Models for DNA Evolution

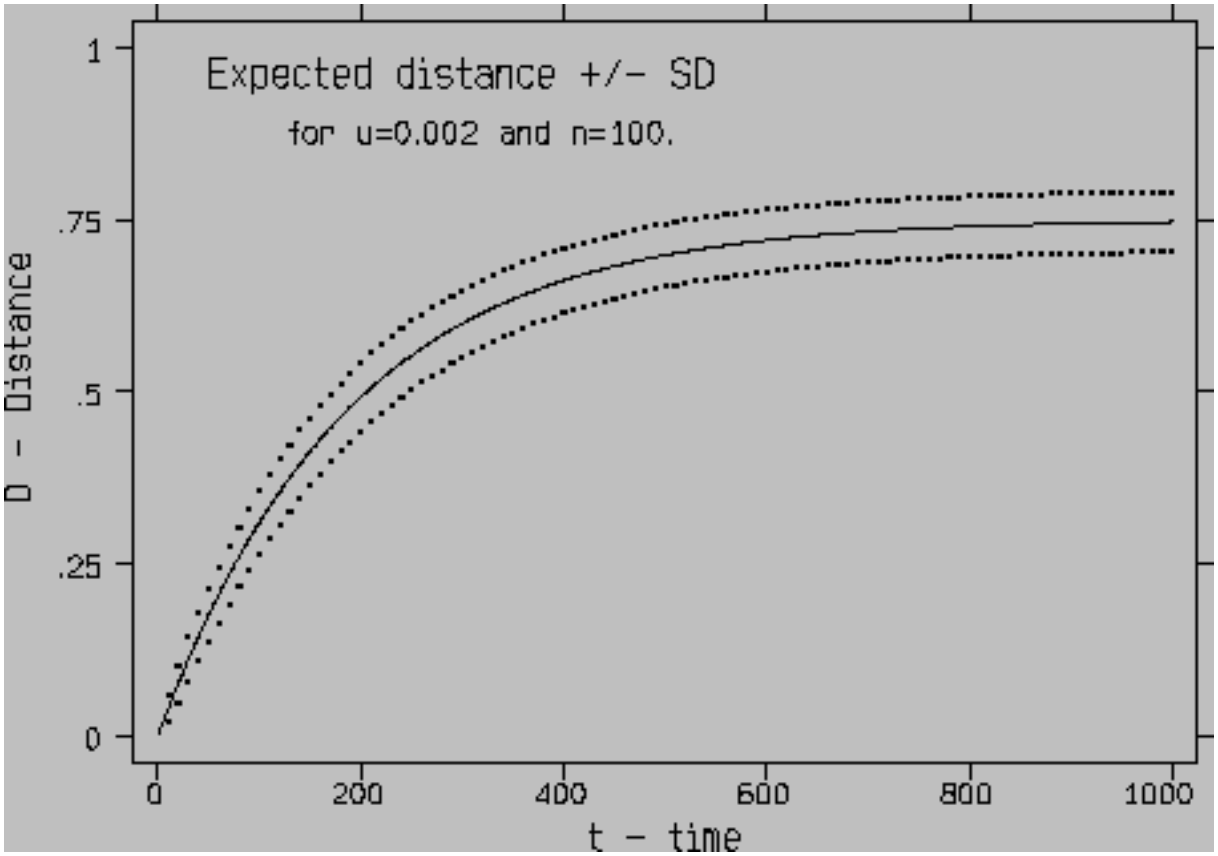
- D. Graur and W. H. Li: Fundamentals of Molecular Evolution

- We were talking about closely related sequences and remotely relates sequences.
- How can we measure the degree of divergence?

