Outline

• Hidden Markov Models: Viterbi algorithm

Reading:
• Jones and Pevzner: Chapters 11.1-11.3
• Lecture notes
CpG Islands

**Question**: Given four nucleotides $\Sigma = \{A, T, C, G\}$, what is the probability of observing dinucleotide $CG$?
CpG Islands

**Question:** Given four nucleotides $\Sigma = \{A, T, C, G\}$, what is the probability of observing dinucleotide $CG$?

$CG \rightarrow C*G \rightarrow TG$

$CG$ is least observed dinucleotide as $C$ is easily methylated and has tendency to mutate into a $T$ afterwards.
CpG Islands

- Methylation is suppressed around promoter regions of genes in a genome. So CG appears at relatively high frequency within these CpG island.
- Finding CpG islands in a genome is an important problem for annotating genes and regulatory regions.
CpG Islands

- Methylation is suppressed around promoter regions of genes in a genome. So CG appears at relatively high frequency within these CpG islands.
- Finding CpG islands in a genome is an important problem for annotating genes and regulatory regions.

**Input:** DNA sequence $x = x_1 x_2 \ldots x_n$

**Output:** $\pi : \{1, \ldots, n\} \to \{\text{yes, no}\}$

<table>
<thead>
<tr>
<th>$x$</th>
<th>CGA</th>
<th>TG</th>
<th>CGA</th>
<th>AAAAAAT</th>
<th>AACGA</th>
<th>TTATATCG</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi$</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
</tbody>
</table>

**Question:** How do we identify CpG islands?
A Related Problem: Fair Bet Casino

• Game is to flip coins, two outcomes:
  Head or Tail

• Two coins: Fair and Biased
  \[
  \Pr(H \mid F) = \Pr(T \mid F) = 1/2 \\
  \Pr(H \mid B) = 3/4, \Pr(T \mid B) = 1/4
  \]

• The crooked dealer changes between Fair and Biased coins with probability 10%
CpG Islands and Fair Bet Casino

CG Islands
Input: DNA sequence \( x = x_1 x_2 \ldots x_n \) where \( x_i \in \{A, T, C, G\} \)
Output: \( \pi : \{1, \ldots, n\} \rightarrow \{\text{yes, no}\} \)

Fair Bet Casino
Input: Coin flips \( x = x_1 x_2 \ldots x_n \) where \( x_i \in \{H, T\} \)
Output: \( \pi : \{1, \ldots, n\} \rightarrow \{F, B\} \)

Question: Given \( x \), what is more likely: \( \pi \) or \( \pi' \)?
Markov Model $\mathcal{M} = (Q, A)$

- Set of states $Q$
  - Markov property:
    \[
    \Pr(Q_i = q_i \mid Q_1 = q_1, \ldots, Q_{i-1} = q_{i-1}) = \Pr(Q_i = q_i \mid Q_{i-1} = q_{i-1})
    \]
- Transition probabilities $A = [a_{ij}]$ on pairs of states
  - Rows sum to 1

**Fair Bet Casino**

$Q = \{F, B\}$

$A = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}$

**Where is the professor?**

$Q = \{\text{Providence, Boston, Beijing}\}$

$A = \begin{pmatrix} 0.5 & 0.4 & 0.1 \\ 0.65 & 0.3 & 0.05 \\ 0.7 & 0.1 & 0.2 \end{pmatrix}$
Hidden Markov Model $\mathcal{M} = (Q, A, \Sigma, E)$

- Set of hidden states $Q$
  - Markov property
- Transition probabilities $A = [a_{ij}]$ on pairs of states
- Set of emitted symbols $\Sigma$
- Emission probabilities $E = [e_{ik}]$ on state-symbol pairs

Two decisions:
1. What symbol should I emit? [emission probabilities $E$]
2. What state should I move to next? [transition probabilities $A$]

Fair Bet Casino

$$Q = \{F, B\}$$

$$A = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}$$

$$\Sigma = \{H, T\}$$

$$E = \begin{pmatrix} 0.5 & 0.5 \\ 0.75 & 0.25 \end{pmatrix}$$
Three Questions

**Question 1:** \( \Pr(\bar{x}, \pi) \)
What is the most probable path \( \pi^* \) that generated observations \( \bar{x} \)?

