7.91 / 7.36 / BE.490
Lecture #4
Mar. 4, 2004

Markov & Hidden Markov Models
for DNA Sequence Analysis

Chris Burge
Organization of Topics

Lecture Object

3/2

Dependence Structure

3/4

Weight Matrix Model

Independence

Hidden Markov Model

Local Dependence

3/9

Energy Model, Covariation Model

Non-local Dependence
Markov & Hidden Markov Models for DNA

• Markov Models for splice sites
• Hidden Markov Models
  - looking under the hood
• The Viterbi Algorithm
• Real World HMMs

See Ch. 4 of Mount
Review of DNA Motif Modeling & Discovery

- WMMs for splice sites
- Information Content of a Motif
- The Motif Finding/Discovery Problem
- The Gibbs Sampler

The Gibbs Sampling Algorithm Multimedia Experience

- Motif Modeling - Beyond Weight Matrices

See Ch. 4 of Mount
Information Content of a DNA/RNA Motif

\[ f_k = \text{freq. of nt } k \text{ at position} \]

Shannon Entropy

\[ H(\mathbf{f}) = -\sum_k f_k \log_2(f_k) \text{ (bits)} \]

Information/position

\[ I(\mathbf{f}) = 2 - H(\mathbf{f}) = 2 + \sum_k f_k \log_2(f_k) = \sum_k f_k \log_2\left(\frac{f_k}{4}\right) \text{ (bits)} \]

Motif containing m bits of info. will occur approximately once per \( 2^m \) bases of random sequence
Variables Affecting Motif Finding

L = avg. sequence length
N = no. of sequences
I = info. content of motif
W = motif width
How is the 5′ss recognized?

U1 snRNA

3′ ………CCAUUCAUAG – 5′

Pre-mRNA 5′…………UUCGUGAGU…………… 3′
RNA Energetics I

Free energy of helix formation derives from:

- base pairing: $G \uparrow > A \uparrow > C \uparrow > U \downarrow$

- base stacking:

Doug Turner’s Energy Rules:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>-1.30</td>
</tr>
<tr>
<td>C</td>
<td>.</td>
<td>.</td>
<td>-2.40</td>
<td>.</td>
</tr>
<tr>
<td>G</td>
<td>.</td>
<td>-2.10</td>
<td>.</td>
<td>-1.00</td>
</tr>
<tr>
<td>T</td>
<td>-0.90</td>
<td>.</td>
<td>-1.30</td>
<td>.</td>
</tr>
</tbody>
</table>
RNA Energetics II

A) \[ \text{N \ p \ N \ p \ N \ p \ N \ p \ N \ p \ N \ p \ N} \]
\[ \text{X} \ | \ | \ | \ | \text{X} \text{X} \]
\[ \text{N \ p \ N \ p \ N \ p \ N \ p \ N \ p \ N \ p \ N} \]

Lots of consecutive base pairs - good

B) \[ \text{N \ p \ N \ p \ N \ p \ N \ p \ N \ p \ N} \]
\[ \text{X} \ | \ | \ | \text{X} \ | \ | \text{X} \]
\[ \text{N \ p \ N \ p \ N \ p \ N \ p \ N \ p \ N \ p \ N} \]

Internal loop - bad

C) \[ \text{N \ p \ N \ p \ N \ p \ N \ p \ N \ p \ N} \]
\[ \text{X} \ | \ | \ | \text{X} \ | \ | \text{X} \]
\[ \text{N \ p \ N \ p \ N \ p \ N \ p \ N \ p \ N \ p \ N} \]

Terminal base pair not stable - bad

Generally A will be more stable than B or C
Conditional Frequencies in 5’ss Sequences

5’ss which have G at +5

<table>
<thead>
<tr>
<th>Pos</th>
<th>-1</th>
<th>+3</th>
<th>+4</th>
<th>+6</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>9</td>
<td>44</td>
<td>75</td>
<td>14</td>
</tr>
<tr>
<td>C</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>18</td>
</tr>
<tr>
<td>G</td>
<td>78</td>
<td>51</td>
<td>13</td>
<td>19</td>
</tr>
<tr>
<td>T</td>
<td>9</td>
<td>3</td>
<td>9</td>
<td>49</td>
</tr>
</tbody>
</table>

