Correlating Expression Data and Genomic Sequence Data

1 Introduction

In general terms, our goal is to find relationships between a gene's expression profile (for now, this means mRNA abundance) and observable patterns in the genomic DNA sequence around that gene. For example, genes that have a match to the pattern `TGACTC` within 1000 bp upstream of their translation start site might tend to show increased mRNA levels under a certain set of conditions. Specific strategies include:

1. Start with a cluster of genes that show similar expression profiles. Search their upstream regions (say, the first 1000 bp upstream of their translation start site) for some sequence "pattern" that occurs more frequently than can be explained by chance alone. Instances of this pattern might be signals that control mRNA levels.

2. Start with a DNA sequence "pattern", and find all genes having a match to the pattern in their upstream region. Determine whether the expression profiles of those genes show a higher degree of similarity than can be explained by chance alone. For instance, taking "pattern" to mean a given 6-mer, it is feasible to do this with all 4096 possible patterns and ∼6200 yeast genes. This identifies patterns that may be correlated with mRNA levels.

3. As in the previous example, start with a sequence pattern and an upstream match to the pattern. Look for a set of experimental conditions (columns) where the genes behave differently depending on whether or not they have an upstream match to the pattern. For instance, you might learn that the site is associated with liver-specific genes.

4. Since a single sequence "signal" may need to occur in concert with other signals to exert a regulatory effect, more complicated strategies may be needed, like the following. Start with a DNA sequence "pattern" and the set of genes with upstream matches to it. Attempt to identify one or more additional patterns such that the subset of genes having upstream matches to both (or all) patterns show similar expression patterns.

But what should we mean by a "pattern" in genomic DNA sequence? The feasibility of studies like the four mentioned above will depend critically on the answer. We need to look at the range of biological mechanisms that influence mRNA levels and see what forms are taken by the relevant "signals" that can be identified from genomic DNA sequence data.

As a brief digression, it might be worthwhile to consider a transcriptome-proteome comparison, i.e., a simultaneous analysis of expression data and alignments of protein sequences. This could be done with two species, e.g., to look for orthologous proteins having different expression patters. It might even be done within a single species, e.g., comparing expression patterns of paralogous proteins to explore the hypothesis that gene duplication followed by a change in expression pattern is a common mode of genome enlargement. I don't know of studies like this, but we're sure to see them in the future. In the mean time, back to DNA sequences.

2 Regulatory Signals in Genomic DNA Sequence

The number of mRNA molecules from a particular gene that are found in a cell under particular conditions depends on the rates of both production and degradation. The latter is frequently influenced by signals in the $3'$ UTR, but I don't know of any studies that relate occurrences of those signals to microarray data.

The more important factor is rate of production, i.e., of transcription. The cell can control this by a variety of mechanisms, only some of which leave an observable signal in genomic sequence data.

- In some organisms, transcription is affected by chemical modification, particularly methylation, of certain nucleotides. Over evolutionary time this can affect the DNA sequence composition, since `CpG` dinucleotides tend to become `TpG` if the `C` is methylated. Hence, "CpG islands", i.e., regions of a few hundred basepairs that don't show a strong depletion of `CpG` dinucleotides, give a DNA-sequence "signature" that one might attempt to correlate with mRNA abundance.

- In some organisms, transcription is affected by chemical modification of certain proteins bound to the DNA sequence, such as acetylation of histones. It is unclear how genes regulated this way might be identified from DNA sequence data. (Perhaps there are subtle nucleosome positioning signals.)

- It seems plausible that transcription frequency is affected by the length of the gene, number of exons, distance to neighboring genes, GC content, and a number of other aspects of genomic layout. However, short-duration changes like those typically measured in microarray experiments aren't made by changing these factors.

- In all probability, the single most influential mechanism for controlling transcription rates over short time periods is the binding of certain proteins, called *transcription factors* to genomic sequence, frequently within a few hundred basepairs upstream of the transcription start site. (Technical note: it is typically much easier to predict the start site for translation than the start site for transcription, so the former is frequently substituted in practice.)

Summary: The vast majority of current work comparing mRNA abundance with genomic sequence patters focuses on transcription factor binding sites. This leads us to the question of what sorts of sequence "patterns" can be used to characterize such sites.

3 Transcription Factor Binding Sites

Nature has invented a variety of "DNA binding domains" that are used by transcription factors, such as "homeodomains" and "zinc fingers." Frequently, the DNA binding domain of a transcription factor recognizes short segments of genomic DNA, e.g., of length 5-8, and generally the sequence motif that it recognizes and binds is somewhat variable.

