CS284A: Representations and Algorithms in Molecular Biology

Scribe Notes on Lectures 3 & 4: Motif Discovery via Enumeration & Motif Representation Using Position Weight Matrix

Joshua Gervin
Based on presentations by Professor Xiaohui Xie on January 14 & 16, 2008

I. Motif Discovery via Enumeration

A. A Model for Motif Discovery (Review from Lecture 2)

We want to identify *biologically significant motifs* in a set S of n sequences, $s_1, s_2, \ldots, s_n$. Each potentially significant motif m_i of length w is associated with a summation variable k_i, which is the total number of sequences from S in which the motif appears. To systematically measure this significance, we must first find the underlying probability p any sequence of length l contains any theoretical motif of length w. With the overriding assumption that the four bases are *uniformly* distributed, or $\left(P(A),P(C),P(G),P(T)\right)=\left(\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4}\right)$, we have calculated a value for p of $1-\left(1-\frac{1}{4^w}\right)^{l-w+1}$. We use p as the probability of success for finding this theoretical motif each time we sample a sequence from set S. For k out of n trials, the probability of success is *binomial*,

$$P(k)=\binom{n}{k}p^k(1-p)^{n-k},$$

$$\text{where } \binom{n}{k}=\frac{n!}{k!(n-k)!},$$

as a motif either is in a sequence or is not. To test the significance of our specific motif m_i, we evaluate a *p-value*, or the probability, based on our distribution, that m_i would appear in *at least* k_i sequences:

$$\sum_{k=k_i}^{n} P(k) = \sum_{k=k_i}^{n} \binom{n}{k} p^k (1-p)^{n-k}.$$

If the p-value is smaller than a chosen significance level,[1] we can say with some confidence that our motif m_i is biologically significant. For large n the binomial distribution is approximated by a normal distribution, and we can map k_i to a new distribution and compute the *z-score* to determine the significance of our motif m_i.

B. Problems with this Model

1. The assumption that the four bases are uniformly distributed in the sequences is not necessarily correct. To be more accurate, we would need to model the *first-order statistics* (i.e., $P(A)$, $P(C)$, $P(G)$, and $P(T)$) of the nucleotide distribution.

2. The model ignores *second-order statistics*. Two bases might be more likely paired together than distributed at random (e.g. $P(GA) \geq P(G)P(A)$). The same could be also said for higher-order statistics.

C. Control Sequences

In order not to rely on the assumption of uniform distribution of bases to measure significance, we can generate a set of *N control sequences,* s_1^o, $s_2^o, \ldots, s_N^o$. The assumption is that our motif of interest m_i is not significant in the control sequences. Now we have two sets of sequences. Each m_i is associated with two values k_i and k_i^o, which correspond with the number of different sequences this motif appears in the sets of sequences S and S^o, respectively. Now to find out if our motif m_i is biologically significant, we choose the appropriate probability distribution for successfully finding a motif in k out of n trials. There are two types to choose from:

1. *The binomial distribution*

If the set S is independent of S^o, we can still model the probability of success $P(k)$ on finding a motif in k out of n trials using the binomial distribution. If $S \subset S^o$ (i.e., the set S is a subset of S^o), choosing the appropriate distribution now depends on the size of both sets and the distribution of our motif m_i in them. If the number N of S^o sequences and the number k_i^o of sequences containing our motif are large compared to the number n of S sequences, then the probability p of randomly picking a sequence with our motif remains essentially unchanged for n trials, and we could still model the probability $P(k)$

using the binomial distribution.[2] For these scenarios the only change we need to make from the model in Part A is to adopt a different underlying probability *p* of success for finding a motif every time we sample a sequence. For *p* we will use the *relative frequency* $\frac{k_i^o}{N}$ our motif m_i is found in the set S^o. This way, when we run *k* trials, we can compare the distributions from both *S* and S^o to see if our motif indeed stands out in *S*. The probability of success on *k* out of *n* trials may be written as

$$P(k) = \binom{n}{k}\left(\frac{k_i^o}{N}\right)^k \left(1 - \frac{k_i^o}{N}\right)^{n-k}.$$

To test the significance of our motif, we calculate the p-value in the same fashion as we did before: $\sum_{k=k_i}^{n} P(k)$. For large *n* we can again map k_i to a normal distribution with mean *np* and variance *np*(*1-p*) and compute the z-score.

2. *The hypergeometric distribution*

If $S \subset S^o$ and if either *N* or k_i^o is not large compared to *n* for a given m_i, the sequence of *n* trials is analogous to *sampling without replacement. The probability p of randomly picking a sequence with our motif changes significantly over n trials*. Hence, we cannot use the binomial distribution, which assumes the same *p* for all trials. The appropriate distribution is *hypergeometric*, where the probability of success on finding a motif in *k* out of *n* trials is

$$P(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}},$$

where $\binom{K}{k}$ is the number of ways of choosing *k* sequences with a motif from the total number *K* of sequences with that motif, $\binom{N-K}{n-k}$ is the number of ways of choosing *n-k* sequences without the motif from the total number *N-K* of sequences without the motif, and $\binom{N}{n}$ is the number of ways of choosing *n* sequences from the total number *N* sequences.

