Pairwise comparison table

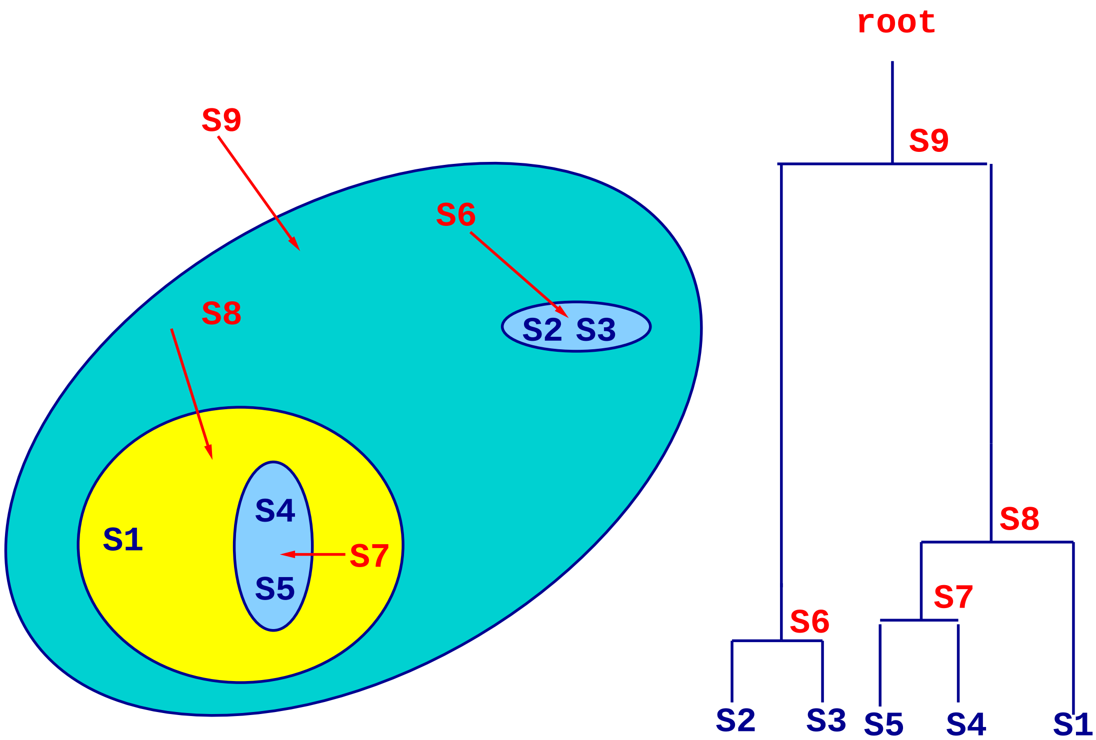
Calculate all pairwise alignment scores
and arrange them in a table

	S1	S2	S3	S4	S5
S1	●	−10	−5	4	−2
S2	−10	●	25	−8	0
S3	−5	25	●	−11	9
S4	4	−8	−11	●	−1
S5	−2	0	9	−1	●

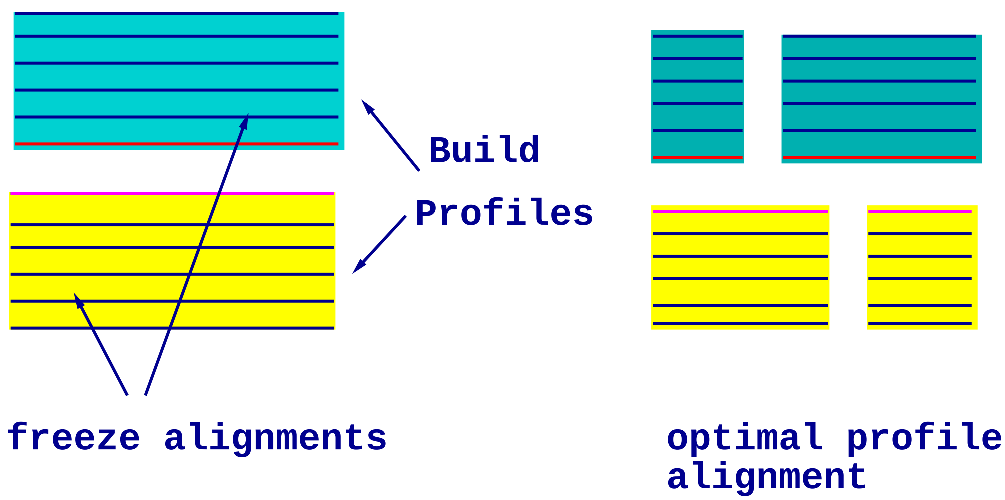
Convert all score into distances ...

