October 5, 2005 - Lecture #3
Notes by Imran Rashid
CSE 527 w/ Prof. Larry Ruzzo

HW #1 – post review of article by Monday, 10/10

Focused on a review of

Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO, Herskowitz I. "The transcriptional program of sporulation in budding yeast." Science. 1998 Oct 23;282(5389):699-705.

The paper sought to establish that microarrays can be used to measure gene expression. While there had already been technical reports describing microarrays, this was one of the first papers to use microarrays in a real biological experiment. In particular, this experiment studied how mRNA expression levels of all 6200 yeast genes change during yeast sporulation, to help give researchers an idea of which genes are involved in sporulation.

Classical biologists had already identified 4 stages of spore formation – Early, Middle, Mid-late, and Late – and had identified various genes and transcription factors associated with these stages.

Chu et al. grew sample yeast colonies, placed them in a sporulation-inducing medium, and then collected samples at seven time points, t = 0, 0.5,2, 5, 7, 9, 11 hours. For each time point, the yeast mRNA expression levels were measures using microarrays which contained all 6200 yeast genes using 2 color slides. A microarray was used for every time t > 0. For each microarray, samples from the test colony and the control at time t = 0 were labeled different colors and both placed on the microarray. The data was then visualized by taking the ratio of expression level against t = 0.

Note that at any time, the yeast colonies are actually in a mixture of states, as they don't all change states at exactly the same time. Even at t = 11 hours, the last time point, less than 20% of the yeast were spores.

The researchers first verified the correctness of the microarrays using Northern Blot analysis, a well understood laboratory technique. They demonstrated that for four well-characterized genes, the expression levels determined by Northern Blot agreed with their microarray data.

From the microarray data, Chu et al. decided that instead of four stages of spore formation, there were seven stages: metabolic, early I, early II, early-mid, middle, mid-late, and late. All of the genes which showed a change in expression level were then placed into one of these seven groups.

The initial clustering was done by hand. The researchers took ~40 prototype genes that were already well understood, and by examining the expression profile, they decided to create the seven groups. For example, if genes X and Y were both involved in digestion and had similar expression profiles, they were clustered into the metabolic group.

Once the 40 prototype genes had been clustered by hand, the rest of the genes (~500 genes induced, ~500 repressed) were placed into one of these groups by using a regression analysis to find the best fit among the seven groups. In other words, they used a supervised clustering technique which leveraged expert knowledge.

The researchers only reported the ratio of mRNA expression against t = 0, they never reported the raw expression levels – why? Ratios may help correct for some non-linearity in the data, e.g. given the same levels, gene X will fluoresce much brighter than gene Y (for some unknown reason – the exact mechanics of microarrays are still poorly understood). By examining raw expression levels, it may seem that gene X was much more prevalent than gene Y – using ratios will correct for that error. Also, ratios have a nice intuitive explanation; biologists can quickly understand if expression levels increase two-fold.

However, if expression levels are low at t =0, than the ratios are very susceptible to the noise in the data.

It is also important to note that microarrays only measure mRNA expression levels. They do not measure the amount of proteins that the mRNAs are encoding. While there is certainly a relationship between mRNA levels and proteins, there are many other factors which determine the amount of each protein present in the cells.

Chu et al. used existing knowledge about the transcription factors to confirm the accuracy of their clustering

- So they've included some extra columns which contain data from other sources.
    - For example, MSE is a transcription factor known to be active in the middle phase. They have a column for "how well does the known sequence that MSE binds match with some sequence in the upstream region near this gene" (with 1 gene on each row). The brighter the blue, the better MSE matches.
    - Similarly, URS1 is a transcription factor known to be active in an early phase, and is given a similarly color-coded column
    - ➔This tends to indicate that they've obtained plausible data.

Summary:
They've gotten a lot of data, for not too much work. They discovered that ~3-10x more genes are involved in sporulation than were previously known. The role of many of these genes in sporulation was later confirmed by direct knockout experiments.

A lot of computation was involved in this experiment (and even more could be done as well). Computational analysis was used for visualization and clustering. It could also have been used for sequence analysis:

- <u>Similarity search:</u>
  Given a sequence, find similar sequences in a given collection of <DNA>

- Motif search:
  Starting with a given collection of <DNA>, are there multiple sequences that all have the same pattern (same motif)?

The class came up with the following critiques of the paper:

Strength	Weakness
Verified many of their results both biologically AND computationally	They didn't repeat the experiment (especially since the experiment only took 12 hours to run). In their defense, microarrays may have simply been too expensive at that time.
Good for hypothesis generation that can then be checked with classical biology techniques. For example, followup with knockout experiments to confirm the role of new genes.	
Assay was done genome-wide (on all 6,200 genes)	
	Very empirical
As an early paper using microarrays, it is a good proof of concept.	
Clustering was performed using expert knowledge.	Should have been more formal about use of expert knowledge.
	While ratios are help with intuitive understanding, a more extensive analysis of the raw data would have been good.
	Should have used later timepoints – less than 20% of the yeast were spores at the final timepoint.

In this paper, the researchers used supervised clustering to group the genes into seven categories. However, it is not altogether obvious that there truly are the seven groups they indicate – the difference between the expression profile of the some of the groups is very small, considering the error in the data. Furthermore, often data simply doesn't cluster well at all – for example, using Principal Component Analysis on the same set of data leads to a mess.