Overview

A. Gene Expression and Regulation

We will cover some basic biological background

B. Measuring Gene Expression: Microarrays

Microarrays are not the only high-throughput experimental method of assessing gene expression, but they are the most important one & will be the focus of today’s lecture; the other methods will not be discussed.

C. Finding Regulatory Motifs

We will discuss the computational problem of motif finding (finding short, inexact words within biological sequences), and its connection to gene regulation.

Gene Expression and Regulation

- A genome is static
  - Every cell in our body has a copy of same genome
- A cell is dynamic
  - Responds to external conditions
  - Most cells follow a cell cycle of division
- Cells differentiate during development
- Gene expression varies according to:
  - Cell type
  - Cell cycle
  - External conditions
  - Location

Examples of the typical external conditions for which cell response might be assessed would include heat, food, poisons. Cell cycle studies might examine gene expression during different stages of cell division, or at cell death such as apoptosis. The topic of gene regulation examines the question of how a cell manages these responses/changes.
Gene regulation can occur at any of the above stages. Today our main focus will be on the transcription phase, because transcription phase regulation is a very efficient method of regulation and is a main regulatory mechanism used by cells. (Nota bene: to this scribe it also appears that transcription phase regulation is simply the best understood and most readily analyzed of the various forms of regulation.) Please note that transcription phase regulation includes not only the on/off control of transcription common to all cells, but also the control of gene splicing found in eukaryotes.

Transcriptional Regulation

- Efficient place to regulate
  - No energy wasted making intermediate products
- However, slowest response time
  - After a receptor notices a change:
    1. Cascade message to nucleus
    2. Open chromatin & bind transcription factors
    3. Recruit RNA polymerase and transcribe
    4. Splice mRNA and send to cytoplasm
    5. Translate into protein

Note: Probably because of the slow response time inherent to transcription phase regulation, many cellular responses to external stimuli are regulated after transcription, especially via protein modification.
Note: Transcription factors are primarily protein structures (with corresponding genes) and their binding to DNA is via general chemical interaction, NOT via Watson-crick base pairing.

For transcription to occur, first the DNA chromatin must be opened to allow access (but this is a relatively large portion that gets opened). The polymerase that actually performs transcription initially binds to a promoter sequence. Promoters are generally somewhat, but not very, specific – some tens of nucleotides just before the start of the gene. Because the attraction between the polymerase and the promoter is not especially strong, it is aided by several additional protein factors bound to “enhancer” regions on the DNA; these additional transcription factors together with the polymerase form a large transcription complex. The level of enhancement provided by these additional
transcription factors can be by orders of magnitude. They can also be quite distant from
the promoter site (up to hundreds of thousands of nucleotides in higher organisms.
Enhancer transcription factors have a positive effect on transcription; other factors known
as suppressors can have a negative effect on transcription, and there is a third class
known as insulators (whose function was not explained).

The chemico-biological operation of enhancers, suppressors and insulators is not well
understood so far. The manner in which they act has not been well characterized, nor are
their generally recognized good methods for determining their locations. To give an idea
of how much remains to be understood: about 1.5% of the human genome has been
labeled as genes (which is believed to be a fairly accurate labeling today). Of these
genes, it is believed that some 50% to 80% have alternative splicings, but very few of
these are actually known. Additionally, analysis of genetic conservation between
mammals shows high conservation of approximately 5% - 8% of the genome, which
suggests that the “dark matter” portion of this beyond the labeled genes (3.5% - 6.5%), is
in fact functional, i.e., regulatory control elements.

In the above example, we see that there are several enhancer regions upstream of this
human heat shock protein that control the logic of when it is expressed and the degree to
which it is expressed. The TATA and CCAAT motifs are called constitutive because
they directly affect the polymerase and therefore always apply. They are also quite
common – the TATA box appears with about half of human genes. The other factors are
generally more distant.
DNA Microarrays

Microarrays are a way of discovering the process of gene regulation in a high throughput manner, and became available in last decade. Microarray can be thought of as a list of all the genes that we want to measure in an organism. The position of a gene in microarray doesn’t have specific meaning. Each location corresponds to a DNA sequence whose quantity we want to measure; more specifically, it has the complement of the sequence whose quantity we want to measure. So we can have a microarray with a location for each of the ~22,000 “known” human genes. So if we have a sample of human mRNAs within a cell, we hybridize the mRNAs on the chip to their complement as shown in the following slide.
Our sample includes a mix of different mRNAs expressed at different levels. We use PCR to make cDNAs from the mRNAs, then soak our chip in the solution with the cDNAs, which then hybridize to our array locations & let us measure the level of expression.

There are two main kinds of arrays: oligo-nucleotide arrays and cDNA arrays. Oligonucleotide arrays are believed more robust by Serafim, but there’s a lot of debate about which is best. Oligo types are made by Affymetrix and maybe others. The oligo types probe a gene by making many, many short sequences for different substrings of each gene. cDNA technology, on the other hand, hybridizes the entire complement of an mRNA.
There are also two general methods of how to take readings – one is taking your sample to be the contents of a cell, the 2nd is to compare expression in a baseline/control group versus the cells of interest. In the latter case use two different labels, typically the control is green & target is red. They are hybridized to the chip together & give a reading of green & red locations (or yellow where they are mixed) showing relative over and under expression of each gene relative to the control sample.