One well-studied transcription factor is called GCN4. Inspection of 22 experimentally confirmed binding sites revealed a typical pattern of `TGACTC`. However, the positions differ in the degree to which the letter seems to be required. For instance, the first `T` and the first `A` (position 3) looked mandatory, while the last `C` was observed in just over half of the examples. Actually, it makes sense that not all binding sites are equivalent; cells need different amounts of the various proteins regulated by a given transcription factor, and nature's solution might be binding sites of different affinity.

A more precise summary of the observed binding sites, one that captures binding-site variability, is provided by the following *frequency matrix*, which gives the number of times that each position was occupied by each possible nucleotide.

A	0	0	22	1	0	5
C	0	0	0	17	0	12
G	0	21	0	2	1	3
T	22	1	0	2	21	2

However, for GCN4, requiring an exact match to the consensus 6-mer, `TGACTC`, does a reasonably good job of identifying binding sites.

GAL4 is another well-studied yeast transcription factor. (We'll hear more about it later in the semester, when we discuss "yeast two hybrid" experiments.) When yeast cells are shifted to galactose (a kind of sugar) from another carbon source, they need to activate a number of genes to process the new food. An important step in this pathway is the activation of the GAL4 transcription factor, which in turn up-regulates a set of genes having upstream GAL4 binding sites. In other words, a microarray experiment involving the switch to galactose should identify one or more clusters of genes with a similar expression profile (up-regulation), for which upstream GAL4 binding sites are over-represented.

GAL4 binding sites have been studied in great detail, including X-ray crystallography of the GAL4-DNA binding complex, identification of functional GAL4 binding sites, and *in vitro* measurements of the binding affinity to various oligonucleotides. To a first approximation, a GAL4 binding sites has reverse-complemented `CGG` triplets at each end, separated by 11 rather arbitrary nucleotides, i.e., CGGN^{11}CCG, using `N` to denote an unspecified nucleotide. More sensitive recognition of these sites can be obtained with a pattern that utilize IUPAC symbols for ambiguous nucleotides to express the weaker requirements at internal positions of the GAL4 binding site. For instance, the nucleotide in position 8 appears to almost always be `C` or `G` (denoted `S`), while position 9 is `A` or `T` (denoted `W`), which suggests the pattern CGGN^{4}SWN^{5}CCG.

Even greater precision can be gained by using a frequency matrix, which can express e.g., the fact that the nucleotide at position 8 is more often `C` than `G`. The following matrix gives the frequency with which the various nucleotides occur at each position in the GAL4 binding site, normalized so that the sum of each column is 10.

A	0	0	0	4	1	1	7	0	5	1	0	2	0	2	0	0	0
C	10	0	1	2	3	5	0	7	0	4	2	5	5	1	9	10	0
G	0	10	9	4	5	3	2	3	0	3	1	1	4	1	1	0	10
T	0	0	0	0	1	1	1	0	5	2	7	2	1	6	0	0	0
position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17

In particular, note that a search for over-represented 6-mers is not expected to work for finding GAL4 binding sites.

4 Using a Frequency Matrix to Predict Binding Sites

Currently, the most popular computational approach for predicting transcription factor binding sites is to use a *position weight matrix* (PWM) derived from the frequency matrix for each transcription factor. To get started, consider a hypothetical frequency matrix for 5-mers:

A	1	6	5	1	4
C	2	1	1	2	1
G	2	1	3	3	4
T	5	2	1	4	1

(We have invented a matrix containing no zeros, as an initial simplification.)

Suppose we want to decide whether a particular 5-mer, say `TAGTA`, matches this pattern. Perhaps the most obvious approach is to add up a score for each letter that is taken from the corresponding place in the table. For `TAGTA` the sum is $5 + 6 + 3 + 4 + 4 = 22$, which is not too much worse that the maximum value of 24 for `TAATA` or `TAATG`. If we take, say, 20 as a threshold score for predicting a match, the 5-mer matches the pattern.

In practice, people use a somewhat different approach, which is less *ad hoc*. The reasoning goes as follows. Let us convert the matrix into something that looks like probabilities by dividing each column by its sum (10 in this case):

A	0.1	0.6	0.5	0.1	0.4
C	0.2	0.1	0.1	0.2	0.1
G	0.2	0.1	0.3	0.3	0.4
T	0.5	0.2	0.1	0.4	0.1

Think of this as specifying a "random" 5-mer that matches the pattern. Thus, in a matching 5-mer, the first letter is `T` with probability 0.5, it is a `C` with probability 0.2, and so on. The probability that a 5-mer matching the pattern happens to be `TAGTA` is $0.5 \times 0.6 \times 0.3 \times 0.4 \times 0.4 = 0.0144$. We want to compare this with the probability that the 5-mer would be generated in a "null model" where each nucleotide is equally likely to occur at each position. (We could use a null model that considered observed nucleotide frequencies.) Here, the probability of any particular 5-mer is $4^{-5} \sim 0.0009765$. Thus, the "odds ratio", determined by dividing the probability that the 5-mer would be generated by our model for the frequency matrix by the probability that it would be generated randomly, is about $0.0144/0.0009765 \sim 14.7$.

