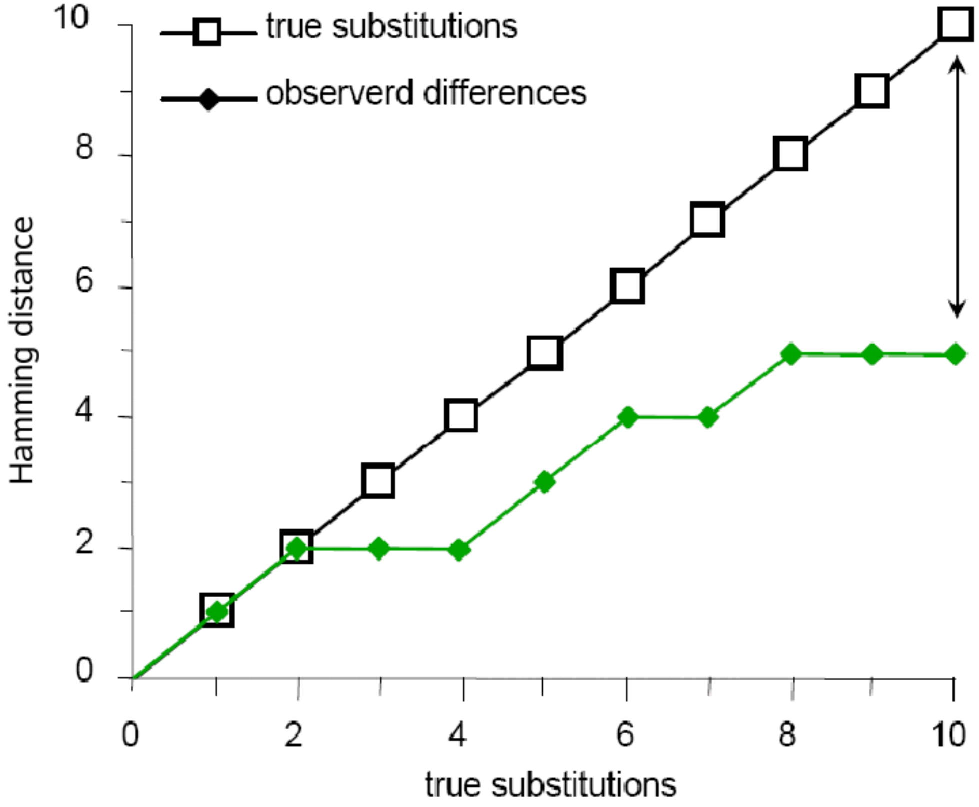
Evolutionary distances



$t = 0$	C	C	A	T	G	C	G
				↓			
$t = 1$	C	C	A	C	G	C	G
			↓				
$t = 2$	C	C	G	C	G	C	G
			↓				
$t = 3$	C	C	C	C	G	C	G
				↓			
$t = 4$	C	C	C	G	G	C	G
		↓					
$t = 5$	C	A	C	G	G	C	G
	↓						
$t = 6$	G	A	C	G	G	C	G
				↓			
$t = 7$	G	A	C	T	G	C	G
							↓
$t = 8$	G	A	C	T	G	C	A
	↓						
$t = 9$	A	A	C	T	G	C	A
			↓				
$t = 10$	A	A	T	C	G	C	A

Evolutionary distances, cont'd

- We are given a multiple alignment and want to obtain pairwise evolutionary distances

- With u as the number of mismatches in an alignment of length n, the Hamming distance per 100 sites is

$$D(u,n) = 100\ \frac{u}{n}$$

- The distance D does not take multiple substitutions into account. As a consequence, pairwise distances are not additivie.

- For any number of mismatches u and alignment lengths n, we have

$$0 <= D <= 100$$

. For example

$$D(u = 0, n = 100) = 0 \quad \text{and} \quad D(u = 75, n = 100) = 75$$

Evolutionary distances, cont'd

- Pairwise evolutionary distances $d(u, n)$ are meant to scale in units of substitutions (per 100 sites) that *most likely* have occured on the evolutionary paths.

- If we assume (as in the Jukes-Cantor model, see below)

  i) that sequence positions are i.i.d. (*independently identically distributed*)

  ii) that nucleotides are uniformly distributed and independently substituted such that the probabilities for nucleotide substitutions are all the same and do not depend on the particular nucleotides

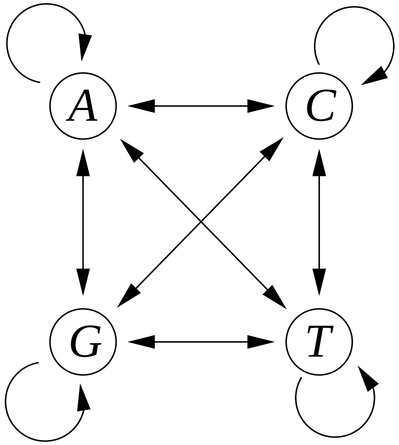
  we require that

  $$d(u = 0, n = 100) = 0 \quad \text{and} \quad d(u = 75, n = 100) = \infty$$

  (the latter follows from the requirement that the evolutionary distance for two random sequences is $d = \infty$)

Markov chains ("time-discrete Markov processes")

1. The states are A, C, G, T
2. A starting distribution of states $\rho^0 = (\rho_A, \rho_C, \rho_G, \rho_T)$
3. Transition probabilites in one "time step" between states $P_{ij} = \mathrm{Pr}(j|i)$



State probabilities depend only on the previous state and not on the past of the chain (Markov property). If transition probabilities don't change in time (homogeneity) the probability of a sequence $x = (x_1, ..., x_L)$ is

$$\mathrm{Pr}(x) = \rho_{x_1} \prod_{i=2}^{L} \mathrm{Pr}(x_i | x_{i-1})$$

Transition probabilities for n steps are obtained from the n-th power of the stochastic one-step transition matrix P, from P^n.