One goal might be to study a particular biological processes, such as a stage of the cell cycle or cellular heat response. Instead, we might do a hypothesis tree study, where cells from yeast or some other organism are treated with many, many different conditions & then the resulting expression levels are data-mined for information regarding the control circuitry. So, given an experiment, we can use the genes as rows & conditions as
columns to assist with visualization or other structuring of the data. It is also possible to go to public websites to find results from hundreds of microarrays for each of various organisms. If we see genes moving in tandem, they may be subject to same control. There are not a lot of transcription factors per organism—maybe 100-200 for yeast and perhaps about the same for humans, so searching for potentially common control factors when common behavior has been observed is a feasible search.

There are two main kinds of analysis performed with microarray data: (i) clustering, where we try to find out what is important with each group, such as a common binding site, and (ii) classification (very important medically).

In classification we start with two kinds of samples, such as from healthy patients versus patients with a certain type of cancer. We do microarray for tens of patients in each sample. Then we use classification techniques from machine learning, such as support vector machines, so each sample is treated as a vector of gene measurements, and try to identify a separating hyper-plane that separates healthy patients from sick. We can then assess its quality for classifying patients. A well known paper showed using this to predict outcomes for leukemia patients.

Today we’ll focus on clustering.
There are several types of clustering; the most commonly used for microarrays is hierarchical clustering & the second is K-means. UPGMA (recall from phylogenies) is a type of hierarchical clustering. In any of these you have a measure of distance between the items you want to cluster (in our case the expression levels), which can be by gene, or you can cluster the conditions, or you can cluster by both genes and conditions.

One important characteristic affected by the definition of inter-cluster distance is the overall time complexity of the algorithm. The above distance metrics can be characterized as follows—the single link method tries to find clusters with a lot of empty space between them; the complete link method tries to keep clusters roughly spherical, and the average link method is the same as the UPGMA method we studied previously. The centroid method is very similar to the average link method (but this scribe believes that it does not take into account the relative sizes of the two clusters as does the UPGMA method).
A LOT of clustering work has been done in the microarray field, and virtually any aspect of clustering analysis can be found with applications to microarrays. One program very commonly used to analyze microarrays is the CLUSTER program, which uses hierarchical clustering and provides good visualization. It is important to note that hierarchical clustering requires $O(N^2 M)$ time, where $N$ is the number of characteristics being clustered and $M$ is the number of measurements, typically the genes and conditions, respectively. So hierarchical clustering can be quite slow.

k-means basically involves finding the “best” centers for a fixed number of clusters. A lot of tricks can be used here, such as merging clusters. The running time is $O(KNM)$, typically much better than hierarchical clustering.
K-Means Algorithm

- Randomly Initialize Clusters

- Assign data points to nearest clusters

- Recalculate Clusters

- Repeat

- Repeat ... until convergence

Time: O(kNM) per iteration
N: #genes
M: #conditions
Similar to K-Means clustering, but much better and more robust, is the “mixture of Gaussians” method. In the slide above, N corresponds to the number of experiments/conditions (M in the previous slides before it).

Because it is hard to learn such a large covariance matrix, simplifications are often done, such as using a diagonal covariance matrix, or even uniform covariance.

A program available from Avery Gasch and Mike Eisen uses mixture of Gaussians to analyze microarray data, although it calls it “fuzzy K-means clustering”, probably to be more readily understood by biologists.
There are several biological databases that label genes according to biological process; we can use these to assess the significance of our clusters: So, for at least some of our datapoints (genes or conditions) we are given labels. We want to assess if our clusters are better at grouping than random chance. If we had 25 different categories, then we would have a value for each category, either + meaning positive for that category, or “-“ for all the others – p labeled plus, and N-p labeled minus. The combinatorial analysis on the “Evaluating Clustering Analysis” slide then gives us a probability that this number of positives in a given cluster could have occurred by chance.

Why treat this as clustering rather than classification? Mostly because our labels may only be describing a small number of the processes/conditions/genes we are interested in.

Finding Regulatory Motifs
The above are two common formulations of the motif discovery problem. This will be discussed more in the next lecture.

For the combinatorial method, consider a possible motif \( M \), let the difference between \( M \) and any particular sequence \( S_i \) be the number of differences between \( M \) and the best match within \( S_i \). Find \( M \) with the minimum sum of distances over all \( S \).

For probabilistic method, we use a profile with each position \( 1:w \) in the motif \( M \) having four probabilities, one for each letter. The fitness of match between the motif profile \( M \) and a given sequence \( S_i \) is then the product of probabilities \( p_1:p_w \) (or the sum of logs of the probabilities). We would then look for the profile with the highest composite match over all sequences.

### Algorithms

- **Probabilistic**
  1. Expectation Maximization:
     - MEME
  2. Gibbs Sampling:
     - AlignACE, BioProspector

- **Exhaustive**
  - CONSENSUS, TEIRESIAS, SP-STAR, MDscan

The probabilistic methods are the most popular in the literature, and in particular the Gibbs sampling methods are the most popular. AlignAce was the first of these, and BioProspector is a second generation of it developed at Stanford (and better of course!). The reference to exhaustive methods above refers to combinatorial methods.
Discrete Formulations

Given sequences $S = \{x^1, \ldots, x^n\}$

- A motif $W$ is a consensus string $w_1\ldots w_K$
- Find motif $W'$ with "best" match to $x^1, \ldots, x^n$

Definition of "best":

\[
d(W, x^i) = \text{min hamming dist. between } W \text{ and a word in } x^i
\]

\[
d(W, S) = \sum_i d(W, x^i)
\]

This slide was only briefly introduced.