5’ss which lack G at +5

<table>
<thead>
<tr>
<th>Pos</th>
<th>-1</th>
<th>+3</th>
<th>+4</th>
<th>+6</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2</td>
<td>81</td>
<td>51</td>
<td>22</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>3</td>
<td>28</td>
<td>20</td>
</tr>
<tr>
<td>G</td>
<td>97</td>
<td>15</td>
<td>9</td>
<td>30</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>2</td>
<td>12</td>
<td>28</td>
</tr>
</tbody>
</table>

Data from Burge, 1998 “Computational Methods in Molecular Biology”
What kind of model could incorporate interactions between positions?
A Markov Model
Terminology

Random Variable (RV):
A quantity which may assume any one of a set of values, each with a definite probability of occurrence

Examples: $X$ = the outcome of rolling a die

$P(X=1) = \frac{1}{6}$, $P(X=2) = \frac{1}{6}$, ... $P(X=6) = \frac{1}{6}$

The craps process: $X_1, X_2, X_3, ...$ successive dice rolls

Stochastic Process:
a random process
or a sequence of Random Variables
What is a *Markov* Model (aka *Markov* Chain)?

**Classical Definition**

A discrete stochastic process $X_1, X_2, X_3, \ldots$ which has the Markov property:

$$P(X_{n+1} = j \mid X_1=x_1, X_2=x_2, \ldots X_n=x_n) = P(X_{n+1} = j \mid X_n=x_n)$$

(for all $x_i$, all $j$, all $n$)

**In words:**

A random process which has the property that the future (next state) is conditionally independent of the past given the present (current state)

Markov - a Russian mathematician, ca. 1922
Inhomogeneous 1st-Order Markov Model

\[ P_{-2}(A | C) = \frac{N_{CA}^{(-3, -2)}}{N_{C}^{(-3)}} \]

\[ S = S_1 S_2 S_3 S_4 S_5 S_6 S_7 S_8 S_9 \]

\[ R = \frac{P(S|+)}{P(S|-)} = \frac{P_{-3}(S_1)P_{-2}(S_2 | S_1)P_{-1}(S_3 | S_2) \cdots P_6(S_9 | S_8)}{P_{bg}(S_1)P_{bg}(S_2 | S_1)P_{bg}(S_3 | S_2) \cdots P_{bg}(S_9 | S_8)} \]
Estimating Parameters for a Markov Model

$$P_{-2}(A | C) = \frac{N_{CA}^{(-3,-2)}}{N_{c}^{(-3)}}$$

What about longer-range dependence?

- k-order Markov model

$$\sim 4^{k+1}$$ parameters / position for Markov model of order k
\[ S = S_1 S_2 S_3 S_4 S_5 S_6 S_7 S_8 S_9 \]

\[
R = \frac{P(S|+)}{P(S|-)} = \frac{P_{-3}(S_1)P_{-2}(S_2 | S_1)P_{-1}(S_3 | S_2) \cdots P_6(S_9 | S_8)}{P_{bg}(S_1)P_{bg}(S_2 | S_1)P_{bg}(S_3 | S_2) \cdots P_{bg}(S_9 | S_8)}
\]

\[ s = \log_2 R \]
WMM vs 1st-order Markov Models of Human 5’ss

Decoy 5’ss

WMM

True 5’ss

WMM 5’ss Score

Markov

Decoy 5’ss

11M 5’ss Score

True 5’ss
Splicing Model I

5' splice site

branch site

3' splice site
Splicing Model II

(A) 5' splice site  Branch point  Polypyrimidine tract  3' splice site
A/CAG GURAGU YNYURAY Y 10-20 YAG

(B) Commitment complex
Complexes across intron
Exon definition complex

Pre-spliceosome complex
Spliceosome

ATP U2 snRNP
ATP U5, U4/U6 snRNPs
m7Gppp
CBC
SRp
ESE
U1 snRNP
U2AF65
SF1/BBP
U2AF55
pol II

