7.91 / 7.36 / BE.490

Lecture #7

May 4, 2004

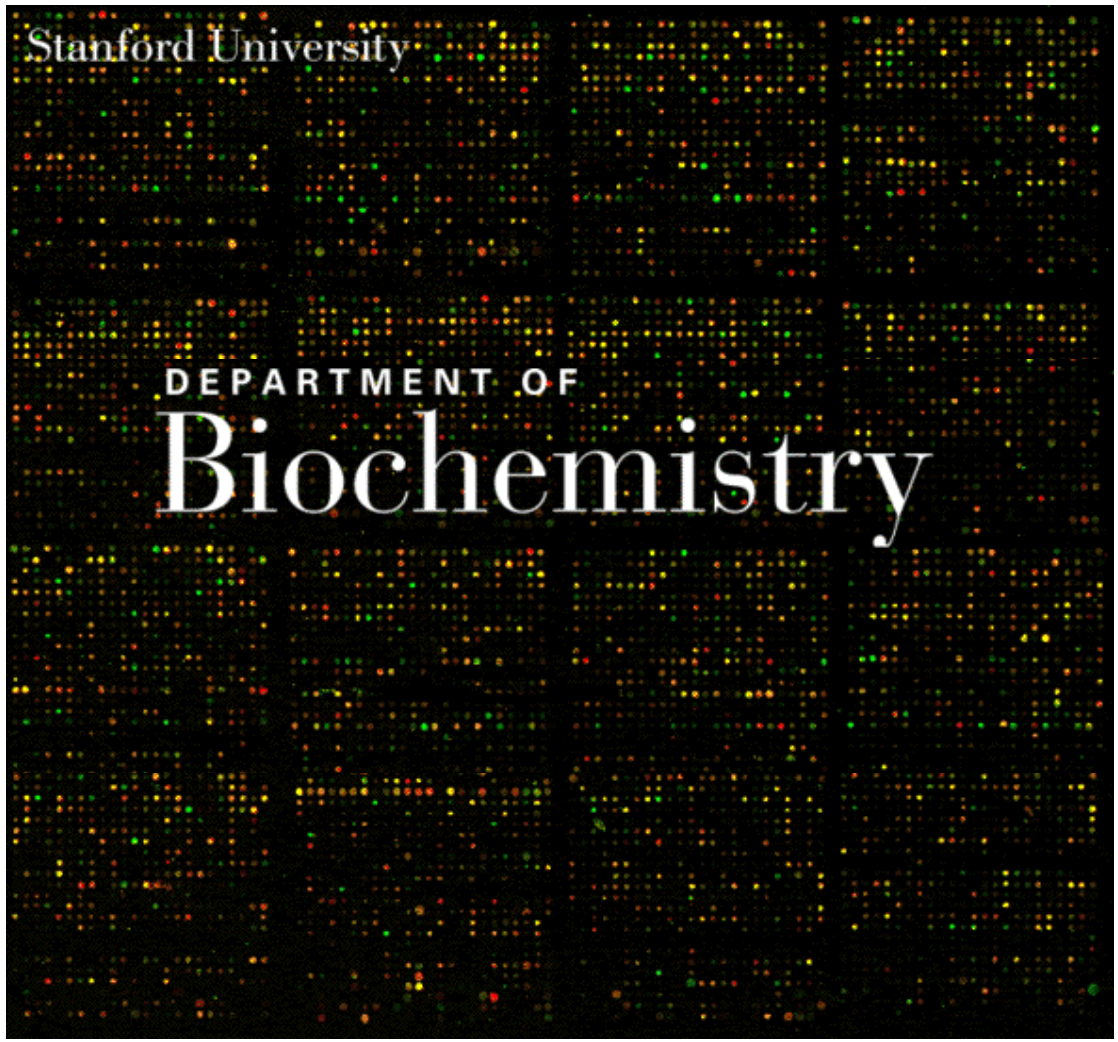
DNA Microarrays & Clustering

Chris Burge

DNA Microarrays & Clustering

- Why the hype?
- Microarray platforms
  - cDNA vs oligo technologies
- Sample applications
- Analysis of microarray data
  - clustering of co-expressed genes
  - some classic microarray papers

Stanford U. Dept. of Biochemistry Web Site



http://cmgm.stanford.edu/biochem/

Why Microarrays?

