Models of Sequence Evolution

A **substitution model** describes the process from which a sequence of characters of a fixed size from some alphabet changes into another set of traits.

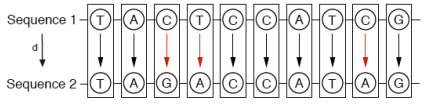
For example, in cladistics, each position in the sequence might correspond to a property of a species
which can either be present or absent. The alphabet could then consist of "0" for absence and "1" for presence.

In phylogenetics, sequences are often obtained by firstly obtaining a nucleotide or protein sequence alignment,

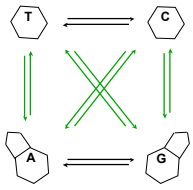
and then taking the bases or amino acids at corresponding positions in the alignment as the characters.

Substitution models are used for a number of things:
•Constructing evolutionary trees in phylogenetics or cladistics.
•Simulating sequences to test other methods and algorithms.

Modeling sequence evolution



substitution scheme



Common assumptions about the evolutionary process

- evolutionary rate is the same for each site
- each site in a sequence evolves independently of the others

Markovian model:

- Continous-time stationary Markov process defined by a rate matrix:

$$\mathbf{Q} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{pmatrix} A & C & G & T \\ q_{AA} & q_{AC} & q_{AG} & q_{AT} \\ q_{CA} & q_{CC} & q_{CG} & q_{CT} \\ q_{GA} & q_{GC} & q_{GG} & q_{GT} \\ q_{TA} & q_{TC} & q_{TG} & q_{TT} \end{pmatrix}$$

- Off diagonal entries

$$q_{ij} > 0$$

specify instantaneous rates

- Diagonal elements $q_{ii} = -\sum_{j \neq i} q_{ij}$,
- equilibrium frequency $\pi_i, i \in \{A, C, G, T\}$

Substitution rate:

$$-\sum_i \pi_i q_{ii} = 1,$$

Reversibility:

$$\pi_i q_{ij} = \pi_i q_{ii} \quad \text{for all } i \neq j \in \{A, C, G, T\}.$$
$$\Rightarrow$$
$$q_{ij} = \pi_j b_{ij} \quad \text{with} \quad b_{ij} = b_{ji}$$

The model has nine free parameters.

Spectral decomposition:

$$\mathbf{Q} = \mathbf{U}\ \mathbf{diag}\,[\lambda_0, \lambda_1, \lambda_2, \lambda_3]\ \mathbf{U}^{-1},$$

where

- $\mathbf{U}$ is the matrix of (right) eigenvectors
- $\mathbf{diag}\,[\lambda_0, \lambda_1, \lambda_2, \lambda_3]$ a diagonal matrix containing the eigenvalues of $\mathbf{Q}$
- $$\lambda_0 = 0 > \lambda_1 \geq \lambda_2 \geq \lambda_3.$$

Transition probabilities:
For a branch of length t (i.e. average number of substitutions) given by $\mathbf{P}(t) = \exp(\mathbf{Q}t)$.

$$\mathbf{P}(t) = \mathbf{U}\ \mathbf{diag}\,[\eta_0, \eta_1, \eta_2, \eta_3]\ \mathbf{U}^{-1},$$

with $\eta_s = e^{\lambda_s t}, s = 0, 1, 2, 3$

and $\eta_0 = 1 \geq \eta_1 \geq \eta_2 \geq \eta_3 > 0$

Jukes Cantor model

$$\mathbf{R}_{JC} = \begin{pmatrix} & A & C & G & T \\ A & -1 & 1/3 & 1/3 & 1/3 \\ C & 1/3 & -1 & 1/3 & 1/3 \\ G & 1/3 & 1/3 & -1 & 1/3 \\ T & 1/3 & 1/3 & 1/3 & -1 \end{pmatrix} \qquad k_{JC} = -\frac{1}{4}\sum r_{ii} = 1$$

