Markov Models & DNA Sequence Evolution

Chris Burge
Review of Markov & HMM Models for DNA

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

Ch. 4 of Mount
CpG Islands

% C+G
CpG Island Hidden Markov Model

\[ P_{gg} = 0.99999 \]
\[ P_{ig} = 0.001 \]
\[ P_{ii} = 0.999 \]
\[ P_{gi} = 0.00001 \]

CpG Island: 0.3 0.3 0.2 0.2
Genome: 0.2 0.2 0.3 0.3
CpG Island HMM II

P_{gg} = 0.99999  \quad P_{ig} = 0.001

Island

P_{ii} = 0.999

P_{gi} = 0.00001

“Transition probabilities”

“Emission Probabilities”

CpG Island:

C \quad G \quad A \quad T

0.3 \quad 0.3 \quad 0.2 \quad 0.2

Genome:

0.2 \quad 0.2 \quad 0.3 \quad 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)

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

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

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

\[ P(O = o_1, ..., o_n) \text{ somewhat difficult to calculate} \]

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

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
“Trellis” Diagram for Viterbi Algorithm

Position in Sequence →

1 ... i i+1 i+2 i+3 i+4 ... L

Hidden States →

I ... A T C G C ... A

Run time for k-state HMM on sequence of length L?
Viterbi Algorithm Examples

What is the optimal parse of the sequence:

• (ACGT)_{10000}

• A_{1000}C_{80}T_{1000}C_{40}A_{1000}G_{60}T_{1000}

Powers of 1.5:

\[
N = 20 \quad 40 \quad 60 \quad 80 \\
(1.5)^N = 3 \times 10^3 \quad 1 \times 10^7 \quad 3 \times 10^{10} \quad 1 \times 10^{14}
\]
What else can you model with HMMs?

Bacterial gene:

Open Reading Frame

Start

Stop
Useful HMMs developed
Parameter Estimation for HMMs

How many parameters for a $k$-state HMM over an alphabet of size 4?

Initial probabilities:
Transition probabilities:
Emission probabilities:
Pseudocounts

Courtesy of M. Yaffe

• If the number of sequences in the training set is both large and diverse, then the sequences in the training set represent a good statistical sampling of the motif....if not, then we have a sampling error!

Correct for this by adding pseudocounts. How many to add?

→ Too many pseudocounts dominate the frequencies... and the resulting matrix won’t work!

→ Too few pseudocounts then we’ll miss many amino acid variations, and matrix will only find sequences that produced the motif!

Add few pseudocounts if sampling is good (robust), and add more pseudocounts if sampling is sparse.

One reasonable approach is to add \( \sqrt{N} \) pseudocounts, where N is the number of sequences...

As N increases, the influence of pseudocounts decreases since N increases faster than \( \sqrt{N} \), but doesn’t add enough at low N.
Dealing With Small Training Sets

<table>
<thead>
<tr>
<th>Position</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Training Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ACCTG</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AGCTG</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ACCCG</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ACCCTG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ACCCCA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GACTG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ACGTA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ACCCTG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CCCCCG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ACATAC</td>
</tr>
</tbody>
</table>

If the true frequency of T at pos. 1 was 10%, what’s the probability we wouldn’t see any Ts in a sample of 10 seqs?

\[
P(N=0) = \frac{10!}{0!10!}(0.1)^0(0.9)^{10} = \sim 35\%\]

So we should add pseudocounts
## Pseudocounts (Ψ counts)

<table>
<thead>
<tr>
<th>Nt</th>
<th>Count</th>
<th>Ψ_count</th>
<th>Bayes_count</th>
<th>ML est.</th>
<th>Bayes est.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>8</td>
<td>+ 1</td>
<td>9</td>
<td>0.80</td>
<td>0.64</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>+ 1</td>
<td>2</td>
<td>0.10</td>
<td>0.14</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>+ 1</td>
<td>2</td>
<td>0.10</td>
<td>0.14</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>+ 1</td>
<td>1</td>
<td>0.00</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td></td>
<td>14</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

The ‘add 1 to each observed count’ rule can be derived analytically from the Bayesian posterior distribution under a Dirichlet prior - see Appendix A of statistics primer for details.
Real World HMMs
Please see the following Web site:  http://www.cbs.dtu.dk/services/TMHMM/

Architecture of TMHMM

Please see figures 1a and 1c of:

