CS 2427 - Algorithms in Molecular Biology

Lecture #5: 25 January 2006

Lecturer: Michael Brudno

Scribe Notes by: Midori Hyndman

1 Today's Topics:

- DNA Mutation
- CpG Islands
- Markov Models
- Hidden Markov Models

2 DNA Mutation

Last class we discussed the UPGMA, a greedy algorithm for building tree-graph representations of sequence evolution. When sequences segments mutate at an uneven rate, the results of this approach were shown to be inadequate.

The rate of mutations in a population is related to the length of time between generations. For example, the rate of mutation for mice is higher that the rate for humans. And, during any given time period, a population of mice will have significantly more generations than will a human population. Given more opportunities for mutations to develop within the population, the mouse genome is more likely to evolve at a faster rate. Following the same reasoning, we would expect to find more mutations among larger populations.

Mutations are generally non-uniformly distributed across the DNA sequence. This is largely due to the fact that genes are under selection. Organisms which are weakened by mutations are less likely to live long enough to reproduce. For this reason, we find fewer mutations in functional regions of the genome, which affect protein synthesis and cell development. In fact, a lack of mutations is used as an indicator for regions of interest: if there are fewer mutations than the neutral estimate in a particular region, there is a high chance that something of interest is going on. The *Neutral Estimate* is the proportion of mutations which we would expect to see in a DNA sequence; the value of this estimate is debated within the community.

One common mutation occurs among neighbouring C (cytosine) and G (guanine) bases on a DNA strand. The CG dinucleotide is typically written as CpG, where the p corresponds to the phosphate molecule which bonds bases on the same strand. This notation distinguishes CpG dinucleotides from C-G base pairs which occurs across two strands of DNA.

2.1 Methylation

When the C (cytosine) occurs in a CpG pair, it has a high chance of being chemically modified by a process known as methylation. Methylation is the replacement of an of a hydrogen atom (H) with a methyl group (CH_3). Accidental or spontaneous deamination converts methyl-C into thymine T. When this reaction occurs in DNA, the mutations tend to persist if the repair mechanisms do not recognize thymine as erroneous, that is unless it affects the function of the gene. This characteristic of the DNA repair mechanism contributes to the lack of CpG dinucleotides in non-functional regions of the genome.

3 CpG islands

In certain stretches of the genome the methylation process is suppressed and we find a higher density of CpG pairs. These regions of the DNA strand are called CpG islands. These segments occur around the promoters of genes. They are typically between a few hundred to a few thousand bases in length. CpG islands control whether or not the gene gets expressed in transcription, that is whether the cell will have a copy of the gene. If the CpG island is no longer present the transcription process could be disrupted and the gene might always be activated or never activated in the transcribed cell.

It is important to note that in CpG islands, the CpG dinucleotides do not dominate the sequence, but there are more than would be expected when looking at the total distribution of CpG nucleotides in the genome. Overall, C and G bases are rarer in the genome than A and T. The distributions are specific to the organism and are heavily influenced by the environmental factors. In the human genome we generally find the following distribution of nucleotides,

$$C/G = 40\% \qquad A/T = 60\%$$

CpG islands are of particular interest, because knowing their location within a DNA strand, assists in finding the promoters of genes. The following two questions motivated the remainder of this discussion:

1. Given a short sequence of DNA, how do we determine whether it comes from a CpG island?
2. Given a long sequence of DNA, if CpG islands exist within the sequence, how do we locate them?

4 Markov Models

Markov Models are used to address the first of the two questions presented above. The approach involves developing two probabilistic generative models, one for CpG islands and one for non-CpG

Islands. Given a segment of DNA, the goal is to determine which of the two models is more likely to have generated the segment.

There are two reasons why Markov Models are a natural choice for DNA segments. A key property of the first-order Markov model is that the probability of a symbol depends only on the previous symbol, and is independent of the symbols found earlier in the sequence. Since we are scanning the sequences for pairs of bases, CpG pairs in particular, therefore, the relative importance of the any nucleotide in the sequence depends the nucleotide immediately preceding it. A second important property of the Markov model is time independence. The specific location of the CpG pairs in the segment is not relevant, we only need to know that they are present in the segment.

