Lecture 11

Correlation of Positions in Sequences

February 10, 2000
Notes: Tammy Williams

This lecture explores the validity of the assumption that the residues appearing at different positions in a sequence are independent. In previous lectures the computations assumed such positional independence. (See Section 7.2.) Here we describe a method to determine the level of dependence among residues in a sequence. By calculating the relative entropy of two models, one modeling dependence and the other modeling independence of positions, we can quantify the validity of the positional independence assumption.

Most of the material for this lecture is from Phil Green's MBT 599C lecture notes, Autumn 1996.

11.1. Nonuniform Versus Uniform Distributions

We will begin with a warmup to the method that still assumes positional independence, and proceed to the dependence question in Section 11.2. Given the genome of an organism, a simple calculation determines the frequency of each nucleotide. It is reasonable to suspect that these frequencies are more informative than the uniform nucleotide distribution, in which the probability of each nucleotide is 0.25. One can compare the frequency distribution (the nonuniform distribution in which the probability of residue r is equal to the frequency of residue r in the genome as a whole) to the uniform distribution.

Example 11.1: This example calculates the relative entropy of the frequency distribution to the uniform distribution for the archaeon *M. jannaschii*. *M. jannaschii* is a *thermophilic* prokaryote, meaning that it lives in extremely high temperature environments such as thermal springs. The frequency distribution of residues for *M. jannaschii* is given in Table 11.1.

A: 0.344
C: 0.155
G: 0.157
T: 0.343

Table 11.1: The frequency distribution of residues in *M. jannaschii*

Notice that the frequencies of residues A and T are very similar but not equal. Likewise, the residues C and G have similar frequencies. When calculating these frequencies, only one strand of DNA was used. (Had both strands been used, base pair complementarity would have ensured that these frequencies would be exactly equal rather than just similar.) Because genes and other functional regions tend to occur on both

strands of DNA equally often, any bias of such a region on one strand over the other (see Section 11.4) is canceled out. This phenomenon, together with the fact that the bases occur in complementary pairs, explains why the frequencies of A's and T's are similar and the frequencies of G's and C's are similar.

Let Q be the uniform probability distribution, and let P be the frequency distribution. Notice that the frequency of residue r is equal to the probability of randomly selecting residue r from the distribution P. How much better does P model the actual genome than Q? More quantitatively, how much more information is there using P rather than Q? Recall from Section 8.3 that the relative entropy is defined as follows:

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

In Example 11.1 for *M. jannaschii*, $D_2(P||Q) = 0.103$. This implies that there are 0.103 more bits of information per position in the sequence by using distribution P over distribution Q. The value 0.103 might seem insignificant, but it means that a sequence of 100 bases has ten bits of extra information when chosen according to distribution P. Suppose a random sequence s of length 100 is selected according to the probability distribution P. Since the relative entropy is the expected log likelihood ratio for s, the sequence s is approximately $2^{10} = 1024$ times more likely to have been generated by P than by Q. The mathematics leading to this observation is flawed, since the log function and expectation do not "commute": that is, for it to be correct we would need the expected log likelihood ratio to equal the log of the expected likelihood ratio, which is not true in general. However, the intuition is helpful.

The next section explores the application of this relative entropy method to the question of dependence of nucleotides.

11.2. Dinucleotide Frequencies

How much dependence is there between *adjacent* nucleotides in a DNA sequence? Since there are four nucleotides, there are 16 possible pairs of nucleotides. To calculate the frequencies of each such pair (i, j) in a sequence, a simple algorithm computes the total number of observed occurrences of i followed immediately by j, and divides by the total number of pairs, which is the length of the sequence minus one. Let P_{ij} be the frequency of the residue i immediately followed by the residue j. In addition let P_i be the frequency of residue i in the single nucleotide distribution. The value $P'_{ij} = \frac{P_{ij}}{P_i P_j}$ gives a score representing the validity of the positional independence assumption. If $P'_{ij} = 1$, then the independence assumption is valid for residue i followed by residue j. (See Definition 7.1.) As the deviation from one increases, the independence assumption becomes less valid.

Example 11.2: Let us return to Example 11.1 involving the organism *M. jannaschii*. The P'_{ij} values are given in Table 11.2 with the residue i indexed by row and the residue j indexed by column, where residue i immediately precedes j in the sequence. Upon examination of the table, one can see that there are sizable deviations from one. For example the pairs (C,C) and (G,G) occur much more often than expected if they were independent, and (A,C), (C,G), and (G,T) occur much less often. Also, the diagonal entries show that two consecutive occurrences of the same residue occur more often than expected. Such repeats of the same residue might result from the slippage of the DNA polymerase during the replication process (see Section 1.5). The DNA polymerase inserts an extra copy of the base or misses a copy while duplicating one of the DNA strands. Even though there is a post-replication repair system to repair mistakes produced by the DNA

polymerase, there is a small chance that the repeats will persist after a copy mistake. (In a similar way, dinucleotide repeats might occur during the replication process.)