Optimal Parse

TMHMM Output for Mouse Chloride Channel CLC6

Graph showing posterior probability along a sequence with segments for transmembrane, inside, and outside regions.
Structure of a Typical Human Gene

5–10 Coding Exons

Promoter
ATG
5’ss
3’ss
Stop
PolyA Signal

5’ UTR

3’ UTR
Genscan Model

Incorporates:

Transcriptional signals
Splicing signals
Translational signals
Composition of exons
Composition of introns
Other gene features

Burge & Karlin, J Mol Biol 1997
Semi-Markov HMM Model
Genscan predictions in human CD4 gene region

Overall: ~75% of exons exactly correct

Burge and Karlin *J. Mol. Biol.* 1997
Genscan, GenomeScan Predictions in Human BRCA1 Region

Please see figures 1 of

DNA Sequence Evolution

**Generation n-1 (grandparent)**

5’ TGGCATGCACCCCTGTAAGTCAATATAAATGGCTACGCCTAGCCCATGCGA 3’
3’ ACCGTACGTGGGACATTCAGTTATATTTACCAGATCGGATCGGGTACGCT 5’

**Generation n (parent)**

5’ TGGCATGCACCCTGTAAGTCAATATAAATGGCTATGCCTAGCCCATGCGA 3’
3’ ACCGTACGTGGGACATTCAGTTATATTTACCAGATCGGATCGGGTACGCT 5’

**Generation n+1 (child)**

5’ TGGCATGCACCCCTGTAAGTCAATATAAATGGCTATGCCTAGCCCATGCGA 3’
3’ ACCGTACGTGGGACATTCAGTTATATTTACCAGATCGGATCGGGTACGCT 5’
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
DNA Sequence Evolution is a Markov Process

No selection case

\( S_n = \text{base at generation } n \)

\( P_{ij} = P(S_{n+1} = j \mid S_n = i) \)

\( \vec{q}^n = (q_A, q_C, q_G, q_T) = \text{vector of prob's of bases at gen. } n \)

Handy relations:

\( \vec{q}^{n+1} = \vec{q}^n P \)

\( \vec{q}^{n+k} = \vec{q}^n P^k \)
Limit Theorem for Markov Chains

\[ S_n = \text{base at generation } n \quad P_{ij} = P(S_{n+1} = j \mid S_n = i) \]

If \( P_{ij} > 0 \) for all \( i, j \) (and \( \sum_j P_{ij} = 1 \) for all \( i \))

then there is a unique vector \( \vec{r} \) such that

\[ \vec{r} = \vec{r}P \quad \text{and} \quad \lim_{n \to \infty} \vec{q} P^n = \vec{r} \quad (\text{for any prob. vector } \vec{q}) \]

\( \vec{r} \) is called the “stationary” or “limiting” distribution of \( P \)

See Ch. 4, Taylor & Karlin, An Introduction to Stochastic Modeling, 1984 for details
Stationary Distribution Examples

2-letter alphabet: R = purine, Y = pyrimidine

Stationary distributions for:

\[
I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad Q = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}
\]

\[
P = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix} \quad 0 < p < 1
\]

\[
P' = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix} \quad 0 < p < 1, \ 0 < q < 1
\]
How does entropy change when a Markov transition matrix is applied?

If limiting distribution is uniform, then entropy increases
  (analogous to 2nd Law of Thermodynamics)

However, this is not true in general (why not?)
How rapidly is the stationary distribution approached?
Assume each nucleotide equally likely to change into any other nt, with rate of change = $\alpha$. Overall rate of substitution = $3\alpha$.

...so if G at $t=0$, at $t=1$, $P_{G(1)} = 1 - 3\alpha$

and $P_{G(2)} = (1 - 3\alpha)P_{G(1)} + \alpha [1 - P_{G(1)}]$.

Expanding this gives $P_{G(t)} = 1/4 + (3/4)e^{-4\alpha t}$.

Can show that this gives $K = -3/4 \ln[1-(4/3)(p)]$

$K$ = true number of substitutions that have occurred, $P$ = fraction of nt that differ by a simple count.

Captures general behaviour...
Literature Discussion Tues. 3/16

Paper #1:


Part 1 - Finding Genes, etc., pp. 241-247
Part 2 - Regulatory Elements, pp. 247-254

Paper #2: