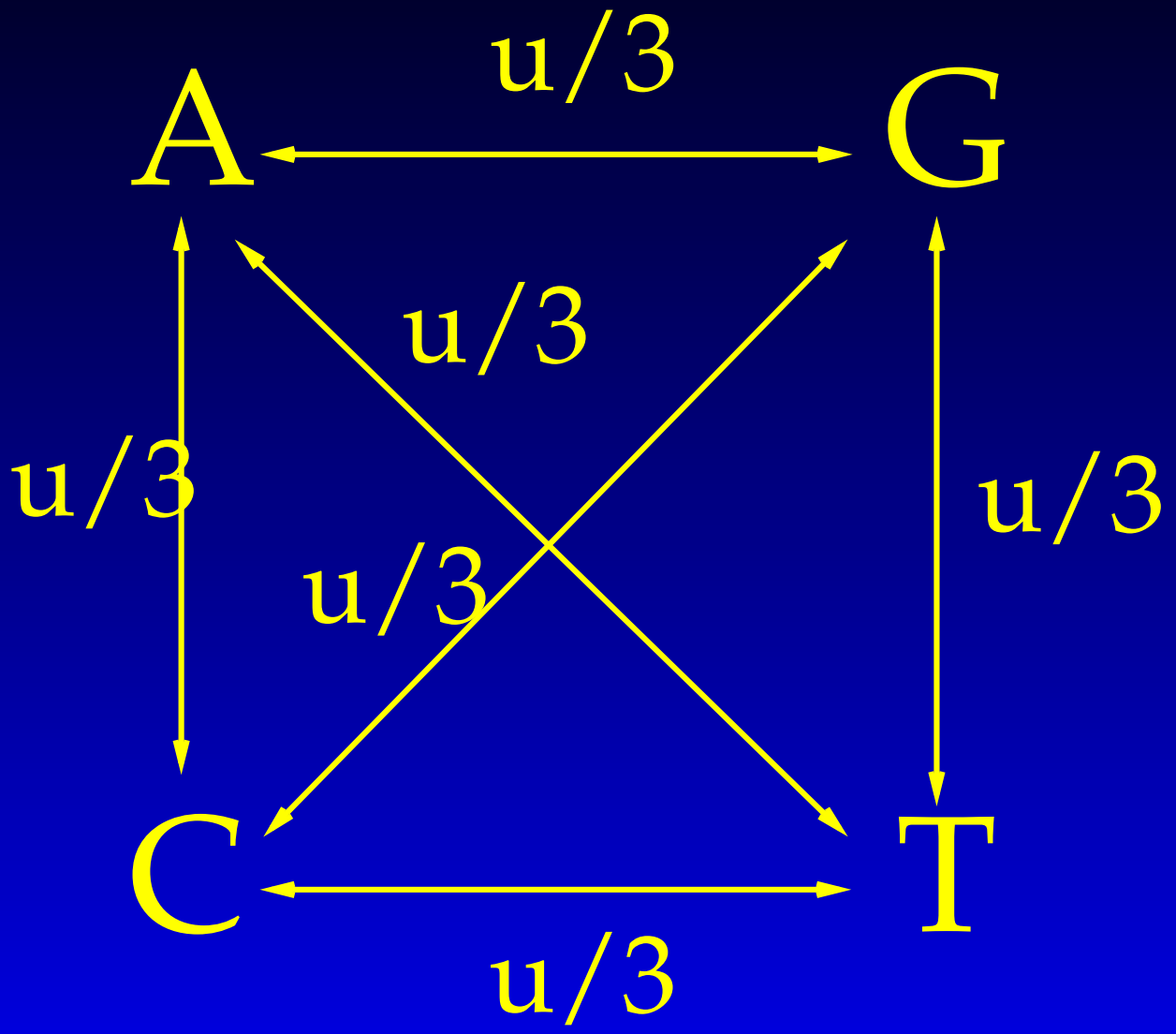
Lecture 24. Phylogeny methods, part 4 (Models of DNA and protein change)