**Question 2:** \( \Pr(\bar{x}) = \sum_{\pi} \Pr(\bar{x}, \pi) \)
What is probability of observations \( \bar{x} \) generated by any path \( \pi \)?

**Question 3:** \( \Pr(\pi_i = s | \bar{x}) \)
What is the probability of observation \( x_i \) generated by state \( s \)?
Three Questions

**Question 1:**
What is the most probable path $\pi^*$ that generated observations $x$?

**Question 2:**
What is probability of observations $x$ generated by any path $\pi$?

**Question 3:**
What is the probability of observation $x_i$ generated by state $s$?
Joint Probability

$$\Pr(\bar{x}, \pi) = \Pr(\bar{x}_1, \pi_1, \ldots, \bar{x}_n, \pi_n) = \Pr(\bar{x}_n, \pi_n | \bar{x}_{n-1}, \pi_{n-1}, \ldots, \bar{x}_1, \pi_1) \cdots \Pr(\bar{x}_2, \pi_2 | \pi_1) \Pr(x_1, \pi_1) \Pr(\pi_1)$$

In practice, $\pi$ is hidden to us.

Goal: find $\pi$ maximizing $\Pr(x, \pi)$
Joint Probability

We are interested in identifying \( \pi^* \) with max prob \( \Pr(\hat{\pi}^* \mid \pi^*) \).

\[
\pi^* \in \Omega^n \\
\sum_{i=1}^{n} \pi_i \xi_i a_{\pi_{i-1}, \pi_i} \\
\pi = \pi_n, \pi_{n-1}, ..., \pi_1
\]

Optimal sub-structure
Joint Probability

\[ x_i = x_1, \ldots, x_c \]
\[ \pi_i = \pi_1, \ldots, \pi_i \]
\[ \overline{x} = \overline{x}_n \]
\[ \overline{\pi} = \overline{\pi}_n \]

\( \nu[s, i] \) denote the prob. \( \Pr(\overline{x}_i, \overline{\pi}_{i-1}, \pi_i = s) \)

prob. of the most probable state path \( \overline{\pi}_{i}^* \) of the first \( i \) observations, \( \overline{x}_i \) with end state \( \pi_i = s \)

\[ \Pr(\overline{x}, \overline{\pi}^*) = \max_{s \in Q} \Pr(\overline{x}_n, \overline{\pi}_{n-1}, \pi_n = s) = \max_{s \in Q} \nu[s, n] \]
Joint Probability

Base case: \( \sum_{i=1}^{\frac{x_1}{T}} v[s, i] = \Pr(x_1 = T, v[F, 1]) \)

\( i \geq 2 \)

\( \sum_{i} v[s, i] = \Pr(x_1, \pi_{i-1}, \pi_i = s) \)

\( = \Pr(x_1, \ldots, x_{i-1}, x_i, \pi_{i-1}, \ldots, \pi_i, \pi_{i-1}, \pi_i = s) \)

\( = \Pr(x_1, \ldots, x_{i-1}, \pi_{i-1}, \pi_i = s | x_1, \pi_{i-1}, \pi_i) \Pr(B) \)

\( = \Pr(x_i, \pi_i = s | \pi_{i-1}) \Pr(B) \)

\( \Pr(A | B) = \frac{\Pr(A, B)}{\Pr(B)} \)
Recurrence

\[ \mathbb{P}(\xi_i) > 1 \]

\[
= \Pr(x_i, \pi_i = s \mid \pi_i - 1, \pi_i^*) \Pr(x_4, \ldots, x_{i-1}, \pi_i^*)
\]

\[= \max_{t \in \mathbb{Q}} \Pr(x_i, \pi_i = s \mid \pi_i - 1 = t) \Pr(x_{i-1}, \pi_{i-2}, \pi_{i-1} = t) \]

\[
\text{cond. independence}
\]

\[= \max_{t \in \mathbb{Q}} \Pr(x_i \mid \pi_i = s) \Pr(x_{i-1} \mid \pi_{i-1} = t) \nu \sum_{t, c-1} \]

