Lecture 9

Relative Entropy and Binding Energy

February 1, 2000
Notes: Neil Spring

Binding energy is a measure of the affinity between two molecules. Because it is an expression of free energy released rather than absorbed, a large negative number conventionally represents a strong affinity, and suggests that these molecules are likely to bind. The binding energy depends on a number of factors such as temperature and salinity, which we will assume are not varying.

This lecture describes a paper of Stormo and Fields [3], which investigates the binding energy between a given DNA-binding protein and various short DNA sequences. In particular, it discusses an interesting relationship between binding energy and log likelihood weight matrices, shedding a new light on the relative entropy.

9.1. Experimental Determination of Binding Energy

Given a DNA-binding protein P, we would like to determine with what binding energy P binds to all possible length n DNA sequences. The "binary" question of whether or not P will bind to a particular DNA sequence oversimplifies a more complicated process: more realistically, P binds to most such sequences, but will occupy preferred sites for a greater fraction of time than others. Binding energies reflect this reality more clearly.

If c is the alphabet size, then one cannot hope to perform all the experiments to measure the binding energy of P with each of the possible c^n sequences of length n. Instead, Stormo and Fields proposed the following experimental method for estimating the binding energy of P with each length n sequence.

1. Choose some good site S of length n.
2. Construct all sequences of length n that differ from S in only one residue. There are $(c-1)n$ such sequences.
3. For each such sequence S', experimentally measure the difference in binding energy between P binding with S and P binding with S'.
4. Record the results in a $c \times n$ matrix G, where $G_{r,j}$ is the change in binding energy when residue r is substituted at position j in S.

Stormo and Fields then make the approximating assumption that changes in energy are additive. That is, the change in binding energy for any collection of substitutions is the sum of the changes in binding

energy of those individual substitutions. With this assumption, one can predict the binding energy of P to any length n sequence $s = s_1 s_2 \cdots s_n$ by the following formula:

$$\sum_{j=1}^{n} G_{s_j,j}.$$

Thus, G is a weight matrix that assigns a score to each sequence s according to the usual weight matrix formula given in Section 8.1.

9.2. Computational Estimation of Binding Energy

Unfortunately, creating the matrix G for every DNA-binding protein in every organism of interest still requires an infeasible amount of experimental work. This motivated Stormo and Fields to ask how to approximate G computationally, given a collection $\mathcal{A}$ of good binding sites for P and a collection $\mathcal{B}$ of nonsites.

Choosing W to be the log likelihood ratio weight matrix for $\mathcal{A}$ with respect to $\mathcal{B}$ assigns the highest scores to the sites in $\mathcal{A}$. Since G also assigns high (negative) scores to the sites in $\mathcal{A}$, there is good reason to expect that W approximates G well (after the appropriate scaling).

Recall from Section 8.3 that the relative entropy $D_2(A||B)$ is the expected score assigned by W to a randomly chosen site. If W approximates G well, the relative entropy $D_2(A||B)$ then approximates the expected binding energy of P to a randomly chosen site. This provides us a new interpretation of relative entropy.

It also provides an estimate of how great we should expect the relative entropy to be for a good collection of binding sites. There is some probability that a good site will appear in the genomic background simply by chance. This probability increases with the size Γ of the genome. If the relative entropy is too small with respect to Γ, the expected binding energy at true sites will be too small, and the protein will spend too much time occupying nonsites.

Stormo and Fields suggest from experience that the relative entropy for binding sites will be close to $\log_2 \Gamma$. A simple scenario suggests some intuition for this particular estimate: Assume a uniform background distribution $B_{r,j} = 0.25$, and assume that the site profile A has a 1 in each column, that is, all sites are identical. This imples that the relative entropy is 2 bits per position (as in two of the columns of Table 8.2), so the total relative entropy is $D_2(A||B) = 2n$. In a random sequence generated according to B, one would expect this site sequence to appear once every 4^n residues. In order for P not to bind to too many random locations in the background, $4^n = 2^{2n} = 2^{D_2(A||B)}$ must be not much less than Γ, so $D_2(A||B)$ must be not much less than $\log_2 \Gamma$.

9.3. Finding Instances of an Unknown Site

This leads us into our next topic. Suppose we are not given a sample $\mathcal{A}$ of known sites. We want to find sequences that are significantly similar to each other, without any *a priori* knowledge of what those sequences look like. A little more precisely, given a set of biological sequences, find instances of a short site that occur more often than you would expect by chance, with no *a priori* knowledge about the site.

Given a collection of k such instances (ignoring, for the moment, how to find them), this induces a profile A as described in Section 7.1. As usual, we compute a profile B from the background distribution.

From A and B, we can compute $D_2(A||B)$ as in Section 8.3, and use that as a measure of how good the collection is. The goal is to find the collection that maximizes $D_2(A||B)$. In particular, if we are looking for unknown *binding* sites, then the argument of Section 9.2 suggests that a relative entropy around $\log_2 \Gamma$ would be encouraging.

A version of the computational problem, then, is to take as inputs k sequences and an integer n, and output one length n substring from each input sequence, such that the resulting relative entropy is maximized. Let us call this the *relative entropy site selection problem*. Unfortunately, this problem is likely to be computationally intractable (Section 6.4):

Theorem 9.1 (Akutsu [1, 2]): The relative entropy site selection problem is NP-complete.

Akutsu also proved that selecting instances so as maximize the sum-of-pairs score (Section 6.2) rather than the relative entropy is NP-complete.

References

[1] T. Akutsu. Hardness results on gapless local multiple sequence alignment. Technical Report 98-MPS-24-2, Information Processing Society of Japan, 1998.

[2] T. Akutsu, H. Arimura, and S. Shimozono. On approximation algorithms for local multiple alignment. In *RECOMB00: Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*, Tokyo, Japan, Apr. 2000.

[3] G. D. Stormo and D. S. Fields. Specificity, free energy and information content in protein-DNA interactions. *Trends in Biochemical Sciences*, 23:109–113, 1998.