1. Example - TATA boxes in E. coli

1.1. Finding motifs

An important factor of gene regulation is regulation of transcription. Transcription is guided by promoter sequences upstream of each gene. Two important regions of the E. coli promoter are located about 10 and 35 bases upstream of the transcription start site; each of these regions is about 6 bases in length. The sequences found in the -10 and -35 regions are different from promoter to promoter, but are always only slight variants of each other. The most commonly occurring sequence in each of these regions is known as the consensus sequence; promoters with -10 and -35 regions of high similarity to the consensus sequences are stronger promoters than promoters with -10 and -35 regions of weak similarity. The consensus sequence of the -10 region is TATAAT, and the consensus for the -35 region is TTGACA. The -10 region is known as the TATA box.

Suppose we had a set of TATA box sequences and we wanted to know if we could construct a motif from them. Data is shown in Table 1.

We want to define a model $\theta_{\text{motif}}$ with parameters $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6)$ that will assign probabilities for finding any particular base at each position in the TATA box.

$$\theta_{\text{motif}} = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6)$$ (1)

Technically, the -10 region in E. coli is called the Pribnow box. The TATA box is the eukaryotic equivalent, located about 25 bases upstream of eukaryotic transcription start sites.
Estimate parameters using the Maximum Likelihood Estimate (MLE); i.e., define each $\theta_i(B)$ to be equal to the frequency of base B in the $i^{th}$ position of the TATA boxes in the data set. The model for our data set is shown in Table 2.

<table>
<thead>
<tr>
<th>Table 1. TATA Box Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>TATGCT</td>
</tr>
<tr>
<td>TATAAT</td>
</tr>
<tr>
<td>GATAAT</td>
</tr>
<tr>
<td>CATGTT</td>
</tr>
<tr>
<td>TATACCT</td>
</tr>
</tbody>
</table>

Here $\theta_4(A) = 0.6$ is the probability that the fourth base is adenine in a real TATA box (given our data). Note the values in each column are normalized to sum to 1; at any position in a TATA box, the base must be either A, C, G, or T.

Now consider any instance of 6 bases of DNA. Denote the generic sequence as follows:

$$x = x_1x_2x_3x_4x_5x_6$$

(2)

Now that we have a model, we can calculate the probability of any particular 6 base-pair sequence of DNA using this model. To do so we will assume that the bases occurring at each position are statistically independent; i.e., each base has no influence over what its neighboring bases can be. This assumption is not necessarily biologically correct, but it is convenient and it allows us to find the probability of the entire sequence by multiplying the probabilities at each position in the sequence.

The probability that a given sequence is a TATA box using our model.
can be written symbolically as follows:

\[ P(x|\theta^{motif}) = P(x_1 x_2 x_3 x_4 x_5 x_6|\theta^{motif}) = \prod_{j=1}^{6} P(x_j|\theta_j^{motif}) \]  

(3)

For the consensus sequence,

\[ P(TATAAT|\theta^{motif}) = 0.6 \times 1.0 \times 1.0 \times 0.6 \times 0.4 \times 1.0 = 0.144 \]  

(4)

Thus we would expect to see the exact consensus TATA box about 14% of the time, but the rest of the time we would expect to see variants. This is in keeping with our data.

1.2. Bayesian Reasoning

There are about 5 million bases and 5000 genes in the E. coli genome. Thus, on average we expect there to be 1 gene in every 1000 bases of E. coli DNA.

Suppose you found a T in the E. coli genome. What is the chance that base is the start of a TATA box?

First we need to make some assumptions:

(1) Our TATA box model is exactly right.
(2) All bases occur equally frequently in non-TATA box DNA
(3) All bases are independent of each other.

The second assumption is used to construct a background model for DNA. We need a background model so we can distinguish what looks like a TATA box from what just looks like regular DNA sequence. Thus with the second assumption our background model is defined as follows:

\[ P(x_j|\theta^{bg}) = \frac{1}{4} \quad x_j \in \{A, C, G, T\} \]  

(5)

Here “bg” is used to indicate the background model. Then the probability of the consensus TATA sequence occurring at random is as follows:

\[ P(TATAAT|\theta^{bg}) = \left(\frac{1}{4}\right)^6 \]  

(6)
1.2.1. Bayes’ Rule

Bayes’ Rule states that for events A and B,

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$ (7)

This rule arises from the definition of joint and conditional probabilities,

$$P(A, B) = P(B|A)P(A) = P(A|B)P(B),$$ (8)

from which Bayes’ Rule follows easily. Here the notation $P(A, B)$ is used to signify the joint probability of events A and B.

From the models $\theta^{\text{motif}}$ and $\theta^{\text{bg}}$ we know

$$P(x_1 = T|m = \theta^{\text{motif}}) = 0.6, \quad (9)$$

and

$$P(x_1 = T|m = \theta^{\text{bg}}) = 0.25, \quad (10)$$

but now we want to go the other way. Specifically, we want to find $P(m = \theta^{\text{motif}}|x_1 = T)$. Using Bayes’ Rule:

$$P(m = \theta^{\text{motif}}|x_1 = T) = \frac{P(x_1 = T|m = \theta^{\text{motif}})P(m = \theta^{\text{motif}})}{P(x_1 = T)}$$ (11)

Note that we are missing two values on the right side of the equation; namely, $P(m = \theta^{\text{motif}})$ (called the prior) and $P(x_1 = T)$ (a boring normalization constant).

Let’s try to find the prior: we know that there are about 5,000 genes and 5,000,000 bases in the E.coli genome. Then

$$P(m = \theta^{\text{motif}}) = \frac{\# \text{TATAs}}{\text{genome size}} = \frac{5,000}{5,000,000} = \frac{1}{1,000} \quad (12)$$

Then it follows that

$$P(m = \theta^{\text{bg}}) = 1 - 0.001 = 0.999 \quad (13)$$

Now to find $P(x_1 = T)$. For any two events A and B,
\[ P(A) = P(A, B) + P(A, \bar{B}) \] (14)

Here \( \bar{B} \) is used to denote the event that \( B \) does not happen.

We can use the definition of conditional probability to expand the right-hand side of the last equation:

\[ P(A) = P(A|B)P(B) + P(A|\bar{B})P(\bar{B}) \] (15)

Now we can use the above equation to find \( P(x_1 = T) \):

\[
P(x_1 = T) = P(x_1 = T|\theta_{motif}^j)P(\theta_{motif}^j) + P(x_1 = T|\theta_{bg}^j)P(\theta_{bg}^j)
\]
\[
= 0.6 \times 0.001 + 0.25 \times 0.999
\]
\[
\approx 0.25
\]

Thus it follows that

\[
P(m = \theta_{motif}^j|x_1 = T) \approx \frac{0.6 \times 0.001}{0.25} \approx 0.00024.
\]

This agrees with our intuition that simply seeing at “T” in the E. coli genome isn’t strong enough evidence for us to conclude that a TATA box starts there. Next we will look at how more evidence accumulates as we see more matching bases. In practice, we also need to be more careful about what we use for the background model.

1.2.2. Log Odds, Log Prior Odds, and Log Posterior Odds

Think about finding instances of motifs in terms of scores for the whole motif instance. For each position \( j \) in the motif, define a score as follows:

\[
s_j(x_j) = \log_2 \frac{P(x_j|\theta_{motif}^j)}{P(x_j|\theta_{bg}^j)} \quad 1 \leq j \leq 6 \quad (16)
\]

Then the score of a given 6 base-pair sequence would be the sum of the scores at each position:

\[
S = \sum_{j=1}^{K=6} s_j(x_j) = \sum_{j=1}^{K=6} \log_2 \frac{P(x_j|\theta_{motif}^j)}{P(x_j|\theta_{bg}^j)} \quad (17)
\]
We can use the fact that the sum of logs is equal to the log of products to rewrite the last equation:

\[
S = \sum_{j=1}^{K=6} \log_2 \frac{P(x_j|\theta_j^{\text{motif}})}{P(x_j|\theta_j^{\text{bg}})} \tag{18}
\]

\[
= \log_2 \frac{\prod_{j=1}^{K=6} P(x_j|\theta_j^{\text{motif}})}{\prod_{j=1}^{K=6} P(x_j|\theta_j^{\text{bg}})} \tag{19}
\]

\[
= \log_2 \left( \frac{P(x|\theta^{\text{motif}})}{P(x|\theta^{\text{bg}})} \right) \tag{20}
\]

The final expression on the right is known as the log likelihood ratio, or simply log odds. The ratio without the logarithm is known as the likelihood ratio or simply odds.