Another way to obtain this value for an arbitrary 5-mer is to divide each entry in the matrix of probabilities by the probability of the letter in a random sequence (i.e., multiply by 4 under our simple assumptions), giving:

A	0.4	2.4	2.0	0.4	1.6
C	0.8	0.4	0.4	0.8	0.4
G	0.8	0.4	1.2	1.2	1.6
T	2.0	0.8	0.4	1.6	0.4

To get the score of a 5-tuple, one multiplies the table entries, e.g., for `TAGTA` we get $2.0 \times 2.4 \times 1.2 \times 1.6 \times 1.6 \sim 14.7$.

To remove the needed to perform multiplications, we can replace each entry in the matrix by its logarithm, getting:

A	-0.916	0.875	0.693	-0.916	0.47
C	-0.223	-0.916	-0.916	-0.223	-0.916
G	-0.223	-0.916	0.182	0.182	0.47
T	0.693	-0.223	-0.916	0.47	-0.916

Now we can add entries instead of multiplying, since $\log(XY) = \log(X) + \log(Y)$. For TAGTA we get $0.693 + 0.875 + 0.182 + 0.47 + 0.47 = 2.993$. The result is the *log odds ratio*, i.e., the logarithm of the above-mentioned odds ratio.

Care must be taken when 0 is an entry of the frequency matrix, since the logarithm of 0 is undefined. A simple-minded approach is to add 1 to every position of the frequency matrix before starting. Of course, there are statistically smarter alternatives.

Such a matrix, containing entries formed by taking the logarithm of the ratio of the probability that the letter occurs in the corresponding position of a "true k-mer" divided by the probability that it occurs randomly, is frequently called a *position weight matrix* (PWM). The usual way of modeling a transcription factor binding site is to determine an appropriate PWM and score threshold.

5 Binding Sites Work in Concert

On reflection, it seems quite implausible that a transcription factor binding site could function in isolation, particularly in multi-cellular organisms. For simplicity, suppose that a necessary and sufficient condition for binding of a given protein to DNA is occurrence of an exact match to a certain 6-mer. Then on average, one of every four genes will have such a match in its upstream 1000 bp, purely by chance. (There are 4096 possible 6-mers.) Moreover, the number of times that a given 6-mer occurs in the human genome is about 3 million, on average, and a cell won't contain enough molecules of the transcription factor to bind all of them. Finally, the number of regulatory patterns among different genes (say, in humans) is far greater than the number of different transcription factors. Thus, it is inadequate to use a simple-minded explanation of regulation, where binding of a single transcription factor to a precisely-defined signal in the genomic sequence regulates a gene's transcription.

The "signal" in the genomic DNA sequence that causes a certain transcriptional profile, e.g., that constitutes necessary and sufficient conditions for a gene to be expressed in muscle tissues, may consist of a cluster of binding sites for certain transcription factors, with certain constraints on the relative order and possibly orientations of the sites. The following are currently active research areas:

- For a particular tissue and developmental stage, be able to predict from genomic DNA sequence whether a given gene is expressed under those conditions. For instance, consider the recent paper:

  > Berman, B. P, Y. Nibu, B. Pfeiffer, P. Tomancak, S. Celnicker, M. Levine, G. Rubin and M. B. Eisen (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *PNAS* **99**, 757-762.

  By looking for 700-bp intervals containing a dense cluster of matches to (PWMs for) five different transcription factor binding sites, the authors were able to accurately identify genes active in the early fruit-fly embryo. The paper

> Krivan, W., and W. W. Wasserman (2001) A predictive model for regulatory sequence directing liver-specific transcription. *Genome Research* **11**, 1559-1566.

had some success at characterizing expression in the liver using 200-bp intervals and four different transcription factors.

- Given upstream regions from a set of co-regulated genes, predict binding sites and combinatorial rules (e.g., spacing constraints) potentially involving several transcription factors that govern the expression pattern. A few recent papers on this topic are available at the class website, but this is a highly technical area.