	A	C	G	T
A	1.13	0.73	1.10	0.94
C	1.03	1.37	0.32	1.11
G	1.05	1.12	1.39	0.71
T	0.83	1.05	1.03	1.14

Table 11.2: The ratios of the observed dinucleotide frequency to the expected dinucleotide frequency (assuming independence) in *M. jannaschii*

Definition 11.3: The *mutual information* of a pair (X, Y) of random variables is

$$I(X;Y) = \sum_x \sum_y Pr(X = x \ \& \ Y = y) log_2 \frac{Pr(X = x \ \& \ Y = y)}{Pr(X = x)Pr(Y = y)}.$$

If the probability distribution P is the joint distribution of X and Y, and Q is the distribution of X and Y assuming independence, then $I(X;Y) = D_2(P||Q)$. By Theorem 8.8, then, $I(X;Y) \geq 0$, with equality if and only if X and Y are independent, since in the equality case $P = Q$.

By setting the random variable X to be the first base and Y to be the second base of a pair, the value $I(X;Y)$ for *M. jannaschii* is 0.03. For a sequence of 100 bases, there are three bits of extra information when the sequence is chosen from the dinucleotide frequency distribution rather than the independence model. Thus, a random sequence s of length 100 generated by a process according to dinucleotide distribution P is eight times more likely to have been generated by P than by the independent nucleotide distribution Q.

11.3. Disymbol Frequencies

A generalization of the dinucleotide frequency is called the *disymbol frequency*, in which the two positions are not restricted to be adjacent. For example, one could study the dependence relationship between pairs of nucleotides separated by ten positions. The extension of methods presented above to this generalized setting is straightforward. Studies have revealed that the mutual information between DNA nucleotides separated by more than one base is lower than for adjacent residues. In fact, for separations of length 2, 3, and 4, the mutual information is an order of magnitude less than for adjacent residues.

11.4. Coding Sequence Biases

A similar application of relative entropy is finding biases in coding sequences. Recall that coding sequences consist of codons that are three consecutive bases: see Section 1.6.2. Do the three positions each have the same distribution as the background distribution? If such statistical features of protein-coding regions are known, they can be exploited by algorithms that locate genes.

In the bacterium *H. influenzae*, the residues A and G are more likely to appear in the first position of codons than in the genomic background. Using an analysis analogous to that used in Sections 11.1 and 11.2,

there are 0.082 bits of information in the first codon position relative to the background distribution for *H. influenzae*. Since most of the *H. influenzae* genome consists of coding regions, it makes little difference if the background distribution is measured genome-wide or coding-region-wide. There are 0.175 bits of information per residue in the first position of codons for *M. jannaschii*.

The total relative entropy for the entire codon is simply the sum of the relative entropies for the three positions. (See Section 8.3.) The number of bits per codon for the organisms *H. influenzae, M. jannaschii, C. elegans*, and *H. sapiens* are 0.12, 0.21, 0.09, and 0.12, respectively. For *H. influenzae* the number of bits of information for the second position is close to zero. In humans there is more information is in the second position.

11.4.1. Codon Biases

Recall from Section 1.6.2 that there are 64 possible mRNA sequences of length three, but there are only 20 amino acids plus the stop codon. Thus, there exist *synonymous codons* that encode the same amino acid. Another statistical clue for locating genes is whether an organism uses synonymous codons equally often or has a bias toward certain codons in its genome.

For example, in *H. influenzae* the codon TTT is used about four times as often as TTC, although both TTT and TTC encode the amino acid phenylalanine. One conjecture as to why this occurs is that the tRNA for TTT is more abundant than the tRNA for TTC. Recall from Section 2.2 that the tRNA carries an amino acid to the ribosome during translation. There is selective pressure on the organism to choose the codon that is most efficiently translated, which would be affected by tRNA abundance.

A similar study investigates if organisms prefer one amino acid over another, since some amino acids such as leucine and isoleucine are chemically similar. (See Section 1.1.1.)

11.4.2. Recognizing Genes

Codon bias can be applied to the problem of recognizing genes in a DNA sequence. Define a score for codon C as follow:

$$score(C) = \log_2 \frac{C_R}{C_B},$$

where C_R is the frequency of codon C in the coding regions and C_B is the frequency of codon C in the background.

The score of a sequence $C_1, C_2, ..., C_n$ of codons is defined to be the sum of the scores of each C_i. When recognizing genes, one facet would be to identify sequences with high scores. Each reading frame must be examined since moving the frame window to the right one position or two positions results in different sequences of codons.

One drawback to this technique is that we must know the coding regions (in order to estimate C_R) before recognizing (those same) genes in the genome. There are simpler methods for finding long coding regions, and once these are known they can be used to estimate C_R and thus used to find more genes.