The Markov model of sequence evolution

Sequence evolution is modeled by a (time-continuous) *Markov process* that acts **independently** on the sites of the sequence.

$$\begin{array}{rcccccc}
X_{t_1} = & \mathbf{A} & \text{T} & \text{C} & \text{G} & \text{C} & \cdots \\
 & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \\
X_{t_2} = & \mathbf{G} & \text{T} & \text{C} & \text{A} & \text{G} & \cdots \\
 & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \\
X_{t_3} = & \mathbf{G} & \text{T} & \text{C} & \text{A} & \text{C} & \cdots \\
 & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \\
X_{t_4} = & \mathbf{A} & \text{G} & \text{C} & \text{A} & \text{G} & \cdots
\end{array}$$

A *Markov process* is a sequence of random variables $(X_t)_{t \geq 0}$ given by a triple $\left(\mathcal{A}, \rho^0, Q\right)$, where $\mathcal{A} = \{1, ..., n\}$ is the set of states (nucleotides or amino acid residues) (X_t) takes, ρ^0 is the initial probability distribution of states ($\rho_i^0 = Pr[X_0 = i]$) and the rate matrix Q as a $n \times n$ matrix with substitution rates (something like transition probabilities for infinitesimal small time steps) between states.

The Markov model of sequence evolution, cont'd

- **Markov property** (the process is memoryless):
  $\Pr[X(t_n) = s | X(t_1) = i_1, X(t_2) = i_2, ..., X(t_{n-1}) = i_{n-1}]$
  $= \Pr[X(t_n) = s | X(t_{n-1}) = i_{n-1}]$

- **Homogeneity**:
  Transition probabilities only depend on the time interval:
  $P_{ij}(t) = \Pr[X_{t+s} = j | X_s = i] \ = \ \Pr[X_t = j | X_0 = i]$

- The time t of the Markov process is measured in units of substitutions

- The transition probablity $P_{ij}(t)$ is the probability that state i changes into state j in time t

- We think of the distribution $\rho(t)$ as a row vector. The evolution of the distribution of states at time s in time t is given by

$$\rho(s)P(t) = \rho(s+t)$$

The Markov model of sequence evolution, cont'd

- **Stationary distribution:**
  π is the *stationary distribution* of the process, if π doesn't change in time:

$$\pi_j = \sum_{i \in \mathcal{A}} \pi_i P_{ij}(t) \quad \text{for all } j$$

$$\pi P(t) = \pi$$

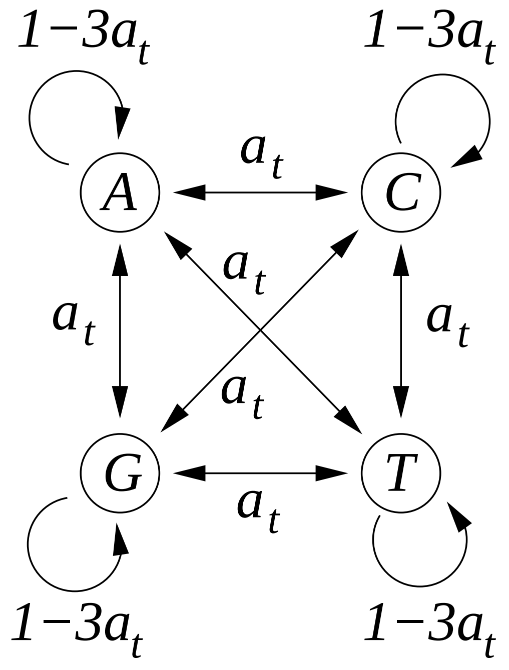
  We say that the process is in equilibrium if the distribution of the process is the stationary distribution π.

  π exists if any state can be reached by any other state.

The transition probability matrix

For nucleotides, the simplest model is the *Jukes–Cantor–model* (1969). The set of states comprises the nucleotides ($\mathcal{A} = \{1, 2, 3, 4\}$). The stationary distribution π of nucleotides is the uniform distribution ($\pi = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$) and the probabilities that any nucleotide is substituted by another any other nucleotide are equal.

Thus, the *transition probability matrix* of the Jukes–Cantor model has the form



$$P(t) = \begin{pmatrix} 1-3a_t & a_t & a_t & a_t \\ a_t & 1-3a_t & a_t & a_t \\ a_t & a_t & 1-3a_t & a_t \\ a_t & a_t & a_t & 1-3a_t \end{pmatrix}$$

The transition probability matrix, cont'd

The transition probability matrix $P(t)$ is a stochastic matrix and has the following properties:

- $P(0) = I, \qquad I$ - identity matrix,
- $P_{ij}(t) \geq 0$ and $\sum_j P_{ij}(t) = 1$,
- $P(s+t) = P(s)P(t)$

The latter equation is called *Chapman–Kolmogorov equation*. E.g. think of $\mathcal{A} = \{1, 2, 3, 4\}$ and the process being in state 1 reaching state t in time $s+t$. The transition probability $P_{14}(s+t)$ is

$$\begin{aligned} Pr[X_{s+t} = 4 | X_0 = 1] &= Pr[X_s = 1 | X_0 = 1] \cdot Pr[X_{s+t} = 4 | X_s = 1] \\ &\quad + Pr[X_s = 2 | X_0 = 1] \cdot Pr[X_{s+t} = 4 | X_s = 2] \\ &\quad + Pr[X_s = 3 | X_0 = 1] \cdot Pr[X_{s+t} = 4 | X_s = 3] \\ &\quad + Pr[X_s = 4 | X_0 = 1] \cdot Pr[X_{s+t} = 4 | X_s = 4] \\ &= \sum_{k \in \mathcal{A}} P_{1k}(s) P_{k4}(t) \end{aligned}$$

Maximum Likelihood and coin tossing

Assume, we have flipped a coin 10 times and got 7 times its head and 3 times its tail. We want to estimate the probability Prob(head), that the head shows up when the coin is flipped?

The likelihood $\mathcal{L}(p)$ is the probability to observe one outcome (of many possible outcomes) of a random experiment (one data set) under the probabibilistic model with its model parameter p.

$\mathcal{L}(p) = \mathrm{Pr}(\mathrm{data}|p) = p^7(1-p)^3$

We think of the likelihood as a function depending on the model parameters. Note that the sum or the integral over the parameter space is not 1!

$\widehat{p} = \mathrm{Prob(head)}$ is determined as the p where $\mathcal{L}$ assumes its maximum.

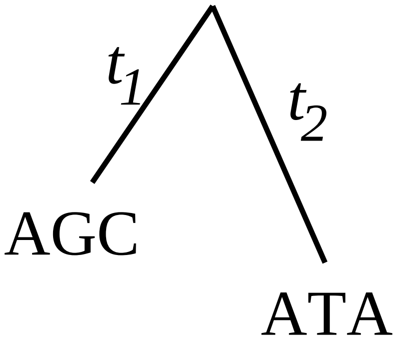
The variance of the estimate depends on the sample size and can be estimated from the likelihood curvature. If the data was generated under the model, the ML estimate of the parameters yields exact or true values for infinite sample sizes.