\[= \max_{t \in \mathbb{Q}} a_{t, s} \nu \sum_{t, i-1} \]

\[= c_{s, i} \nu \sum_{t, i-1} \]

\[= c_{s, i} \cdot \max_{t \in \mathbb{Q}} a_{t, s} \nu \sum_{t, i-1} \]
\[ v[s_i] = \begin{cases} a_{015} e_{s_i, x_{1i}}, \\ e_{s_i, x_i} \max_{t \in Q} v[t, c-1] a_{e_t} \end{cases} \]

if \( c = 1 \)

if \( c > 1 \)

Size: \( \frac{|Q|}{n} \)

\( 0(\frac{|Q|^2}{n}) \) Viterbi alg.
Alignment vs. Decoding Problem

Valid directions in the **alignment problem**.

Valid directions in the **decoding problem**.

IQ case
Viterbi Algorithm

- Finds path $\pi^*$ with maximum $Pr(x, \pi^*)$

- Dynamic Programming algorithm

- Runs in $O(\#\text{edges}) = O(n|Q|^2)$
Viterbi Algorithm – Numerical Issues

Value of products can become extremely small, leading to underflow

\[ v[s, i] = \begin{cases} 
  a_{0,s} \cdot e_{s, x_1}, & \text{if } i = 1, \\
  e_{s, x_i} \max_{t \in Q} \{v[t, i-1]a_{t,s}\}, & \text{if } i > 1. 
\end{cases} \]
Viterbi Algorithm – Numerical Issues

Value of products can become extremely small, leading to underflow

\[ v[s, i] = \begin{cases} a_{0,s} \cdot e_{s,x_1}, & \text{if } i = 1, \\ e_{s,x_i} \max_{t \in Q} \{v[t, i - 1]a_{t,s}\}, & \text{if } i > 1. \end{cases} \]

Use logarithms!

\[ \log(v[s, i]) = \begin{cases} \log(a_{0,s}) + \log(e_{s,x_1}), & \text{if } i = 1, \\ \log(e_{s,x_i}) + \max_{t \in Q} \{\log(v[t, i - 1]) + \log(a_{t,s})\}, & \text{if } i > 1. \end{cases} \]
Fair Bet Casino: Example

\[ X = \begin{array}{cccccccccccccc}
1 (T) & 2 (H) & 3 (T) & 4 (H) & 5 (H) & 6 (H) & 7 (T) & 8 (H) & 9 (H) & 10 (H) & 11 (T) \\
F & 0.25 & \cdot & 1 & \cdot & & & & & & & \\
B & 0.75 & \cdot & & & & & & & & & \\
\end{array} \]

\[ a_{0,F} = a_{0,B} = \frac{1}{2} \]

\[ Q = \{ F, B \} \]

\[ A = \begin{pmatrix}
0.9 & 0.1 \\
0.1 & 0.9 \\
\end{pmatrix} \]

\[ \Sigma = \{ H, T \} \]

\[ E = \begin{pmatrix}
0.5 & 0.5 \\
0.25 & 0.75 \\
\end{pmatrix} \]

\[ v[F, 1] = a_{0,F} \cdot e_{F,T} = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} \]

\[ v[B, 1] = a_{0,B} \cdot e_{B,T} = \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{8} \]

\[ v[s, i] = \begin{cases}
\frac{a_{0,s} \cdot e_{s,x_1}}{e_{s,x_i} \max_{t \in Q} \{ v[t, i-1]a_{t,s} \}} & \text{if } i = 1, \\
\frac{a_{0,s} \cdot e_{s,x_i}}{e_{s,x_i} \max_{t \in Q} \{ v[t, i-1]a_{t,s} \}} & \text{if } i > 1.
\end{cases} \]

\[ v[F, i] = \begin{cases}
\max \{ 0.5 \cdot 0.25 \cdot 0.9, 0.8 \cdot 0.125 \cdot 0.1 \} \\
\end{cases} \]

\[ \text{v} = 11 \]
\[ Pr(\bar{x}) = \sum_{\pi} Pr(x_1, \pi_1, \ldots, x_n, \pi_n) = \sum_{(\pi_2, \ldots, \pi_n)} \prod_{i=2}^{n} \mathbb{P}(x_i | \pi_i, \pi_{i-1}, \pi) \]

