Lecture Outline: Phylogeny Reconstruction using Distance Methods

1. **Summary of Models of Molecular Evolution**

   (a) The standard form for the rate matrix of a time-reversible continuous-time Markov model of DNA substitution is the following, where the base order is A, C, G, T and the dots on the main diagonal represent the negative row-sum of the off-diagonal elements. The most general model is the general time-reversible model (GTR).

$$Q_{\text{GTR}} = \begin{pmatrix} \cdot & s_{AC}\pi_C & s_{AG}\pi_G & s_{AT}\pi_T \\ s_{AC}\pi_A & \cdot & s_{CG}\pi_G & s_{CT}\pi_T \\ s_{AG}\pi_A & s_{CG}\pi_c & \cdot & s_{GT}\pi_T \\ s_{AT}\pi_A & s_{CT}\pi_C & s_{GT}\pi_G & \cdot \end{pmatrix}$$

   (b) The six parameters $s_{ij} = s_{ji}$ for each pair of bases appear symmetrically in the matrix across the main diagonal.

   (c) The four parameters $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$ are the *stationary distribution* and represent both the long-run relative frequency of the bases but also the probabilities of the bases at any given time.

   (d) There are only nine free parameters as the stationary distribution satisfies the constraint $\sum_i \pi_i = 1$.

   (e) Some models also define $\pi_R = \pi_A + \pi_G$ for the sum of purine probabilities and $\pi_Y = \pi_C + \pi_T$ for the sum of pyrimidine probabilities.

   (f) The expected number of substitutions per unit time is the weighted average of the negative main diagonal elements.

$$\begin{aligned} \mu &= \pi_A(s_{AC}\pi_c + s_{AG}\pi_G + s_{AT}\pi_T) + \pi_C(s_{AC}\pi_A + s_{CG}\pi_G + s_{CT}\pi_T) \\ &\quad + \pi_G(s_{AG}\pi_A + s_{CG}\pi_C + s_{GT}\pi_T) + \pi_T(s_{AT}\pi_A + s_{CT}\pi_C + s_{GT}\pi_G) \\ &= 2(\pi_A\pi_C s_{AC} + \pi_A\pi_G s_{AG} + \pi_A\pi_T s_{AT} + \pi_C\pi_G s_{CG} + \pi_C\pi_T s_{CT} + \pi_G\pi_T s_{GT}) \end{aligned}$$

   (g) Often the matrix Q is rescaled so that one unit of time represents one expected substitution (per site) by dividing each element by μ which reduces the number of free parameters by one.

   (h) There are many standard models that are special cases of the GTR model. These are:

Jukes-Cantor (JC69)

$$Q_{\text{JC69}} = \mu^{-1}\begin{pmatrix} \cdot & 1 & 1 & 1 \\ 1 & \cdot & 1 & 1 \\ 1 & 1 & \cdot & 1 \\ 1 & 1 & 1 & \cdot \end{pmatrix}$$

Kimura 2-parameter (K80)

$$Q_{\text{K80}} = \mu^{-1}\begin{pmatrix} \cdot & 1 & \kappa & 1 \\ 1 & \cdot & 1 & \kappa \\ \kappa & 1 & \cdot & 1 \\ 1 & \kappa & 1 & \cdot \end{pmatrix}$$

Felsenstein (F81)

$$Q_{\text{F81}} = \mu^{-1}\begin{pmatrix} \cdot & \pi_C & \pi_G & \pi_T \\ \pi_A & \cdot & \pi_G & \pi_T \\ \pi_A & \pi_C & \cdot & \pi_T \\ \pi_A & \pi_C & \pi_G & \cdot \end{pmatrix}$$

Felsenstein (F84)

$$Q_{\text{F84}} = \mu^{-1}\begin{pmatrix} \cdot & \pi_C & (1+\kappa/\pi_R)\pi_G & \pi_T \\ \pi_A & \cdot & \pi_G & (1+\kappa/\pi_Y)\pi_T \\ (1+\kappa/\pi_R)\pi_A & \pi_C & \cdot & \pi_T \\ \pi_A & (1+\kappa/\pi_Y)\pi_C & \pi_G & \cdot \end{pmatrix}$$