Joe Felsenstein

Department of Genome Sciences and Department of Biology

The Jukes-Cantor model (1969)



the simplest symmetrical model of DNA evolution

Transition probabilities under the Jukes-Cantor model

- All sites change independently
- All sites have the same stochastic process working at them
- Make up a fictional kind of event, such that when it happens the site changes to one of the 4 bases chosen at random (equiprobably)
- Assertion: Having these events occur at rate $\frac{4}{3}u$ is the same as having the Jukes-Cantor model events occur at rate u
- The probability of none of these fictional events happens in time t is $\exp(-\frac{4}{3}ut)$
- No matter how many of these fictional events occur, provided it is not zero, the chance of ending up at a particular base is $\frac{1}{4}$.

Jukes-Cantor transition probabilities, cont'd

Putting all this together, the probability of changing to C, given the site is currently at A, in time t is

$$\text{Prob}\,(C|A,t) = \frac{1}{4}\left(1 - e^{-\frac{4}{3}ut}\right)$$

while

$$\text{Prob}\,(A|A,t) = e^{-\frac{4}{3}t} + \frac{1}{4}\left(1 - e^{-\frac{4}{3}ut}\right)$$

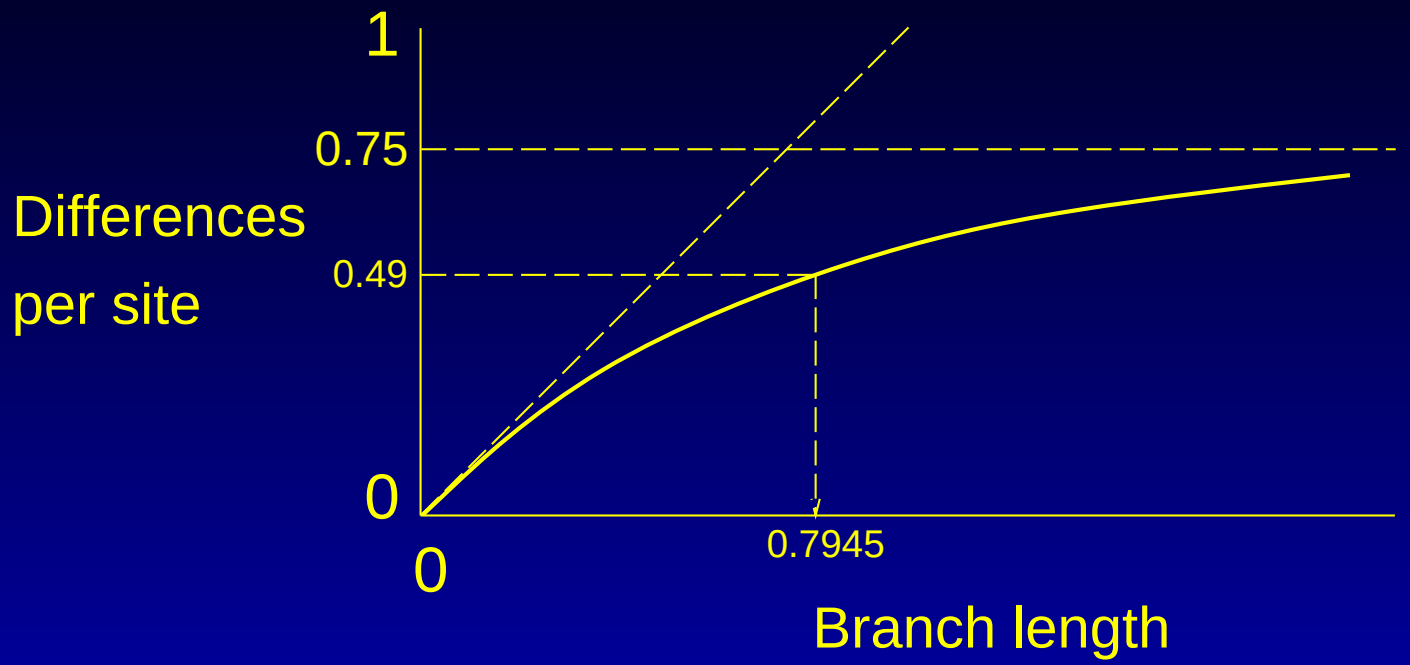
or

$$\text{Prob}\,(A|A,t) = \frac{1}{4}\left(1 + 3e^{-\frac{4}{3}ut}\right)$$

so that the total probability of change is

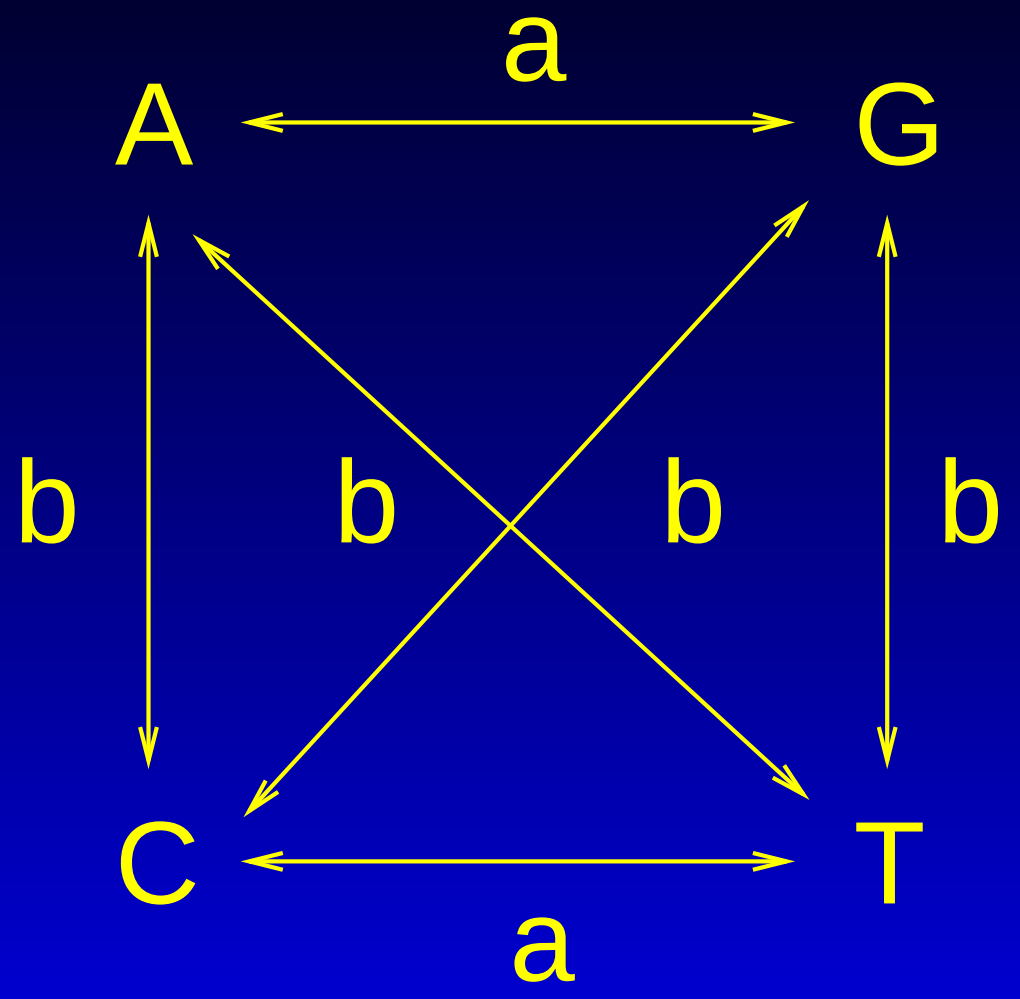
$$\text{Prob}\,(\text{change}|t) = \frac{3}{4}\left(1 - e^{-\frac{4}{3}ut}\right)$$

Fraction of sites different, Jukes-Cantor



after branches of different length, under the Jukes-Cantor model

Kimura's (1980) K2P model of DNA change,



which allows for different rates of transitions and transversions,

Motoo Kimura

Motoo Kimura, with family in Mishima, Japan in the 1960's

Transition probabilities for the K2P model

with two kinds of events:

I. At rate α, if the site has a purine (A or G), choose one of the two purines at random and change to it. If the site has a pyrimidine (C or T), choose one of the pyrimidines at random and change to it.