Evolutionary distances with Maximum Likelihood

We think of the observed alignment $\mathcal{D}$ as the outcome of the Markovian evolution.

Consider the following alignment $\mathcal{D}$:

A G C
A T A



We assume that the process is in equilibrium.

The *likelihood* to observe the alignment $\mathcal{D}$ (the data) with distance $t = (t_1 + t_2)$ given the Markov model $\mathcal{M}$ then is

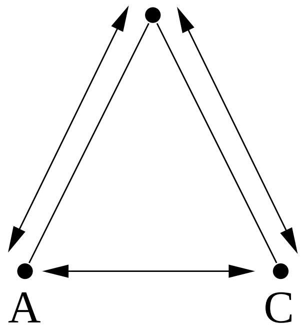
$$\Pr(\mathcal{D}|t, \mathcal{M}) = \sum_{i \in \mathcal{A}} \pi_i P_{iA}(t_1) P_{iA}(t_2) \cdot \sum_{i \in \mathcal{A}} \pi_i P_{iG}(t_1) P_{iT}(t_2) \cdot \sum_{i \in \mathcal{A}} \pi_i P_{iC}(t_1) P_{iA}(t_2)$$

Evolutionary distances with Maximum Likelihood, cont'd

The Markov process is called *reversible*, if the evolution of state i into state j in time t is modelled by the same process as the evolution of state i into state j in time t:

$$\pi_i P_{ij}(t) = \pi_j P_{ji}(t) \qquad \text{for all} \;\; i, j, t$$

(*detailed balance equations*)



We assume the time-reversible Jukes-Cantor model and apply the Chapman-Kolmogorov equations:

$$\begin{aligned}\text{Pr}(\mathcal{D}|t,\mathcal{M}) &= \pi_A P_{AA}(t) \cdot \pi_G P_{GT}(t) \cdot \pi_C P_{CA}(t) \\ &= \pi_A P_{AA}(t) \cdot \pi_T P_{TG}(t) \cdot \pi_A P_{AC}(t)\end{aligned}$$

If the process is reversible and if we are given a pairwise alignment, we are ignorant about the location of the root node.

Evolutionary distances with Maximum Likelihood, cont'd

Consider $\text{Pr}(\mathcal{D}|t,\mathcal{M})$ as likelihood function depending on the distance t as model parameter:

$$\log \mathcal{L}(t) = \log \text{Pr}(\mathcal{D}|t,\mathcal{M})$$

The evolutionary distance is estimated as distance $\hat{t}$ where the likelihood function assumes its maximum.

If sequences have evolved according to the evolutionary model $(\mathcal{M}, t)$, and if we have infinitely many samples (alignment columns) of the outcome of this evolution, the evolutionary distance can be exactly reestimated by Maximum Likelihood (ML), i.e. the ML distance estimator is consistent. For finite sample sizes, ML estimates $\hat{t}$ are normally distributed around the 'true' value for t.

We have to evaluate the likelihood function and thus the transition probabilities for different times or distances t. This is achieved by means of the rate matrix...

The rate matrix

The *rate matrix* Q of a time-continuous Markov process provides an infinitesimal description of the process.

We assume that the probability transition matrix $P(t)$ of a time continuous Markov process is continuous and differentiable at any $t > 0$. I.e. the limit

$$Q := \lim_{t \searrow 0} \frac{P(t) - I}{t}$$

exists. Q is known as the *rate matrix* or the *generator* of the Markov chain. For very small time periods $h > 0$, transition probabilities are approximated by

$$\begin{aligned} P(h) &\approx I + hQ \\ P_{ij}(h) &\approx Q_{ij} \cdot h, \qquad i \neq j. \end{aligned}$$

From the last equation we see, that the entries of Q may be interpreted as substitution rate.

The rate matrix, cont'd

From the Chapman-Kolmogorov equation we get

$$\begin{aligned}
\frac{d}{dt}P(t) &= \lim_{h \searrow 0} \frac{P(t+h) - P(t)}{h} \\
&= \lim_{h \searrow 0} \frac{P(t)P(h) - P(t)I}{h} \\
&= P(t) \lim_{h \searrow 0} \frac{P(h) - P(0)}{h} \\
\frac{d}{dt}P(t) &= P(t)Q
\end{aligned}$$

Under the initial condition $P(0) = I$ the differential equation can be solved and yields (as in the one–dimensional case)

$$P(t) = \exp(tQ) = \sum_{k=0}^{\infty} \frac{Q^k t^k}{k!}.$$

Transition probabilities for any $t > 0$ are computed from the matrix Q.

The rate matrix, cont'd

Recall, that for very small h we have $P(h) \approx I + hQ$.

Q has the following properties:

- $Q_{ij} \geq 0 \quad \text{for} \quad i \neq j$
- $Q_{ij} \geq 0,\ i \neq j \ \Rightarrow \ Q_{ii} \leq 0$
- $\sum_j Q_{ij} = 0,\ Q_{ii} = -\sum_{j \neq i} Q_{ij}$

Further,

- π is stationary distribution if $\quad \pi Q = 0$
- the process is reversible, if $\quad \pi_i Q_{ij} = \pi_j Q_{ji} \quad$ for all $\ i, j$

The rate matrix, cont'd

The rate matrix of the Jukes-Cantor model is

$$Q = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}.$$

where $\alpha \geq 0$.

Due to the simple structure of Q, exp(tQ) can be calculated analytically. The transition probability matrix is

$$P(t) = \begin{pmatrix} 1-3a_t & a_t & a_t & a_t \\ a_t & 1-3a_t & a_t & a_t \\ a_t & a_t & 1-3a_t & a_t \\ a_t & a_t & a_t & 1-3a_t \end{pmatrix},$$

where

$$a_t = \frac{1 - \exp(-4\alpha t)}{4}$$

The rate matrix, cont'd

If we assume the stationary distribution, Q summarizes all model parameters of the Markov process, since $\pi Q = 0$. Clearly, Q can be multiplied with a factor and the distribution π doesn't change. In other words: The model parameters hold substitution rates. And rates hold the information how many substitutions per time unit one expects.

The rate matrix can be calibrated to *PAM (percent accepted mutations)*–units. 1 PAM is the time (or evolutionary distance) where one substitution event per 100 sites is expected to have occured.