Goal: \( O(n \log n) \)
\[ \Pr(\bar{x}) = \max_{\pi} \sum_{\pi} \Pr(\bar{x}, \pi) \geq \Pr(\bar{x}, \pi^*) \]
\[ \overline{\chi_i} = \chi_1, \ldots, \chi_i \]
\[ \overline{\pi_i} = \pi_1, \ldots, \pi_i \]

\[ \Pr(\overline{\chi_i}, \overline{\pi_i} = S) = \Pr(\chi_1, \ldots, \chi_i, \overline{\pi_i} = S) = \left\{ \begin{array}{l} E[S_{i, i}] \end{array} \right\} \]

\[ = \sum_{\overline{\pi_{i-1}}} \Pr(\overline{\chi_i} | \overline{\pi_{i-1}}, \pi_i = S) \]

\[ \Pr(\overline{\chi}) = \Pr(\chi_2, \ldots, \chi_n) = \sum_{S \in Q} \Pr(\chi_1, \ldots, \chi_n, \pi_n = S) \]

\[ = \sum_{S \in Q} E[S_{i, n}]. \]
\[ E[s, i] = \Pr (x_1, \pi_i = s) = a_{0s} \cdot \varepsilon_s, x_1 \]

Step \( i > 1 \)

\[ E[s, i] = \Pr (\bar{x}_i, \pi_i = s) = \Pr (x_1, \ldots, x_i, \pi_i = s) \]

\[ = \sum_{(\pi_2, \ldots, \pi_{i-1})} \Pr (x_1, \ldots, x_{i-1}, x_i, \pi_2, \ldots, x_{i-1}, \pi_i = s) \]

\[ = \sum_{\bar{\pi}_{i-2}} \Pr (x_1, \ldots, x_{i-2}, x_i, x_{i-2}, \pi_2, \ldots, \pi_{i-2}, \pi_i = s) \cdot \varepsilon_s, x_i \]
\[
\sum_{\pi_{i-1}} \Pr\left( x_1, \ldots, x_{i-1}, \pi_1, \ldots, \pi_{i-2}, \pi_{i-1} = t, \pi_i = s \right) a_{s, x_i} = \\
\sum_{t \in Q} \left( \sum_{\pi_{i-2}} \Pr( x_1, \ldots, x_{i-1}, \pi_1, \ldots, \pi_{i-2}, \pi_{i-1} = t) \right) a_{t, s} e_{s, x_i} \\
= e_{s, x_i} \sum_{t \in Q} \mathbb{I}[t, i-1] a_{t, s}
\[ e[s_i] = \begin{cases} a_0, & \text{if } i = 1, \\ e[s, x_i] \sum_{t \in q} e[t, i-1] a_t, & \text{if } i > 1. \end{cases} \]
\[ P_r(\bar{\tau}_i = s \mid \bar{x}) = \frac{P_r(\bar{x}, \tau_i = s)}{P_r(\bar{x})} \]

we know how to compute \( P_r(\bar{x}) = \sum_{\text{seq} \in S, n} \)

(forward algo.)

\[ P_r(\bar{x}, \tau_i = s) = P_r(x_1, \ldots, x_i, x_{i+1}, \ldots, x_n, \tau_i = s) \]

\[ = P_r(x_2, \ldots, x_i, \tau_i = s) \cdot P_r(x_{i+1}, \ldots, x_n \mid \tau_i = s) \]

\[ = \sum_{S \in \text{seq}} P_r(x_{i+1}, \ldots, x_n \mid \tau_i = s) \]
\[ b_{\pi(n)} = \Pr(\pi_{n+1}, \ldots, \pi_n | \pi_n = s) \]

\[ c_{\pi(n)} : \quad b_{\pi(n)} = \Pr(\pi_{n+1}, \ldots, \pi_n | \pi_n = s) \]

\[ \Omega = \text{sample space} = \Pr(\Omega | \pi_n = s) \]

\[ b_{\pi(s)} = 1 \quad \forall s \in \Omega \]

\[ \frac{\Pr(\Omega, \pi_n = s)}{\Pr(\pi_n = s)} = \frac{1}{\Pr(\pi_n = s)} \]
Step: $i > 1$.