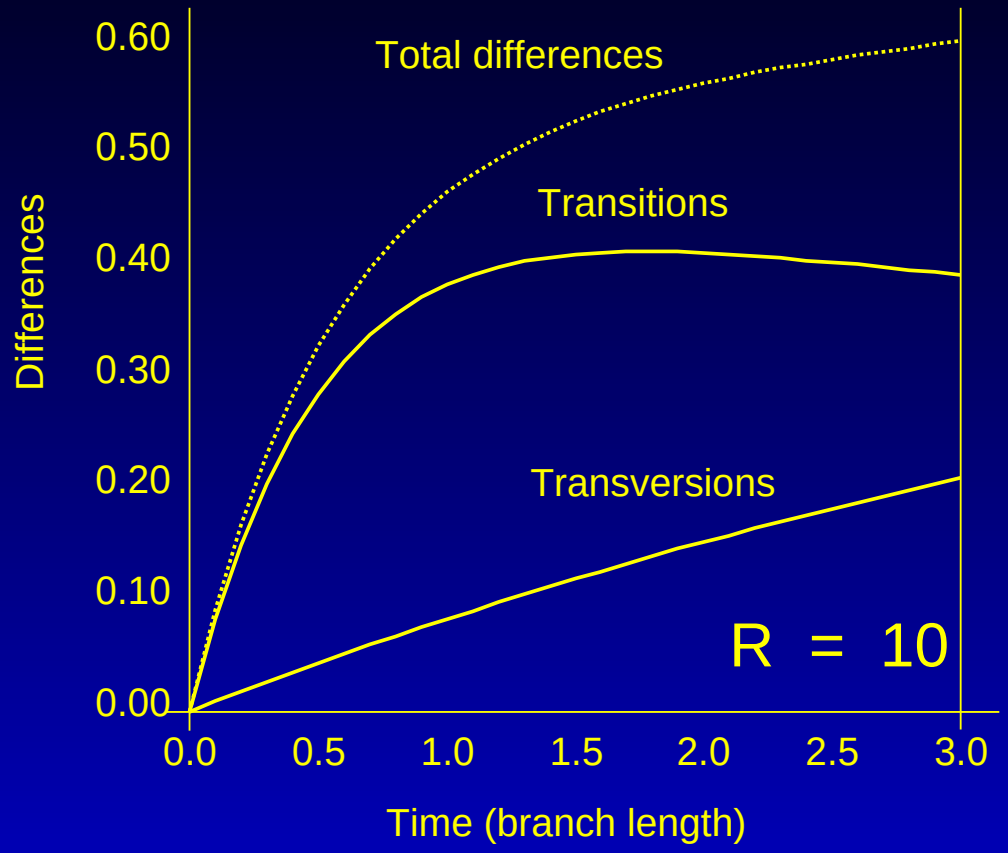
II. At rate β, choose one of the 4 bases at random and change to it.

By proper choice of α and β one can achieve the overall rate of change and T_s/T_n ratio R you want. For rate of change 1, the transition probabilities (*warning: terminological tangle*).

$$\begin{aligned} \text{Prob (transition}|t) &= \tfrac{1}{4} - \tfrac{1}{2}\exp\left(-\tfrac{R+\frac{1}{2}}{R+1}t\right) + \tfrac{1}{4}\exp\left(-\tfrac{2}{R+1}t\right) \\ \text{Prob (transversion}|t) &= \tfrac{1}{2} - \tfrac{1}{2}\exp\left(-\tfrac{2}{R+1}t\right). \end{aligned}$$

