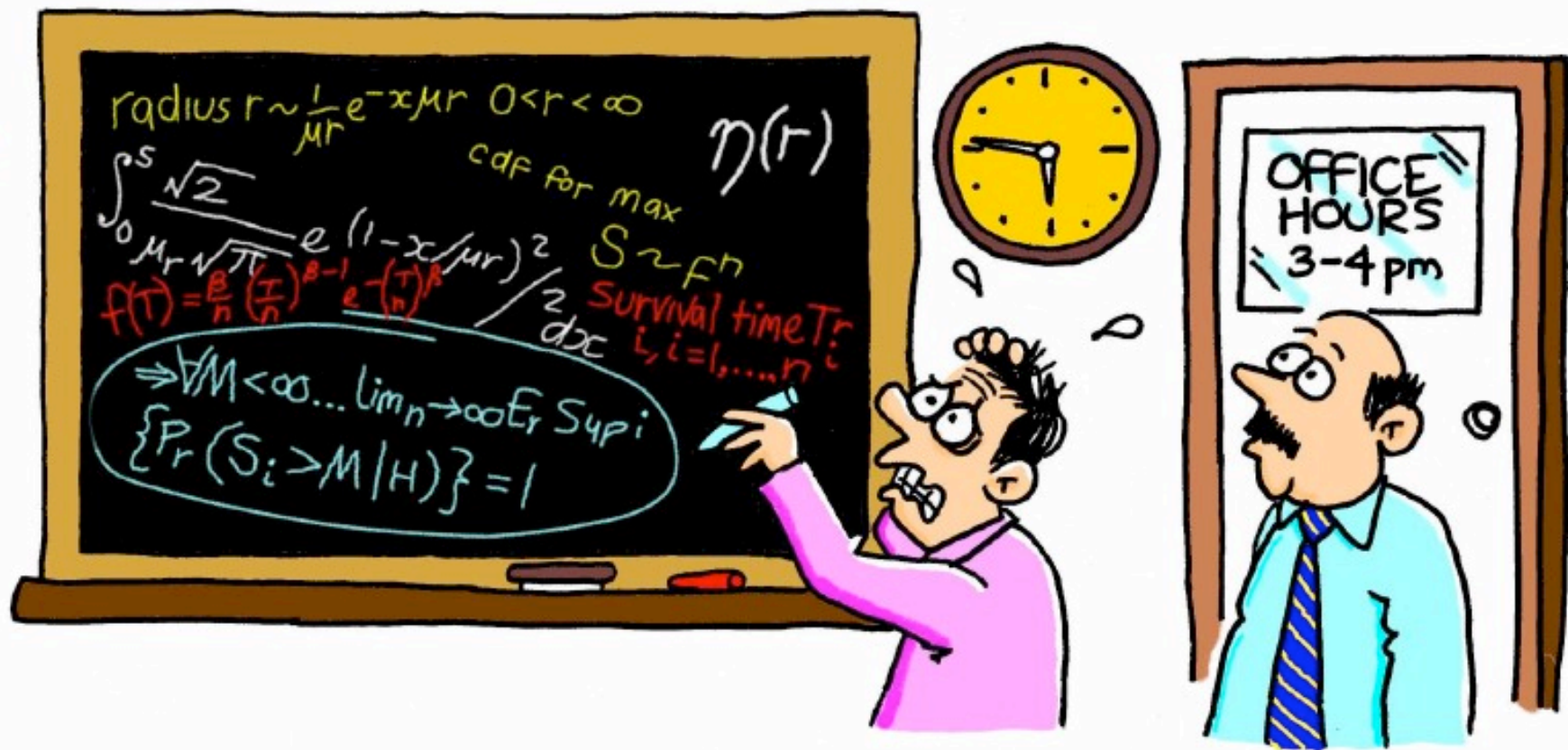


Recall patient 1 infected patients 8 then 11 then 7 then 5 and then 2. Patient 5 infected patient 6. Patient 2 infected patient 3. Patient 8 infected patient 9.



Probabilistic Models of Molecular Evolution

Idea: Model the molecular events that drive the changes seen in nucleotide and amino acid sequences.



He watched patiently as his student battled to try and calculate "a snowball's chance in hell".