- Changes in gene expression are important in many biological contexts:
  - Development
  - Cancer
  - Other Diseases
  - Environmental Adaptation

- DNA microarrays provide a high throughput way to study these changes.

What's new?
… progression to chip technology

- Hybrid detection
  - radioactive labeling
  - fluorescent labeling
- Solid support for sample fixation
  - Southern blots, Northern blots, etc.

- Main advantage of microarrays is scale
  - Probes are attached to solid support
  - Efficient robotics
  - Bioinformatic analysis
- Parallel measurement of thousands of genes at a time

Array Platforms

- cDNA arrays (spotted arrays)
  - Probes are PCR products from cDNA libraries or clone collections
  - May be printed on glass slides (e.g., P. Brown lab, Stanford), OR
  - May be printed on nylon membranes (e.g., Millennium)
  - Spots are 100-300 μm in size and about the same distance apart
  - ~30,000 cDNAs can be fit onto the surface of a microscope slide

- Oligonucleotide arrays
  - 20-25 mers synthesized onto silicon wafers *in situ* or printed onto glass slides by: photolithography (Affymetrix) or ink-jet printing (Rosetta/Agilent)
  - Presynthesized oligos can also be printed onto glass slides

- Other technologies (e.g., bead arrays attached to optical fibers)

cDNA Arrays I - Overview

Please See

Duggan, DJ, M Bittner, Y Chen, P Meltzer, and JM Trent. "Expression Profiling using cDNA Microarrays." *Nat Genet.* 21, no. 1 Suppl (January 1999): 10-4.

cDNA Arrays II - Printing

1. Templates for genes of interest obtained and amplified by PCR

2. After purification and quality control, aliquots of ~5 nl printed on coated glass microscope slide using high speed robot

Please See

Duggan, DJ, M Bittner, Y Chen, P Meltzer, and JM Trent. "Expression Profiling using cDNA Microarrays." *Nat Genet.* 21, no. 1 Suppl (January 1999): 10-4.

cDNA Arrays III - Labeling, Hybing

1. Total RNA from test and reference samples is fluorescently labeled with Cy5/Cy3 dye using a single round of reverse transcription

2. Pooled fluorescent targets are hybridized to the clones on the array

Please See

Duggan, DJ, M Bittner, Y Chen, P Meltzer, and JM Trent. "Expression Profiling using cDNA Microarrays." *Nat Genet.* 21, no. 1 Suppl (January 1999): 10-4.

cDNA Arrays IV - Scanning

1. Laser excitation of hybridized targets - emission spectra measured using a scanning confocal laser microscope
2. Monochrome images (from scanner) are imported into software in which images are pseudo-colored and merged
3. Data analyzed as normalized ratio (Cy3/Cy5) - gene expression increase or decrease relative to reference sample

Please See

Duggan, DJ, M Bittner, Y Chen, P Meltzer, and JM Trent. "Expression Profiling using cDNA Microarrays." *Nat Genet.* 21, no. 1 Suppl (January 1999): 10-4.

cDNA Arrays Oligo Arrays

Please See

Schulze, A, and J Downward. "Navigating Gene Expression using Microarrays --A Technology Review." *Nat Cell Biol.* 3, no. 8 (August 2001): E190-5.

Oligo Arrays I - Light-directed printing

- Synthetic linkers modified with photochemically removable protecting groups attached to substrate and direct light through a photolithographic mask to specific areas on the surface to produce localized photodeprotection.

- Chemical coupling occurs at those sites that were illuminated in the preceding step. Next, light is directed to different regions and cycle is repeated.
- Current versions now exceed one million probes per array.

Please See

Lipshutz, RJ, SP Fodor, TR Gingeras, and DJ Lockhart. "High Density Synthetic Oligonucleotide Arrays." *Nat Genet.* 21, no. 1 Suppl (January 1999): 20-4.

Oligo Arrays II - Other types of printing

Bubble-Jet printing technology
for covalent attachment of DNA

Please See

Okamoto, T, T Suzuki, and N Yamamoto. "Microarray Fabrication with Covalent Attachment of DNA using Bubble Jet Technology." *Nat Biotechnol.* 18, no. 4 (April 2000): 438-41.