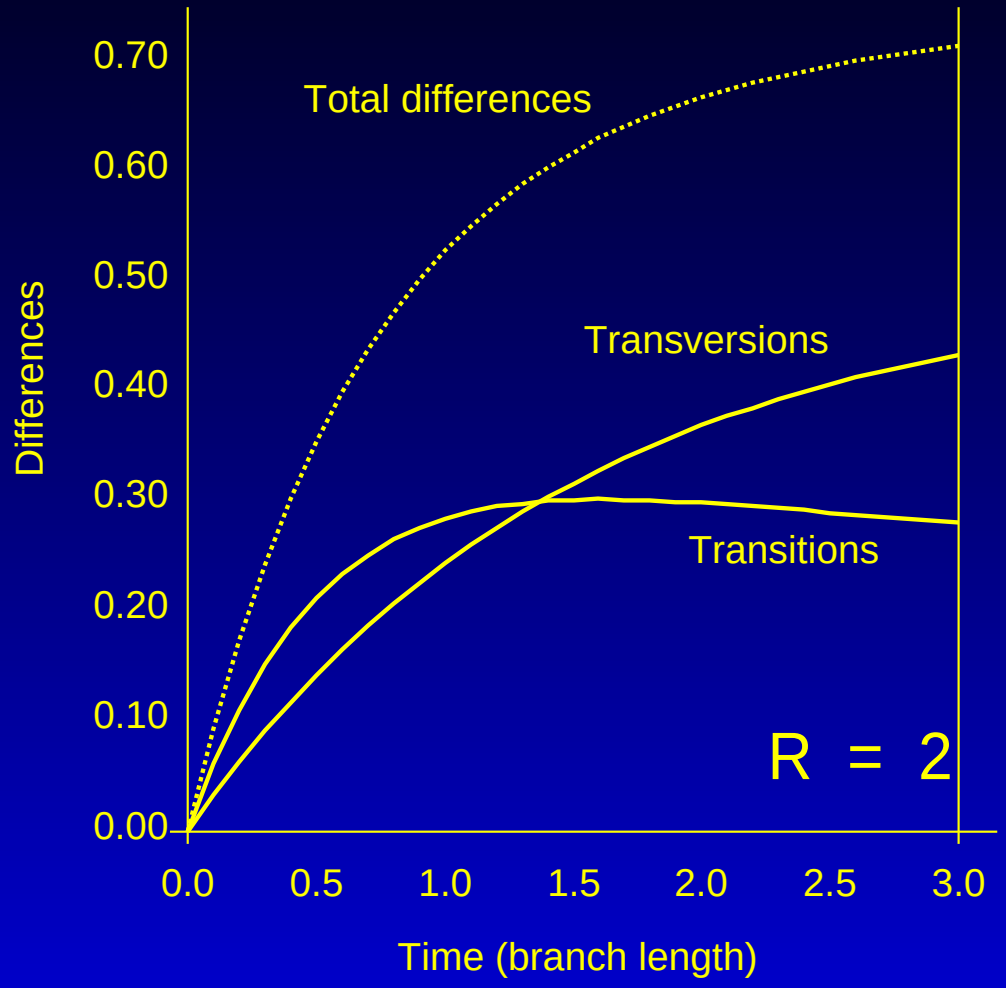
(the transversion probability is the sum of the probabilities of both kinds of transversions).

Transitions, transversions expected



in different amounts of branch length under the K2P model, for $T_s/T_n = 10$

Transitions, transversions expected



in different amounts of branch length under the K2P model, for $T_s/T_n = 2$

Other commonly used models include:

Two models that specify the equilibrium base frequencies (you provide the frequencies $\pi_A, \pi_C, \pi_G, \pi_T$ and they are set up to have an equilibrium which achieves them), and also let you control the transition/transversion ratio:

The **Hasegawa-Kishino-Yano (1985) model**:

to : from :	A	G	C	T
A	$-$	$\alpha\pi_G + \beta\pi_G$	$\alpha\pi_C$	$\alpha\pi_T$
G	$\alpha\pi_A + \beta\pi_A$	$-$	$\alpha\pi_C$	$\alpha\pi_T$
C	$\alpha\pi_A$	$\alpha\pi_G$	$-$	$\alpha\pi_T + \beta\pi_T$
T	$\alpha\pi_A$	$\alpha\pi_G$	$\alpha\pi_C + \beta\pi_C$	$-$

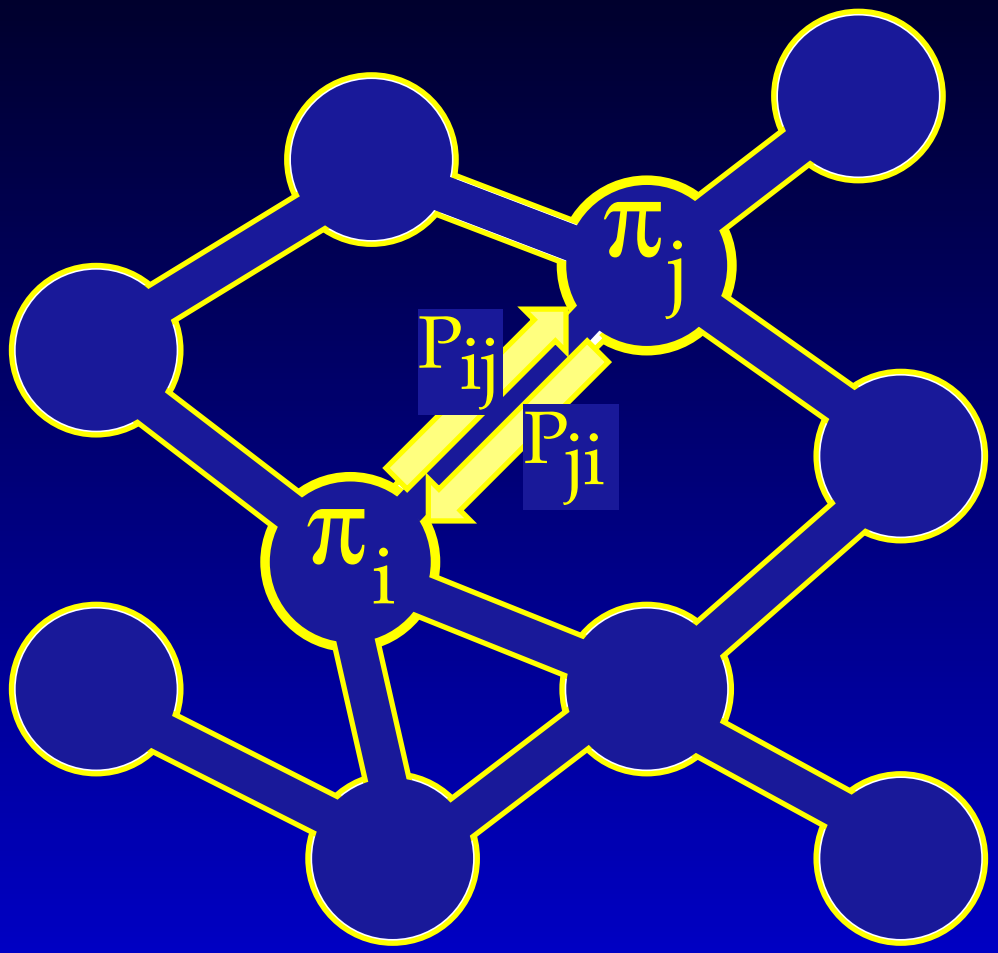
My F84 model

to : from :	A	G	C	T
A	$-$	$\alpha\pi_G + \beta\frac{\pi_G}{\pi_R}$	$\alpha\pi_C$	$\alpha\pi_T$
G	$\alpha\pi_A + \beta\frac{\pi_A}{\pi_R}$	$-$	$\alpha\pi_C$	$\alpha\pi_T$
C	$\alpha\pi_A$	$\alpha\pi_G$	$-$	$\alpha\pi_T + \frac{\beta\pi_T}{\pi_Y}$
T	$\alpha\pi_A$	$\alpha\pi_G$	$\alpha\pi_C + \beta\frac{\pi_C}{\pi_Y}$	$-$

where $\pi_R = \pi_A + \pi_G$ and $\pi_Y = \pi_C + \pi_T$ (The equilibrium frequencies of purines and pyrimidines)

Both of these models have formulas for the transition probabilities, and both are subcases of a slightly more general class of models, the **Tamura-Nei model** (1993).

