Ab initio methods

- Use information embedded in the genomic sequence *exclusively* to predict the gene structure.
- Find structure $G$ representing gene boundaries + internal gene structure which maximizes the probability $P(G|\text{genomic sequence})$.
- Hidden Markov models are the predominant generative method for modeling the problem.
Ab-initio methods

- **Advantages**
  - Intuitive, natural modeling
  - Prediction of ‘novel’ genes, *i.e.*, with no a priori known cDNA or protein evidence

- **Caveats**
  - Not effective in detecting alternatively spliced forms, interleaved or overlapping genes
  - Difficulties with gene boundary identification
  - Potentially large number of false positives with over-fitting
A simple example: CpG Islands

CpG nucleotides in the genome are frequently methylated. (Write CpG not to confuse with CG base pair)

C → methyl-C → T

Methylation often suppressed around genes, promoters → CpG islands

(c) Devika Subramanian, 2009
Example: CpG Islands

In CpG islands, CG is more frequent than in the rest of the genome.
GC content

- Genes tend to be rich in GC content
- GC-rich genomes are more thermodynamically stable
GC content of the human genome: mean 41%
CpG islands
Two problems

- Given a short DNA sequence, does it come from a CpG island or not?
  - Is this part of a CpG island or not?
- How to find the CpG islands in a long sequence?
Generative models

Models generate sequences of strings in the A, T, C, G alphabet. Model parameters are tuned to reflect characteristics of CpG and non CpG islands.
We are interested in predicting weather \((w)\), which can be either be sunny \((s)\) or rainy \((r)\). \(s\) and \(r\) are values of the random variable \(w\).

The weather on a given day depends only on the weather on the previous day.

\[
P(w_t \mid w_{t-1}, \ldots, w_1) = P(w_t \mid w_{t-1})
\]

This is the Markov property.
Markov process example

- We have knowledge of the transition probabilities between sunny and rainy days.

\[
\begin{bmatrix}
0.5 & 0.5 \\
0.1 & 0.9
\end{bmatrix}
\]

Rows of the transition matrix sum to 1.

- We know the initial probabilities of s and r.

(c) Devika Subramanian, 2009
Generating weather sequences

- Lets say we start with a sunny day.
- Now we consult our transition matrix and find that \( P(w_t | w_{t-1} = s) = [0.9 \ 0.1] \). It is more likely that the next day will be sunny too.
- We repeat this process, flipping coins biased by the probability \( P(w_t | w_{t-1}) \) to get a sequence representing weather for a consecutive set of days.
Generating sequences

\[
\begin{pmatrix}
0.9 & 0.1 \\
0.5 & 0.5
\end{pmatrix}
\]
Predictions

- Suppose day 1 is rainy. We will represent this as a vector of probabilities over the two values.

\[ p(1) = [0 \ 1]; \]

- How do we predict weather on day 2 given \( p(1) \) and the transition probabilities \( P \)?

- From \( P \), we can see that the probability of day 2 being sunny is .5, and for being rainy is 0.5

\[ p(1) * P = [0.5 \ 0.5]; \]
Probability of a sequence

- Given a Markov model specified by an initial state probability vector $\pi$, and a transition probability matrix $P$, what is the probability of observing the sequence “rrrrrrrs”?

$$P(X = rrrrrrrs) = \pi(r)P(r \mid r)P(r \mid r)P(r \mid r)P(r \mid r)P(r \mid r)P(s \mid r)$$

$$= \pi(r) \prod_{t=2..7} P(x_t \mid x_{t-1}) = (0.5)^7$$

$\pi = [0.5 \ 0.5]$;

$P = \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix}$
Which weather pattern is more likely?

- Given a transition model

\[
\begin{pmatrix}
  s & r \\
  0.9 & 0.1 \\
  r & 0.5 & 0.5
\end{pmatrix}
\]

- And an initial state distribution: [0.5 0.5]

- And two sequences: rrrrrrs and sssssssr

Which is more likely, given the model?
Comparing likelihoods

\[ P(X = rrrrrrs \mid Model) = \pi(r)[P(r \mid r)]^5 P(s \mid r) = (0.5)^7 \]

\[ P(X = ssssssr \mid Model) = \pi(s)[P(s \mid s)]^5 P(r \mid s) = 0.5 \times (0.9)^5 \times 0.1 \]
Markov models (summary)