1. Feng−Doolitle: $D=-\log(S-S_{rand})/(S_{max}-S_{rand})$
2. Model based distances

   log det formular
   maximum likelihood


S9
S6
S8
S2 S3
S4
S1
S7
S5
root
S9
S8
S7
S6
S2
S3
S5
S4
S1

Progressive Profile Alignment



ClustalW

Thompson Higgins Gibson 1994

Put together some good ideas:

Distances of pairs of sequences are based on a full stochastic model of sequence evolution (... to be discussed soon)

The guide tree is computed by a valid method of phylogenetic tree reconstruction
Saitou and Nei 1987

The multiple alignment is put together by progressive profile alignment

Sequence weighting is applied
Gaussian branching processes
Altschul Carroll Lipman 1989

Contributions from biology:

Score matrices influence the alignment results.

The score matrix used to score pairwise alignments is chosen on the basis of the evolutionary distance of the sequences.

Different matrices for closely related sequences and remote pairs of sequences.

The hydrophobic core of a protein is more conserved than its surface.

Position specific gap-open profile penalties are multiplied by a modifier that is a function of the residues observed at the position.

Hydrophobic residues (which are more likely to be in the well conserved core of the protein) give higher gap penalties than hydrophilc residues (which are more likely to be on the water accesible and less conserved surface of the protein)

Loops on the surface of a protein are often missing in other members of a protein family.

Gap open penalties are also decreased if the position is spanned by a consecutive strech of five or more hydrophilic residues.

Insertions and deletions are rare events, but once they occur, they are propagated and show up in many family members at the same position.

Both gap-open and gap-extend penalties are increased if there are no gaps in a column but gaps occur nearby in the partial alignments. This tries to enforce that gaps show up at the same position.

A pairwise alignment whispers, a full multiple alignment shouts out loud.

In the progressive alignment stage, if the score of a profile alignment is low, the guide tree may be adjusted on the fly to defer the low scoring alignment until later when more profile information has been accumulated.

...along the same lines:

Early decissions might be wrong since there was little profile information available at this time.

Remove the first sequence from the alignment and realign it to the almost full alignment using sequence to profile alignment.