As shown in Figure 1, Markov Model can be represented by a directed graph where the states are nodes and the edges are labeled with the transitional probability. In this case, the states of the system are the four letters in the DNA alphabet and the edges represent the probability that a certain base will follow another base in the segment.

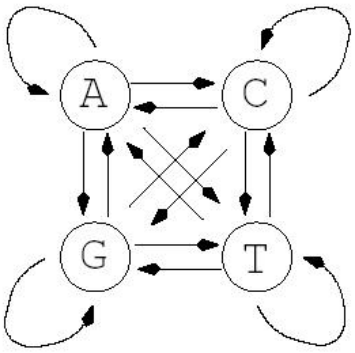

Figure 1: Markov Model represented as a directed graph. The states correspond to the symbols from the DNA alphabet and the edges represent the transition to the next nucleotide in the DNA segment. Generally the edges are labeled with transition probabilities.

We add two additional states to our sequence to represent the beginning (b) and end (e) of the segment. The graph is shown in Figure 2. The transitions from the beginning state to each of the nucleotides are assigned equally, in this case. Adjusting the transition matrices to accommodate the ending state, is a straightforward process.

4.1 Classifying CpG Islands

To determine whether a segment belongs to a CpG island, we begin by developing two models which capture the different the distributions of CpG pairs: one for CpG islands and one for non-CpG Islands. Within CpG islands, the probability that a C will follow a G will be much higher than it will in other areas of the genome. Assuming we have human DNA where the putative CpG islands have been labeled. Using this we can derive the Markov models: one to describe transition probabilities within the CpG islands (M_{CpG}), and a second to describe transition probabilities in the remainder of the sequence (M_{GEN}).

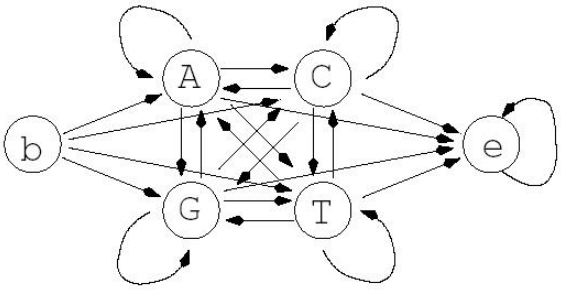

Figure 2: Markov Model represented as a directed graph augmented with two silent states corresponding to the beginning (b) and the end (e) of the DNA segment.

Instead of using the graphical model, we can represent transitional probabilities in matrix form. In Figure 3 we show the transition probability matrix for a CpG islands segment, and in Figure 4 is the transition probability matrix for a non-CpG island.[1] The row are ordered A, C, G, T from top to bottom and the columns are ordered similarly from left to right. For example, in the second row the numbers indicate the frequency with which a C is followed by each of the four nucleotides. Note the higher transition probability for C-¿G in the CpG island transition matrix. The probabilities sum to one along the rows, since at each time step we move to the next nucleotide in the DNA segment. On a side note, transitional probabilities for DNA sequences are strongly organism related. For a general idea of the scale of difference, the probabilities for humans are similar to those for mice, but are quite different from those of flies.

```
# Transition matrix:
.1795 .2735 .4255 .1195
.1705 .3665 .2735 .1875
.1605 .3385 .3745 .1245
.0785 .3545 .3835 .1815
.2495 .2495 .2495 .2495
.0000 .0000 .0000 .0000
```

Figure 3: The transition probability matrix for a CpG islands segment. The row are ordered A, C, G, T from top to bottom and the columns are ordered similarly from left to right.

Given a DNA segment and Markov models for CpG islands and non-CpG islands, we can estimate which of the two models is more likely to have generated the sequence. To do so we shift to probabilistic notation.

