Lecture 8

Relative Entropy

January 27, 2000
Notes: Anne-Louise Leutenegger

8.1. Weight Matrices

A *weight matrix* is any $c \times n$ matrix W that assigns a score to each sequence $s = s_1 s_2 \cdots s_n$ according to the formula $\sum_{j=1}^{n} W_{s_j,j}$. The log likelihood ratio matrix described at the end of Section 7.2, and illustrated in Table 7.3, is an example of a weight matrix.

In computing log likelihood ratios, we often take $B_{r,j}$ to be the "background" distribution of residue r in the entire genome, or a large portion of the genome. That is, $B_{r,j}$ is the frequency with which residue r appears in the genome as a whole. In this case, $B_{r,j}$ is independent of j, that is, $B_{r,j} = B_{r,j'}$ for all j and j'. Note, however, that this does not mean that $B_{r,j} = 0.25$ in the case of nucleotides. Although this *uniform* distribution is a fair estimate for the nucleotide composition of *E. coli*, it is not for other organisms. For instance, the nucleotide composition for the archaeon *M. jannaschii* is approximately $B_{A,j} = B_{T,j} = 0.34$ and $B_{C,j} = B_{G,j} = 0.16$.

8.2. A Simple Site Example

Example 8.1: As a simpler example of a collection of sites than the CRP binding sites of Table 7.1, Table 8.1 shows eight hypothetical translation start sites. For this example, we will assume a uniform background distribution $B_{r,j} = 0.25$. Table 8.2(a) shows the site profile matrix, and Table 8.2(b) the log likelihood ratio weight matrix, for this example. As illustrations of the log likelihood ratio calculations,

ATG
ATG
ATG
ATG
ATG
GTG
GTG
TTG

Table 8.1: Eight Hypothetical Translation Start Sites

A	0.625	0	0
C	0	0	0
G	0.25	0	1
T	0.125	1	0

(a)

A	1.32	$-\infty$	$-\infty$
C	$-\infty$	$-\infty$	$-\infty$
G	0	$-\infty$	2
T	-1	2	$-\infty$

(b)

	0.701	2	2

(c)

Table 8.2: (a) Profile, (b) Log Likelihood Weight Matrix, and (c) Positional Relative Entropies, for the Sites in Table 8.1, with Respect to Uniform Background Distribution

$W_{T,2} = \log_2 \frac{A_{T,2}}{B_{T,2}} = \log_2 \frac{1}{0.25} = \log_2 4 = 2$, and $W_{G,1} = \log_2 \frac{0.25}{0.25} = 0$, meaning both distributions have the same frequency for G in position 1.

8.3. How Informative is the Log Likelihood Ratio Test?

The next question to ask is how informative is a given weight matix W for distinguishing between sites and nonsites. If the distributions for sites and nonsites were identical, then every entry in the weight matrix would be 0, and it would be totally uninformative.

Definition 8.2: A *sample space* S is the set of all possible values of some random variable s.

Definition 8.3: A *probability distribution* P for a sample space S assigns a probability $P(s)$ to every $s \in S$, satisfying

1. $0 \leq P(s) \leq 1$, and
2. $\sum_{s \in S} P(s) = 1$.

In our application, the sample space is the set of all length n sequences. The site profile A induces a probability distribution on this sample space according to Equation (7.1), as does the nonsite profile B.

Definition 8.4: Let P and Q be probability distributions on the same sample space S. The *relative entropy* (or "information content", or "Kullback-Leibler measure") of P with respect to Q is denoted $D_b(P||Q)$ and is defined as follows:

$$D_b(P||Q) = \sum_{s \in S} P(s) \log_b \frac{P(s)}{Q(s)}.$$

By convention, we define $P(s)\log_b \frac{P(s)}{Q(s)}$ to be 0 whenever $P(s) = 0$, in agreement with the fact from calculus that $\lim_{x\to 0} x\log x = 0$.

Since $\log \frac{P(s)}{Q(s)}$ is the log likelihood ratio, $D_b(P||Q)$ is a weighted average of the log likelihood ratio with weights $P(s)$.

Definition 8.5: The *expected value* of a function $f(s)$ with respect to probability distribution P on sample space S is

$$E(f(s)) = \sum_{s\in S} P(s)f(s).$$

In these terms, the relative entropy is the expected value of $LLR(P, Q, s)$ when s is picked randomly according to $P(s)$. That is, it is the expected log likelihood score of a randomly chosen site.

Note that when P and Q are the same distribution, the relative entropy will be zero. In general, the relative entropy measures how different the distributions P and Q are. Since we want to be able to distinguish between sites and nonsites, we want the relative entropy to be large, and will use relative entropy as our measure of how informative the log likelihood ratio test is.

When the sample space is all length n sequences, and we assume independence of the n positions, it is not difficult to prove that the relative entropy satisfies

$$D_b(P||Q) = \sum_{j=1}^{n} D_b(P_j||Q_j),$$

where P_j is the distribution P imposes on the jth position and Q_j is the distribution Q imposes on the jth position.