- We need a global alignment, and we assume it is correct.
- Naive answers:
  - Percentage of identical positions in the alignment.
  - Edit distance
  - Global alignment score

Percentage of identical positions in the alignment:

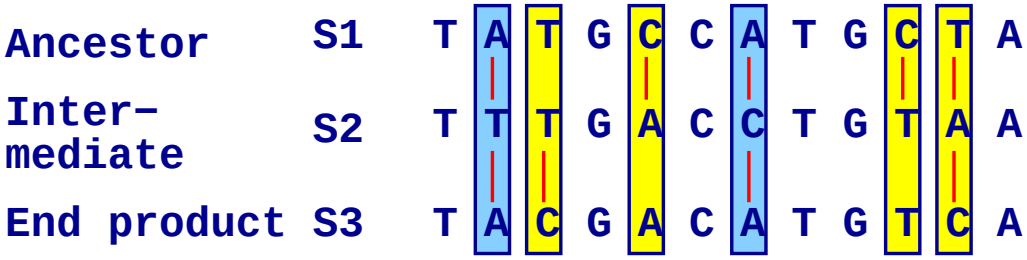
- Let us play evolution again. We start with a comon ancestor, and then we let the sequence evolve along two independent branches.
- In both branches we randomly choose positions and randomly mutate them. We do this with a rate of 0.002 mutations per unit of time.
- The average percentage of identical positions decreases with time.
- Starting in 100% and converging to 25%.
- Even unrelated sequences have in average 25% positions that are identical.


Expected distance +/- SD
for u=0.002 and n=100.
D - Distance
1
.75
.5
.25
0
0
200
400
600
800
1000
t - time

Problem

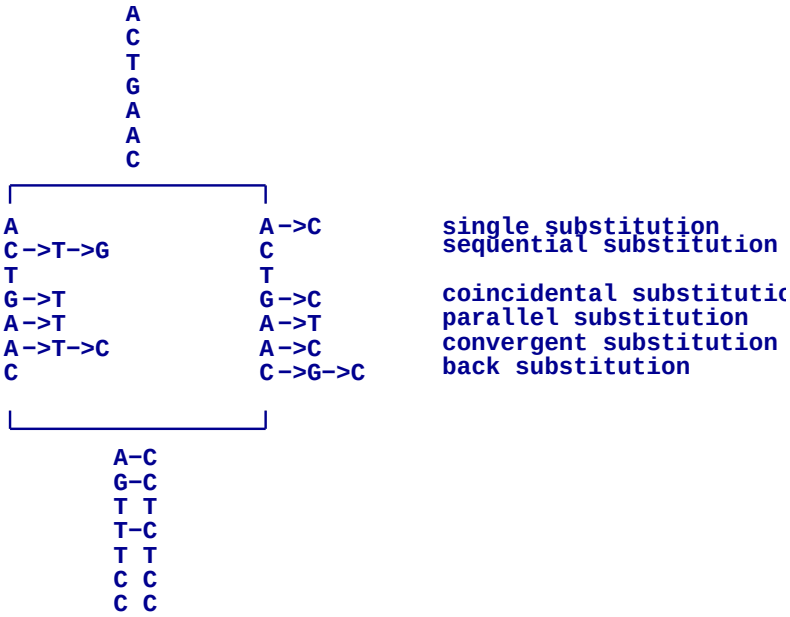
- Early mutations have a significant impact on the percentage identity.
- Late mutations almost do not have any additional effect.

The problem of multiple mutations

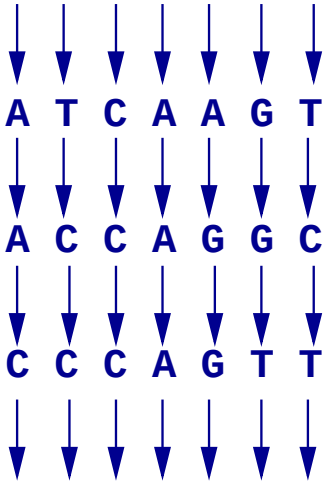


Total number of mutations: 9

Number of differences between S1 and S2: 4


A
C
T
G
A
A
C
A
C->T->G
T
G->T
A->T
A->T->C
C
A->C
C
T
G->C
A->T
A->C
C->G->C
single substitution
sequential substitution
coincidental substitutio
parallel substitution
convergent substitution
back substitution
A-C
G-C
T T
T-C
T T
C C
C C