Let $X = x_1 x_2 \ldots x_L$ be the sequence of states which we will classify. We would like to determine the probability that a Markov chain, $s_1, s_2, \ldots, s_L$ will pass through the given states. First a

[1] These transition matrices were extracted from *Biological sequence analysis* by R. Durbin, s. Eddy, A. Krogh, and G. Mitchison (1998).

```
# Transition matrix:
.2995 .2045 .2845 .2095
.3215 .2975 .0775 .3015
.2475 .2455 .2975 .2075
.1765 .2385 .2915 .2915
.2495 .2495 .2495 .2495
.0000 .0000 .0000 .0000
```

Figure 4: The transition probability matrix for a non-CpG islands segment. The row are ordered A, C, G, T from top to bottom and the columns are ordered similarly from left to right.

reminder of the two fundamental properties of Markov chains.

Transition probabilities are independent of time:

$$P(s_L = j | s_{L-1} = k) \quad = \quad a_{jk}$$

Transition probabilities have a one-step memory:

$$P(s_L = x_L | s_1 = x_1, s_2 = x_2, \ldots, s_{L-1} = x_{L-1}) \quad = \quad P(s_L = x_L | s_{L-1} = x_{L-1})$$

The probability of observing a sequence $X = x_1 x_2 \ldots x_L$, given the Markov model M is described as,

$$\begin{aligned} P(X|M) &= P(s_1 = x_1, s_2 = x_2, \ldots, s_L = x_L) \\ &= P(s_L = x_L | s_{L-1} = x_{L-1}) \ldots P(s_2 = x_2 | s_1 = x_1) P(s_1 = x_1) \\ &= P(s_1 = x_1) \prod_{i=2}^{L} a_{x_{i-1} x_i} \end{aligned}$$

To avoid inhomogeneity in the equation, we incorporate the beginning states, b, with the appropriate transition probabilities $P(s_1 = t) = a_{0t}$. This simplifies the notation to the following,

$$P(X|M) \quad = \quad \prod_{i=1}^{L} a_{x_{i-1} x_i}$$

The log-odds ratio, $S(X)$, is used to determine which of the two models is more likely to have generated the given segment X, (M_{CpG}) or (M_{GEN}). We classify sequences as CpG islands if the log odds are higher than some predetermined threshold.

$$S(x) \quad = \quad \frac{log P(x|M_{CpG})}{P(x|M_{GEN})}$$

4.2 Expectation Maximization

If we do not have complete confidence the initial data used to create our Markov models, we can implement an iterative approach to improve the classification. Given a set of DNA sequences and a general idea of which of the sequences are CpG islands, based on biological intuition, we use this to derive two approximation, $\widetilde{M}_{CpG}$ and $\widetilde{M}_{GEN}$, which are used for initialization. We proceed in two steps:

1. Using $\widetilde{M}_{CpG}$ and $\widetilde{M}_{GEN}$, we calculate the log-odds and classify the each of the sequences from our set.
2. Using the new classification results, we update our models, $\widetilde{M}_{CpG}$ and $\widetilde{M}_{GEN}$

Repeating this process, we are guaranteed to converge to a local maximum likelihood classification. This general approach is known as EM (Expectation Maximization).

5 Hidden Markov Models

Hidden Markov Models (HMMs) are used to approach the second question presented earlier in the lecture. Given a long sequence of DNA, we would like to find all the existing CpG islands within the sequence.

A naive approach involves sliding a window across the sequence and evaluating the log-odds at every location. Unfortunately, this simple approach is inadequate. The appropriate size for the window is a difficult determine and, more importantly, the boundaries of CpG Islands are not fixed in length.

A more specific approach involves HMMs. In this example, we can think of an HMM as merging the two Markov models from the previous section, M_{CpG} and M_{GEN}. A graphical representation is illustrated in Figure 5. As we read the states of the sequence we either transition to a symbol within the CpG Island or to a symbol outside of the CpG Island.

A more formal representation incorporates the probabilities between the groups of states into the eight-state model, shown in Figure 6. The state A^+ corresponds to being in a CpG island and reading A, whereas, A^- corresponds to being outside a CpG island and reading A. The edges shown for the state A^+ exist for all states in the model, but are not illustrated for the sake of clarity. The transition probabilities are shown in the matrix in Figure 7.

