1 Estimation of $\sigma(a, b)$:

- Our goal is to estimate the optimal sigma parameter, where:
  \[
  \sigma(a, b) = \log \left( \frac{P_1(a, b)}{P_0(a)P_0(B)} \right)
  \]

  We have already learned about the estimation of $P_0(a)$, so the remaining challenge is the estimation of $P_1(a, b)$.

- The naive approach would be to take many aligned proteins with a common ancestor, and to calculate $P_1(a, b)$ empirically. There is a problem with this method - We need to have $\sigma$ in order to align the proteins and determine they are derived from the same ancestor, but we need the alignment in order to calculate $\sigma$.

- We will use the Hennikof & Hennikof (1992) method to estimate $\sigma$: 

![Diagram showing the twilight zone and a common ancestor](image-url)
What the Hennikofs did was to decide that 62% is the magic number, meaning that proteins with 62% of similarity have a common ancestor. Next, they estimated the BLOSUM62 matrix:

\[ BLOSUM62 = \sigma_{62}(a,b) = \lfloor \log \left( \frac{P_{62}(a,b)}{P_0(a)P_0(b)} \right) \rfloor \]

where \( P_{62}(a,b) \) is the frequency of seeing the letter \( a \) in front of the letter \( b \) inside blocks that have 62% similarity. They used the basic Edit Distance method to align the 62% similar blocks, and in those alignments they empirically measured the \( P_{62}(a,b) \) probability.

2 Markov Model (MM):

- **Definition**: A Markov Model is a stochastic model used to model randomly changing systems, where it is assumed that future states depend only on the current state not on the events that occurred before it. We will discuss two types of Markov Models: Markov Chain and HMM (Hidden Markov Model).

- **Definition**: A hidden Markov model is a stochastic mathematical model, in which the states system being modeled is assumed to be a Markov process with unobserved (hidden) states.

- We will portray the Croupier problem, to help us understand the HMM. Imagine you are going to a casino, and the dealer has one regular fair die, and another biased die. Whenever he is under pressure of losing, he uses the biased die. We only see a die rolling, without knowing which one was used.

- **Definition**: Process - A series of random variables \( X_1, X_2, \ldots \)

- **Definition**: Markov Chain is a process with the Markovian property (with no hidden states).
• We will observe the probability of a series of states, and decompose it according to the chain rule

\[ P(X_1, X_2, ..., X_n) = P(X_1) P(X_2 | X_1) P(X_3 | X_1, X_2) ... = P(X_1) \prod_{i=2}^{n} P(X_i | X_1, ..., X_{i-1}) \]

• The Markov Property:
  for every \( i, j \)
  \[ P(X_{i+1} | X_1, ..., X_i) = P(X_{i+1} | X_i) \]
  \( X_{i+1} \perp X_{i-1} | X_i \)

\( X_i \) separates all the states before it from the ones after it:

• Under the Markov property assumption:

\[ P(X_1, ..., X_n) = P(X_1) \prod_{i=2}^{n} P(X_i | X_{i-1}) \]

• Markov Model:

\( (X_1, ..., X_n) - \) The states series.
\[ \sum = \{S_1, ..., S_k\} - \) The states set.
\[ \{X_1, ..., X_n\} \in \{S_1, ..., S_k\} \]

• We can model the process as:

\[ X_1 \sim \vec{P}_0 \]

where \( \vec{P}_0 \) is the initial distribution for receiving the first state.

And with a transition matrix of size \( k \times k \) we get:

\[ [P_t]_{i,j} = P(X_t = s_j | X_{t-1} = S_i) \]

This is the probability of transitioning from state \( S_i \) to state \( S_j \).

\( t \) - The current location along the series of states.

• For our needs we assume that the transition matrix is **Homogeneous**, meaning that it does not change along the run over the series

\[ \forall t_1, t_2 \ [P_{t_2}] = [P_{t_1}] \]

We mark the transition matrix value with \( \tau \)

\[ \forall t \ [\tau]_{i,j} = [P]_{i,j} = P(X_t = S_j | X_{t-1} = S_i) \]
• We will now model the **Croupier problem** with the transition matrix.

In the Markov Model the observations are our actual states. So in the Croupier example we observe the sequence of die used. In the case of Hidden Markov Models, as we will see later, we only observe the emissions of each state while the actual states are hidden. In the Croupier problem we see the die rolls while the die used is hidden from us.

• We have the following parameters:

  - \( \tilde{P}_0 \), with size of \( k \times 1 \) (initial probability to get each state)

    \[
    e.g: \ P_0(B) = 0.3, \ P_0(F) = 0.7
    \]

  - \( \tau \), with size of \( k \times k \) (probability to get from one state to another)

    \[
    \tau = \begin{array}{cc}
    \text{biased} & \text{fair} \\
    0.9 & 0.1 \\
    0.05 & 0.95
    \end{array}
    \]

  - \( (X_1, ..., X_n) \), The states series - The sequence of die used

    \[
    e.g: \ (X_0 = F, ..., X_n = B)
    \]

• And if we portray the problem as an HMM, The states series is hidden \( (H_1, ..., H_n) \), and we only observe the series of emissions made by the dice, according to the emission probability table:
3 CpG Islands

- CpG are regions of DNA where a Cytosine is followed by a Guanine in the linear sequence of bases along its $5' \rightarrow 3'$ direction. CpG Island is a region with high frequency of CpG.

$$5' \rightarrow A \rightarrow T \rightarrow T \rightarrow \underbrace{C \rightarrow G}_{\text{CpG}} \rightarrow G \rightarrow C \rightarrow T \rightarrow A \rightarrow 3'$$

- A methyl group can be added to Cytosines in CpG by the process of methylation:

$$CpG \rightarrow C^m pG$$

- And methylated Cytosine can change to a Thymine in the process of deamination:

$$C^m pG \rightarrow TpG$$

- We know that there is a 80% chance of a CpG to be methylated, thus there is a significant chance of seeing a different sequence than the original CpG. Because of that we would like to find CpG Islands, areas of the DNA that have high concentration of CpG. Inside an island: \( \%CpG \text{ in CGI} \geq 2.6\% \) (and this will be the definition of a CpG Island).

Outside an island: \( P(CpG) \ll P(C) P(G) \) (a small chance to see a CpG).

- We will model biological sequences with two non-Hidden MM, one assuming the sequence is a CpG island \( (H_1) \), and the second assuming it is not an island \( (H_0) \). There is no transition between the two for now, as they are different MM’s.
- Transition matrix outside a CGI ($H_0$)

\[
\begin{array}{cccc}
A & C & G & T \\
\tau^- &=& 0.3 & 0.2 & 0.29 & 0.21 \\
\end{array}
\]

- Transition matrix inside a CGI ($H_1$)

\[
\begin{array}{cccc}
A & C & G & T \\
\tau^+ &=& 0.18 & 0.27 & 0.43 & 0.12 \\
\end{array}
\]

- Problem: Determining if a sequence is $\overset{H_1}{\text{CpG Island}}$ or $\overset{H_0}{\text{not}}$, by building an optimal classifier.

- Solution: We will use the likelihood ratio. We'll calculate the likelihood of the sequence according to $H_0$ and $H_1$, and determine if the ratio is bigger than a predefined threshold.

By observing a sequence $\vec{x}$ of nucleotides, we can calculate the probability of the sequence being a CpG island the following way:

\[
P^+(\vec{x}) = P(X_1, ..., X_n) = \prod_{i=2}^{n} \tau^+_{X_{i-1},X_i} = P^+(X_1) \prod_{i=2}^{n} P^+(X_i \mid X_{i-1})
\]

We can do the same thing assuming the sequence is not a CpG island:

\[
P^-(\vec{x}) = P(X_1, ..., X_n) = \prod_{i=2}^{n} \tau^-_{X_{i-1},X_i} = P^-(X_1) \prod_{i=2}^{n} P^-(X_i \mid X_{i-1})
\]

- Then in order to determine between our two hypotheses regarding the sequence we can calculate the log likelihood ratio of the two to decide between our two hypotheses

\[
\log\left(\frac{P^+(\vec{x})}{P^-(\vec{x})}\right) = \log\left(\frac{P^+(X_1)}{P^-(X_1)}\right) + \sum_{i=2}^{n} \log\left(\frac{P^+(X_i \mid X_{i-1})}{P^-(X_i \mid X_{i-1})}\right) = \sum_{i=2}^{n} \log\left(\frac{\tau^+_{X_{i-1},X_i}}{\tau^-_{X_{i-1},X_i}}\right) = \sum \beta(X_{i-1}, X_i)
\]

Where $\beta$ is the optimal classifier

\[
\begin{array}{cccc}
A & C & G & T \\
\beta = \log(\tau^+) &=& -0.74 & 0.42 & 0.58 & -0.8 \\
\end{array}
\]