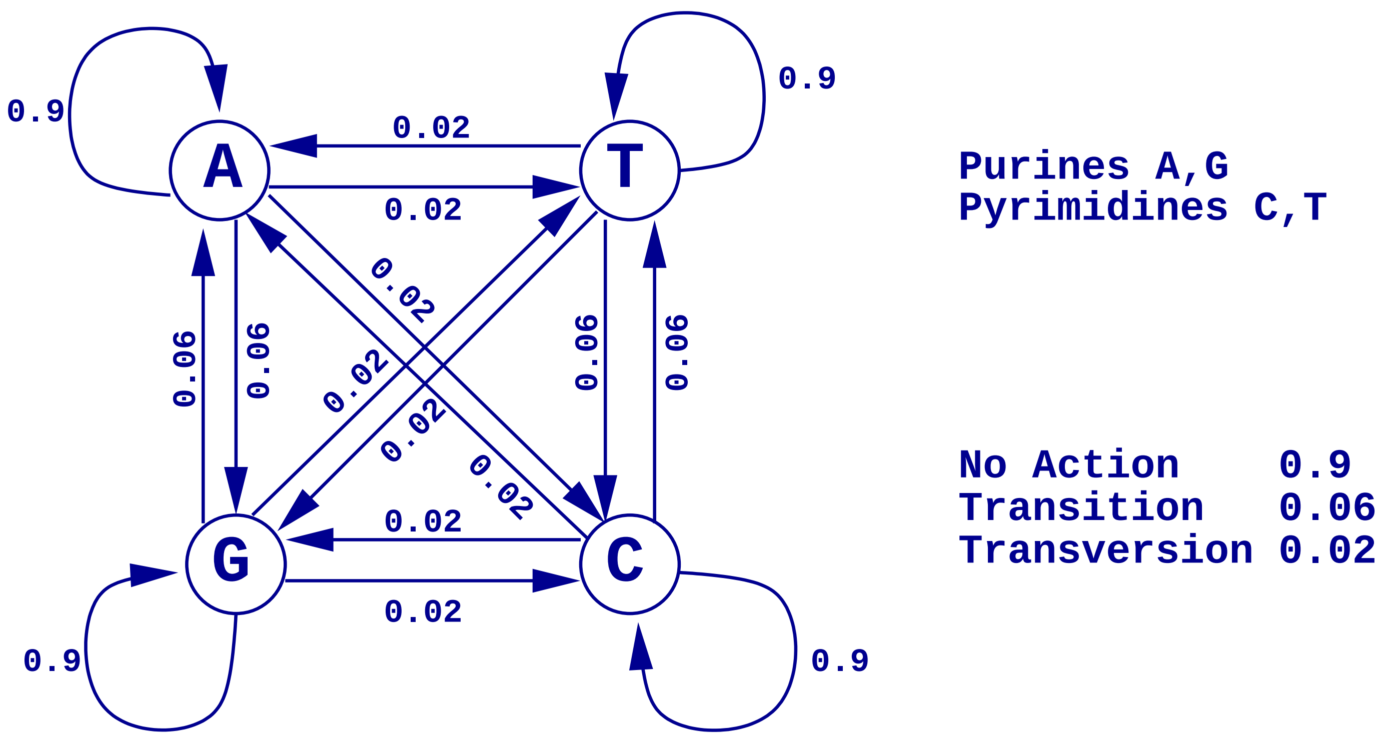
Continue with the second, third ,... sequence.

Iterate for some time.

(not implemented in ClustalW)

MARKOV CHAINS

- For a basic introduction, see: Ross, S (1997) A first course in probability (Prentice Hall)
- For details on the resolvent and related topics, see ISDS Discussion paper 00-25


A
T
G
C
0.9
0.02
0.06
Purines A,G
Pyrimidines C,T
No Action 0.9
Transition 0.06
Transversion 0.02

AAAAAAAAAAAAATTTTTTTTTTTTTTTTT
CCCCCCCTTTTTTTTTTCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCTTTTTTTTTTTTTTTTTTTTTCC
CCCCCCCCCCCCCCCCCCCCCCCCCGGAAA
AGGGGGGGAAAAAAAAAAAAAAAAAAGGGG
GGGGGGGGGGGGGGTTTTCCCCCCGGGGGG
GGGAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAGGCCCAGGGGGCC
CCCCCCCAGGGGGGGGGGGGGGGGGAAAGG
TTTTTTTTTTTTTTCCCGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGCCCCCCCCCCC
CCCCCCCCCCCTTTTTTTTTTTTTTTTTAA
AAAAAAAAAAAAAAAAAGGGGGGGGGGGGG
GAAAAAAAAAAAAAGGGGGGAAAAAAAAAA
AAAGGTTTTTTTTTAAAAAAAAAAAAAAAA
AAAAAAAAAACCCCCCCCCC

This experiment defines a stochastic process. Let us denote the first letter that is generated X_0, the second X_1, and so on. The stochastic process is then given by the sequences of random variables:

$$X_0, X_1, X_2, \ldots$$

A realization might be

$$X_0 = A, \quad X_1 = A, \quad X_2 = G, \quad \ldots.$$

We assume that we continue doing the experiment forever. Hence there is no last variable X_n.