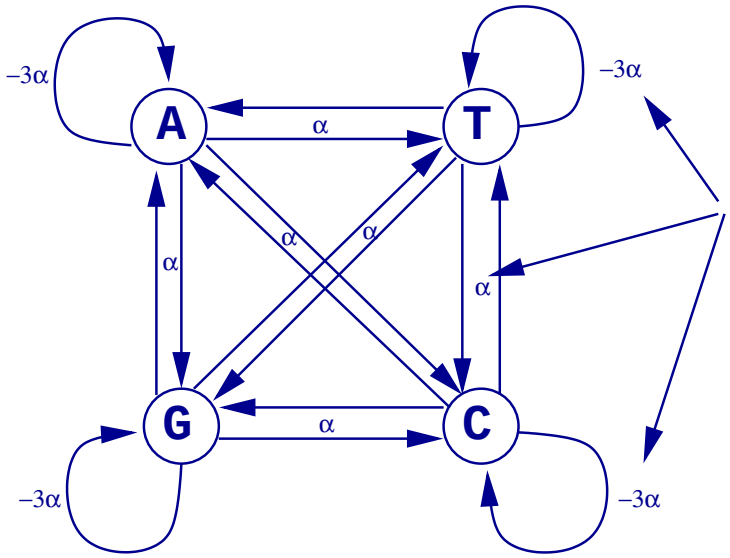
- We want to model the evolution of sequences by a set of independent but identical continuous time Markov chains for each position in the sequence.



The same Markov chain is operating on all of the positions of the evolving sequence

- We can estimate the most likely time parameter t from the observed differences between the two sequences.

The Jukes Cantor model



- $\{A = 1, C = 2, G = 3, T = 4\}$
- Assume we start the chain in state 1 (in an 'A')
- After a short period of time h, the probability that the chain is still in its initial state is

$$p_{11}(h) = 1 - 3\alpha h,$$

while for any of the other states, the probability that chain is in it is

$$p_{1j}(h) = \alpha h \quad j \neq 1.$$

- Let us shift to some arbitrary time point t:
- We are interested in the probability that the chain is in state 1 at $t+h$.
- What can happen in the short time period $[t, t+h]$?
  - Nothing, the chain was in state 1 at time ‘‘t’’ and stayed there.
  - The chain was in a different state at time t and changed to state 1 during the interval.

  We have

$$p_{11}(t+h) = p_{11}(t)(1-3\alpha h) + (1-p_{11}(t))\alpha h$$

- We have

$$\frac{p_{11}(t+h)-p_{11}(t)}{h} = -p_{11}(t)\,3\alpha + (1-p_{11}(t))\,\alpha$$

- Since the right hand side does not depend on h any more, taking the limit $h \to 0$ is trivial and we have the first order linear differential equation:

$$\frac{d\,p_{11}(t)}{d\,t} = -4\alpha p_{11}(t) + \alpha.$$

The solution is:

$$p_{11}(t) = \frac{1}{4} + \left(p_{11}(0) - \frac{1}{4}\right) e^{-4\alpha t}.$$

