Lecture 7

Finding Instances of Known Sites

January 25, 2000
Notes: Elisabeth Rosenthal

With this lecture we begin a study of how to identify functional regions from biological sequence data. This includes the problem of how to identify relatively long functional regions such as genes, but we begin instead with the problem of identifying shorter functional regions.

A *site* is a short sequence that contains some signal, that signal often being recognized by some enzyme. Examples of nucleotide sequence sites include the following:

1. origins of replication, where DNA polymerase initially binds (Section 1.5),
2. transcription start and stop sites (Section 1.6.1),
3. ribosome binding sites in prokaryotes (Section 2.2),
4. promoters, or transcription factor binding sites (Section 2.3), and
5. intron splice sites (Section 2.5).

We will further subdivide the problem of identifying sites into the problems of finding instances of a known site, and finding instances of unknown sites. We begin with the former. What makes all these problems interesting and challenging is that instances of a single site will generally not be identical, but will instead vary slightly.

7.1. How to Summarize Known Sites

Suppose that we have a large sample $\mathcal{A}$ of length n sites, and a large sample $\mathcal{B}$ of length n nonsites. Given a new sequence $s = s_1 s_2 \cdots s_n$ of length n, is s more likely to be a site or a nonsite? If we can derive an efficient way to determine this, we can screen an entire genome, testing every length n sequence, and thereby generate a "complete" list of candidate sites (excepting sequences where the test gives the wrong answer).

To illustrate, the *cyclic AMP receptor protein* (CRP) is a transcription factor (see Section 2.3) in *E. coli*. Its binding sites are DNA sequences of length approximately 22. Table 7.1, taken from Stormo and Hartzell [2], shows just positions 3–9 (out of the 22 sequence positions) in 23 *bona fide* CRP binding sites.

The "signal" in Table 7.1 is not easy to detect at first glance. Notice, though, that in the second column T predominates and in the third column G predominates, for example. Our first goal is to capture the most relevant information from these 23 sites in a concise form. (This would clearly be more important if we

TTGTGGC
TTTTGAT
AAGTGTC
ATTTGCA
CTGTGAG
ATGCAAA
GTGTTAA
ATTTGAA
TTGTGAT
ATTTATT
ACGTGAT
ATGTGAG
TTGTGAG
CTGTAAC
CTGTGAA
TTGTGAC
GCCTGAC
TTGTGAT
TTGTGAT
GTGTGAA
CTGTGAC
ATGAGAC
TTGTGAG

Table 7.1: Positions 3–9 from 23 CRP Binding Sites [2]

A	0.35	0.043	0	0.043	0.13	0.83	0.26
C	0.17	0.087	0.043	0.043	0	0.043	0.3
G	0.13	0	0.78	0	0.83	0.043	0.17
T	0.35	0.87	0.17	0.91	0.043	0.087	0.26

Table 7.2: Profile for CRP Binding Sites Given in Table 7.1

were given thousands of sites rather than just 23.) In order to do this, suppose that the sequence residues are from an alphabet of size c. Consider a $c \times n$ matrix A where $A_{r,j}$ is the fraction of sequences in $\mathcal{A}$ that have residue r in position j. Table 7.2 shows the 4×7 matrix A for the CRP sites given in Table 7.1. Such a matrix is called a *profile*. The profile shows the distribution of residues in each of the n positions. For example, in column 1 the residues are quite mixed, in column 2, T occurs 87% of the time, etc.

7.2. Using Probabilities to Test for Sites

An alternative way to think of $A_{r,j}$ is in terms of probability. Let $t = t_1 t_2 \cdots t_n$ be chosen randomly and uniformly from $\mathcal{A}$. Then $A_{r,j} = \Pr(t_j = r \mid t \in \mathcal{A})$. In words, this says, "$A_{r,j}$ is the probability that the j-th residue of t is the residue r, given that t is chosen randomly from $\mathcal{A}$." For instance, $A_{T,2} = \Pr(t_2 = T \mid t \in \mathcal{A}) = 0.87$.

For the time being, we will make the following *Independence Assumption*: which residue occurs at position j is independent of the residues occurring at other positions. In other words, residues at any two different positions are uncorrelated. Although this assumption is not always realistic, it can be justified in some circumstances. The first justification is that it keeps the model and resulting analysis simple. The second justification is its predictive power in some (but admittedly not all) situations.

The independence assumption can be made precise in probabilistic terms:

Definition 7.1: Two probabilistic events E and F are said to be *independent* if the probability that they both occur is the product of their individual probabilities, that is, $\Pr(E \ \& \ F) = \Pr(E) \cdot \Pr(F)$.