Reversibility



The General Time-Reversible model (GTR)

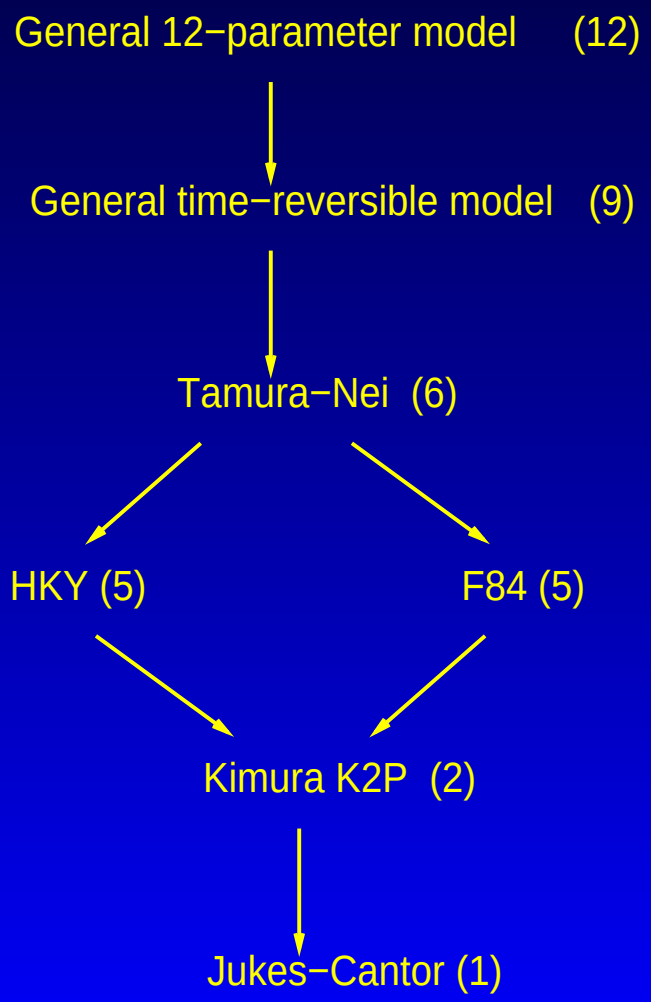
It maintains “detailed balance" so that the probability of starting at (say) A and ending at (say) T in evolution is the same as the probability of starting at T and ending at A:

to : from :	A	G	C	T
A	$-$	$\alpha\pi_G$	$\beta\pi_C$	$\gamma\pi_T$
G	$\alpha\pi_A$	$-$	$\delta\pi_C$	$\epsilon\pi_T$
C	$\beta\pi_A$	$\delta\pi_G$	$-$	$\upsilon\pi_T$
T	$\gamma\pi_A$	$\epsilon\pi_G$	$\upsilon\pi_C$	$-$

And there is of course the **general 12-parameter model** which has arbitrary rates for each of the 12 possible changes (from each of the 4 nucleotides to each of the 3 others). (Neither of these has formulas for the transition probabilities, but those can be done numerically.)

Relation between models

There are many other models, but these are the most widely-used ones. Here is a general scheme of which models are subcases of which other ones:



Rate variation among sites

- In reality, rates of evolution are not constant among sites.
- Fortunately, in the transition probability formulas, rates come in as simple multiples of times
- Thus if we know the rates at two sites, we can compute the probabilities of change by simply, for each site, multiplying all branch lengths by the appropriate rate
- If we don't know the rates, we can imagine averaging them over a distribution of rates. Usually the Gamma distribution is used
- In practice a discrete histogram of rates approximates the integration
- (For the Gamma it seems best to use Generalized Laguerre Quadrature to pick the rates and frequencies in the histogram).
- Also, there are actually autocorrelations with neighboring sites having similar rates of change.
- This can be handled by Hidden Markov Models, which we cover later.