$$b[s, i] = \sum_{t \in Q} a_{s, t} \cdot e_t \cdot x_{i+1} \cdot b[t, i+1]$$
$$\mathbb{E}[s_i] = \sum_{t \in Q} 1, \quad \text{if } c = n,$$

$$\sum_{t \in Q} a_{s_i + t, x_{t+1}} b[t, y], \quad i \neq 1 \leq i \leq n.$$ 

$$\Pr(\tau_i = s \mid \bar{x}) = \frac{\Pr(\bar{x}, \tau_i = s)}{\Pr(\bar{x})} = \frac{\sum_{\bar{z} \in \mathbb{Z}_s, i} f[\bar{z}, c]}{\sum_{s \in Q} f[\bar{z}, n]}$$

1. Run forward to compute $f[O(10^2 n)]$
2. Run backward to compute $b[O(10^2 n)]$
Posterior Decoding

\[ \pi^* = \pi^n \]

\[ \pi_i = \text{arg max}_{s \in Q} Pr(\pi_i = s | \vec{x}) \]

Viterbi Decoding: \( \pi^* = \text{arg max} Pr(\vec{x}, \vec{\pi}) \)
\( P(A) \)

\( P(A) \)

\( P(A, B) \)

\( P(C | A, B) \)
10X Genomics: Synthetic Long Reads

Genome indexing by partitioning and molecular barcoding

Genome

partitioning

barcoding short reads of each partition

50–75 molecules per droplet

<50 kb, coverage ~0.1X

100 bp

2016–17
$R_j = \{ r_i | \forall i, r_i \text{ contains barcode } j \}$: Paired-reads possessing barcode $j$

Sort linked-reads and calculate distances between them

$$D_j = [d_i | \forall i, d_i : \text{distance between } r_i \text{ and } r_{i+1}, r_i < r_{i+1}]$$

Define distances $d_i$ as intra- or inter- long molecules

$$\Sigma_j = [d_1, d_2, d_3, ..., d_{500}, d_{501}, d_{502}, ..., d_{1001}, d_{1002}, d_{1003}, ...]$$

$$Q_j = [\text{I, I, I, ...}, \text{O}, \text{I, ...}, \text{I, O, I, ...}]$$

I: intra long molecule, O: inter long molecules
HMM

\[ \Sigma_i = [d_1, d_2, d_3, \ldots, d_{500}, d_{501}, d_{502}, \ldots, d_{1001}, d_{1002}, d_{1003}, \ldots] \]

\[ Q_j = [I, I, I, \ldots, I, O, I, \ldots, I, O, I, \ldots] \]

\[ \ell_{j1}, \ell_{j2}, \ell_{j3} \]

I: intra long molecule, O: inter long molecules
Copy Number Variation

- Different individuals may have different number of copies of segments of genome.
- These variants are associated with various diseases: autism, schizophrenia, cancer
Measuring Copy Number Variants

Comparative Genomic Hybridization (CGH)

Chromosome CGH provides "cytogenetic" resolution ~ 10 Mb

Resolution of array CGH depends on spacing and length of clones
Segmentation and Copy Number Calling

Divide genome into segments of equal copy number

Genomic position

Input

Output

Genomic position

Log$_2$(R/G)

Deletion

Amplification
Segmentation and Copy Number Calling

Divide genome into segments of equal copy number

Input: $X_i = \log_2 \frac{T_i}{R_i}$, clone $i = 1, ..., N$
Output: Assignment $s(i) \in \{S_1, ..., S_K\}$ where $S_i$ represent copy number states
Summary

• Markov property – Current state depends only on previous state
• Hidden Markov Models: states are not given only emitted symbols
• Viterbi algorithm: Find the most likely sequence of states given a set of observations

Reading:
• Jones and Pevzner: Chapters 11.1-11.3
• Lecture notes