Given Q, one expects $E = \sum_i \pi_i \sum_{j \neq i} Q_{ij} = -\sum_i \pi_i Q_{ii}$ substitution events per time unit.

The Jukes–Cantor rate matrix Q is calibrated to PAM-units by setting $E = \frac{1}{100} \Leftrightarrow -4 \cdot \frac{1}{4} \cdot -3\alpha = \frac{1}{100} \Leftrightarrow \alpha = \frac{1}{300}$.

Evolutionary distances with Maximum Likelihood

Again, consider the log likelihood of the alignment $\mathcal{D}$: $\begin{array}{ccc} \mathsf{A} & \mathsf{G} & \mathsf{C} \\ \mathsf{A} & \mathsf{T} & \mathsf{A} \end{array}$

We had

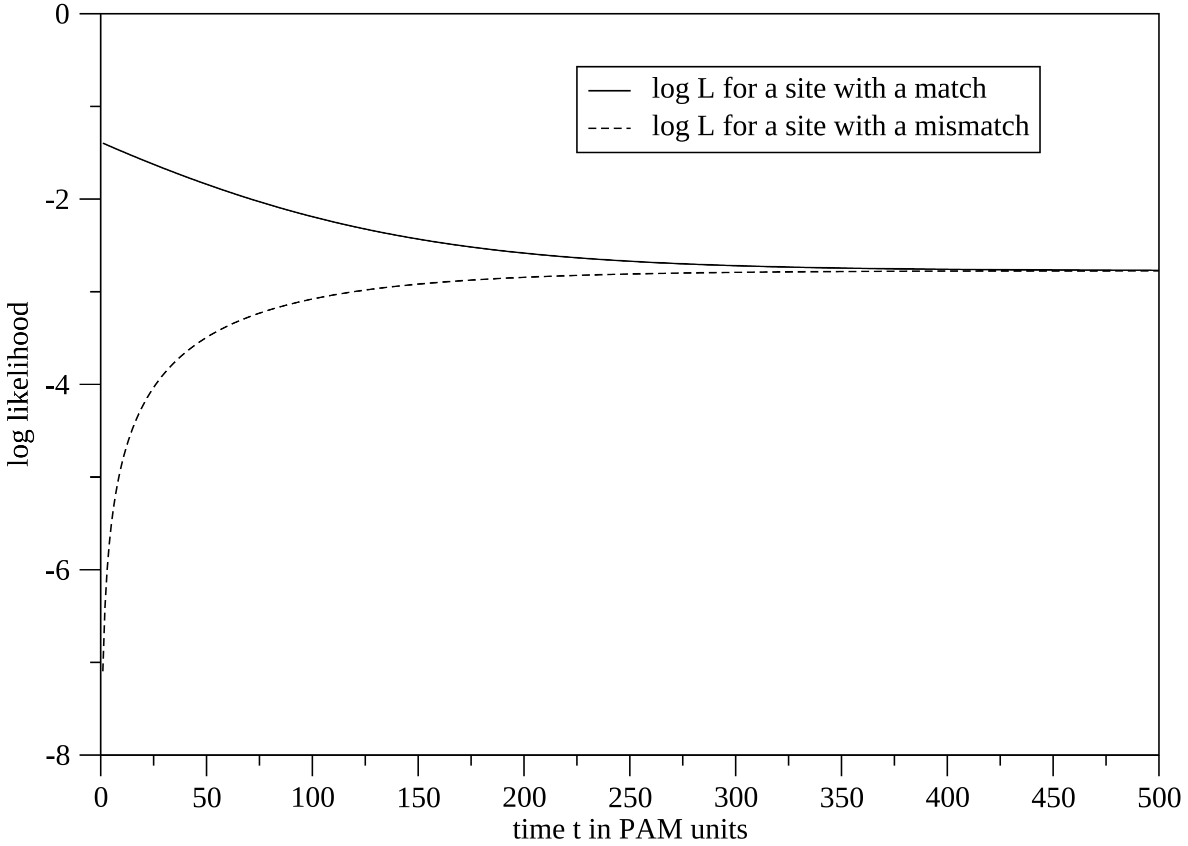
$$\log \mathcal{L}(t) = \log(\pi_A P_{AA}(t)) + \log(\pi_G P_{GL}(t)) + \log(\pi_C P_{CA}(t))$$

with the Jukes-Cantor model:

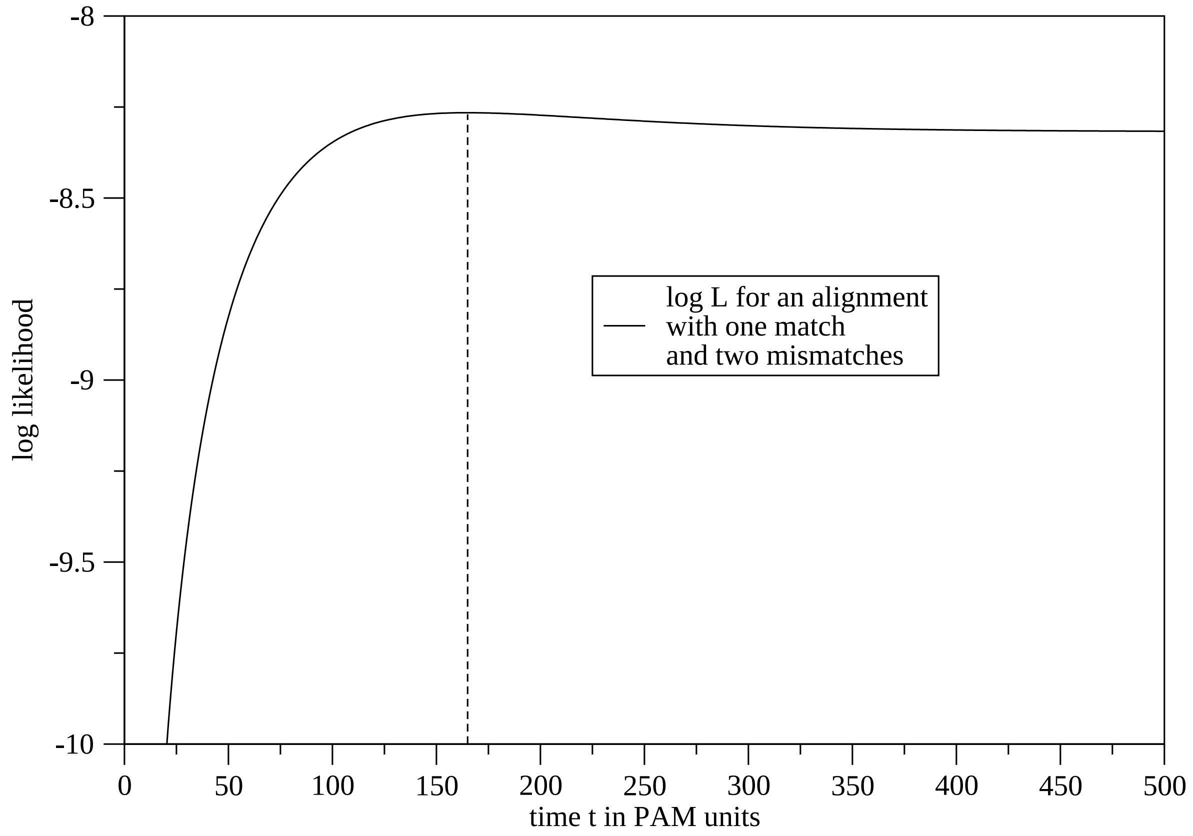
$$\log \mathcal{L}(t) = \log \left(\frac{1}{4} \cdot \Big(1 - \frac{3}{4} \cdot (1 - \exp(\frac{-4}{300}\ t)) \Big) \right) + 2 \cdot \log \left(\frac{1}{4} \cdot \frac{1 - \exp(\frac{-4}{300}\ t)}{4} \right)$$