Commercially Available Microarrays

Please See

Lipshutz, RJ, SP Fodor, TR Gingeras, and DJ Lockhart. "High Density Synthetic Oligonucleotide Arrays." *Nat Genet.* 21, no. 1 Suppl (January 1999): 20-4.

cDNA vs Oligo Arrays

- Requirements:
  - purified DNA vs sequence info alone
- Reproducibility
- Cost
- Hybridization specificity / probe size
- Applications

Some applications of microarrays

- Temporal order of gene expression program (cell cycle)
- Effect of perturbations of the cellular environment on gene expression (e.g., medium, temperature, drugs, etc.)
- Differential gene expression in different pathological conditions / tissue types
- Identification of genes / exon-intron structures
- Mutation analysis
- Mapping binding sites of transcription factors

Microarray Data Analysis - Normalization & Clustering

- Normalization
  - use all genes in sample, OR
  - use designated unchanging subset of genes
  - measure variance of normalizing set
  - use to generate expected variance, confidence intervals
  - use CIs to define up- and down-regulated genes

What is clustering?

- A way of grouping together data samples that are ***similar*** in some way - according to criteria of your choice

- A form of ***unsupervised learning*** – generally don't have examples of how the data *should* be grouped together

- So, a method of ***data exploration*** – a way of looking for patterns or structure in the data that are of interest

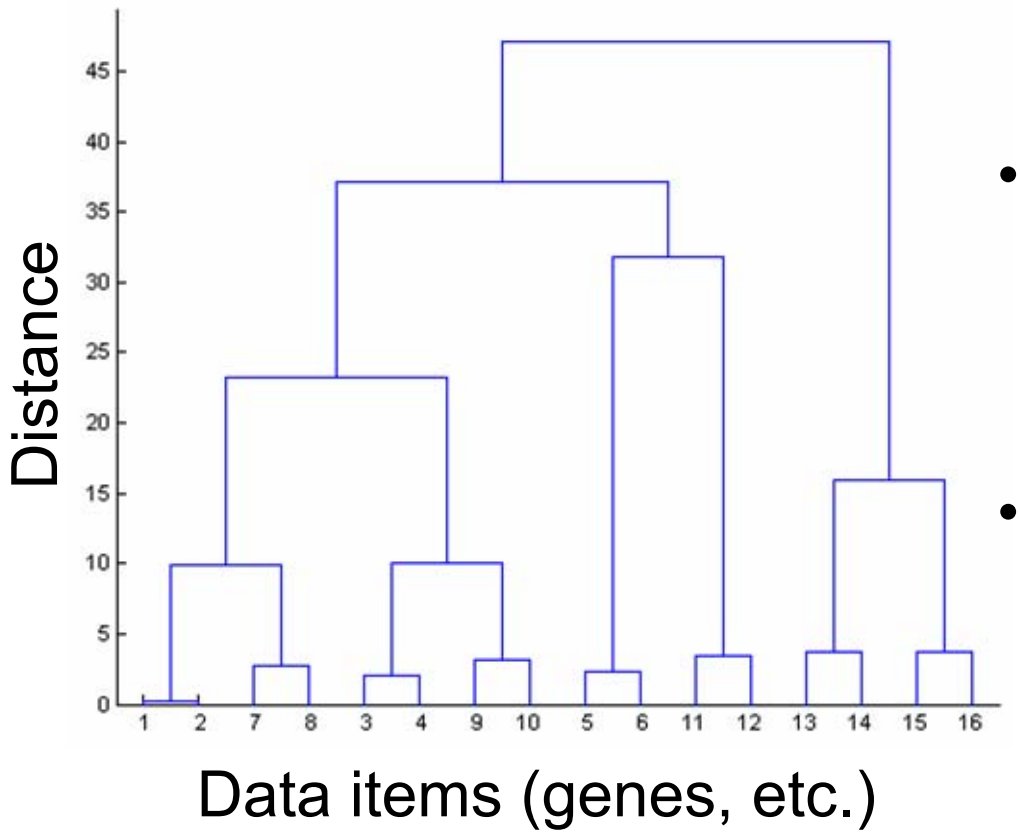
Why cluster?