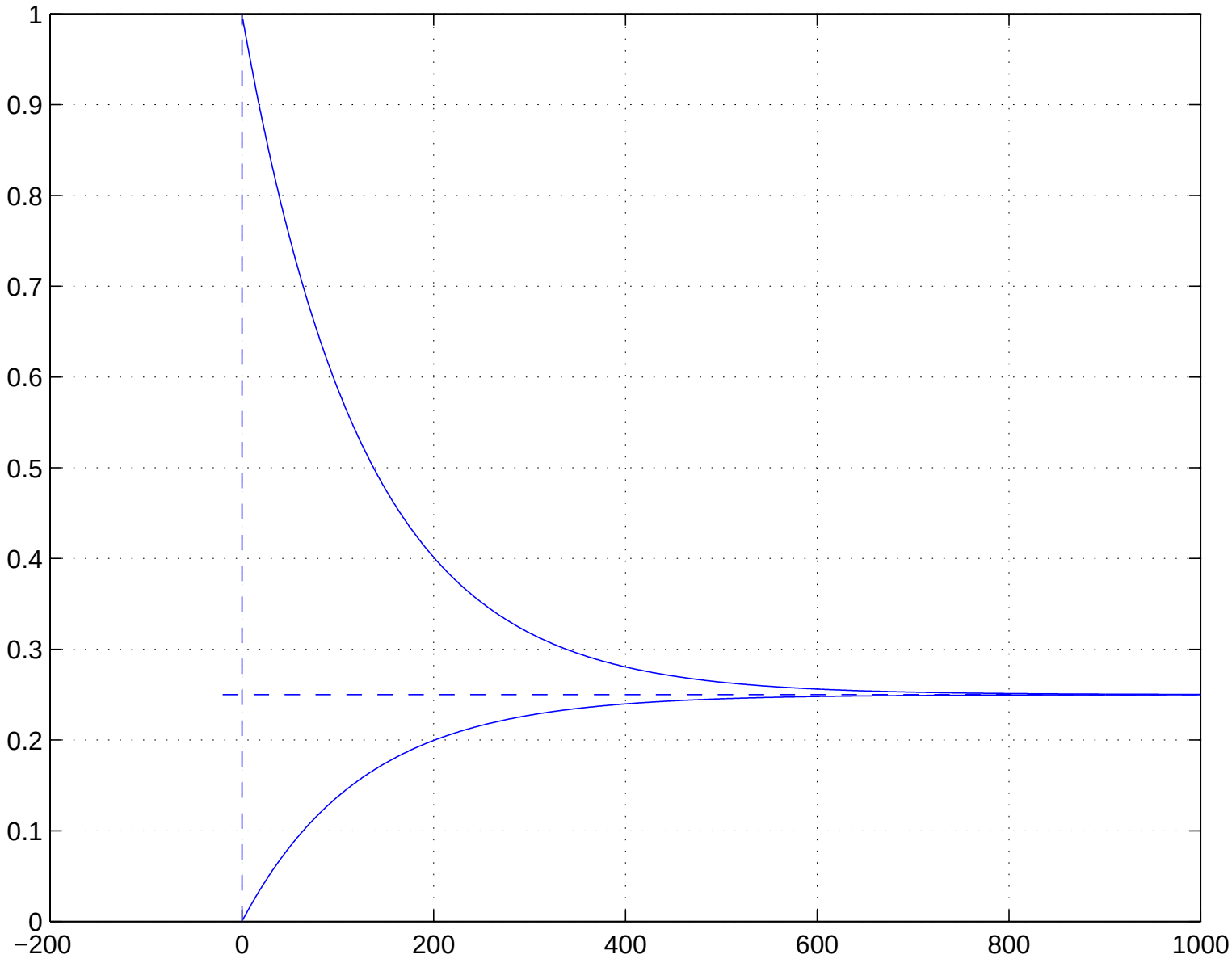
- Since $p_{11}(0) = 1$ and more generally $p_{ii}(0) = 1$ we get

$$p_{ii}(t) = \frac{1}{4} + \frac{3}{4}\, e^{-4\alpha\, t}$$

- And since the computation actually does not depend on the initial state at all, we also have for $i \neq j$

$$\begin{aligned} p_{ij}(t) &= \frac{1}{4} + \left(p_{ij}(0) - \frac{1}{4} \right) e^{-4\alpha t} \\ &= \frac{1}{4} - \frac{1}{4}\, e^{-4\alpha\, t} \end{aligned}$$

- Note, that we get an explicit formula for long range transition probabilities, that can be computed for each (i, j)-pair separately. In the Chapman-Kolmogorov equation, we have matrix products instead. Hence all the $p_{ij}(1)$ are needed to calculate $p_{11}(t)$


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
−200
0
200
400
600
800
1000

- The stationary distribution of the Jukes Cantor model is the uniform distribution.
- The model satisfies detailed balance. hence the model is time reversible.

- It is completely unclear what the unit of time $t = 1$ is?
- There are two strategies to overcome this problem:
  - Calibrate the time, such that $P[X_t \neq X_{t+1}] = 0.01$. One percent change in one unit of time. This always works.
  - Get rid of the time parameter and replace it by the expected number of substitution events. This is appropriate for simple models.