While using this distribution to test the significance of our particular motif m_i, we assign k_i^o to the value *K*. Like before we calculate the p-value using the summation $\sum_{k=k_i}^{n} P(k)$. We cannot compute a z-score here, as a normal distribution does not approximate a hypergeometric distribution for large *n*.

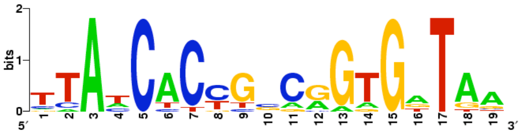
II. Representation of a Motif Using a Position Weight Matrix

A. What is a Position Weight Matrix?

Motifs are hardly ever represented accurately by a unique consecutive sequence of A's, C's, G's and T's. Instead, we create a *position weight matrix* (PWM) to represent the frequencies of each base at each position in the motif:

G	0	1.0	0	0	0.7	1.0	0	0	0.4	0.8
A	0.4	0	1.0	0	0	0	1.0	0	0	0
T	0.6	0	0	1.0	0.3	0	0	1.0	0.4	0.2
C	0	0	0	0	0	0	0	0	0.2	0

Sometimes a position weight matrix is represented by a *sequence logo*, where the height of the letters representing the nucleotides correlates with the frequency that base is found in *n* different sequences containing the motif:



From the example above, position 1 is said to be *degenerate*; there is no single nucleotide that represents the motif here. On the other hand position 3 is said to be *stringent* because the motif is well represented by adenosine.

B. Mathematical Representation of a Position Weight Matrix

The position weight matrix for a motif of width *w* can be expressed as

$$\theta = \begin{bmatrix} \theta_{11} & \theta_{21} & \cdots & \theta_{w1} \\ \theta_{12} & \theta_{22} & \cdots & \theta_{w2} \\ \theta_{13} & \theta_{23} & \cdots & \theta_{w3} \\ \theta_{14} & \theta_{24} & \cdots & \theta_{w4} \end{bmatrix},$$

where each row *j* represents A, C, G, or T, and each column *i* represents one position of the motif, and is normalized:

$$\sum_{j=1}^{4} \theta_{ij} = 1$$

for all i = 1, 2,…*w*. For example θ_{23} is the *relative frequency* that guanine is found in position 2 of the motif.

C. Likelihood of a Sequence

If all the relative frequencies θ_{ij} are given for the position weight matrix θ, we can measure the probability of generating a sequence $S = (s_1, s_2,\ldots, s_w)$. This is also known as the *likelihood* $L(\theta)$ of the sequence. For example we can use a position weight matrix of width $w = 3$ to calculate likelihood of the sequence GGG. It is simply the product of three relative frequencies θ_{13}, θ_{23}, and θ_{33}.

Generalizing this using mathematics, we find the likelihood of a sequence $S = (s_1, s_2,\ldots, s_w)$ given θ_i is

$$L(\theta) = P(S \mid \theta) = \prod_{i=1}^{w} \sum_{j=1}^{4} \theta_{ij} I(s_i = j),$$

$$\text{where } I(s_i = j) = \begin{cases} 1 & \text{if } s_i = j \\ 0 & \text{if not} \end{cases}$$

Let us briefly go over a few syntax elements. First of all, the expression $P(S|\theta)$ represents a *conditional probability*: We are asking, "What is the likelihood of sequence *S* given the condition that the position weight matrix is θ?" Secondly, the $\prod$ (i.e., capital pi) notation means we take the *product* of the associated terms. Finally, for convenience we converted the alphabetical string (A, C, G, T) into a numerical one (1, 2, 3, 4). These numbers are represented by the variable *j* in the above expression.

Other ways of expressing the likelihood $L(\theta)$ are

$$
\begin{aligned}
L(\theta) = P\left(S \mid \theta\right) &= \prod_{i=1}^{w} P\left(s_i \mid \theta_i\right) \\
&= \prod_{i=1}^{w} \theta_{i,s_i} \qquad .
\end{aligned}
$$

The conditional probability $P(s_i|\theta_i)$ is the probability of generating a nucleotide element s_i given its relative frequency θ_i.

We can expand this idea further and measure the likelihood for a set of sequences S_1, S_2,…, S_n given θ. Since we are assuming each sequence S_k is generated independently from θ, this probability is simply the product of the relative frequencies $\theta_{i,s_{ki}}$ representing each nucleotide element s_{ki}:

$$
\begin{aligned}
L(\theta) = P\left(S_1, S_2, ..., S_n \mid \theta\right) &= \prod_{k=1}^{n} P\left(S_k \mid \theta\right) \\
&= \prod_{k=1}^{n} \prod_{i=1}^{w} \theta_{i,s_{ki}}
\end{aligned}
$$

Note that the syntax $P(S_1, S_{2,\ldots,}S_n|\theta)$ represents a *joint probability*—the probability of generating sequences S_1, S_2,…, S_n —as well as a conditional probability—the probability given θ.

D. Using Maximum Likelihood to Estimate the Positional Weight Matrix θ

Often times we want to construct a position weight matrix θ of length w from observed sequence data. For a set of sequences S_1, S_2,…, S_n represented by the same θ, our strategy is to maximize the likelihood $L(\theta)$ over all possible values of θ_{ij}. This could be done by setting the partial derivative $\frac{\partial L(\theta)}{\partial \theta_{ij}}$ equal to zero and solving for θ_{ij}; however, it is much easier to take the partial derivative with respect to the *log-likelihood function* (i.e., the logarithm of the likelihood) and set it to zero

$$
\frac{\partial \log L(\theta)}{\partial \theta_{ij}} = 0
$$

because the product associated with the likelihood $L(\theta)$ turns into a sum. Note that there are only $3w$ and not $4w$ parameters for which we need to solve, since if we figure out θ_{i1}, θ_{i2}, and θ_{i3}, we can use the relation $\sum_{j=1}^{4} \theta_{ij} = 1$to give us θ_{i4}.

Using this method on a set of sequences S_1, S_2,…, S_n, all with the same θ, we can derive an expression for the relative frequency

$$\theta_{ij} = \frac{n_{ij}}{n},$$

which is simply the *absolute frequency* of each nucleotide j for every column i, divided by the total number of sequences n.

Often times it is much harder to solve for the position weight matrix θ. It is quite likely within a set of n given sequences S_1, S_2,…, S_n that only some sequences contain the motif, and thus only this subset can generate the weight matrix θ. The problem is we do not know which sequences form this subset. Let us assume the rest of the "*non-motif*" (also called *background*) sequences form a subset generated from a single distribution (i.e., from a second position weight matrix θ^0 made up of identical columns of $p^o = (p^o_A, p^o_C, p^o_G, p^o_T) = (p^o_1, p^o_2, p^o_3, p^o_4)$. The likelihood $L(\theta, \theta^0)$ for this set of sequences S_1, S_2,…, S_n is now

$$L(\theta,\theta^o) = P\left(S_1,S_2,...,S_n \mid z,\theta,\theta^o\right) = \prod_{k=1}^{n}\left[z_k P\left(S_k \mid \theta\right) + \left(1-z_k\right)P\left(S_k \mid \theta^o\right)\right],$$

$$\text{where } z_k = \begin{cases} 1 \text{ if } S_k \text{ is generated by } \theta \\ 0 \text{ if } S_k \text{ is generated by } \theta^o \end{cases}.$$

The problem of not knowing if a sequence S_k belongs to the motif (θ) or the background model (θ^0) can now be expressed mathematically as not knowing which value 0 or 1 to use for the binary function z_k associated with each S_k. Fortunately, we can remove z from the equation by integrating the likelihood $L(\theta, \theta^0)$ over all possible events z:[3]

$$P\left(S_1,S_2,...,S_n \mid \theta,\theta^o\right) = \sum_z P\left(S_1,S_2,...,S_n \mid z,\theta,\theta^o\right)P(z).$$

After integration, we are left with

$$L(\theta,\theta^o) = P\left(S_1,S_2,...,S_n \mid \theta,\theta^o\right) = \prod_{k=1}^{n}\left[P\left(z_k\right)P\left(S_k \mid \theta\right) + \left(1-P\left(z_k\right)\right)P\left(S_k \mid \theta^o\right)\right].$$

We may be fortunate to know the probability $P(z_k = 1)$ for the set of sequences S_1, S_2,…, S_n. Representing this probability as the constant α, the likelihood of the set may now be written as

$$L(\theta,\theta^o) = P\left(S_1,S_2,...,S_n \mid \theta,\theta^o\right) = \prod_{k=1}^{n}\left[\alpha P\left(S_k \mid \theta\right) + \left(1-\alpha\right)P\left(S_k \mid \theta^o\right)\right]$$

.

Having successfully expressed the likelihood as a function of 3*w* independent variables $\theta_{i,s_{ki}}$ and 3 independent variables $\theta^o_{i,s_{ki}}$, we can now use our strategy of solving for $\theta_{i,s_{ki}}$ and $\theta^o_{i,s_{ki}}$ when the likelihood is at a maximum. However, setting the partial derivatives of the log-likelihood function equal to zero is too difficult a task because the likelihood $L(\theta, \theta^o)$ in this case is simply not just a product of the independent variables. We will implement the *EM Algorithm* next lecture to solve this *maximum likelihood estimation problem*.

[1] Wikipedia, "P-value," http://en.wikipedia.org/wiki/P-value.

[2] The *relative frequency* $\frac{k_i^o}{N}$ the motif is found in the set S^o must also not be close to 0 or 1.

[3] In general we can calculate a *marginal probability* from a conditional or joint probability by removing one of the variables using integration

$$P(X) = \sum_Y P(X,Y) = \sum_Y P(X \mid Y)P(Y),$$

where we take the sum over all possible events *Y*. From R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis,* Cambridge University Press, 2006, p. 6.