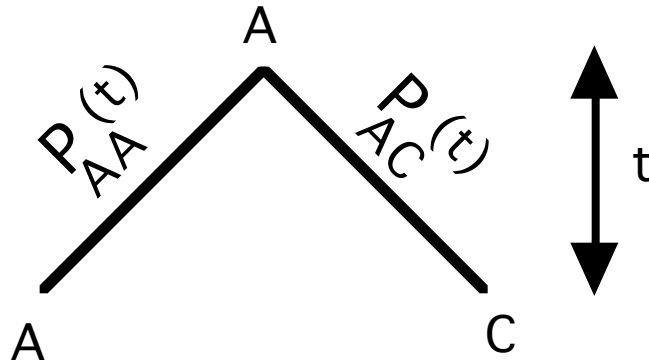


Probability Models

- Molecular evolution is reasonably well modeled by finite state continuous time Markov chains.
- More controversial are assumptions of stationarity and site-to-site independence which underlie many phylogenetic methods.
- The evolution of a single “site” is modeled over the state space of an n-letter alphabet.

Probability Models

- Define $X(t)$ = letter at time t ,
- $P_{ij}(t) = P(X(t+s) = j | X(s) = i)$ (stationarity)



- Matrix notation $\mathbf{P}(t)$ (note: rows add to 1)



Probability Models

- Need to assume:

$$\lim_{t \rightarrow 0} P_{ij}(t) = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$



Markov Process Theory

- Chapman-Kolmogorov

$$P_{ik}(s+t) = \sum_{j=1}^n P_{ij}(s)P_{jk}(t) \quad \text{for } t, s \geq 0.$$

In matrix notation

$$\mathbf{P}(s+t) = \mathbf{P}(t)\mathbf{P}(s)$$



Markov Process Theory

- $\mathbf{P}(t)$ is continuous and differentiable:

$$\mathbf{Q} = \mathbf{P}'(0)$$

- Since , $\sum_{j=1}^n \mathbf{P}_{ij}(t) = \mathbf{1}$

$$\sum_{j=1}^n q_{ij} = 0 \quad (\text{i.e. rows of } \mathbf{Q} \text{ sum to } 0).$$



Markov Process Theory

- \mathbf{Q} is called the generator matrix of the process.

The infinitesimal description of the process (h small):

$$P[X(h)=j|X(0)=i] = q_{ij}h + o(h) \quad \text{for } i \neq j$$

$$P[X(h)=i|X(0)=i] = 1 + q_{ii}h + o(h)$$

$$\mathbf{Q} = \mathbf{P}'(0) \quad \mathbf{P}'(t) = \mathbf{P}(t) \mathbf{Q}$$

$$\mathbf{P}(t) = \exp(\mathbf{Q}t) = \mathbf{I} + \sum_{k=1}^{\infty} \frac{\mathbf{Q}^k t^k}{k!} \quad (\text{find by diagonalizing } \mathbf{Q})$$



Markov Process Theory

- When the Markov chain is irreducible (all states communicate) then there is a limiting distribution which is independent of the initial state:

$$\lim_{t \rightarrow \infty} P_{ij}(t) = \pi_j > 0$$

- The limiting distribution is found by solving:

$$\mathbf{0} = \boldsymbol{\pi} \mathbf{Q}$$

$$\pi_j(1 - q_{jj}) = \sum_{i \neq j} \pi_i q_{ij} \quad \text{for } j = 0, \dots, n$$



Markov Process Theory

- For a stationary process $\pi \mathbf{P}(t) = \pi$ for all t .
- For a time-reversible process

$$\pi_i P_{ij}(t) = \pi_j P_{ji}(t)$$



Probability Models for Molecular Evolution

- $\mathbf{Q} = \mathbf{P}'(0)$ the generator matrix gives the infinitesimal description of the process.
- The Jukes-Cantor (1969) model: 4-letter alphabet with all substitutions equally likely

$$\mathbf{Q} = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{bmatrix}$$



Probability Models (Jukes-Cantor)

$$P_{ii}(t + h) = (1 - 3\alpha h)P_{ii}(t) + \alpha h(1 - P_{ii}(t)) + o(h)$$

$$P'_{ii}(t) = -4\alpha P_{ii}(t) + \alpha$$

subject to $P_{ii}(0)=1$

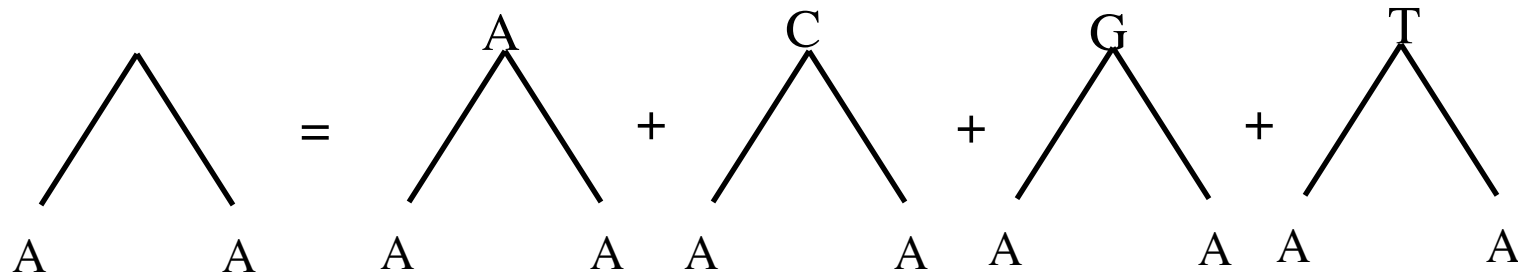
- For Jukes-Cantor

$$P_{ij}(t) = \begin{cases} 0.25(1 + 3\exp(-4\alpha t)) & i = j \\ 0.25(1 - \exp(-4\alpha t)) & i \neq j \end{cases}$$

$$\pi_A = \pi_C = \pi_G = \pi_T = 0.25$$

Probability Models (Jukes-Cantor)

- If two sequences evolved from a common ancestor



Then the chance that $j=k$ is

$$[0.25(1+3\exp(-4\alpha t))]^2 + 3[0.25(1-\exp(-4\alpha t))]^2 \\ = 0.25(1+3\exp(-8\alpha t))$$

$\theta = P[\text{two sequences differ at a particular site}]$

$$= 0.75(1-\exp(-8\alpha t))$$



Probability Models (Jukes-Cantor)

- The expected number of substitutions per site from ancestor to descendent over time t is $3\alpha t$.
- The expected number of substitutions per site for two sequences that evolved from a common ancestor over time t is $D = 6\alpha t = -0.75 \text{Ln}(1 - 4\theta/3)$.

Pooling the information from all of the K sites gives the estimate

$$\hat{D} = -0.75 \text{Ln}(1 - 4\hat{\theta} / 3) \quad \text{where}$$

$\hat{\theta}$ is the proportion of sites that differ between the two sequences.



Probability Models (Jukes-Cantor)

- \hat{D} estimates the expected number of substitutions per site. It is called the “Jukes-Cantor distance” between the two sequences.
- Using the delta-method...

$$\text{Var}(\hat{D}) \approx D'(\theta)^2 \text{Var}(\hat{\theta}) = \frac{\theta(1-\theta)}{K(1-4\theta/3)^2}$$



Kimura's (1980) Model

- Two parameter model. The rate for transitions may be more common than for transversions ($\kappa = \alpha/\beta > 1$):

$$\mathbf{Q} = \begin{matrix} & A & C & G & T \\ \begin{matrix} A \\ C \\ G \\ T \end{matrix} & \begin{bmatrix} -\alpha - 2\beta & \beta & \alpha & \beta \\ \beta & -\alpha - 2\beta & \beta & \alpha \\ \alpha & \beta & -\alpha - 2\beta & \beta \\ \beta & \alpha & \beta & -\alpha - 2\beta \end{bmatrix} \end{matrix}$$



Kimura's (1980) Model

- Solving $\mathbf{0} = \boldsymbol{\pi}\mathbf{Q}$ gives the limiting values

$$\pi_A = \pi_C = \pi_G = \pi_T = 0.25$$

$$P_{ij}(t) = \begin{cases} \frac{1}{4}(1 + \exp(-4\beta t) + 2\exp(-2(\alpha + \beta)t)) & i = j \\ \frac{1}{4}(1 + \exp(-4\beta t) - 2\exp(-2(\alpha + \beta)t)) & \text{TRANSITION} \\ \frac{1}{4}(1 - \exp(-4\beta t)) & \text{TRANSVERSION} \end{cases}$$



Felsenstein's (1981) Model

- Four parameter model. The stationary probabilities may differ:

$$\mathbf{Q} = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} -\mu(1 - \pi_A) & \mu\pi_C & \mu\pi_G & \mu\pi_T \\ \mu\pi_A & -\mu(1 - \pi_C) & \mu\pi_G & \mu\pi_T \\ \mu\pi_A & \mu\pi_C & -\mu(1 - \pi_G) & \mu\pi_T \\ \mu\pi_A & \mu\pi_C & \mu\pi_G & -\mu(1 - \pi_T) \end{bmatrix}$$

$$P_{ij}(t) = \begin{cases} \pi_j + (1 - \pi_j) \exp(-\mu t) & i = j \\ \pi_j (1 - \exp(-\mu t)) & i \neq j \end{cases}$$



Felsenstein's (1981) Model

- Solve $\mathbf{P}(t) = \exp(\mathbf{Q}t)$

$$P_{ij}(t) = \begin{cases} \pi_j + (1 - \pi_j)\exp(-\mu t) & i=j \\ \pi_j(1 - \exp(-\mu t)) & i \neq j \end{cases}$$



Hasegawa-Kishino-Yano (1985) Model

- Five parameter model. Differing stationary probabilities and possible transition-transversion bias.

$$Q = \begin{bmatrix} -\mu(\kappa\pi_G + \pi_W) & \mu\pi_C & \mu\kappa\pi_G & \mu\pi_T \\ \mu\pi_A & -\mu(\kappa\pi_T + \pi_V) & \mu\pi_G & \mu\kappa\pi_T \\ \mu\kappa\pi_A & \mu\pi_C & -\mu(\kappa\pi_A + \pi_W) & \mu\pi_T \\ \mu\pi_A & \mu\kappa\pi_C & \mu\pi_G & -\mu(\kappa\pi_C + \pi_V) \end{bmatrix}$$

where $\pi_W = \pi_C + \pi_T$ and $\pi_V = \pi_A + \pi_G$



Hasegawa-Kishino-Yano (1985) Model

- This model is “time reversible”

$$\pi_i P_{ij}(t) = \pi_j P_{ji}(t)$$

- The most general time reversible model GTR_4 for a 4-letter alphabet has 9 parameters (8 identifiable).



Goldman & Yang's (1994) Codon Model

- 61 letter alphabet of the “sense” codons
- 60 stationary probabilities + transition rate + transversion rate + synonymous/non-synonymous ratio
- Instantaneous rates requiring more than one base change = 0
- This is a time-reversible model



Protein sequence evolution

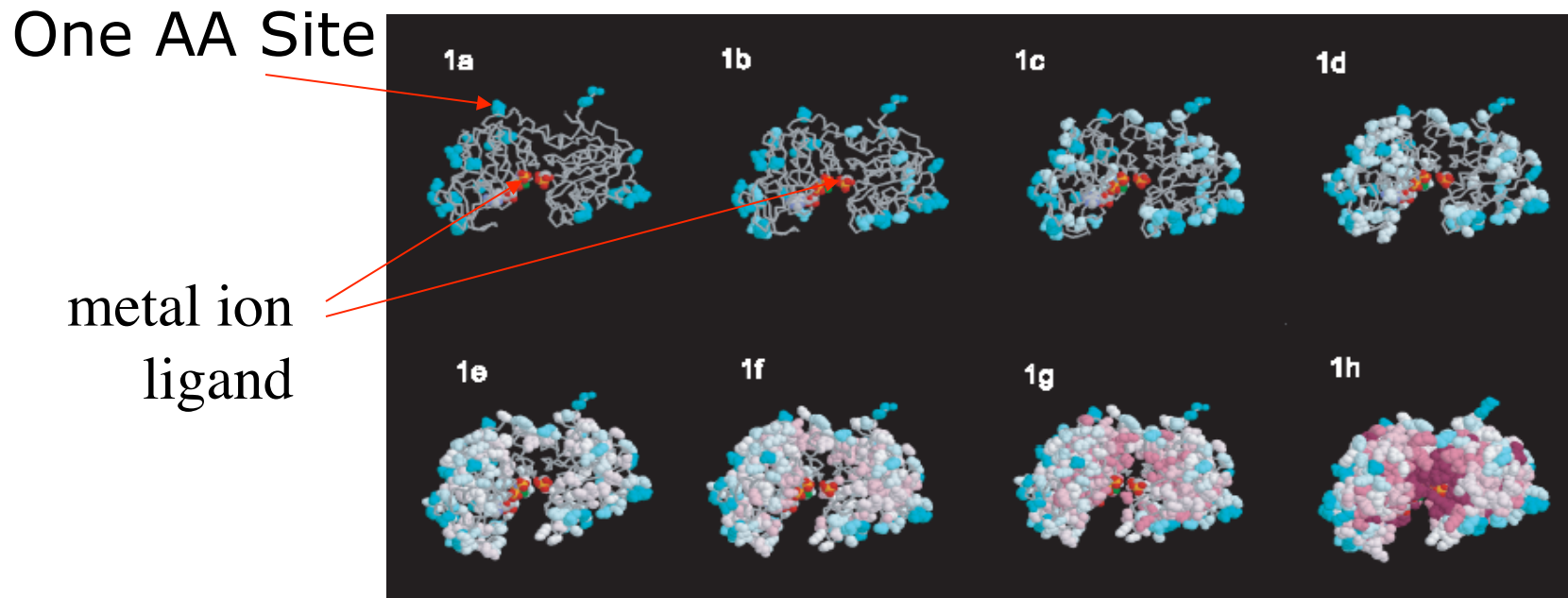
- Works with 20-letter alphabet of amino acids
- Biochemical properties such as size, charge, hydrophobicity often form classes to generalize Kimura's model
- Empirical substitution models based on large databases [e.g. PAM (1979) JTT (1994), WAG (2001)]