Some Questions:

- If we know that $X_n = A$, what do we expect X_{n+1} to be?
- If we know that $X_{n-3} = T$, what do we expect X_{n+1} to be?
- If we know that $X_{n-3} = T$ and that$X_n = A$, what do we expect X_{n+1} to be?

- In general, past outcomes contain information on future outcomes.
- The older an outcome is the less it affects the future.
- But, if we know the present state (character) X_n then all past states (X_{n-3}) have absolutely no influence on future outcomes (X_{n+1}).
  (That is the way the experiment is designed)
- Or in other words:
  The future is independent of the past given the present.
- Stochastic processes with this property are called Markov Processes.

- The experiment is driven by the conditional probabilities $P[X_{n+1} = s | X_n = x]$. These probabilities are called transition probabilities.
- In the experiment the process could be in four different states: A,T,G or C. In general the set of possible states of a Markov Process is called its state space.
- If the state space consists of a finite or countable number of states the process is called Markov Chain.

- We assume that for all n and m

$$P[X_{n+1} = j | x_n = i] = P[X_{m+1} = j | x_m = i]$$

holds. ''We do not dream up a different experiment for each step, but use the same conditional probabilities for all of them.''

- If the state space is finite, we can enumerate the states by numbers $1, 2, \ldots n$ and summarize all transition probabilities in a $n \times n$ matrix $P = (p_{ij})$. Where

$$p_{ij} = P[X_{n+1} = j | x_n = i]$$

This matrix is called transition matrix

- Since the entries are probabilities $p_{ij} \geq 0$ and $\sum_j p_{ij} = 1$ hold.

- In our example we have

$$P = \begin{pmatrix} 0.9000 & 0.0200 & 0.0600 & 0.0200 \\ 0.0200 & 0.9000 & 0.0200 & 0.0600 \\ 0.0600 & 0.0200 & 0.9000 & 0.0200 \\ 0.0200 & 0.0600 & 0.0200 & 0.9000 \end{pmatrix}$$

- We need to specify with which state we want to start the chain:
- This can be done in a deterministic way by naming the state explicitly.
- Or it can be done in a stochastic way by choosing the initial state randomly.
- Let μ_i^0 denote the probability that the chain starts in state i
- The vector $(\mu_1^0, \ldots, \mu_n^0)$ is called the start distribution.

- By definition $P[X_0 = i] = \mu_i^0$ holds. We write $X_0 \sim \mu^0$. But ...
- ...what are the distributions of X_1, X_2 or X_n?

- X_1 first:

  Assume we start in state 1 ('A'): The probability of this event is

  $P[X_0 = A] = \mu_1^0$

  Now assume we go to state 2 ('C'): The probability is

  $P[X_0 = A]P[X_1 = C | X_0 = A]$

  In total the probability of having a 'C' in the second step is

$$
\begin{aligned}
& \sum_{l \in \{A,T,C,G\}} P[X_0 = l]P[X_1 = C | X_0 = l] \\
= & \sum_i \mu_i^0 p_{i2} =: \mu_2^1
\end{aligned}
$$

- Or more general, using matrix and vector notation:

$$\mu^1 = \mu^0 P \quad \text{and} \quad X_1 \sim \mu^1$$

- What about the distribution of X_2?

- $X_2 \sim ?$

$$\begin{aligned} P[X_2 = l] &= \sum_k P[X_1 = k] P[X_2 = l | X_1 = k] \\ &= \mu^1 P \\ &= (\mu^0 P) P = \mu^0 P^2 \end{aligned}$$

- Or more general for X_n:

$$X_n \sim \mu^n = \mu^0 P^n$$

- $P(n) = P^n$ is called n-step transition matrix.

- Chapman-Kolmogorov equation:

$$\begin{aligned} p_{ij}(m+n) &= \sum_k p_{ik}(m) p_{kj}(n) \\ P(n+m) &= P(n) P(m) \end{aligned}$$