- Cluster genes (rows)
  - Measure expression at multiple time-points, different conditions, etc.
  - Similar expression patterns may suggest similar functions of genes
- Cluster samples (columns)
  - e.g., expression levels of thousands of genes for each tumor sample
  - Similar expression patterns may suggest biological relationship among samples

Hierarchical Agglomerative Clustering

- Start with each data point in separate cluster
- Keep merging most similar pairs of data points/clusters until all form one big cluster
- Called bottom-up or agglomerative method

Hierarchical Clustering II

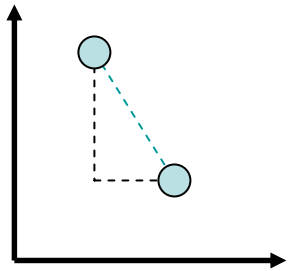


- This produces a binary tree or ***dendrogram***
- The final cluster is the root and each data item is a leaf
- The heights of the bars indicate how close the items are

How do we define “similarity”?

- The goal is to group together “similar” data –
  but how to define similarity/distance between points (or clusters)?
- In general, depends on what we want to find or emphasize in the data - **clustering is an art**
- The similarity measure is often more important than the clustering algorithm used

Euclidean distance



$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

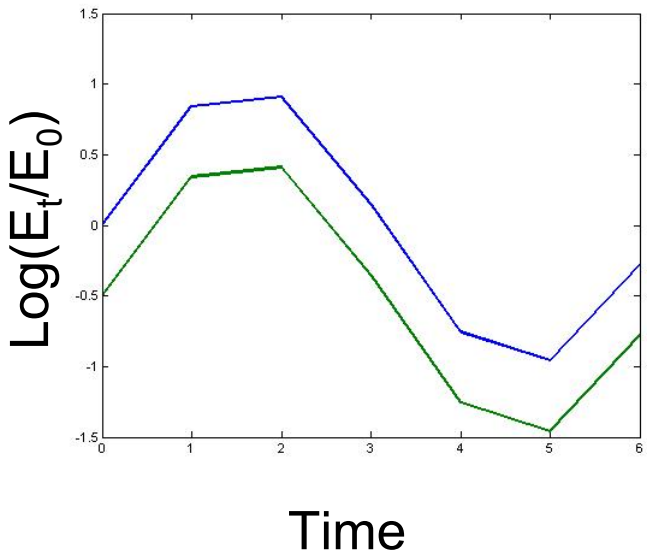
- Here *n* is number of dimensions in the data vector

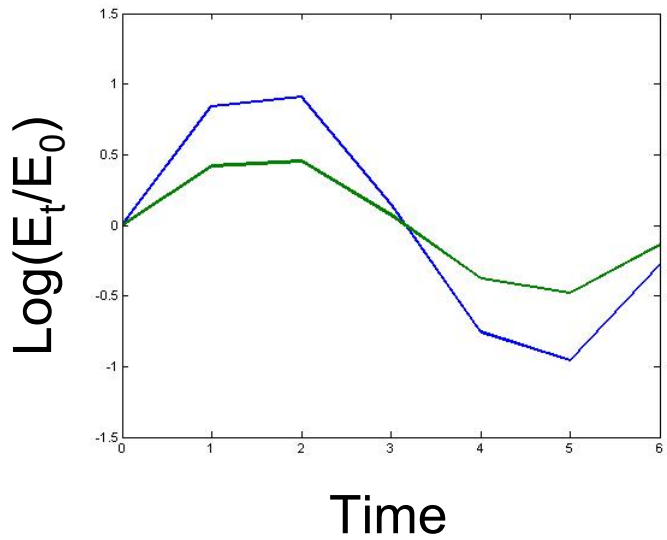
  For instance:

  - Number of time-points/conditions (when clustering genes)
  - Number of genes (when clustering samples)

Correlation

- We might care more about the overall shape of expression profiles more than the actual magnitudes
- That is, we want to consider genes similar when they go "up" and "down" together





Pearson or Product-Moment Correlation

$$\rho(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i$$

$$\bar{y} = \frac{1}{n}\sum_{i}^{n} y_i$$

Product of corresponding terms in vector, using difference from mean rather than value, and normalizing by the product of the standard deviations.