When $b = 2$, the relative entropy is measured in "bits". This will be the usual case, unless specifically stated otherwise.

Continuing Example 8.1, Table 8.2(c) shows the relative entropies $D_2(P_j||Q_j)$ for each nucleotide position j separately. For instance, looking at position 2, residues A, C, and G do not contribute to the relative entropy (see Table 8.2(a)). Residue T contributes $1 \cdot W_{T,2} = 2$ (see Tables 8.2(a) and (b)). Hence, $D_2(P_2||Q_2) = 2$. This means that there are 2 bits of information in position 2. If the residues were coded with 0 and 1 so that 00 = A, 01 = C, 10 = G, and 11 = T, only 2 bits (11) would be necessary to encode the fact that this residue is always T. Position 3 has the same relative entropy of 2. For position 1, the relative entropy is 0.7 so there are 0.7 bits of information, indicating that column 1 of Table 8.2(a) is more similar to the background distribution than columns 2 and 3 are. The total relative entropy of all three positions is 4.7.

Example 8.6: Let us now modify Example 8.1 to see the effect of a nonuniform background distribution. Consider the same eight translation start sites of Table 8.1, but change the background distribution to $B_{A,j} = B_{T,j} = 0.375$, $B_{C,j} = B_{G,j} = 0.125$. The site profile matrix remains unchanged (Table 8.2(a)). The new weight matrix and relative entropies are given in Table 8.3.

Note that the relative entropy of each position has changed and, in particular, the last two columns no longer have equal relative entropy. The site distribution in position 2 is now more similar to the background distribution than the site distribution in position 3 is, since G is rarer in the background distribution. Thus, the relative entropy of position 3 is greater than that of position 2. An interpretation of $D_2(P_3||Q_3) = 3$ is that the residue G is $2^3 = 8$ times more likely to occur in the third position of a site than a nonsite. The total relative entropy of all three positions is 4.93.

A	0.737	$-\infty$	$-\infty$
C	$-\infty$	$-\infty$	$-\infty$
G	1	$-\infty$	3
T	-1.58	1.42	$-\infty$

(b)

	0.512	1.42	3

(c)

Table 8.3: (b) Log Likelihood Weight Matrix, and (c) Positional Relative Entropies, for the Sites in Table 8.1, with Respect to a Nonuniform Background Distribution

0.12	1.3	1.1	1.5	1.2	1.1	0.027

Table 8.4: Positional Relative Entropy for CRP Binding Sites of Tables 7.1 – 7.3

Example 8.7: Finally, returning to the more interesting CRP binding sites of Table 7.1, the seven positional relative entropies are given in Table 8.4. Note that 1.5 (middle position) is the highest relative entropy and corresponds to the most biased column (see Table 7.2). The value 0.027 (last position) is the lowest relative entropy because the distribution in this last position is the closest to the uniform background distribution (see Table 7.2).

8.4. Nonnegativity of Relative Entropy

In these examples, the relative entropy has always been nonnegative. It is by no means obvious that this should be, since it is the expected value of the log likelihood ratio, which can take negative values. For instance, why should the expected value of the last column of Table 7.3 be positive (0.027, according to Table 8.4)? The following theorem demonstrates that this must, indeed, be the case.

Theorem 8.8: For any probability distributions P and Q over a sample space S, $D_b(P||Q) \geq 0$, with equality if and only if P and Q are identical.

Proof: First, it is true that $\ln x \leq x - 1$ for all real numbers x, with equality if and only if $x = 1$. The reason is that the curve $y = \ln x$ is concave downward, and its tangent at $x = 1$ is the straight line $y = x - 1$. Thus, $\ln \frac{1}{x} = \ln(x^{-1}) = -\ln x \geq 1 - x$. In the following derivation, we will use this inequality with $x = \frac{Q(s)}{P(s)}$:

$$\begin{aligned}
D_b(P||Q) &= \sum_{s\in S} P(s)\log_b \frac{P(s)}{Q(s)} \\
&= \frac{1}{\ln b}\sum_{s\in S} P(s)\ln\frac{P(s)}{Q(s)} \\
&\geq \frac{1}{\ln b}\sum_{s\in S} P(s)(1-\frac{Q(s)}{P(s)}) \\
&= \frac{1}{\ln b}\sum_{s\in S}(P(s)-Q(s)) \\
&= \frac{1}{\ln b}(\sum_{s\in S} P(s)-\sum_{s\in S} Q(s)) \\
&= 0,
\end{aligned}$$

since $\sum_{s\in S} P(s) = \sum_{s\in S} Q(s) = 1$, by Definition 8.3. Note that the relative entropy is equal to 0 if and only if $x = Q(s)/P(s) = 1$ for all $s \in S$, that is, P and Q are identical probability distributions. $\square$