Evolutionary distances with Maximum Likelihood, cont'd

The log likelihood functions for the single alignment columns and JC69:



Evolutionary distances with Maximum Likelihood, cont'd



The Maximum Likelihood estimate $\widehat{t} = 165$ PAM is the value for t where where $\log \mathcal{L}(t)$ is maximal. The variance of the estimate is huge because i) the small sample size, ii) the large distance. Variances can be computed from the second derivative of $\log \mathcal{L}(\widehat{t})$.

The Jukes-Cantor correction

The Hamming distance $D = \frac{100 \cdot u}{n}$ (u-mismatches, n- sequence length) for the distance between two DNA sequences ignores the putative occurence of multiple substitutions. The Jukes-Cantor correction d provides a formula for the evolutionary distance d of two DNA sequences, i.e. $d(u,n)$ holds the number of substitutions which are expected to have occured per 100 sites.

The probability p to observe that a nucleotide is not substituted after time t is

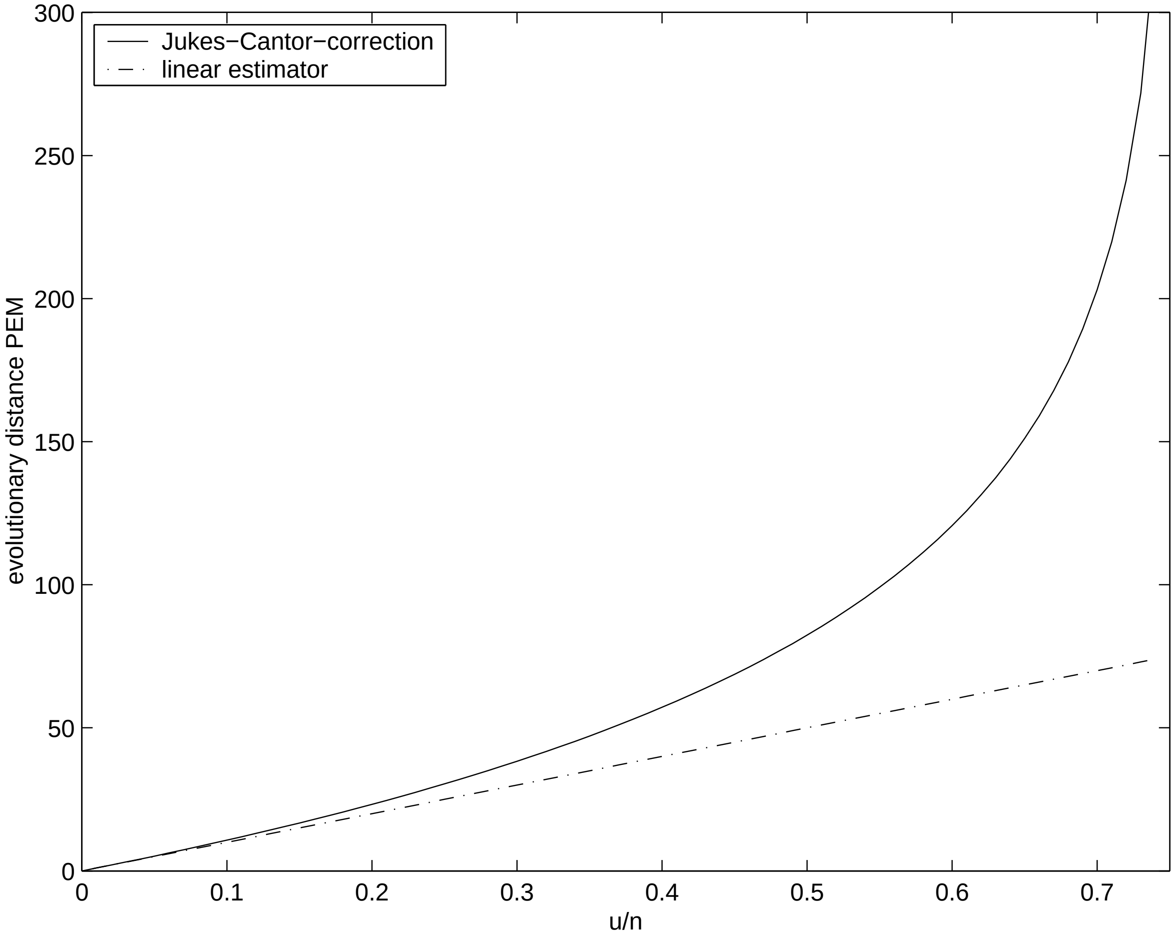
$p = \sum_i \pi_i P_{ii}(t) = 4 \cdot \frac{1}{4}(1 - 3a_t) = 1 - \frac{3}{4}(1 - \exp(-4\alpha t)) = \frac{1+3\exp(-4\alpha t)}{4}$