5.1 Emission Probabilities

Since the states of the system no longer uniquely identify the DNA nucleotides we need to introduce a second layer of states to our model (a, c, t, g). These are the observed states; they correspond to the symbols we will read from the sequence. Since an HMM is a generative model, these are the known states emitted from the hidden states. The edges which connect this second layer of states to the eight hidden states are labeled with emission probabilities. In this example, the emission probabilities are quite simple. For example, if we are in the state A^+, the corresponding

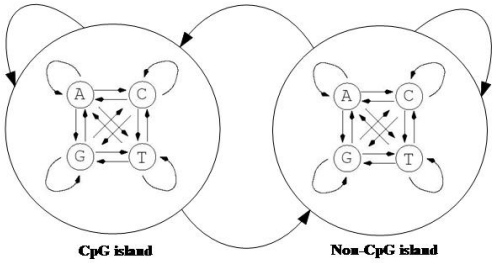

Figure 5: An eight-state Hidden Markov Model. The graph is fully connected, however, for simplicity only the edges from state A^+ are shown. The hidden states which correspond to being in a CpG island are on the left, and those which correspond to being outside a CpG island are on the right.

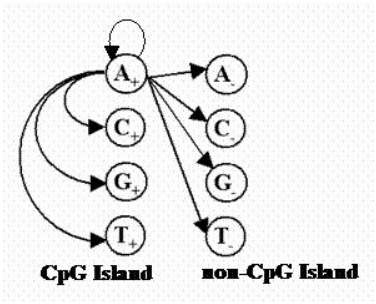

Figure 6: An eight-state Hidden Markov Model. The graph is fully connected, however, for simplicity only the edges from state A^+ are shown. The hidden states which correspond to being in a CpG island are on the left, and those which correspond to being outside a CpG island are on the right.

DNA sequence will have an C. therefore we will be in known state c with a probability of one, and a, g, and t with a probability of zero. The transition matrix is shown below in Figure 8.

If we read an A in our DNA segment we know we are in state a, however, we would like to determine whether we are in the hidden state A^+ or A^-. This process is known as decoding. Given our DNA segment, there are multiple paths through the HMM which could correspond to this segment. For example, the paths $A^+C^-C^+$, $A^+C^-C^-$ and $A^-C^-C^-$ would all generate the segment $ACTGT$.

In the following lecture we will discuss an algorithm for finding the most probably path through a HMM, the Viterbi Algorithm, and the probability that a HMM model generated a specified path.

```
# Transition matrix, probability to change from +island to -island (and vice versa) is 10E-4
#     0         A+        C+        G+        T+         A-        C-        G-        T-
0   0.0000000 0.0725193 0.1637630 0.1788242 0.0754545  0.1322050 0.1267006 0.1226380 0.1278950
A+  0.0010000 0.1762237 0.2682517 0.4170629 0.1174825  0.0035964 0.0054745 0.0085104 0.0023976
C+  0.0010000 0.1672435 0.3599201 0.2679840 0.1838722  0.0034131 0.0073453 0.0054690 0.0037524
G+  0.0010000 0.1576223 0.3318881 0.3671328 0.1223776  0.0032167 0.0067732 0.0074915 0.0024975
T+  0.0010000 0.0773426 0.3475514 0.3759440 0.1781818  0.0015784 0.0070929 0.0076723 0.0036363
A-  0.0010000 0.0002997 0.0002047 0.0002837 0.0002097  0.2994005 0.2045904 0.2844305 0.2095804
C-  0.0010000 0.0003216 0.0002977 0.0000769 0.0003016  0.3213566 0.2974045 0.0778441 0.3013966
G-  0.0010000 0.0001768 0.0002387 0.0002917 0.0002917  0.1766463 0.2385224 0.2914165 0.2914155
T-  0.0010000 0.0002477 0.0002457 0.0002977 0.0002077  0.2475044 0.2455084 0.2974035 0.2075844
```

Figure 7: The transition probability matrix for the HMM hidden states in the HMM.

```
# Emission probabilities:
#   a c g t
0   0 0 0 0
A+  1 0 0 0
C+  0 1 0 0
G+  0 0 1 0
T+  0 0 0 1
A-  1 0 0 0
C-  0 1 0 0
G-  0 0 1 0
T-  0 0 0 1
```

Figure 8: The emission probabilities for the observed states in the HMM.