HKY85

$$Q_{\text{HKY85}} = \mu^{-1}\begin{pmatrix} \cdot & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & \cdot & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & \cdot & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & \cdot \end{pmatrix}$$

Tamura-Nei (TN93)

$$Q_{\text{TN93}} = \mu^{-1}\begin{pmatrix} \cdot & \pi_C & \kappa_R\pi_G & \pi_T \\ \pi_A & \cdot & \pi_G & \kappa_Y\pi_T \\ \kappa_R\pi_A & \pi_C & \cdot & \pi_T \\ \pi_A & \kappa_Y\pi_C & \pi_G & \cdot \end{pmatrix}$$

Model	Stationary Distribution	Rates	# Parameters
JC69	Uniform	Equal	1
	$\pi = (0.25, 0.25, 0.25, 0.25)$	$s_{AC} = s_{AG} = s_{AT} = s_{CG} = s_{CT} = s_{GT}$	
K80	Uniform	Transitions $\neq$ Transversions	2
	$\pi = (0.25, 0.25, 0.25, 0.25)$	$s_{AG} = s_{CT};\ s_{AC} = s_{AT} = s_{CG} = s_{GT}$	
F81	Flexible	Equal	4
	$\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$	$s_{AC} = s_{AG} = s_{AT} = s_{CG} = s_{CT} = s_{GT}$	
HKY85 F84	Flexible	Transitions $\neq$ Transversions	5
	$\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$	$s_{AG} = s_{CT};\ s_{AC} = s_{AT} = s_{CG} = s_{GT}$	
TN93	Flexible	Two Transition rates, Transversions	6
	$\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$	$s_{AG};\ s_{CT};\ s_{AC} = s_{AT} = s_{CG} = s_{GT}$	
GTR	Flexible	Flexible	9
	$\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$	$s_{AC};\ s_{AG};\ s_{AT};\ s_{CG};\ s_{CT};\ s_{GT}$	

(i) The probability transition matrix has the form $P(t) = \{p_{ij}(t)\} = \mathrm{e}^{Qt}$ where $p_{ij}(t)$ equals the probability of being in state j after t time units given that the process begins in state i.

(j) The spectral decomposition of Q is $Q = V\Lambda V^{-1}$ where

i. V is a 4×4 matrix whose columns are eigenvectors of Q;

ii. Λ is a 4×4 diagonal matrix whose diagonal elements are eigenvalues of Q;

iii. V^{-1} is the 4×4 matrix such that $VV^{-1} = V^{-1}V = I$.

(k) The probability transition matrix has the form $P(t) = V\mathrm{e}^{\Lambda t}V^{-1}$.

(l) If the eigenvalues of Q are $\lambda_1 = 0,\ \lambda_2,\ \lambda_3,\ \lambda_4 < 0$, then

$$p_{ij}(t) = \pi_j + c_2^{(ij)}\mathrm{e}^{\lambda_2 t} + c_3^{(ij)}\mathrm{e}^{\lambda_3 t} + c_4^{(ij)}\mathrm{e}^{\lambda_4 t}$$

where

$$c_2^{(ij)} + c_3^{(ij)} + c_4^{(ij)} = \begin{cases} 1 - \pi_j & \text{when } i = j \\ -\pi_j & \text{when } i \neq j \end{cases}$$

so that

$$p_{ij}(0) = \begin{cases} 1 & \text{when } i = j \\ 0 & \text{when } i \neq j \end{cases}$$

(m) The probability transtion matrix of the GTR model does not have an algebraic form and must be computed numerically.

