Chapter 4: Hidden Markov Models

4.1 Introduction to HMM

Overview

- Markov models of sequence structures
- Introduction to Hidden Markov Models (HMM)
- HMM algorithms; Viterbi decoder

Durbin chapters 3-5
The Challenges

- Biological sequences have modular structure
  - Genes → exons, introns
  - Promoter regions → modules, promoters
  - Proteins → domains, folds, structural parts, active parts

How do we identify informative regions?
- How do we find & map genes
- How do we find & map promoter regions

Mapping Protein Regions

- Kinase Domain
- ABL Model
- ATP Binding Domain
- SH2 Domain
- SH3 Domain
- SH2 Contact
- SH3 Contact
- Active components
- Interfaces
- Activation Domain
- ATP Binding Site
Statistical Sequence Analysis

- Example: CpG islands indicate important regions
  - CG (denoted CpG) is typically transformed by methylation into TG
  - Promoter/start regions of gene suppress methylation
  - This leads to higher CpG density
  - How do we find CpG islands?

- Example: active protein regions are statistically similar
  - Evolution conserves structural motifs but varies sequences

- Simple comparison techniques are insufficient
  - Global/local alignment
  - Consensus sequence

- The challenge: analyzing statistical features of regions

Review of Markovian Modeling

- Recall: a Markov chain is described by transition probabilities
  - $\pi(n+1)=A\pi(n)$ where $\pi(i,n)=\text{Prob}\{S(n)=i\}$ is the state probability
  - $A(i,j)=\text{Prob}\{S(n+1)=j|S(n)=i\}$ is the transition probability

- Markov chains describe statistical evolution
  - In time: evolutionary change depends on previous state only
  - In space: change depends on neighboring sites only
From Markov To Hidden Markov Models (HMM)

- Nature uses different statistics for evolving different regions
  - Gene regions: CpG, promoters, introns/exons...
  - Protein regions: active, interfaces, hydrophobic/philic...
- How can we tell regions?
  - Sample sequences have different statistics
  - Model regions as Markovian states emitting observed sequences...
- Example: CpG islands
  - Model: two connected MCs one for CpG one for normal
  - The MC is hidden; only sample sequences are seen
  - Detect transition to/from CpG MC
  - Similar to a dishonest casino: transition from fair to biased dice

Hidden Markov Models

- HMM Basics
  - A Markov Chain: states & transition probabilities $A=[a(i,j)]$
  - Observable symbols for each state $O(i)$
  - A probability $e(i,X)$ of emitting the symbol $X$ at state $i$
Coin Example

- Two states MC: \{F,B\} F=fair coin, B=biased
- Emission probabilities
  - Described in state boxes
  - Or through emission boxes
- Example: transmembrane proteins
  - Hydrophilic/hydrophobic regions

\[
e(F, H) = 0.5 \\
e(F, T) = 0.5 \\
e(B, H) = 0.9 \\
e(B, T) = 0.1
\]

HMM Profile Example (Non-gapped)

A State per Site
How Do We Model Gaps?

- Gap can result from “deletion” or “insertion”
  - Deletion = hidden delete state
  - Insertion = hidden insert state

Profile HMM

- Profile alignment
  - E.g., What is the most likely path to generate ACATATC?
  - How likely is ACATATC to be generated by this profile?
In General: HMM Sequence Profile

HMM For CpG Islands

<table>
<thead>
<tr>
<th>CpG generator</th>
<th>Regular Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ A G C T</td>
<td>+ A G C T</td>
</tr>
<tr>
<td>A 0.180 0.274 0.420 0.120</td>
<td>A 0.3 0.205 0.285 0.21</td>
</tr>
<tr>
<td>G 0.170 0.360 0.270 0.180</td>
<td>G 0.322 0.298 0.079 0.302</td>
</tr>
<tr>
<td>C 0.160 0.330 0.370 0.125</td>
<td>C 0.248 0.246 0.298 0.208</td>
</tr>
<tr>
<td>T 0.070 0.350 0.380 0.180</td>
<td>T 0.177 0.230 0.290 0.293</td>
</tr>
</tbody>
</table>
Modeling Gene Structure With HMM

- Genes are organized into sequential functional regions
- Regions have distinct statistical behaviors

HMM Gene Models

- HMM “state” $\rightarrow$ region; Markov transitions between regions
- Emission \{A,C,T,G\}; regions have different probabilities

\[\begin{align*}
A &= .29 \\
C &= .31 \\
G &= .04 \\
T &= .36
\end{align*}\]
Computing Probabilities on HMM