- We fix a single position in the sequences. Let us assume the ancestor state is 1 ('A').
- The probability that the two sequences are identical after a time period t of evolution in both branches is:

$$I_A(t) = P(t)_{11}^2 + P(t)_{12}^2 + P(t)_{13}^2 + P(t)_{14}^2$$

which can be calculated as

$$I_A(t) = \frac{1}{4} + \frac{3}{4} e^{-8\,\alpha\,t}$$

- Since all states are equal likely in the ancestor and the above calculation can be done in the same way for any ancestral state, we have for the probability of identical positions unconditional on the ancestor:

$$I(t) = I_A(t) = I_G(t) = I_C(t) = I_T(t).$$

- The probability of identity after t is

$$I(t) = \frac{1}{4} + \frac{3}{4} e^{-8\,\alpha\,t}$$

- The probability of difference after t is hence

$$D(t) = 1 - I(t) = \frac{3}{4}\left(1 - e^{-8\,\alpha\,t}\right)$$

- This gives us the equation

$$8\,\alpha\,t = -\log\left(1 - \frac{4}{3} D(t)\right).$$

- We have

$$8\,\alpha\,t = -\log\left(1 - \frac{4}{3}D(t)\right).$$

- We can observe $D = D(t)$. It is just the relative frequency of non matching positions in the alignment.
- We do not know anything about the unit of time and the rate α.

- What ever the unit of time is, let $K(t)$ be the expected number of mutations during a time period of t.
- For $t = 1$, these are 3α mutations per branch (linear approximation).
- Hence we have for an observed relative mismatch frequency D

$$K = -\frac{3}{4} \log \left(1 - \frac{4}{3} D \right).$$

- K ist the Jukes-Cantor distance of the sequences.