(n) The TN93 model has the following spectral decomposition, $Q = V\Lambda V^{-1}$, with

$$Q_{\text{TN93}} = \mu^{-1} \begin{pmatrix} \cdot & \pi_C & \kappa_R\pi_G & \pi_T \\ \pi_A & \cdot & \pi_G & \kappa_Y\pi_T \\ \kappa_R\pi_A & \pi_C & \cdot & \pi_T \\ \pi_A & \kappa_Y\pi_C & \pi_G & \cdot \end{pmatrix}$$

$$V = \begin{pmatrix} 1 & \frac{-1}{\pi_R} & \frac{\pi_G}{\pi_R} & 0 \\ 1 & \frac{1}{\pi_Y} & 0 & \frac{\pi_T}{\pi_Y} \\ 1 & \frac{-1}{\pi_R} & \frac{-\pi_A}{\pi_R} & 0 \\ 1 & \frac{1}{\pi_Y} & 0 & \frac{-\pi_T}{\pi_Y} \end{pmatrix}$$

$$\Lambda = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -(\kappa_R\pi_R + \pi_Y) & 0 \\ 0 & 0 & 0 & -(\kappa_Y\pi_Y + \pi_R) \end{pmatrix}$$

$$V^{-1} = \begin{pmatrix} \pi_A & \pi_C & \pi_G & \pi_T \\ -\pi_A\pi_Y & \pi_C\pi_R & -\pi_G\pi_Y & \pi_T\pi_R \\ 1 & 0 & -1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix}$$

(o) The probability transition matrix is $P_{\text{TN93}} = V\mathrm{e}^{\Lambda t}V^{-1}$ which equals

$$\begin{pmatrix} \pi_A + \frac{\pi_A\pi_Y}{\pi_R}e_2 + \frac{\pi_G}{\pi_R}e_3 & \pi_C - \pi_C e_2 & \pi_G + \frac{\pi_G\pi_Y}{\pi_R}e_2 - \frac{\pi_G}{\pi_R}e_3 & \pi_T - \pi_T e_2 \\ \pi_A - \pi_A e_2 & \pi_C + \frac{\pi_C\pi_R}{\pi_Y}e_2 + \frac{\pi_T}{\pi_Y}e_4 & \pi_G - \pi_G e_2 & \pi_T + \frac{\pi_T\pi_R}{\pi_Y}e_2 - \frac{\pi_T}{\pi_Y}e_4 \\ \pi_A + \frac{\pi_A\pi_Y}{\pi_R}e_2 - \frac{\pi_A}{\pi_R}e_3 & \pi_C - \pi_C e_2 & \pi_G + \frac{\pi_G\pi_Y}{\pi_R}e_2 + \frac{\pi_A}{\pi_R}e_3 & \pi_T - \pi_T e_2 \\ \pi_A - \pi_A e_2 & \pi_C + \frac{\pi_C\pi_R}{\pi_Y}e_2 - \frac{\pi_C}{\pi_Y}e_4 & \pi_G - \pi_G e_2 & \pi_T + \frac{\pi_T\pi_R}{\pi_Y}e_2 + \frac{\pi_C}{\pi_Y}e_4 \end{pmatrix}$$

where $e_2 = \mathrm{e}^{-t}$, $e_3 = \mathrm{e}^{-(\kappa\pi_R+\pi_Y)t}$, and $e_4 = \mathrm{e}^{-(\kappa\pi_Y+\pi_R)t}$.

2. Distance Estimation

(a) Consider estimating pairwise distances under the TN93 model.
Assume that we have estimates of all parameters, π, κ_R, and κ_Y.
Then we would want to find the time t that would maximize the probability of the observed data.
If we observed a site AC, the probability of this would be

$$\pi_A P_{AC}(t)$$

as a function of t.

(b) The likelihood of the data would be

$$L(t) = \prod_i \prod_j (\pi_i P_{ij}(t))^{n_{ij}}$$

where n_{ij} is the number of times that pattern ij is observed.