Site Heterogeneity

- Simple model: some sites are “invariable” with probability p while others are free to change.
- Generalization: sites have a random non-negative rate R . Often R is taken to follow a gamma distribution.
- Independence of sites keeps the calculation of the log-likelihood tractable.
- Can also have R depend on site specific information unrelated to letter occupying site.

Influence of Spatial Location



More Conserved

(Pan, 2006 suggests $Q^* = f(r) Q$ with f monotone in distance r)



Dealing with Other Issues

- The covarion model. Rates within a site can vary over time. Uses theory from Markov Modulated Markov chains
- Autocorrelation models. Rates at neighboring sites are correlated
- Hidden Markov Models (e.g. rates determined by random process based on previous sites)



Model-based distances

- Define $Y(t)$ = letter at time t for a second sequence that evolved from a common ancestor of X over time t .
- For a time-reversible model:

$$\begin{aligned} P[X(t)=i, Y(t)=j] &= \sum_k \pi_k P_{ki}(t) P_{kj}(t) = \sum_k \pi_i P_{ik}(t) P_{kj}(t) \\ &= \pi_i P_{ij}(2t) \end{aligned}$$



Model-based distances

- Since the generator matrix gives the infinitesimal description of the process, we can see that the rate of change is given by

$$\lim_{\Delta \rightarrow 0} P[X(t + \Delta) \neq X(t)] / \Delta = \sum_i -q_{ii} \pi_i$$

- Thus, the expected number of substitutions per site for two sequences that evolved from a common ancestor over time t is

$$D = 2t \sum_i -q_{ii} \pi_i$$



Model-based distances

For the models above:

$$D = 6\alpha t \quad (\text{Jukes-Cantor})$$

$$D = 2(\alpha + 2\beta)t \quad (\text{Kimura, 1980})$$

$$D = 2\mu t \left(1 - \sum_i \pi_i^2\right) \quad (\text{Felsenstein, 1981})$$

$$D = 2\mu t \left[(\kappa - 1)(\pi_A \pi_G + \pi_C \pi_T) + 1 - \sum_i \pi_i^2 \right] \quad (\text{HKY, 1985})$$

To estimate D we can invert the estimate for the number of differences as we did for the Jukes-Cantor model above.



Model-based distances

- In the Felsenstein 1981 model:

$$P[X(t) = Y(t)] = \sum_k \pi_k P_{kk}(2t)$$

$$= \sum_k \pi_k [\pi_k + (1 - \pi_k) \exp(-2\mu t)]$$

$$\theta = P[X(t) \neq Y(t)] = 1 - \sum_k \pi_k [\pi_k + (1 - \pi_k) \exp(-2\mu t)]$$

$$= B(1 - \exp(D / B)) \quad \text{where } B = (1 - \sum_i \pi_i^2).$$

Solving ... $D = -B \ln(1 - \theta / B)$



Model-based distances

- Finally, pooling across K sites

$$\hat{D} = -B \text{Ln}(1 - (\hat{\theta} / B)),$$

which has approximate variance (for fixed B)

$$\text{Var}_B(\hat{D}) \approx D'(\theta)^2 \text{Var}(\hat{\theta}) = \frac{\theta(1 - \theta)}{K(1 - (\theta / B))^2}.$$

Notice that

$B = \text{P}[\text{two unrelated sequences differ at a site}]$

$\theta = \text{P}[\text{two related sequences differ at a site}]$



Calculating the Likelihood

- Suppose model is time-reversible, sites are independent, and evolution has reached the equilibrium state.
- Time reversibility implies that the likelihood of a tree does not depend on how it is rooted.
- Independence implies the likelihood, L , can be put in terms of the products of likelihoods for each site:

$$L = \prod_{j=1}^K L_j \quad \text{Or} \quad \log(L) = \sum_j \log(L_j)$$

Calculating the likelihood for a single site

Sequence # site j

1C.....