There are u mismatches among n sites. That is, we observe $p = 1 - \frac{u}{n}$. Calibration to PAM–units and setting $t = d$ yields

$$1 - \frac{u}{n} = \frac{1 + 3\exp(-4d/300)}{4}$$

Jukes–Cantor correction, cont'd

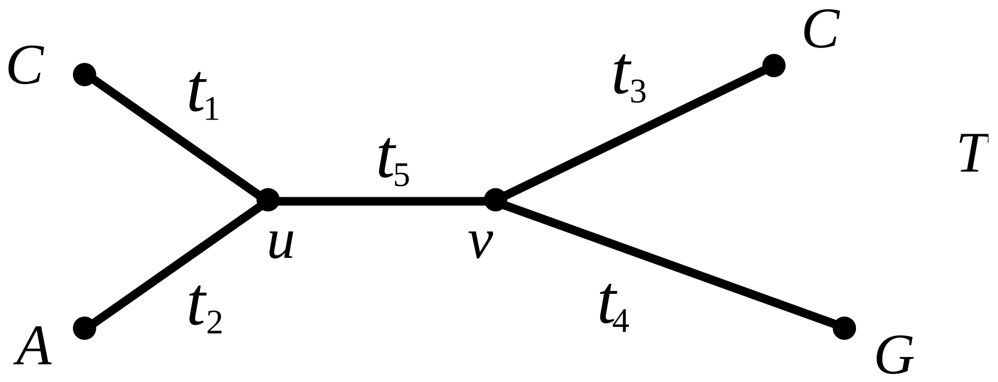
$$d = -\frac{300}{4} \ln\left(1 - \frac{4u}{3n}\right) \text{ PAM}$$



If $\frac{u}{n} \geq 0.75$, d is not defined.

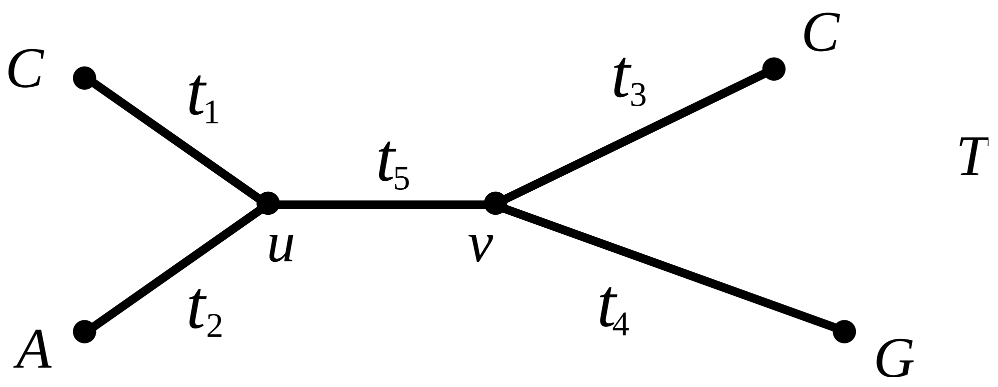
Maximum Likelihood Trees

Consider one site $\mathcal{D}_s$ of an alignment with the states A,C,C,T $\in \mathcal{A}$. We consider a particular tree topology $\mathcal{T}$ with edge lengths $\vec{t} = (t_1, ..., t_5)$ and label the leaves with the states of the alignment.



We want to compute the likelihood to observe the states under this tree $(T, \vec{t})$ and the Markov model Q. Reversibility implies that the likelihood does not depend on the position of a root node.

Maximum Likelihood Trees, cont'd



We choose node u as root node. First assume that we know states at internal nodes u and v and that both of them are C. Then

$$\mathcal{L}(T,\vec{t},Q \mid \mathcal{D}_s,[C,C]) = \pi_C P_{CC}(t_1)P_{CA}(t_2)P_{CC}(t_5)P_{CC}(t_3)P_{CG}(t_4)$$

Because we do not know states at internal nodes

$$\mathcal{L}(T,\vec{t},Q \mid \mathcal{D}_s) = \sum_{i\in\mathcal{A}} \pi_i P_{iC}(t_1)P_{iA}(t_2) \sum_{j\in\mathcal{A}} P_{ij}(t_5)P_{jC}(t_3)P_{jG}(t_4)$$

Note that we have 4^n summands for n internal nodes.

Maximum Likelihood Trees, cont'd

Recursive definition of the likelihood

We want to apply a dynamic programming strategy to compute the likelihood. The algorithm requires a rooted tree which is traversed from the leaves to the root (as the Sankoff algorithm does).

Felsenstein (1981) defines the conditional likelihood

$$\mathcal{L}_k(w)$$

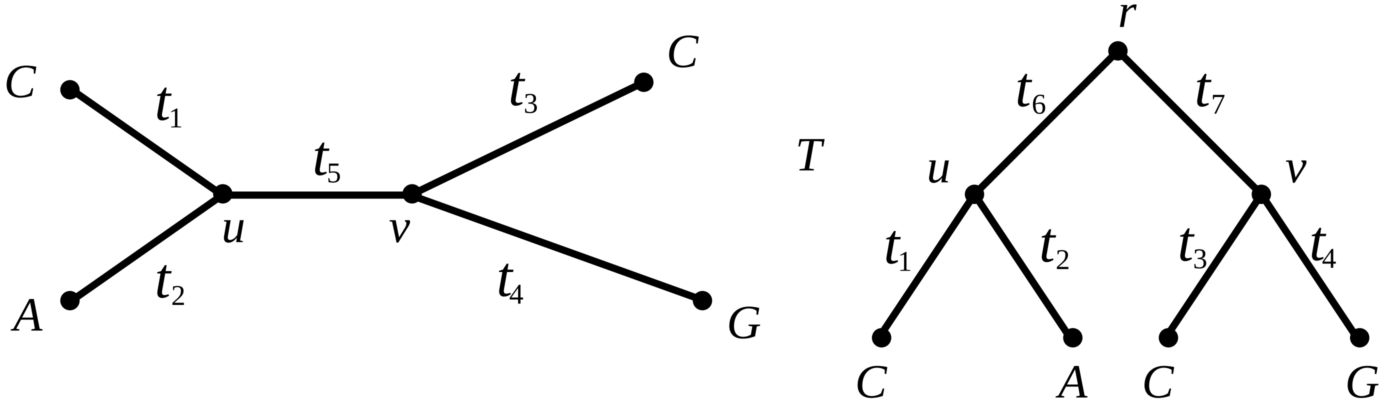
as the likelihood of the subtree rooted at node w, given that node w has state $k \in \mathcal{A}$.