(c) In practice, it is easier to work with the natural logarithm of this expression.

$$\ell(t) = \sum_i \sum_j n_{ij} \left(\ln \pi_i + \ln P_{ij}(t)\right)$$

(d) For a pair of sequences, you can summarize the number of observed site patterns in a 4×4 matrix.

```
                        10          20          30          40          50
                         +           +           +           +           +
swan        GTG ACC TTC ATC AAC CGA TGA CTA TTT TCC ACT AAC CAT AAA GAT ATC GG
osprey      ATG ACA TTC ATC AAC CGA TGA CTA TTC TCA ACC AAC CAC AAA GAC ATT GG
```

Above are the first 50 bases of the *cytochrome oxidase I* mitochondrial gene from swan and osprey.
Here is a summary of the count data.

		Osprey				
		A	C	G	T	Total
Swan	A	16	0	0	0	16
	C	3	9	0	1	13
	G	1	0	6	0	7
	T	0	4	0	10	14
	Total	20	13	6	11	50

3. **General Estimation Method:**
   (a) Use observed data to estimate parameters of the Q matrix.
   (b) Use maximum likelihood to estimate the distance given the observed data.

4. **Example: Jukes-Cantor**

   There are no parameters to estimate from the data.

   We find $\mu = 3$.

   The probability transition matrix has p for $i \neq j$ and $1 - 3 * p$ for $i = j$ where

$$\begin{aligned} p &= \left(\frac{1}{4} - \frac{1}{4}\mathrm{e}^{-\frac{4}{3}t}\right) \\ 1 - 3p &= \left(\frac{1}{4} + \frac{3}{4}\mathrm{e}^{-\frac{4}{3}t}\right) \end{aligned}$$

   If we let n be the total number of sites and x be the number of sites with a change, then the likelihood is

$$L(t) = (1 - 3p)^{n-x} p^x$$

   To maximize this, take logarithms, take the derivative, set equal to 0 and solve. This leads to

$$\hat{p} = \frac{x}{3n}$$

   which makes sense as x/n is the observed fraction of changes and $1/3$ of these, on average in the long run, would go to each of the three possible other bases. By solving

$$\frac{x}{3n} = \left(\frac{1}{4} - \frac{1}{4}\mathrm{e}^{-\frac{4}{3}t}\right)$$

   for t we find the estimated distance

$$\hat{t} = -\frac{3}{4}\ln\left(1 - \frac{4x}{3n}\right)$$

   Note that this formula only is valid when $x/n < 3/4$. Otherwise, the estimated distance is ∞.

   For our example, $x = 9$ and $n = 50$, so $\hat{p} = 9/(3 \times 50) = 0.06$ and $\hat{t} = 0.2058$.

5. **Example: TN93 model**

Estimate the stationary distribution by the observed frequencies. With 50 bases per species, there are 100 observed bases. The estimate is $\pi = (0.36, 0.26, 0.13, 0.25)$.

The two parameters κ_R and κ_Y can be estimated by considering the expected proportions of transitions in purines and pyrimidines and the expected proportion of transversions.

The probability of a transition in a purine after t time units is

$$\pi_A P_{AG}(t) + \pi_G P_{GA}(t) = 2\pi_A\pi_G(1 + \frac{\pi_Y}{\pi_R}e_2 - \frac{1}{\pi_R}e_3)$$

The probability of a transition in a pyrimidine after t time units is

$$\pi_C P_{CT}(t) + \pi_T P_{TC}(t) = 2\pi_C\pi_T(1 + \frac{\pi_R}{\pi_Y}e_2 - \frac{1}{\pi_Y}e_4)$$

The probability of a transversion after t time units is

$$\pi_A P_{AC}(t) + \pi_A P_{AT}(t) + \pi_C P_{CA}(t) + \pi_C P_{CG}(t) + \pi_G P_{GC}(t) + \pi_G P_{GT}(t) + \pi_T P_{TA}(t) + \pi_T P_{TG}(t)$$

which simplifies to

$$2(1 - e_2)(\pi_A\pi_C + \pi_A\pi_T + \pi_C\pi_G + \pi_G\pi_T)$$

The observed proportions of these three types of site patterns are $S_R = 1/50 = 0.02$, $S_Y = 5/50 = 0.10$, and $V = 3/50 = 0.06$.