A pioneer of protein evolution

Margaret Dayhoff, about 1966

Models of amino acid change in proteins

There are a variety of models put forward since the mid-1960's:

1. Amino acid transition matrices
   - Dayhoff (1968) model. Tabulation of empirical changes in closely related pairs of proteins, normalized. The PAM100 matrix, for example, is the expected transition matrix given 1 substitution per position.
   - Jones, Taylor and Thornton (1992) recalculated PAM matrices (the JTT matrix) from a much larger set of data.
   - Jones, Taylor, and Thurnton (1994a, 1994b) have tabulated a separate mutation data matrix for transmembrane proteins.
   - Koshi and Goldstein (1995) have described the tabulation of further context-dependent mutation data matrices.
   - Henikoff and Henikoff (1992) have tabulated the BLOSUM matrix for conserved motifs in gene families.
2. Goldman and Yang (1994) pioneered codon-based models (see next screen).

Approaches to protein sequence models

Making a model for protein sequence evolution (a not−very−practical approach)

1. Use a good model of DNA evolution.
2. Use the appropriate genetic code
3. When an amino acid changes, accept it with a probability that declines as the amino acids become more different
4. Fit this to empirical information on protein evolution
5. Take into account variation of rate from site to site
6. Take into account correlation of rate variation in adjacent sites
7. How about protein structure? Secondary structure? 3−D struncture?

References

Barry, D., and J. A. Hartigan. 1987. Statistical analysis of hominoid molecular evolution. *Statistical Science* **2:** 191-210. **[Early use of full 12-parameter model]**

Dayhoff, M. O. and R. V. Eck. 1968. *Atlas of Protein Sequence and Structure 1967-1968.* National Biomedical Research Foundation, Silver Spring, Maryland. **[Dayhoff's PAM modelfor proteins]**

Goldman, N., and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* **11:** 725-736 . **[codon-based protein/DNA models]**

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* **22:** 160-174. **[HKY model]**

Henikoff, S. and J. G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences, USA* **89:** 10915-10919. **[BLOSUM protein model]**

References

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. *Computer Applcations in the Biosciences (CABIOS)* **8:** 275-282. **[JTT model for proteins]**

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1994a. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry* **33:** 3038-3049. **JTT membrane protein model]**

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1994b. A mutation data matrix for transmembrane proteins. *FEBS Letters* **339:** 269-275 . **[JTT membrane protein model]**

Jukes, T. H. and C. Cantor. 1969. Evolution of protein molecules. pp. 21-132 in *Mammalian Protein Metabolism*, ed. M. N. Munro. Academic Press, New York. **[Jukes-Cantor model]**

Kimura, M. 1980. A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16:** 111-120. **[Kimura's 2-parameter model]**

References

Koshi, J. M. and R. A. Goldstein. 1995. Context-dependent optimal substitution matrices. *Protein Engineering* **8:** 641-645. **[generating other kinds of protein model matrices]**

Lanave, C., G. Preparata, C. Saccone, and G. Serio. 1984. A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution* **20:** 86-93. **[General reversible model]**

Lockhart, P. J., M. A. Steel, M. D. Hendy, and D. Penny. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. *Molecular Biology and Evolution* **11:** 605-612. **[The LogDet distance for correcting for changing base composition]**

Tamura, K. and M. Nei. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution* **10:** 512-526. **[Tamura-Nei model]**

How it was done

This projection produced

- using the `prosper` style in LaTeX,
- using Latex to make a `.dvi` file,
- using `dvips` to turn this into a Postscript file,
- using `ps2pdf` to mill it into a PDF file, and
- displaying the slides in Adobe Acrobat Reader.

Result: nice slides using freeware.