At a leaf node l we have

$$\mathcal{L}_k(l) = \begin{cases} 1 & \text{if the leaf has state } k \\ 0 & \text{else} \end{cases}$$

Maximum Likelihood Trees, cont'd

For ease of illustration, we now insert a root node r at the internal edge such that $t_5 = t_6 + t_7$.



The conditional likelihood at the node r is

$$\mathcal{L}_k(r) = \Big(\textstyle\sum_{i \in \mathcal{A}} P_{ki}(t_6)\mathcal{L}_i(u) \Big) \cdot \ \Big(\textstyle\sum_{i \in \mathcal{A}} P_{ki}(t_7)\mathcal{L}_i(v) \Big)$$

r is already the root of the tree. Thus

$$\mathcal{L}(T, \vec{t}, Q \mid \mathcal{D}_s) = \sum_{i \in \mathcal{A}} \pi_i \mathcal{L}_i(r)$$

Maximum Likelihood Trees, cont'd

Note that the number of summands in the likelihood function now is linear in the number of internal nodes.

Sites are modeled independently of each other. The likelihood to observe an alignment $\mathcal{D}$ with n sites is the product over the site likelihoods

$$\mathcal{L}(T, \vec{t}, Q \mid \mathcal{D}) = \prod_{s=1}^{n} \mathcal{L}(T, \vec{t}, Q \mid \mathcal{D}_s)$$

Accordingly, the log likelihood is a sum over the site log likelihoods.

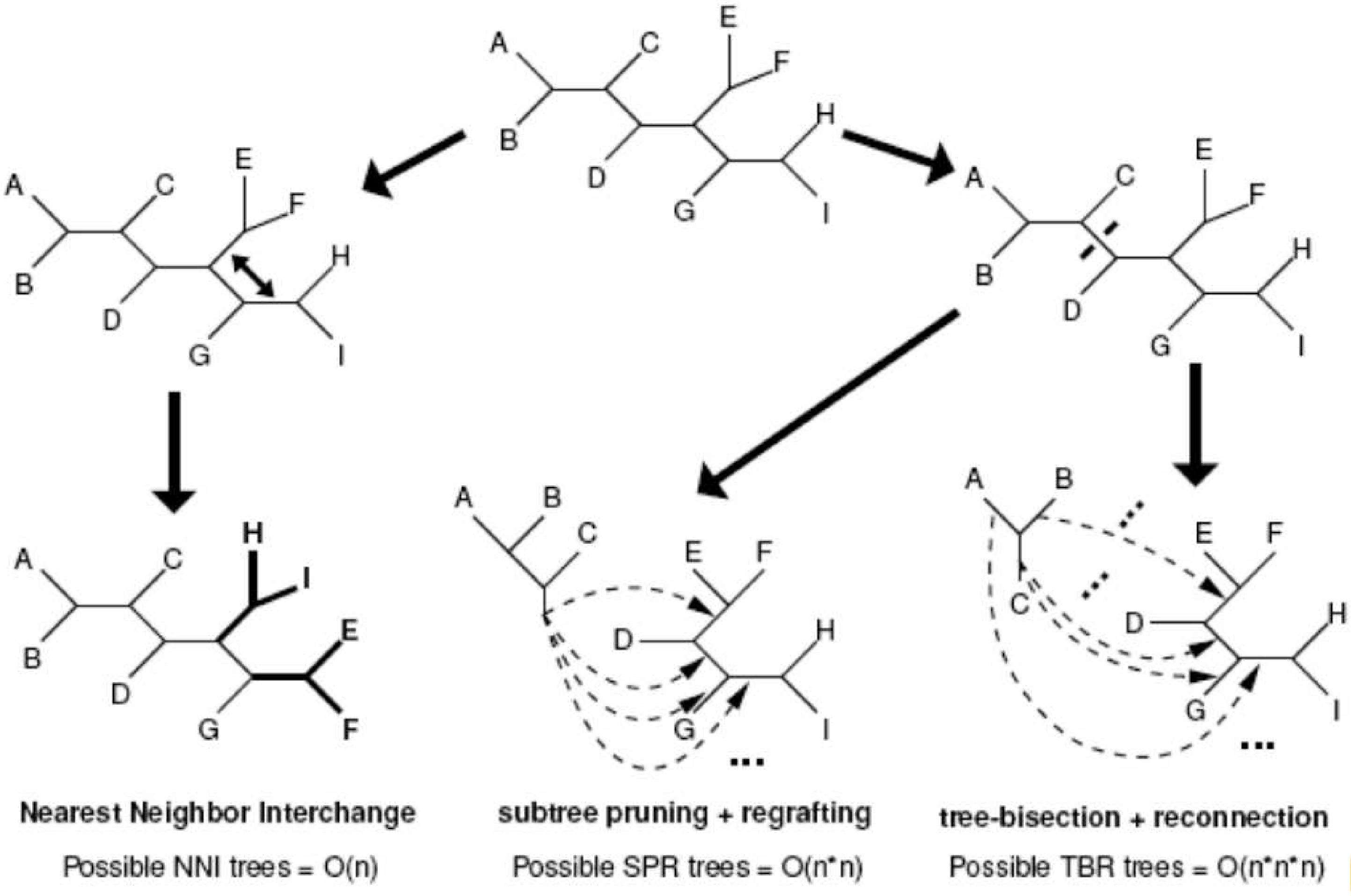
The likelihood $\mathcal{L}(T|\mathcal{D})$ to observe the alignment D under the tree T depends on the model parameters, the edge lengths $\vec{t}$ and the rate matrix elements in Q. In order to compute the likelihood one has to numerically optimize over $\vec{t}$ and the rate matrix Q (for a rate matrix Q with more parameters than the JC69-Q).

A *Maximum Likelihood Tree* $\widehat{\mathcal{T}}$ is the one with the largest likelihood $\mathcal{L}(T|\mathcal{D})$ among all possible tree topologies.

Heuristics to search the tree space

As discussed in the Maximum Parsimony section, the tree space is enormous. If it's not possible to examine all possible tree topologies, heuristic methods to search the tree space are applied.

Start with some 'good' tree (for example a Neighbor Joining tree) ...