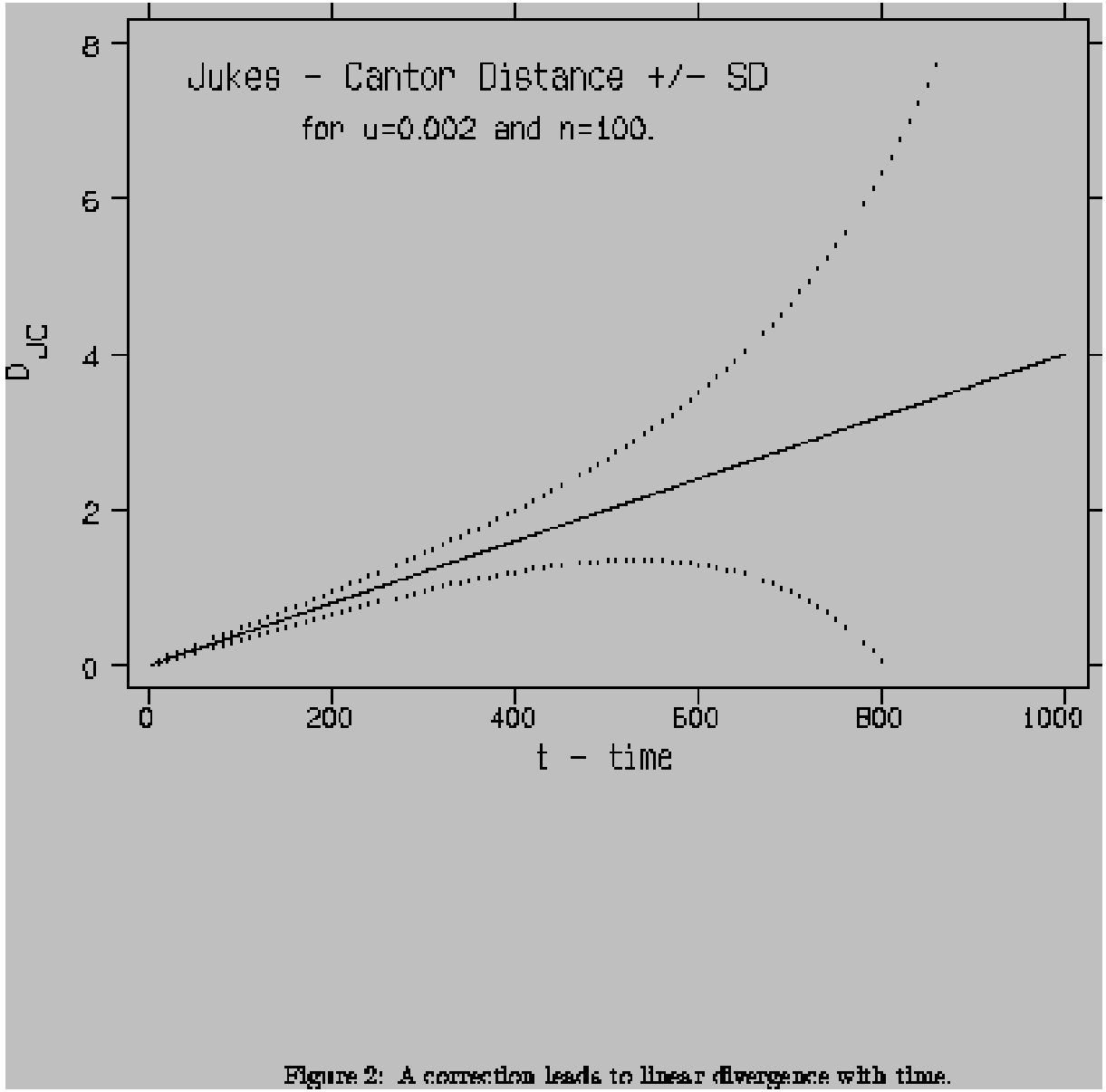
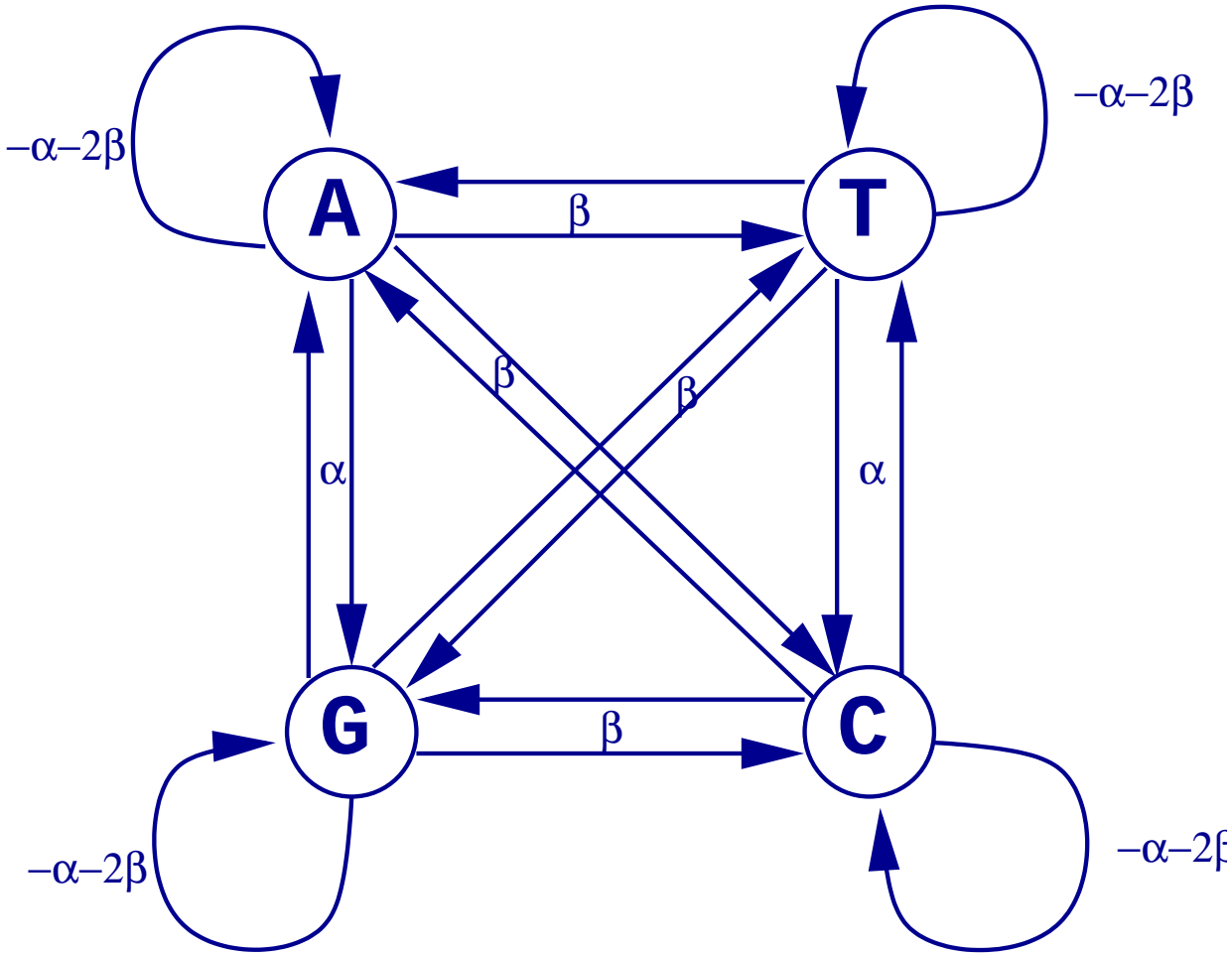