The log prior odds is defined similarly:

\[
\text{log prior odds} = \log_2 \frac{P(\theta^{\text{motif}})}{P(\theta^{\text{bg}})} \tag{21}
\]

Finally, the log posterior odds is defined as follows:

\[
\text{log posterior odds} = \log_2 \frac{P(\theta^{\text{motif}}|x)}{P(\theta^{\text{bg}}|x)} \tag{22}
\]

There are better than even odds that \( x \) is a real TATA box when

\[
P(\theta^{\text{motif}}|x) > P(\theta^{\text{bg}}|x) \iff \frac{P(\theta^{\text{motif}}|x)}{P(\theta^{\text{bg}}|x)} > 1 \tag{23}
\]

\[
\iff \log_2 \left( \frac{P(\theta^{\text{motif}}|x)}{P(\theta^{\text{bg}}|x)} \right) > 0 \tag{24}
\]

Thus \( P(\theta^{\text{motif}}|x) > P(\theta^{\text{bg}}|x) \) when the log posterior odds is strictly positive. We can use Bayes’ Rule to expand the last inequality:
\[ P(\theta_{\text{motif}}|x) > P(\theta_{\text{bg}}|x) \iff \log_2 \left( \frac{P(x|\theta_{\text{motif}})P(\theta_{\text{motif}})}{P(x|\theta_{\text{bg}})P(\theta_{\text{bg}})} \right) > 0 \]  

(25)

\[ \iff \log_2 \left( \frac{P(x|\theta_{\text{motif}})P(\theta_{\text{motif}})}{P(x|\theta_{\text{bg}})P(\theta_{\text{bg}})} \right) > 0 \]  

(26)

\[ \iff \log_2 \frac{P(x|\theta_{\text{motif}})}{P(x|\theta_{\text{bg}})} + \log_2 \frac{P(\theta_{\text{motif}})}{P(\theta_{\text{bg}})} > 0 \]  

(27)

\[ \iff \log \text{odds} + \log \text{prior odds} > 0 \]  

(28)

In our example, this final inequality becomes

\[ \log \text{odds} + \log_2 \left( \frac{1}{1,000} \right) > 0, \]

from which we can see that we need

\[ \log \text{odds} > 10 \text{ bits} \]

in order to have a better than 50% chance that the match is a real TATA box. Thus with our models a given sequence needs a log odds score \( S \) of more than 10 bits in order to be considered a real TATA box. Let’s find the log odds value for the consensus sequence:

\[ \log_2 \left( \frac{P(\text{TATAAT}|\theta_{\text{motif}})}{P(\text{TATAAT}|\theta_{\text{bg}})} \right) = \log_2 \left( \frac{0.144}{4^{-6}} \right) \approx 9 \text{ bits} \]

Thus, even if you see TATAAT in E. coli, it is still more likely that the sequence is just regular DNA and not a real TATA box. Ideally we would like better posterior odds than \( \frac{P(\theta_{\text{motif}}|x)}{P(\theta_{\text{bg}}|x)} > 1 \), e.g., we might want

\[ \frac{P(\theta_{\text{motif}}|x)}{P(\theta_{\text{bg}}|x)} > 1000, \]

in order to confidently believe we had found a real occurrence of a TATA box. We would need a much larger and more conserved motif to get this large of a posterior odds; the log odds score would need to be about 10 bits higher.
1.3. Classical approach

We can also use e-values and p-values to help us determine if an observed pattern can be considered a real instance of our motif or a chance occurrence under our background model.

1.3.1. E-values

An e-value is the expected number of matches to a particular pattern in a database, and it is calculated using the background model of the data.