Splicing Model I
Splicing Model II

(A)
A Recent Model of Human Pre-mRNA Splicing

ESEs are short motifs that enhance recognition of adjacent splice sites in both constitutive and alternatively spliced exons - precise sequence requirements not well characterized.
Hidden Markov Models

aka HMMs

A later development, developed in E. E. for applications to voice recognition
Markov Models

Sequence alignment

<table>
<thead>
<tr>
<th>N</th>
<th>F</th>
<th>L</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>F</td>
<td>L</td>
<td>S</td>
</tr>
<tr>
<td>N</td>
<td>K</td>
<td>Y</td>
<td>L</td>
</tr>
<tr>
<td>Q</td>
<td>W</td>
<td>-</td>
<td>T</td>
</tr>
</tbody>
</table>

PURPLE POSITION REPRESENTS ALIGNMENT IN COLUMN
ORANGE POSITION REPRESENTS INSERT IN COLUMN
GREEN POSITION REPRESENTS DELETE IN COLUMN

Hidden Markov model for sequence alignment

![Graph showing the hidden Markov model with states and transitions](image)

- **Match state**
- **Insert state**
- **Delete state**
- **Transition probability**

*Courtesy of M. Yaffe*
Markov and Hidden Markov Models

Hidden

Genome

Island

Observable

Hidden Markov
CpG Islands
CpG Island Hidden Markov Model

- $P_{gg} = 0.99999$
- $P_{ig} = 0.001$
- $P_{ii} = 0.999$
- $P_{gi} = 0.00001$

### CpG Island Values
- C: 0.3
- G: 0.3
- A: 0.2
- T: 0.2

### Genome Values
- A: 0.2
- C: 0.2
- G: 0.3
- T: 0.3
CpG Island HMM II

- Transition probabilities:
  - Genome: \( P_{gg} = 0.99999 \)
  - Island: \( P_{ii} = 0.999 \)

- Emission Probabilities:
  - CpG Island: [0.3, 0.3, 0.2, 0.2]
  - Genome: [0.2, 0.2, 0.3, 0.3]
Want to infer

Observe

But HMM is written in the other direction (observable depends on hidden)
Inferring the Hidden from the Observable (Bayes’ Rule)

\[ P(H = h_1, h_2, ..., h_n \mid O = o_1, o_2, ..., o_n) \]

\[ = \frac{P(H = h_1, ..., h_n, O = o_1, ..., o_n)}{P(O = o_1, ..., o_n)} \]

\[ = \frac{P(H = h_1, ..., h_n)P(O = o_1, ..., o_n \mid H = h_1, ..., h_n)}{P(O = o_1, ..., o_n)} \]

\( P(O = o_1, ..., o_n) \) somewhat difficult to calculate

But notice:

\[ P(H = h_1, ..., h_n, O = o_1, ..., o_n) > P(H = h'_1, ..., h'_n, O = o_1, ..., o_n) \]

implies \( P(H = h_1, ..., h_n \mid O = o_1, ..., o_n) > P(H = h'_1, ..., h'_n \mid O = o_1, ..., o_n) \)

so can treat \( P(O = o_1, ..., o_n) \) as a constant
Finding the Optimal “Parse”
(Viterbi Algorithm)

Want to find sequence of hidden states \( H^{opt} = h_1^{opt}, h_2^{opt}, h_3^{opt}, \ldots \) which maximizes joint probability:

\[
P(H = h_1, \ldots, h_n, O = o_1, \ldots, o_n)
\]

(optimal “parse” of sequence)

Solution:

Define \( R_i^{(h)} = \) probability of optimal parse of the subsequence 1..i ending in state h

Solve recursively, i.e. determine \( R_2^{(h)} \) in terms of \( R_1^{(h)} \), etc.

A. Viterbi, an MIT BS/Meng student in E.E. - founder of Qualcomm
Run time for k-state HMM on sequence of length L?