Gene Regulation and Motif Discovery
Scribe notes for lecture 2 of CS284A

Notes by James Foulds (Jimmy),
based on a lecture by Xiaohui Xie

October 4, 2008

1 Gene Regulation

The task of identifying regulatory motifs is an important problem in computational biology. In this section we first introduce the relevant biological concepts. A basic method for solving the motif discovery problem is described in Section 2.

1.1 The Central Dogma

Genetic information is stored in living organisms in DNA, RNA and protein molecules. The primary structure of these molecules is a linear chain, and hence they can be viewed as sequences of letters. This sequential biological information can be transferred between DNA, RNA and protein molecules. The so-called "central dogma of molecular biology" describes which types of transfers are allowed: information cannot be transferred from protein to either protein, DNA or RNA, but all other transfers are allowed.

Under normal conditions, only three transfers typically occur: DNA to DNA (*DNA replication*), DNA to mRNA (*transcription*), and RNA to protein (*translation*). This article is about transcription, where DNA information is copied to *messenger RNA* (mRNA), a type of RNA which encodes the information required to create a protein product.

1.2 Transcriptional Regulation

Genes (the functional parts of our DNA) can be divided into coding regions and non-coding regions. Coding regions contain the genetic information that is transcribed into RNA. In the non-coding regions, there are binding sites which affect the transcription process. When DNA-binding proteins bind to these sites, they influence the transcription process. When proteins bind to an *activator* binding site, the rate of transcription is increased, while proteins binding to *repressor* binding sites reduce or prevent transcription. The term *transcriptional regulation* refers to such processes that control the rate of transcription. The DNA-binding proteins involved in transcription regulation are called *transcription factors*.

1.3 Regulatory Motifs

Transcription factors are capable of binding to specific parts of DNA, determined by the DNA sequence information. Hence, strings in the DNA sequence that encode these *transcription factor binding sites* can be directly related to gene regulation. Such strings are called *motifs*. Biologists are interested in understanding gene regulation, and hence are interested in identifying these motifs.

It is known that motifs are generally short (roughly 6 — 20 binding pairs). Motifs are located somewhere within the vicinity of the coding region, usually *upstream* (earlier in the sequence).

2 Regulatory Motif Discovery

To introduce the motif discovery problem, we will consider an example scenario. A scientist is studying gene regulation in yeast. Using a gene chip microarray, she has recorded the level of expression of each gene, in terms of the number of copies of the mRNA sequences, at 15 regular intervals throughout a single cell cycle. Thus, for each gene she has a *profile*, which can be visualized as a graph of expression level over time, with 15 data points.

Using a clustering algorithm such as k-means, she has identified a set of genes that have similar profiles. The scientist hypothesizes that the similar patterns of expression that these genes share may be triggered by a common transcription factor binding site; i.e. the genes share a common motif. To

test this hypothesis, she needs to find candidate motifs, and evaluate the significance of the observed motif occurences.

The problem can now be stated formally as follows: We have a set S containing N sequences of length L over the alphabet $A = \{A, G, T, C\}$. Each sequence represents the sequence data in a fixed-size window upstream from the coding region of a gene. The goal is to identify subsequences that occur in significantly more of the N sequences than we would expect to occur by random chance, i.e. are *overrepresented*. If a string is overrepresented, this is evidence that it may be involved in the regulatory processes for the set of genes.

2.1 Enumeration

A very simple method for finding motifs is to simply consider all subsequences of a given length k (referred to as k-mers), and find the one that occurs in the most elements of S. Since there are 4 letters in our alphabet, there are 4^k k-mers to consider. For each k-mer, we simply iterate over the N sequences, and add one to a count for that k-mer for each sequence that containts that k-mer as a subsequence. If we assume that all k-mers are equally likely to occur, the k-mer with the highest count value is the most likely to be a motif. We can visualize this as constructing a histogram of the occurrences of each k-mer in distinct sequences, and simply selecting the k-mer with the largest value in the histogram.

Although we only consider the simplest form of the enumeration strategy here, the method can easily be extended to consider degenerate patterns, such as by combining smaller overrepresented words into a more flexible motif description. In practice, the assumption that all k-mers are equally likely to occur is not particularly accurate, and overrepresentation must be assessed with respect to a more sophisticated statistical background model, which could for instance be a Markov model derived from the count data.

2.2 Measuring Significance

It is important to evalute potential motifs with respect to statistical significance — how surprised are we that a k-mer m occured in j of the N sequences? If it is very unlikely that this result was observed due to random chance, this supports the hypothesis that m is a true motif.

Here, for simplicity we will rely upon the unrealistic assumption that each letter is equally likely: $(p(A), p(G), p(C), p(T)) = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. Let us first consider one sequence by itself: how likely is a particular k-mer to occur (at least once) in a sequence s of length L?

Since all k-mers are equally likely, and there are 4^k of them, it is easy to see that the probability that a specific k-mer occurs in a sequence of length k is $\frac{1}{4^k}$. The probability that it does not occur in such a sequence is hence $1 - \frac{1}{4^k}$. There are $L - k + 1$ subsequences of length k in the sequence s, so the probability that it does not occur at all in the sequence is $(1 - \frac{1}{4^k})^{L-k+1}$, and hence the probability p that it *does* occur in s is one minus that, i.e.

$$p = 1 - (1 - \frac{1}{4^k})^{L-k+1} .$$

The test "does a sequence $s_i \in S$ contain a subsequence m" can be viewed as a Bernoulli trial (i.e. similar to a weighted coin toss) with probability p of success. The probability distribution of repeated independent Bernoulli trials is binomial, i.e. the probability of j out of the N sequences containing the word m can be described by

$$P(j) = \binom{N}{j} p^j (1-p)^{N-j} .$$

To evaluate the significance of our observation, we need to compute a p-value, which represents the probability of the occurence of an event at least as extreme as the observed result, under the assumption that the null hypothesis is true (all letters are equally likely). This can be evaluated by summing the probabilities of all of the outcomes where the number of occurrences is j or higher, i.e.

$$p \text{ value} = \sum_{i=j}^{N} P(i) = \sum_{i=j}^{N} \binom{N}{i} p^i (1-p)^{N-i} .$$

If the p value is very small, this is evidence that the null hypothesis is false, as it indicates that the observed result is very unlikely to occur given the null hypothesis. Normally, before looking at the data we decide upon a significance level α, which is the highest value of p for which we will reject the null hypothesis. The value $\alpha = 0.05$ is most often used in the scientific literature, corresponding to a five percent chance that the result was observed given the null.

If the p value is less than α, the result is sufficiently unlikely given the null hypothesis for us to conclude that the null hypothesis is probably false. This gives credence to the alternative hypothesis, that the word m is a true motif that is somehow involved in regulation.

References

P. Baldi and S. Brunak. Bioinformatics: The Machine Learning Approach (second edition, pp. 320 — 321). MIT Press (2001).

P. D'haeseleer. How does DNA sequence motif discovery work? Nature Biotechnology – 24, 959 — 961 (2006)

Wikipedia, Central dogma of molecular biology.
`http://en.wikipedia.org/wiki/Central_dogma` , retrieved 10.04.08

Wikipedia, P-value. `http://en.wikipedia.org/wiki/P-value` , retrieved 10.04.08

X. Xie. Motif Discovery Algorithms. Lecture slides available at
`http://www.ics.uci.edu/~xhx/courses/CS284A/lectures/CS284A_Lecture2.pdf`