Jukes Cantor model

$$\mathbf{P}_{JC}(d) = \begin{pmatrix} & A & C & G & T \\ A & p_0(d) & p_1(d) & p_1(d) & p_1(d) \\ C & p_1(d) & p_0(d) & p_1(d) & p_1(d) \\ G & p_1(d) & p_1(d) & p_0(d) & p_1(d) \\ T & p_1(d) & p_1(d) & p_1(d) & p_0(d) \end{pmatrix}$$

$$p_1(d) = \frac{1}{4}\left(1 - Exp\left[-4d/3\right]\right)$$
$$p_0(d) = \frac{1}{4}\left(1 + 3Exp\left[-4d/3\right]\right)$$

Jukes-Cantor (JC, nst=1): Equal base frequencies, all substitutions equally likely (PAUP* rate classification: aaaaaa, PAML: aaaaaa)* (Jukes and Cantor 1969)

Felsenstein 1981 (F81, nst=1): Variable base frequencies, all substitutions equally likely (PAUP*: aaaaaa, PAML: aaaaaa)** (Felsenstein 1981)

Kimura 2-parameter (K80, nst=2): Equal base frequencies, variable transition and transversion frequencies (PAUP*: abaaba, PAML: abbbba) (Kimura 1980)

Hasegawa-Kishino-Yano (HKY, nst=2): Variable base frequencies, variable transition and transversion frequencies (PAUP*: abaaba, PAML: abbbba) (Hasegawa et. al. 1985)

Tamura-Nei (TrN): Variable base frequencies, equal transversion frequencies, variable transition frequencies (PAUP*: abaaea, PAML: abbbbf) (Tamura Nei 1993)

Kimura 3-parameter (K3P): Variable base frequencies, equal transition frequencies, variable transversion frequencies (PAUP*: abccba, PAML: abccba) (Kimura 1981)

Transition Model (TIM): Variable base frequencies, variable transitions, transversions equal (PAUP*: abccea, PAML: abccbe)

Transversion Model (TVM): Variable base frequencies, variable transversions, transitions equal (PAUP*: abcdbe, PAML: abcdea)

Symmetrical Model (SYM): Equal base frequencies, symmetrical substitution matrix (A to T = T to A) (PAUP*: abcdef, PAML: abcdef) (Zharkikh 1994)

General Time Reversible (GTR, nst=6): Variable base frequencies, symmetrical substitution matrix (PAUP*: abcdef, PAML: abcdef) (e.g., Lanave et al. 1984, Tavare 1986, Rodriguez et.

An informative website that you may find useful is the FindModel website hosted by Los Alamos National Laboratory (FindModel Matrices)

http://hcv.lanl.gov/content/hcv-db/findmodel/matrix/all.html.

This site displays color-coded rate matrices for a select number of substitution models.

Codon Models

To take advantage of the genetic code, Goldman and Yang proposed a model in which the accounted for mutations within a codon. A substitution at any site would depend on the rest of the codon and whether the substitution would change the produced amino acid.

$$Q_{ij} = \begin{cases} 0 & \text{if 2 or 3 of the pair } (i_1,j_1),(i_2,j_2),(i_3,j_3) \text{ are different} \\ \mu\pi_j \cdot \exp(-d_{aa_i,aa_j}/V) & \text{if exactly 1 of the pairs } (i_1,j_1),(i_2,j_2),(i_3,j_3) \text{ is different, and that difference is a transversion} \\ \mu\kappa\pi_j \cdot \exp(-d_{aa_i,aa_j}/V) & \text{if exactly 1 of the pairs } (i_1,j_1),(i_2,j_2),(i_3,j_3) \text{ is different, and that difference is a transition} \end{cases} \quad (3)$$

where i and j are sense codons consisting of three nucleotides such that i = i1,i2,i3, j = j1,j2,j3 and i ¹ j. Furthermore, k is a constant which accounts for transition/transversion bias, d_{aa_i,aa_j} is the distance between amino acids coded by i and j, and V is a measure of sequence variability

Amino Acid Substitutions

Empirical substitution models

In contrast to DNA substitution models, amino acid replacement models have concentrated on the empirical approach.

