ICS 284a: Algorithms for Computational Biology
Notes on Lecture 2
Gene Regulation and Motif Discovery

Todd Johnson
based on presentation by Xiaohui Xie

January 13, 2008

1 Gene Regulation

Much of the study of molecular biology is concerned with the processes by which genetic codes are transformed into living organisms. As such, the various processes of gene regulation, transcription, and translation must for the basis of any investigation in the field.

1.1 The Central Dogma

The most fundamental idea in molecular biology, often called *the central dogma* is the two-step process which leads from DNA to protein. The idea is that DNA is used as a template to produce RNA, via the process of *transcription*, and then RNA is used to produce protein, via the process of *translation*.

The central dogma has been modified over time by the discovery of ancillary processes and constrtucts, such as microRNA, reverse transcriptase, and alternative splicing; nonetheless, it is still crucial to our understanding of molecular biology.

1.2 Transcriptional Regulation

Proteins called *transcription factors* (TFs) bind to regions upstream of genes, and either promote or inhibit gene transcription. Amino acids in the TFs recognize particular patterns in the DNA, and bind to them. Then, the presence of the protein can affect transcription. If the transcription factor's effect is to inhibit transcription, it may simply physically block transcription.

If the effect is to enhance transcription, it may act to recruit the molecular machinery necessary to begin transcription. One factor which complicates this picture is that amino acid-DNA recognition is not straight-forward.

Three types of TFs are zinc finger, helix-turn-helix, and leucine zipper.

1.3 Regulatory Motifs

Sequence motifs are patterns of nucleotides which are common in a genome, or appear to be evolutionarily conserved. Because of their prevalence, these sequences are often believed to have biological significance.

In particular, *regulatory motifs* are patterns which are recognized, and in some cases bound, by transcription factors, transcriptional coactivators, and other transcriptional regulators. These patterns are not transcribed, but are nevertheless necessary for synthesis of proteins.

Transcription factor binding sites are the particular regulatory sequences to which transcription factors bind.

2 Regulatory Motif Discovery

Given the importance of transcriptional regulation, one natural question is, "how can we identify regulatory motifs in sequenced genomes?" One way to solve this problem is to collect many sequences which are upstream of the start codons for genes, and therefore are believed to be likely places for finding regulatory motifs, and to then attempt to find subsequences which appear in many of these. If we can find a string which appears many more times than random chance says it should, then we may have found a good candidate regulatory motif.

This, then, gives us a new problem: given a set of sequences, how can we find the subsequences which appear many times. One method is *enumeration*.

2.1 Enumeration

The idea behind enumeration is to restrict ourselves to a fixed size of potential motif, and to list all of the possible motifs of that length, and simply count the number of different sequences in which each appears.

More formally, we suppose that we have a set S of N sequences, $s_1, s_2, \ldots, s_N$, each of length l. We choose a length $w < l$, and create a list of M possible motifs with length w, called $m_1, m_2, \ldots, m_M$. If we consider every possible motif with length w, then $M = 4^w$.

Now, for each sequence s_i and each motif m_j we create an *indicator variable* z_{ij}. Then, if m_j appears as a substring inside of s_i we set z_{ij} to 1, and otherwise we set it to 0. If motif m_j appears more than once inside of s_i, z_{ij} is still only set to 1.

Next, we associate with motif m_j a *summation variable* k_j, where $k_j = \sum_{i=1}^{N} z_{ij}$, the total number of different sequences in which m_j appears.

Now we can say that the most likely regulatory motif, m^*, is the one with the largest summation variable, k^*. This process can be thought of as drawing a histogram of the number of appearances of each motif, and then ordering the motifs with the largest values in the histogram, and conjecturing that this represents the order in which each motif is likely to be biologically conserved.

2.2 Measuring Significance

Assuming that we have found the most likely motif, as in the above example, can we go further to say how confident we are that this sequence is, in fact, biologically conserved, and therefore significant? Put another way, can we say how likely (or, hopefully, unlikely) it is that the motif m^* would appear in k^* out of n sequences of length l simply by chance? If this probability of chance is very low, then we can be confident that the motif is genuinely interesting.

To begin with, let us ask: What is the probability that, for a given sequence s_i and motif m_j, $z_{ij} = 0$? This is just the probability that the motif does not appear inside the sequence. We can compute that by first computing the probability that the motif does not match the first w characters of s_i, and then repeat this probability as many times as there are sets of w characters which m_j could match.

The probability that m_j is not matched is just one minus the probability that it is matched, which is one over the total number of strings of length w. That is, the probability that the first w characters are not equal to m_j is $1 - \frac{1}{4^w}$.

Then, since there are $l-w+1$ strings of length w inside a string of length l, the probability that m_j doesn't appear inside s_i at all is $(1 - \frac{1}{4^w})^{l-w+1}$. If l is much larger than w, and if w is large enough that $\frac{1}{4^w}$ is much less than 1, then this can be approximated by $\frac{l-w+1}{4^w}$.

Now that we have the probability that a motif of length w does not appear inside of a given sequence of length l, we know that the probability p that the motif *is* present is $p = 1 - (1 - \frac{1}{4^w})^{l-w+1}$. With this, we can also compute the probability that the motif appears in exactly k out of

N sequences. The test "does sequence m_j appear in sequence s_i" can be viewed as a Bernouli trial with probability p of success, and we know that the probability of success on k out of N Bernouli trials is governed by the binomial distribution

$$p(k_j = k) = \binom{N}{k} p^k (1-p)^{N-k}$$

.

In order to test the significance of our motif m^* appearing in k^* sequences, we really want to evaluate the *p-value* of this occurance. For this, we need the probability that m^* appears in *at least* k^* sequences. If we use the shorthand $p(k)$ for $p(k_j = k)$, this value is given by $p(k^*) + p(k^* + 1) + \cdots + p(N)$, or

$$\sum_{k=k^*}^{N} p(k) = \sum_{k=k^*}^{N} \binom{N}{k} p^k (1-p)^{N-k}$$

.

If N becomes large, this binomial distribution is well approximated by a Gaussian distribution with the same mean and variance as the binomial distribution. The mean of a binomial distribution with parameters N and p is Np, and the variance is $Np(1-p)$. With this knowledge in hand, we can map the value of k^* into a new distribution, but one which is now Gaussian with mean 0 and unit variance. We do this by computing the *z-score* of k^*, given by $\frac{k^* - Np}{\sqrt{Np(1-p)}}$.

This saves us from having to compute the sum from k^* to N, because the process of mapping data to a Gaussian distribution with 0 mean and variance of 1 gives a standard, well-understood way of measuring significance. The process of integrating over this distribution has been done before, and is available in tables of z-scores and their associated cumulative probabilities.

Therefore, once we have a z-score, we can immediately make statements of the form "this outcome would occur purely by chance one time in X", and if X is large enough (or $\frac{1}{X}$ is small enough), then we believe that having observed this outcome is signficant. We do this by taking our z-score to a table and looking up the value associated with our z-score in that table. In such a table, we will find the cumulative probability of our z-score. This number represents the probabilities of all of the more-likely events occuring, summed together. In our case, this means the total probability of finding fewer than k^* copies of the motif. If this number is very high, then the probability that we have observed k^* copies simply by chance is very low, and we can say with some confidence that the motif m^* is probably biologically significant.