- Path = a sequence of states
  - E.g., X=FFBBBF
  - Path probability: 0.5(0.6)^2 0.4(0.2)^3 0.8 = 4.608 x 10^-4
- Probability of a sequence emitted by a path: p(S|X)
  - E.g., p(HHHHH|FFBBBF) = p(H|F)p(H|F)p(H|B)p(H|B)p(H|F) = (0.5)^3(0.9)^3 = 0.09
- Note: usually one avoids multiplications and computes logarithms to minimize error propagation

The Three Computational Problems of HMM

- Decoding: what is its most likely sequence of transitions & emissions that generated a given observed sequence?
- Likelihood: how likely is an observed sequence to have been generated by a given HMM?
- Learning: how should transition and emission probabilities be learned from observed sequences?
The Decoding Problem: Viterbi’s Decoder

- Input: an observed sequence $S$
- Output: a hidden path $X$ maximizing $P(S|X)$

**Key Idea (Viterbi):** map to a dynamic programming problem
- Describe the problem as optimizing a path over a grid
- DP search: (a) compute “price” of forward paths (b) backtrack
- Complexity: $O(m^2n)$ ($m=$number of states, $n=$ sequence size)

Viterbi’s Decoder

- $F(i,k) =$ probability of the most likely path to state $i$
generating $S_1…S_k$
- Forward recursion:
  \[
  F(i,k+1) = e(i,S_{k+1}) \cdot \max_j \left\{ F(j,k) a(i,j) \right\}
  \]
- Backtracking: start with highest $F(i,n)$ and backtrack
- Initialization: $F(0,0)=1, F(i,0)=0$
Example: Dishonest Coin Tossing

What is the most likely sequence of transitions & emissions to explain the observation: \( S = \text{HHHHHH} \)

\[
\begin{align*}
\Pr(F,H) &= 0.5 \\
\Pr(F,T) &= 0.5 \\
\Pr(B,H) &= 0.9 \\
\Pr(B,T) &= 0.1 \\
\end{align*}
\]

\[
\begin{array}{c|cccccc}
B & H & H & H & H & H & H \\
\hline
F & 1 & 0.25 & 0.18 & 0.05 & 0.02 & 0.02 \\
\hline
\text{Start} & 0 & 0.09 & 0.06 & 0.05 & 0.02 & 0.02 \\
\end{array}
\]

\[
\begin{align*}
\text{Total} &= 0.5 \times (0.9)^2 \times (0.3)^2 \times (0.36)^2 \\
\end{align*}
\]

Example: CpG Islands

Given: observed sequence CGCG what is the likely state sequence generating it?
Computational Note

- Computing probability products propagates errors
- Instead of multiplying probabilities add log-likelihood
- Define \( f(i,k) = \log F(i,k) \)

\[
f(i,k+1) = \log e(i,S_{k+1}) + \max_j \{ f(j,k) + \log a(i,j) \}
\]

- Or, define the weight \( w(i,j,k) = \log e(i,S_{k+1}) + \log a(i,j) \)

To get the following standard DP formulation

\[
f(i,k+1) = \max_j \{ f(j,k) + w(i,j,k) \}
\]

Example

- What is the most likely sequence of transitions & emissions to explain the observation: \( S = H H H H H H \)
- (using base 2 log)

\[
\begin{array}{c|c|c}
\text{B} & -2.32 & -0.15 = -2.47 \\
& -0.32 - 1 = -1.32 & \\
& -1.32 - 0.15 = -1.47 & \\
\text{F} & -0.74 - 1 = -1.74 & \\
\end{array}
\]

\[
\begin{array}{c|c|c|c|c|c}
\text{S} & \text{F} & \text{B} & \text{F} & \text{B} & \text{F} \\
-2 & -1.15 & & & & \\
-1.74 & -1.47 & & & & \\
-1.32 & -2.47 & & & & \\
\end{array}
\]

\[
f(i,k+1) = \max_j \{ f(j,k) + w(i,j,k) \}
\]
Concluding Notes

- Viterbi decoding: hidden pathway of an observed sequence
- Hidden pathway explains the underlying structure
  - E.g., identify CpG islands
  - E.g., align a sequence against a profile
  - E.g., determine gene structure
  - ...

- This leaves the two other HMM computational problems
  - How do we extract an HMM model, from observed sequences?
  - How do we compute the likelihood of a given sequence?