Dayhoff and coworkers developed a model of protein evolution which resulted in the development of a set of widely used replacement matrices (Dayhoff et al. 1978). In the Dayhoff approach, replacement rates are derived from alignments of protein sequences that are at least 85% identical; this constraint ensures that the likelihood of a particular mutation being the result of a set of successive mutations is low.

One of the main uses of the Dayhoff matrices has been in database search methods where, for example, the matrices P(0.5), P(1) and P(2.5) (known as the PAM50, PAM100 and PAM250 matrices) are used to assess the significance of proposed matches between target and database sequences.

However, the implicit rate matrix has been used for phylogenetic applications

Amino Acid Substitutions

PAM Matrix

In the definition of mutation the matrix M implies certain amount of mutation (measured in PAM units). A 1-PAM mutation matrix describes an amount of evolution which will change, on the average, 1% of the amino acids. In mathematical terms this is expressed as a matrix M such that

$$\sum f_{aa}\left(1 - M_{aa}\right) = 0.01$$

The diagonal elements of M are the probabilities that a given amino acid does not change, so (1-Mii) is the probability of mutating away from i.

Amino Acid Substitutions

Dayhoff matix

Dayhoff et al. (1978) presented a method for estimating the matrix M from the observation of 1572 accepted mutations between 34 superfamilies of closely related sequences. Their method was pioneering in the field. A Dayhoff matrix is computed from a 250-PAM mutation matrix, used for the standard dynamic programming method of sequence alignment. The Dayhoff matrix entries are related to M_{250} by

$$D_{ij} = 10\log\frac{M_{ij}^{250}}{f_i}$$

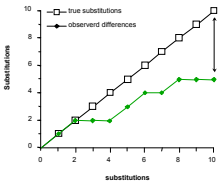
Doublet Model of Sequence Evolution

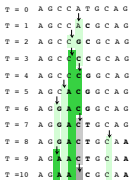
Schöniger and von Haeseler, 1994

	AA	AC	AG	AU	CA	CC	CG	CU	GA	GC	GG	GU	UA	UC	UG	UU
AA	*	π_{AC}	π_{AG}	π_{AU}	π_{CA}	-	-	-	π_{GA}	-	-	-	π_{UA}	-	-	-
AC	π_{AA}	*	π_{AG}	π_{AU}	-	π_{CC}	-	-	-	π_{GC}	-	-	-	π_{UC}	-	-
AG	π_{AA}	π_{AC}	*	π_{AU}	-	-	π_{CG}	-	-	-	π_{GG}	-	-	-	π_{UG}	-
AU	π_{AA}	π_{AC}	π_{AG}	*	-	-	-	π_{CU}	-	-	-	π_{GU}	-	-	-	π_{UU}
CA	π_{AA}	-	-	-	*	π_{CC}	π_{CG}	π_{CU}	π_{GA}	-	-	-	π_{UA}	-	-	-
CC	-	π_{AC}	-	-	π_{CA}	*	π_{CG}	π_{CU}	-	π_{GC}	-	-	-	π_{UC}	-	-
CG	-	-	π_{AG}	-	π_{CA}	π_{CC}	*	π_{CU}	-	-	π_{GG}	-	-	-	π_{UG}	-
CU	-	-	-	π_{AU}	π_{CA}	π_{CC}	π_{CG}	*	-	-	-	π_{GU}	-	-	-	π_{UU}
GA	π_{AA}	-	-	-	π_{CA}	-	-	-	*	π_{GC}	π_{GG}	π_{GU}	π_{UA}	-	-	-
GC	-	π_{AC}	-	-	-	π_{CC}	-	-	π_{GA}	*	π_{GG}	π_{GU}	-	π_{UC}	-	-
GG	-	-	π_{AG}	-	-	-	π_{CG}	-	π_{GA}	π_{GC}	*	π_{GU}	-	-	π_{UG}	-
GU	-	-	-	π_{AU}	-	-	-	π_{CU}	π_{GA}	π_{GC}	π_{GG}	*	-	-	-	π_{UU}
UA	π_{AA}	-	-	-	π_{CA}	-	-	-	π_{GA}	-	-	-	*	π_{UC}	π_{UG}	π_{UU}
UC	-	π_{AC}	-	-	-	π_{CC}	-	-	-	π_{GC}	-	-	π_{UA}	*	π_{UG}	π_{UU}
UG	-	-	π_{AG}	-	-	-	π_{CG}	-	-	-	π_{GG}	-	π_{UA}	π_{UC}	*	π_{UU}
UU	-	-	-	π_{AU}	-	-	-	π_{CU}	-	-	-	π_{GU}	π_{UA}	π_{UC}	π_{UG}	*