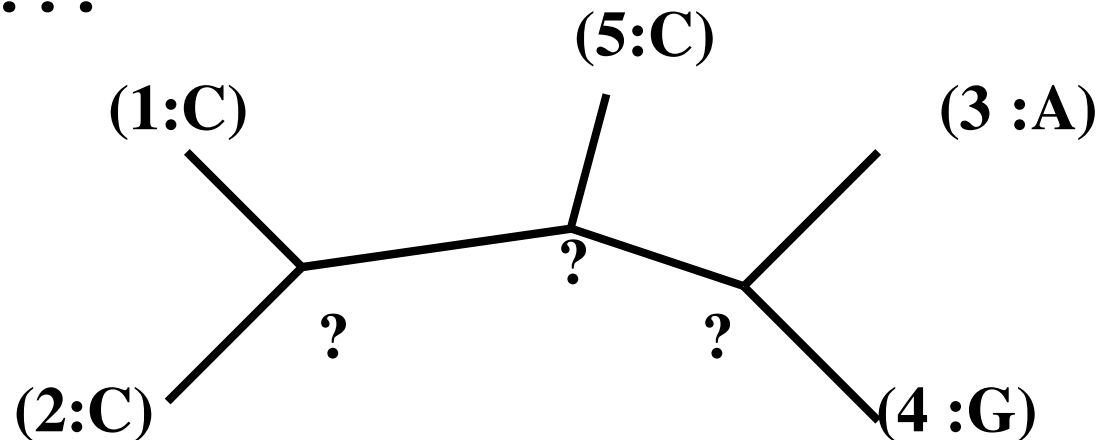
2C.....

3A.....

4G.....

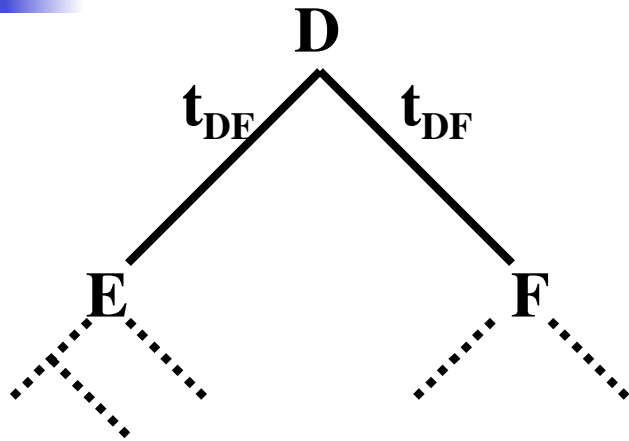
5C.....

Need to find L_j for a candidate tree:





Felsenstein's Peeling Algorithm



Define the conditional likelihood of the subtree descending from node **D** given the nucleotide i is at **D** recursively by:

$$L(i \text{ at } \mathbf{D}) = \left(\sum_k P_{ik}(t_{DE})L(k \text{ at } E) \right) \left(\sum_m P_{im}(t_{DF})L(m \text{ at } F) \right)$$

The likelihood at site j is then its likelihood at the root:

$$L_j = \sum_i P(i \text{ at root})L(i \text{ at root}) = \sum_i \pi_i L(i \text{ at root})$$



Calculating the Likelihood

- The log likelihood for the tree is found by summing over sites.
- For n taxa, we need just one function call for each of the $n-2$ internal nodes
- Calculating the likelihood for a fixed tree goes fast. However finding the tree that optimizes this criteria goes slow:

Remember, there are $\prod_{k=1}^{n-2} (2k-1)$ possible unrooted tree topologies to consider.



Which Model to Choose? – The LR Statistic

- Often a null model is a special case of an alternative model.
- Take $LR = 2$ (maximum log likelihood under alternative – maximum log likelihood under null)
- If the null model is true then LR has an approximate chi-square distribution with d.f. = difference in number of parameters.