Figure 2: A correction leads to linear divergence with time.

- Note, that the Jukes-Cantor model does not specify the rate of mutations explicitly. It only specifies that the rate α is the same for all mutations.
- No attempt is made to estimate α.
- One can obtain more general models if one assumes different rates for different types of mutations.
- The expected number of mutations is then estimated from separate counts of different types of mismatches.

The Kimura model



- The Jukes-Cantor model is for DNA.
- What is different when dealing with proteins?
- ... more is needed, more is possible.
- Now, we aim for estimating rates explicitly.
- The models reflect the similarities of amino acids.
- They are important for estimating the degree of divergence ...
- ... and they are the basis of alignment scores for proteins.

- The 20 amino acids are quite different.
- Some are big some are small.
- Some are polar others are not.
- Some are hydrophilic others are hydrophobic.
- etc.

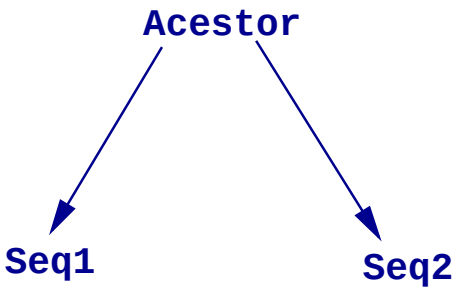
- These properties are highly influential on the fold and the function of a protein.
- It makes a much bigger difference to replace a small hydrophobic amino acid by a large hydrophilic one, than replacing it by another small hydrophobic amino acid.
- The chances that the first mutation is accepted by natural selection is less than for the second one.
- How can this be incorporated into the model?
- How can we asses amino acid similarities?

- Similar amino acids are more often replaced by each other then dissimilar amino acids.
- Dayhoff et al. 1978: Reverse this relation:
- We measure the similarity of amino acids by observing how often they are replaced by each other.
- Available data: sequence alignments.

Counting pairs of aligned amino acids

- Given a set of reliable pairwise alignments.
- For each pair of amino acids (i, j) we can count how often we observe amino acid i in the first sequence and aligned to it amino acid j in the second sequence.
- How can these observations be interpreted from a models perspective?

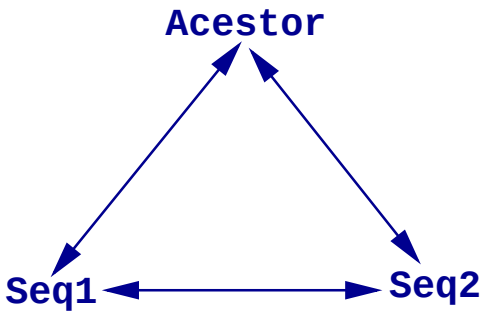
How can the differences between to present time sequences be modelled by a Markov chain?