Let $n$ be the number of places you can have a match in your database. This is 5,000,000 for the E. coli genome.

Let $M$ be the total number of matches. This is a random variable, because we consider the database to be a random sequence in this approach.

Thus the expected number of matches to the consensus TATA sequence according to our background model (denoted $E(M)$) is $(\frac{1}{4})^6 \times 5,000,000$. Why is this true?

For each position $i$ in the genome define $M_i$ as follows:

$$M_i = \begin{cases} 1 & \text{if there is a match at position } i \\ 0 & \text{otherwise} \end{cases} \quad (29)$$

All of the $M_i$ together constitute 5 million random variables. These are statistically dependent! Consider the consensus sequence (TATAAT). A match at the first position means there cannot be a match at the second position, because a match at the first position requires an A at the second position, and a match at the second position requires a T at the second position.

We want to find $E(M)$. First observe that the following stochastic equation is true:

$$M = \sum_{i=1}^{n} M_i \quad \text{where } n=5,000,000 \quad (30)$$

Thus it follows that...
This last step uses an important fact in statistics that expectations sum even if the variables are not statistically independent.

In our example, \( E(M_i) = 4^{-6} \) for all \( i \). Then

\[
E(M) = 5,000,000 \times 4^{-6} = 1,220
\]

Thus we would expect to see TATA boxes occurring 1,220 times at random under the null hypothesis. This is not a good e-value. Ideally e-values are very small. If our motif were of length 12 we would have an e-value of 0.3, which is starting to get significant. If the motif were of length 18 we would have an e-value of 0.000073, which is very good. Of course, all of our e-values depend on our null model describing the background data exactly.

Similar calculations can be made to answer questions such as, “How many matches with a score greater than a given score \( S_0 \) are expected?”

To solve this, calculate the probability of getting a score greater than \( S_0 \) at any particular spot, and multiply by the database size (here 5,000,000).

1.3.2. P-values

A p-value is the probability of finding one or more matches to a particular pattern in a database; it is also calculated using the background model of the data. P-values are related to e-values. Consider the definition of e-value based on the “textbook” definition of expectation:

\[
E(M) = 0 \times P(M = 0) + 1 \times P(M = 1) + \ldots + n \times P(M = n) + \ldots
\]

Now consider what happens for small e-values.
If the e-value is small, then $P(M=1)$ is small,
$P(M=2)$ is really small,
$P(M=3)$ is really really small,
etc.

Thus for small e-values $P(M > 1)$ is negligible, and subsequently the e-value and the p-value are approximately the same. The exact calculation of p-values is hard, because the variables are dependent. We can use Poisson approximation or extreme value theory to help us calculate p-values. For details, see *Introduction to Computational Biology: Maps, Sequences and Genomes* by Michael Waterman.

### 1.4. Methods for Constructing the Motif Model

Now let us consider how we estimated the parameters $\theta_{\text{motif}}$ from the training data, and what other methods we might have used.

In position $j = 2$, the training data had 5 A’s and none of any other base. Thus we estimated $P(x_2 = A | \theta) = 1$; i.e., $\theta_2(A) = 1$. This turns out to be what is called a maximum likelihood estimate (MLE).

The Maximum Likelihood Principle requires that we find a model $\theta$ which maximizes the probability of the data observed. Thus in MLE we are finding $\theta$ to maximize $P(\text{data} | \theta)$. For example, using our model $\theta = \theta_{\text{motif}}$,

$$P\left(\text{training data} \mid \theta_2\right)$$

is maximized if

\[
\begin{align*}
\theta_2(A) &= 1 \\
\theta_2(C) &= 0 \\
\theta_2(G) &= 0 \\
\theta_2(T) &= 0,
\end{align*}
\]

(34)

and

$$P\left(\text{training data} \mid \theta_1\right)$$

is maximized if

\[
\begin{align*}
\theta_1(A) &= 0 \\
\theta_1(C) &= \frac{1}{5} \\
\theta_1(G) &= \frac{3}{5} \\
\theta_1(T) &= \frac{1}{5},
\end{align*}
\]

(35)

Proof of the above statements will be given in the next lecture.