A
B
C
D
E
F
G
H
I
...
Nearest Neighbor Interchange
Possible NNI trees = O(n)
subtree pruning + regrafting
Possible SPR trees = O(n*n)
tree-bisection + reconnection
Possible TBR trees = O(n*n*n)

Heuristics to search the tree space, cont'd

A fast and widely used heuristic to reduce the tree search space is *Quartet Puzzling* (Strimmer, v. Haeseler 1996, see also `http://www.tree-puzzle.de/`). The optimal tree for all subsets of sequences consisting only of four sequences (=quartet) is computed. Subsequently, the quartet trees are combined into a larger tree for all sequences.

Note that heuristics may get stuck in local optima of the likelihood landscape. The heuristic tree search procedure possibly should be repeated several times (with different initializations or starting points).

Evolutionary Markov processes

Müller and Vingron (2000) have summarized the properties of a Markov process being that describes the substitution process at a site of a molecular sequence. A π–EMP has the following properties:

- (X_t) is time homogeneous.

  $P_{ij}(t) = \text{Prob}[X_{s+t} = j | X_s = i] = \text{Prob}[X_t = j | X_0 = i]$.

- (X_t) is stationary w.r.t. π.
  $\pi_j = \sum_i \pi_i P_{ij}(t), \; \pi = \pi P(t) \quad \forall \; t$.

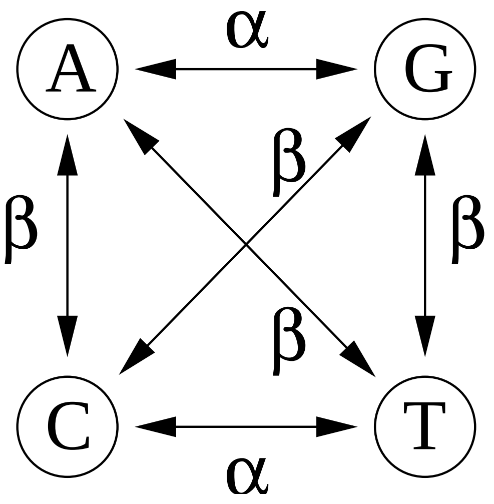
- (X_t) is reversible. $\pi_i P_{ij}(t) = \pi_j P_{ji}(t)$.

Evolutionary Markov processes, cont'd

The assumptions of the Jukes-Cantor model for the evolution of a DNA sequence are simplistic regarding substitution rates and the stationary distribution.

The *Kimura 2-parameter model* takes into account that transitions ($A \leftrightarrow G$ and $C \leftrightarrow T$) are more frequently observed than transversions.

$$Q_{\text{K2P}} = \begin{pmatrix} . & \alpha & \beta & \beta \\ \alpha & . & \beta & \beta \\ \beta & \beta & . & \alpha \\ \beta & \beta & \alpha & . \end{pmatrix}$$



Normally, the ML estimate $\widehat{\alpha}$ is larger than $\widehat{\beta}$.

The stationary distribution π is still the uniform distribution..

Evolutionary Markov processes, cont'd

The *Felsenstein 81 model* has one parameter for a substitution rate, but three parameters for a non-uniform nucleotide distribution:

$$Q_{F81} = \begin{pmatrix} \cdot & \pi_C & \pi_A & \pi_G \\ \pi_T & \cdot & \pi_A & \pi_G \\ \pi_T & \pi_C & \cdot & \pi_G \\ \pi_T & \pi_C & \pi_A & \cdot \end{pmatrix}$$

The *GTR* model is the most general time reversible model for nucleotide sequence evolution with 9 parameters (if one does not care about calibration 8 parameters)

$$Q_{\mathsf{GTR}} = \begin{pmatrix} \cdot & \alpha\pi_C & \beta\pi_A & \gamma\pi_G \\ \alpha\pi_T & \cdot & \delta\pi_A & \epsilon\pi_G \\ \beta\pi_T & \delta\pi_C & \cdot & \pi_G \\ \gamma\pi_T & \epsilon\pi_C & \pi_A & \cdot \end{pmatrix}$$

Empirical models of amino acid evolution

The number of model parameters specifying transitions between amino acids amounts to 209. This large number of parameters cannot be estimated from a single alignment of homologous amino acid sequences. Therefore the empirical approach has become generally accepted. The rate matrix is estimated by considering a large set of aligned sequences from a database and the obtained fixed parameter set is supposed to apply to other datasets.

Dayhoff proposed her pioneering and prominent model of amino acid replacement in the 1970ies from which she derived the PAM family of amino acid similarity matrices. The model is based on global alignments of closely related sequences and the reconstruction of phylogenetic trees followed by the estimation of ancestral sequences. Within the trees she counts the frequency of residues and residue pairs which are used to set up the 1-step transition matrix $P(1)$ of a time-discrete Markov chain. Transition matrices for larger evolutionary distances are obtained from multiples of $P(1)$, for example $P(250) = P(1)^{250}$, that is by extrapolating the observed replacement frequencies between close sequences.

Empirical models of amino acid evolution, cont'd

Similarity scores in the PAM similarity matrices for pairs of amino acids (i, j) are defined as a log likelihood ratio. For example, in the PAM250 similarity matrix,

$$S_{ij}(250) := \log \frac{\pi_i P_{ij}(250)}{\pi_i \pi_j}$$

The nominator is the probability that the residues have diverged from an ancestral residue according to Dayhoff's evolutionary model. The denominator is the probability to observe two residues by chance. The score is positive if the pair (i, j) frequently occurs in the alignments that were used to estimate transition probabilities of the Markov model.

Other empirical models of amino acid evolution are the VT models of Müller and Vingron (2000) and the WAG model of Wheelan and Goldman (2001).

Maximum Likelihood vs. Maximum Parsimony

- Compared to parsimony, Markov models take all possible evolutions into account (there is a small probability for each possible evolution)

- MP trees and ML trees are the same for well conserved alignments, that is, if the probability of change is very small

- We can estimate the variance of real valued parameters with ML

- One can test evolutionary hypothesis with Likelihood Ratio Tests and ask questions like:
  - Did the sequences evolve like a molecular clock and can thus be used to infer divergence times (in physical time units) ?
  - Were the substitution rates different for different nucleotide pairs?
  - Was some gene subject to positive selection in some lineage?

Summarizing probabilistic methods, keywords to remember:

- Time-continuous Markov Models:
  - stationary distribution
  - reversibility (detailed balance eq.)
  - rate matrix exponential
- Likelihood concept and Likelihood as objective function
- Jukes-Cantor correction
- Maximum Likelihood trees
- PAM matrices: the one-step transition probability matrix and the PAM series of similarity matrices