Evolution operates from ancestors to descendants ... anf that is what should be modelled.

However, we can observe the process only indirectly by compairing descendants.

For a time reversible model this is no problem:



The differences between Seq1 and Seq2 can be modelled by a single time reversible model.

- We like a good model of protein evolution to meet the following requirements:
  - The transition matrix has strictly positive entries. (Every mutation is possible without intermediate steps.)
  - The model is in equilibrium. (The overall distribution of nucleotides (amino acids) does not change with time.)
  - The model is time reversible.
- We want these properties for mathematical reasons (simplicity).
- From a biological point of view, they are more or less ok.

Symmetry of the observations

- Deciding which of the sequences is the first and which is the second sequence in an alignment is completely arbitrary. Hence, we should not distinguish between observing i in the first and j in the second or j in the first and i in the second sequence.
- For example in
  MLKEVAKSHH
  MKHEVKHSKH
  we count the (H, K) pair 3 times.
- We can summarize the relative pair frequencies

$$m_{ij} = \frac{\#\text{positions where } i \text{ is aligned to } j}{\Sigma \text{ Length of alignment}}$$

  in a 20 by 20 matrix M_{emp}.
- Due to the symmetry of the observations, M_{emp} is symmetrical too.

Dayhoff's calculations

- Derive transition probabilities from M_{emp}.
- Following the original paper we treat mismatch and match observations separately.
  Mismatches first: assume $i \neq j$:

$$P_{ij} = P[i \text{ mutates}]P[i \to j \mid i \text{ mutates}].$$

- We want to estimate the term on the left, we have data for both terms on the right.

Mutability

- First calculate $P[i \text{ mutates}]$ the mutability of the amino acid i.
- This term can be estimated by

$$m_i = \frac{\sum_j M_{ij}}{\sum_{j,k,k \neq j} M_{kj}}.$$

- $P[i \to j \mid i \text{ mutates}]$ can be estimated by

$$\frac{M_{ij}}{\sum_{k \neq i} M_{ik}}$$

- The diagonal entries of P are consequently

$$P_{ii} = 1 - m_i$$

.

Questions

- Which alignments should be used?
- To which time point t do our observations belong?
- What is the unit of time?
- What is a good time point for deriving a model based score function for protein alignment?

Calibration and PAM Distance

- The time point $t = 1$ corresponds to 1% expected mismatch positions in the observed alignments.

$$P[X_t \neq X_{t+1}] = 0.01$$

- This unit of time is called 1 PAM
  "Point Accepted Mutations"
- 2 PAM correspond to the effect the Markov chain has, if it runs twice as long. In general this results in less than 2% expected mismatch positions, since with some small but positive probability one of the already changed positions mutates a second time.

Dayhoff's data

- Dayhoff et al. only used closely related alignments in the range of 0 to 17 PAM.
- They treated all this data in the same way. Hence they ignored the small differences in the degree of divergence.
- Having M_{emp}, they calculated transition matrices $P(t)$ as described above ... this also gives a rate matrix Q and a stationary distribution π.
- The stationary distribution π reflects the relative frequency of amino acids in the data.

Symmetry and time reversibility

- Since, M_{emp} is symmetrical, the resulting Markov chain is always time reversible
- This is another argument in favor of time reversible models.
- Even if evolution is not a reversible process, we do not have observations that would allow us to distinguish between directions.

Calibration continued

- The expected number of mismatch positions for $t = 0$ is 0. It is then continuously growing with t.
- Hence, there must be a t that corresponds to 1 PAM.
- This time point can be calculated efficiently by diagonalisation of the transition matrices $P(t)$.
- Dayhoff et al. made use of the linear approximation

$$P(t) = I + tQ$$

for small t and calibrated by transforming the mutabilities:

$$m_i \to m_i/100\pi_i.$$