- States: \( S = \{s_1, \ldots, s_N\} \), \( N \) states
- Transition probability:
  - \( a_{ij} = P(X_{t+1} = s_j \mid X_t = s_i) \), \( i, j \) in \([1..N]\)
- Initial state probability
  - \( p_{i} = P(X_1 = s_i) \), \( i \) in \([1..N]\)

Model generates sequences of states from \( S \), and we can compute how likely a sequence is given the model.
Markov models for CpG islands

A state for each of the four letters A, C, G, and T in the DNA alphabet

From a set of known CpG islands, and non CpG islands, estimate the transition probabilities

\[
a_{st}^+ = \frac{c_{st}^+}{\sum_{t'} c_{st'}^+}
\]

<table>
<thead>
<tr>
<th>+</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>.180</td>
<td>.274</td>
<td>.426</td>
<td>.120</td>
</tr>
<tr>
<td>C</td>
<td>.171</td>
<td>.368</td>
<td>.274</td>
<td>.188</td>
</tr>
<tr>
<td>G</td>
<td>.161</td>
<td>.339</td>
<td>.375</td>
<td>.125</td>
</tr>
<tr>
<td>T</td>
<td>.079</td>
<td>.355</td>
<td>.384</td>
<td>.182</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>-</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>.300</td>
<td>.205</td>
<td>.285</td>
<td>.210</td>
</tr>
<tr>
<td>C</td>
<td>.322</td>
<td>.298</td>
<td>.078</td>
<td>.302</td>
</tr>
<tr>
<td>G</td>
<td>.248</td>
<td>.246</td>
<td>.298</td>
<td>.208</td>
</tr>
<tr>
<td>T</td>
<td>.177</td>
<td>.239</td>
<td>.292</td>
<td>.292</td>
</tr>
</tbody>
</table>
Using the model

- To use the model for classification of a given sequence, calculate the log-odds ratio.
- Is the sequence more likely to come from a CpG island or a non-CpG region?

\[
\frac{P(x \mid \text{CpG})}{P(x \mid \text{nonCpG})} > 1
\]

\[
\log \frac{P(x \mid \text{CpG})}{P(x \mid \text{nonCpG})} > 0
\]

Log-odds ratio
The log-odds ratio

\[ S(x) = \log \frac{P(x|CpG)}{P(x|\text{nonCpG})} = \sum_{i=1}^{L} \log \frac{a_{x_{i-1}x_{i}}^{+}}{a_{x_{i-1}x_{i}}^{-}} \]
Given a short sequence $x$, does it come from CpG island (Yes-No question)?

Decision rule: if $S(x) > 0$ then CpG else non-Cpg
How to locate CpG islands?

- Given a DNA sequence, find the CpG islands in it, if any.

  Approach: Calculate the log-odds score for a window of \( w \) nucleotides around every base in the sequence. Predict as CpG islands, those with a positive log-odds score.

- Problem: What should the size of the window \( w \) be? Predictions are sensitive to choice of \( w \).
The occasionally dishonest casino

A casino uses a fair coin most of the time, but occasionally they switch to a loaded coin. You can’t see which coin they are using, just the results of the flips (heads and tails) are visible.

Hidden state

Observables
Generating coin flips

- Start in one of the states, F or L (i.e., pick a fair or loaded coin to start with) (initial probabilities).
- Move to the next state (F or L), based on the transition probabilities. Generate an h or t based on the emission probabilities of that state.
- Repeat above step.
Generating flips (take 2)

State sequence: FFFL  (unobserved)
Obs sequence : hht    (observed)
Hidden Markov Models

- $S = \{s_1, \ldots, s_N\}$, $N$ states
- $O = \{o_1, \ldots, o_M\}$, $M$ observation symbols
- $a_{ij} = P(S_{t+1}=s_j | S_t=s_i)$, $i,j$ in $[1..N]$; transition probabilities
- $b_i(k) = P(E_t=o_k | S_t=s_i)$, $k$ in $[1..M], i$ in $[1..N]$; emission probabilities
- $\pi_i = P(S_1=s_i)$, $i$ in $[1..N]$; initial state probabilities

$\lambda = (A, B, \pi)$ specifies the HMM model
Dishonest casino as an HMM

- $N = 2$, $S = \{F, L\}$
- $M = 2$, $O = \{h, t\}$
- $A = \begin{bmatrix} F & L \\ F & 0.95 & 0.05 \\ L & 0.10 & 0.90 \end{bmatrix}$
- $B = \begin{bmatrix} h & t \\ F & 0.5 & 0.5 \\ L & 0.1 & 0.9 \end{bmatrix}$
- $\pi = [1 \ 0]$