Equating these observed frequencies with the expected proportions gives three linear equations in three unknowns, which can be solved for e_2, e_3, and e_4.

When these expressions are plugged in, we can solve for κ_R, κ_Y, and t.

Finally, the distance is found by multiplying this time t by the expected number of substitutions per unit time, μ.

The equations are:

$$\begin{aligned} \kappa_R &= \frac{a_R - \pi_Y b}{\pi_R b} \\ \kappa_Y &= \frac{a_Y - \pi_R b}{\pi_Y b} \end{aligned}$$

where

$$\begin{aligned} a_R &= -\ln\left(1 - \frac{\pi_R S_R}{2\pi_A\pi_G} - \frac{V}{2\pi_R}\right) \\ a_Y &= -\ln\left(1 - \frac{\pi_Y S_Y}{2\pi_C\pi_T} - \frac{V}{2\pi_Y}\right) \\ b &= -\ln\left(1 - \frac{V}{2\pi_R\pi_Y}\right) \end{aligned}$$

The estimated distance is

$$d = 2b(\pi_A\pi_G\kappa_R + \pi_C\pi_T\kappa_Y + \pi_R\pi_Y)$$

For our data, the estimates are $d = 0.2231$, $\kappa_R = 1.85$, and $\kappa_Y = 8.24$.

6. **Example: GTR**

Using GTR, there are no closed-form formulas for transition probabilities, so all calculatons need to be done numerically by computer. We can estimate the stationary distribution with observed proportions and estimate the time and 6 other rate parameters by maximum likelihood. Sparing the details, here are the estimates: $d = 0.2140$, $s_{AC} = 1.530$, $s_{AG} = 0.967$, $s_{AT} = 0$, $s_{CG} = 0$, $s_{CT} = 4.169$, and $s_{GT} = 0$.

7. **UPGMA**

(a) UPGMA is an algorithm that will produce a rooted ultrametric tree from pairwise distance data.

(b) If the data matches such a tree exactly, the algorithm will recover the tree.

(c) Here is the algorithm:

i. Find the i and j with the smallest distance D_{ij}.

ii. Create a new group (ij) which has $n_{(ij)} = n_i + n_j$ members.

iii. Connect i and j on the tree to a new node (ij). Give the edges connecting i to (ij) and j to (ij) each length so that the depth of group (ij) is $D_{ij}/2$.

iv. Compute the distance between the new group and all other groups except i and j by using

$$D_{(ij),k} = \left(\frac{n_i}{n_i + n_j}\right) D_{ik} + \left(\frac{n_j}{n_i + n_j}\right) D_{jk}$$

v. Delete columns and rows corresponding to i and j and add one for (ij). If there are two or more groups left, go back to the first step.

8. **Neighbor-Joining**

(a) Neighbor-joining produces an unrooted tree from pairwise distance data.

(b) If the data matches such a tree exactly, the algorithm will recover the tree.

(c) Here is the algorithm:

i. For each leaf, compute $u_i = \sum_{j \neq i} D_{ij}/(n-2)$.

ii. Choose the i and j for which $D_{ij} - u_i - u_j$ is smallest.

iii. Join i and j to a new node with lengths $(D_{ij} + u_i - u_j)/2$ to node i and $(D_{ij} + u_j - u_i)/2$ to node j.

iv. Compute the distance to the new node (ij) and the other groups as

$$D_{(ij),k} = \frac{D_{ik} + D_{jk} - D_{ij}}{2}$$

v. Delete columns and rows corresponding to i and j and add one for (ij). If there are three or more groups left, go back to the first step. Otherwise, connect the two remaining nodes with their distance.