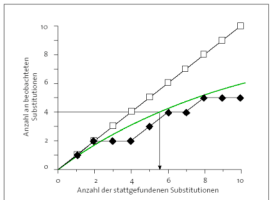
WHY?

multiple hits





Correcting for multiple hits



Jukes Cantor model

$$\mathbf{P}_{JC}(d) = \begin{pmatrix} & A & C & G & T \\ A & p_0(d) & p_1(d) & p_1(d) & p_1(d) \\ C & p_1(d) & p_0(d) & p_1(d) & p_1(d) \\ G & p_1(d) & p_1(d) & p_0(d) & p_1(d) \\ T & p_1(d) & p_1(d) & p_1(d) & p_0(d) \end{pmatrix}$$

$$p_1(d) = \frac{1}{4}\left(1 - Exp\left[-4d/3\right]\right)$$

$$p_0(d) = \frac{1}{4}\left(1 + 3Exp\left[-4d/3\right]\right)$$

Jukes Cantor model

$$obs(d) = \frac{3}{4} - \frac{3}{4}Exp[-4d/3]$$

obs(d) can be estimated from the number of observed different pairs of positions n_1 between two aligned sequences of length *l*.
Solving

$$\frac{n_1}{l} = \frac{3}{4} - \frac{3}{4}Exp[-4d/3]$$

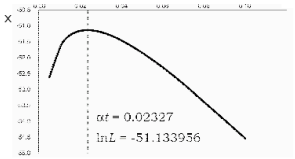
leads to **Jukes Cantor correction**:

$$d = -\frac{3}{4}Log\left[1 - \frac{4}{3}\frac{n_1}{l}\right]$$

Maximum Likelihood II

The Likelihood of sequence s evolving to s' in time t:

$$L(t|s \to s') = \prod_{i=1}^{m}\left(\Pr(s_i) \cdot P_{s_i s'_i}(t)\right)$$



Likelihood surface for two sequences under JC69:

GAAGTCCTGAGAAATAAACTGCACACACTGG
GGAGTCCTGAGAAATAAACTGCACACACTGG

Example by P.O. Lewis

Rate heterogeneity among sites

Modelled by a nonnegative random variable R

- R has density $f(r)$.
- mean
$$\mathbb{E}(R) = 1$$
- moment-generating function
$$\phi(x) = \mathbb{E}(\exp(Rx))$$
- The relative rate r at a position is then given by a realisation of R.

Transition probabilities for a position

$$\mathbf{P}(t|R = r) = \exp(\mathbf{Q}rt).$$

Then, the unconditional transition probability matrix is

$$\begin{aligned}\mathbf{P}(t) &= \int_0^\infty \mathbf{P}(t|R = r)f(r)dr \\ &= \mathbf{U}\ \mathbf{diag}\,[\eta_0, \eta_1, \eta_2, \eta_3]\ \mathbf{U}^{-1},\end{aligned}$$

with eigenvalues

$$\eta_s = \phi(\lambda_s t) = \int_0^\infty e^{\lambda_s r t} f(r)dr.$$

Specifying the distribution

rate heterogeneity is modelled by a Γ-distribution

- density
$$f(r) = \frac{\alpha^\alpha}{\Gamma(\alpha)} r^{\alpha-1} e^{-\alpha r},$$
- and eigenvalues
$$\phi(\lambda_s t) = \left(\frac{\alpha - \lambda_s t}{\alpha}\right)^{-\alpha}.$$