October 6, 2004 - Lecture #3
Notes by Michael Panitz
CSE 527 w/ Prof. Larry Ruzzo

CSE527 emailing list: Several messages have been sent to the emailing list, so if you didn't get them, you're not subscribed. Sign up at (http://www.cs.washington.edu/527), or (if you thought you had signed up, but haven't gotten email) send email to the prof (ruzzo@cs.washington.edu) and/or TA (kasiaw@cs).

HW#2: This assignment will be posted shortly on the website.
The essence of the assignment is to find a paper on microarrays, read it, and critique it. There are a list of references on the website. Also, one can re-critique the paper that was covered in this lecture (by Chu, et al) if one would like.

- Consider both the computational and biological aspects
- What questions did they ask?
- What approaches did they use?
- Critique it – what's good? What's bad? What would you do differently?
- Sources – PubMed tends to be good. Lotsa Comp.Sci people go looking on "CiteSeer", which tends to produce narrow results.

The general plan for the next several lectures is to cover the computational aspects of microarrays, then move on to sequencing.

<Finishing the slides from the previous lecture>
There are lotsa variations

- Note that when using commercial microarray technologies, one will often get different results from different platforms/products, owing in part to their different fabrication methods.
    - <It wasn't clear how much variation, but it was clear that this would be a factor in reproducing someone else's results>
- WITHIN a platform, you tend to get consistent results

<Moving on to the slides for Lecture #3>
The paper (by Chu, et al) is old – about 6 years old

The paper sought to establish that microarrays can be used to measure gene expression. Specifically, by measuring gene expression in yeast cells, as they (slowly) transform from normal cells into spores. The level of expression at a given time point was measured relative to the level of expression at the first time point.

Budding yeast

- grows a small bud, rather than splitting wholesale, like most eukaryotes

- clarification: in the presence of sugar and oxygen, this yeast raises bread
- in the presence of sugar but NO oxygen, it ferments the sugar

---

As a long-term survival strategy (in the absence of needed nutrients), the yeast cell can transform into a spore.

- The spore consists of a clump of 4 cells, surrounded by a tough outer wall
- each of these spore cells is haploid, and is created in a process called meiosis

---

Meiosis is the process of splitting a normal cell into 4 reproductive cells

- The cell first duplicates it's DNA (it starts w/ a pair of any given chromosome, and copies both, for a total of 4 chromosomes)
- It then does crossover mutations between the 2 pairs of chromosomes
- It then divides the 4 chromosomes into 2 pairs
- It then divides each pair, and into separate cells, leaving each of the 4 new cells with 1 chromosome each.

One of the difficulties of trying to measure gene expression as the yeast cells sporulate is that you start with a collection (of lots and lots) of cells, expose them to conditions that induce sporulation, then try to measure what's going on, *even though different cells sporulate at different rates*.

- At any given point in time, there's actually a mixture of cells
    - There were charts demonstrating that they had measured this
    - At the end of the 12 hour time period of the experiment, only 20% of the cells had sporulated.

Side note: PCR amplification:

- One problem w/ measuring gene expression is that some genes may not be expressed much (a couple copies per cell), but may still have an important effect on the overall process.
- One way to try to get around this is to do PCR amplification in order to clone lotsa copies of <everything>
- If possible, this is best avoided, since it introduces more noise into the experiment
- Further, how can one use PCR to "amplify everything"?
    - One way is to use a mixture of primers, choose carefully, and hope for the best
    - Another way is to take advantage of the very common poly-A tail that gets added to all eukaryotic mRNAs(at the 3' end), by using a poly-T primer.
        - This has a bias in favor of the 3' end of the mRNA, since stuff generally goes 5'→3', it gets increasingly tough to amplify stuff that's further upstream (towards the 5' end)

Did a Northern blot electrophoresis test to verify that the microarray was measuring signal, not noise.

- Northern blots have been used for ~30 years, and are both well established and reliable

---

The experimenters tried to figure out if their microarray measured a substantial change in gene expression.
To do this, they picked a semi-arbitrary delta, beyond which they declared that "this gene's expression has changed significantly)

Roughly 1/3 of the genes were affected

Note: In the pictures in the slides, they color-coded expression/suppression – red meaning that the gene had been induced at that time point, green meaning that it had been suppressed. They went back and sorted the rows so that all the green was on top, the reds on the bottom, and there's this nice progression (L to R) as you go down the chart.

In this particular experiment, they used lots of existing biological data to confirm their data.

Previously, biologists had divided up sporulation into 4 separate phases ; after examining a small(ish) number of hand-picked genes, they decided that it would be better to divide up the sporulation process into 7 separate phases.

- The genes were hand-picked based on prior biological knowledge
- The 7 phases were decided on based on data from this experiment

They tried to mechanically correlate the remaining (1000 or so) genes to one of these 7 categories.

- One the one hand this is nice ; on the other hand, we can sort any collect of 1,000 numbers into X buckets – so does this really mean anything?
- 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 from the microarray.

- What they're doing (essentially) is clustering genes by hand, then seeing if the other genes support this. The term "supervised clustering" was repeatedly used to describe this.

"Summary 1" slide:

They've gotten a lot of data, for not too much work.

"Summary 2" slide
What computation could be used here?

Different search types:
- Similarity search:
  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)?

Clustering – can a program be written to pick out common groups without prior knowledge?
- In this paper, would an automatic algorithm have grouped sporulation into 7 phases? More? Less?

The class then critiqued the paper in various ways:

Strength	Neutral	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)
Raw data is available on their website • Including pictures from the microarray • Thus one might actually redo some of the calculations		
	They triggered sporulation	

	by placing yeast in a nitrogen poor environment. Do other conditions trigger sporulation? If so, would they have lead to different patterns of gene expression?	
Assay was done genome-wide (on all 6,200 genes)		
		No statistical work was done (Both statistical work, and repeats, need to be done to get a paper published today)
First paper on the subject		
Provided a foundation for new research		
		Relative Measurements: If the experiment concluded that there was a 2x-10x increase, how do you know that's significant? What if there simply wasn't any in the mix to begin with (and thus, adding a small amount (in absolute terms) produces a huge % increase
		The clustering of gene expression into 7 groups was supervised – would automatic procedures have produced better results?
One could probably do further technical nit-picking: everything in the experiment seemed to be either a reasonable choice, or a clever choice, but were they the best?		

Principal Component Analysis
Find a vector through the data that maximizes the spread