Under the independence assumption, the probability that a randomly chosen site has a specified sequence $r_1 r_2 \cdots r_n$ is determined by Definition 7.1 as follows:

$$\begin{aligned} \Pr(t = r_1 r_2 \cdots r_n \mid t \text{ is a site}) &= \Pr(t_1 = r_1 \ \& \ t_2 = r_2 \ \& \ \cdots \ \& \ t_n = r_n \mid t \text{ is a site}) \\ &= \prod_{j=1}^{n} \Pr(t_j = r_j \mid t \text{ is a site}) \\ &= \prod_{j=1}^{n} A_{r_j,j}. \end{aligned} \tag{7.1}$$

For example, suppose we want to know the probability that a randomly chosen CRP binding site will be TTGTGAC. By using Equation (7.1) and Table 7.2,

$$\Pr(t = \text{TTGTGAC} \mid t \text{ is a site}) = (.35)(.87)(.78)(.91)(.83)(.83)(.3) = 0.045.$$

Although this probability is small, it is the largest probability of any site sequence, because each position contains the most probable residue.

Now form the $c \times n$ profile B from the sample $\mathcal{B}$ of nonsites in the same way. Using the profiles A and B, let us return to the question of whether a given sequence s is more likely to be a site or nonsite. In order to do this, we define the likelihood ratio.

Definition 7.2: Given the sequence $s = s_1 s_2 \cdots s_n$, the *likelihood ratio*, denoted by $LR(A, B, s)$, is defined to be

$$\frac{\Pr(t = s \mid t \text{ is a site})}{\Pr(t = s \mid t \text{ is a nonsite})} = \frac{\prod_{j=1}^{n} A_{s_j,j}}{\prod_{j=1}^{n} B_{s_j,j}} = \prod_{j=1}^{n} \frac{A_{s_j,j}}{B_{s_j,j}}.$$

A	0.48	−2.5	$-\infty$	−2.5	−0.94	1.7	0.061
C	−0.52	−1.5	−2.5	−2.5	$-\infty$	−2.5	0.28
G	−0.94	$-\infty$	1.6	$-\infty$	1.7	−2.5	−0.52
T	0.48	1.8	−0.52	1.9	−2.5	−1.5	0.061

Table 7.3: Log Likelihood Weight Matrix for CRP Binding Sites

To illustrate, let $\mathcal{B} = \{A, C, T, G\}^7$, the set of all length seven sequences. The corresponding profile B has $B_{r,j} = 0.25$ for all r and j. Then for $s =$ TTGTGAC,

$$LR(A, B, s) = \frac{\prod_{j=1}^{n} A_{s_j,j}}{\prod_{j=1}^{n} B_{s_j,j}} = \frac{0.045}{(0.25)^7} = 732.$$

To test a sequence s, compare $LR(A, B, s)$ to a prespecified constant "cutoff" L, and declare s more likely to be a site if $LR(A, B, s) \geq L$.

If n is not small and some entries in A and B are small, then the likelihood ratio may be intractably large or small, causing numerical problems in the calculation. To alleviate this, we define the log likelihood ratio.

Definition 7.3: Given the sequence $s = s_1 s_2 \cdots s_n$, the *log likelihood ratio*, denoted by $LLR(A, B, s)$, is defined to be

$$\log_2 LR(A, B, s) = \log_2 \prod_{j=1}^{n} \frac{A_{s_j,j}}{B_{s_j,j}} = \sum_{j=1}^{n} \log_2 \frac{A_{s_j,j}}{B_{s_j,j}}.$$

The corresponding test of s is that s is more likely to be a site if $LLR(A, B, s) \geq \log_2 L$.

To test for sites, it is convenient to create a scoring matrix W whose entries are the log likelihood ratios, that is, $W_{r,j} = \log_2 \frac{A_{r,j}}{B_{r,j}}$. Table 7.3 shows the weight matrix for the example CRP samples $\mathcal{A}$ and $\mathcal{B}$ we have been discussing. In order to compute $LLR(A, B, s)$, Definition 7.3 says to add the corresponding scores from W: $LLR(A, B, s) = \sum_{j=1}^{n} W_{s_j,j}$.

A technical difficulty arises when an entry $A_{r,j}$ is 0, because the corresponding entry $W_{r,j}$ is then $-\infty$. If the residue r cannot possibly occur in position j of any site for biological reasons, then there is no problem. More often, though, this is a result of having too small a sample $\mathcal{A}$ of sites. In this case, there are various "small sample correction" formulas, which replace $A_{r,j}$ by a small positive number (see, for example, Lawrence *et al.* [1]), but we will not discuss them here.

References

[1] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262:208–214, 8 October 1993.

[2] G. D. Stormo and G. W. Hartzell III. Identifying protein-binding sites from unaligned DNA fragments. *Proceedings of the National Academy of Science USA*, 86:1183–1187, 1989.