If we have some prior knowledge of what the training data set should look like, we can introduce it into our model $\theta$ using Bayes’ Rule.
\[ P(\theta|\text{data}) = \frac{P(\text{data}|\theta)P(\theta)}{P(\text{data})} \]  
(36)

Here \(P(\theta)\) is the prior density. Note that

\[ P(\text{data}) = \int_{0}^{1} P(\text{data}|\theta)P(\theta)d\theta \]  
(37)

is a constant. The indices of the integral are specific to the parameter \(\theta\) we are using.

In place of the MLE, we may use either of the following estimates:

1. **Maximum a posteriori (MAP)** — find \(\theta\) to maximize \(P(\theta|\text{data})\)
2. **Mean Posterior Estimate (MPE)** — estimate \(\theta\) with the average posterior value:
   \[ \hat{\theta} = \int_{0}^{1} \theta P(\theta|\text{data})d\theta \]

The Mean Posterior Estimate is generally more robust.

Let’s use a simple example to demonstrate the use of these principles.

Suppose we are performing a series of coin tosses. Let \(H\) denote heads and \(T\) denote tails.

Suppose we had the following data:
\[
\text{data} = \text{HTTTHTT}
\]

Let \(\theta\) be the probability of heads. Using the Maximum Likelihood Estimate, we would find the frequency of heads in our observed data and assign \(\theta\) this value; e.g., \(\hat{\theta} = \frac{2}{7}\). If we wanted to use the MAP Estimate or MPE, we would need a prior density on \(\theta\). For example, we could use a uniform distribution or a beta distribution to describe the prior density.

Beta distributions are defined on \([0,1]\) as follows:

\[ P(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_{0}^{1} \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta}, \quad \text{for } \alpha, \beta > 0 \]  
(38)

What does this function look like? If \(\alpha = \beta = 11\), the plot loosely resembles a normal distribution with center 0.5. If \(\alpha = 10, \beta = 2\), the plot becomes skewed with mode and mean closer to 1. Note that the uniform distribution is a special case of the beta distribution, with \(\alpha = \beta = 1\).

Now suppose in our data set we had 5 heads and no tails. To get a MAP estimate, we want to maximize \(P(\theta|\text{data})\).
\[ P(\theta|\text{data}) = \frac{P(\text{data}|\theta)P(\theta)}{P(\text{data})} \]  

But \( P(\text{data}) \) is a constant. Thus to maximize \( P(\theta|\text{data}) \) we need to maximize

\[ P(\text{data}|\theta)P(\theta) = \theta^{\alpha-1}(1-\theta)^{\beta-1} \int_0^1 \theta^{n-1}(1-\theta)^{m-1}d\theta. \]  

The denominator on the right side of the equation is a constant as well. Thus to maximize \( P(\theta|\text{data}) \) we only need to maximize \( \theta^{n+\alpha-1}(1-\theta)^{m+\beta-1} \) for the MAP estimate, and we want to maximize \( \theta^n(1-\theta)^m \) for the MLE.

For the MLE:

\[ \hat{\theta} = \frac{n}{n+m} \]  

and for MAP estimate:

\[ \hat{\theta} = \frac{n + \alpha - 1}{n + \alpha + m + \beta - 2} \]  

for \( \alpha, \beta > 1 \)

Here \( \hat{\theta} \) is the \( \theta \) which maximizes. Thus, we see that the MLE and MAP estimates differ by the addition of “pseudocounts” \( \alpha \) and \( \beta \) to the counts \( n \) and \( m \) of actual occurrences of heads and tails in the data. Pseudocounts are used as a supplement in constructing motifs to counteract the possibility that the training data set is too small; e.g., they are often used to avoid counts of 0 for events which occur only rarely and could be absent in the training data.

For the MPE,

\[ \hat{\theta} = \frac{n + \alpha}{n + m + \alpha + \beta} \]

This will not be proven, but can be found on page 320 of the text. There you will also find the more general treatment of MPE’s for distributions over more than two values; e.g., for DNA bases (4 values) or amino acids (20 values). The prior in the text is called a Dirichlet prior. It is more general than a Beta prior.