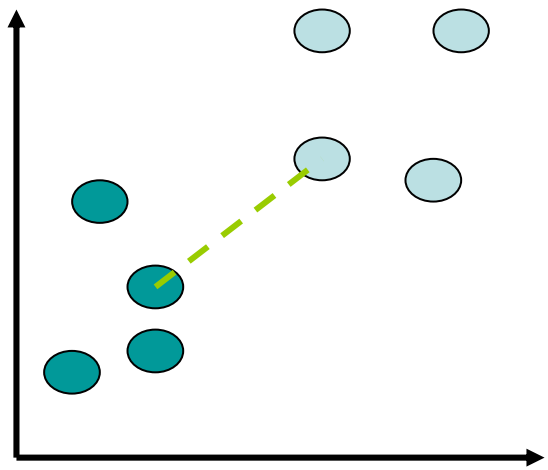
- Always between –1 and +1
- Invariant to scaling and shifting (adding a constant) of the expression values

Linkage in Hierarchical Clustering

- We already know about distance measures between data items, but what about between a data item and a cluster or between two clusters?
- We just treat a data point as a cluster with a single item, so our only problem is to define a ***linkage*** method between clusters
- As usual, there are lots of choices…

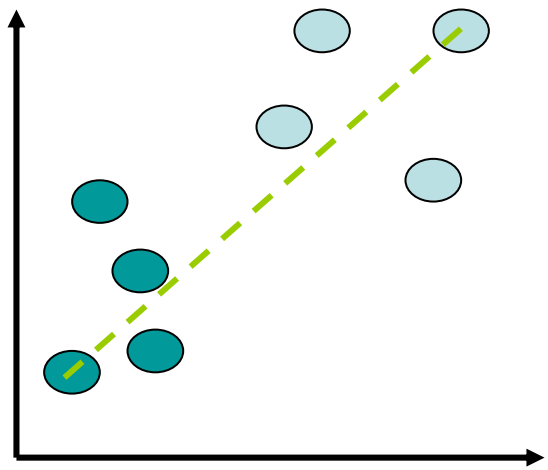
Single (Minimum) Linkage

- The minimum of all pairwise distances between points in the two clusters
- Tends to produce long, “loose” clusters



Complete (Maximum) Linkage

- The maximum of all pairwise distances between points in the two clusters
- Tends to produce very tight clusters



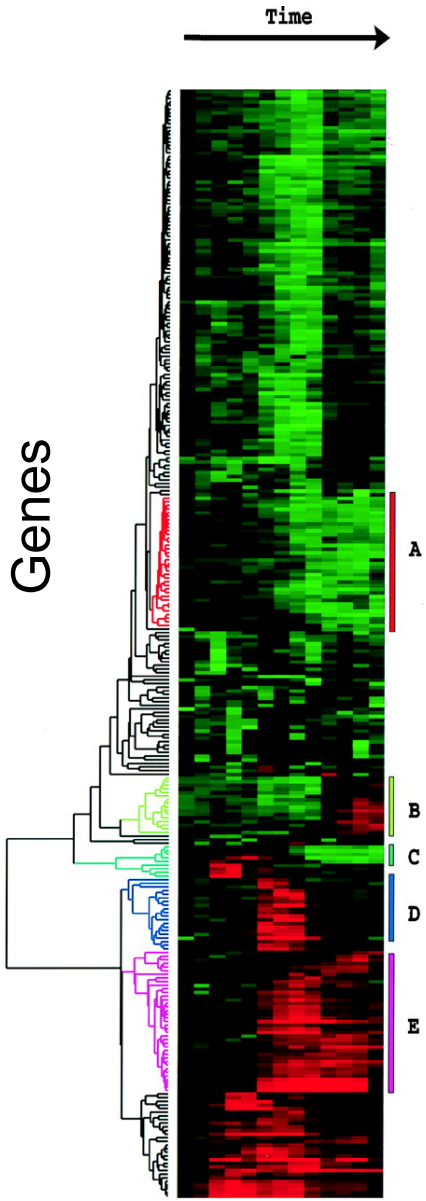
Average Linkage

- M. Eisen's cluster program defines average linkage as follows:
  - Each cluster c_i is associated with a mean vector μ_i which is the mean of all the data items in the cluster
  - The distance between two clusters c_i and c_j is then defined as $d(\mu_i, \mu_j)$

- This is somewhat non-standard – this method is usually referred to as centroid linkage and average linkage is defined as the average of all pairwise distances between points in the two clusters

Eisen et al., PNAS 1998

Hierarchical Clustering Examples



Clustering 8600 human genes based on time course of expression following serum stimulation of fibroblasts

(A) cholesterol biosynthesis

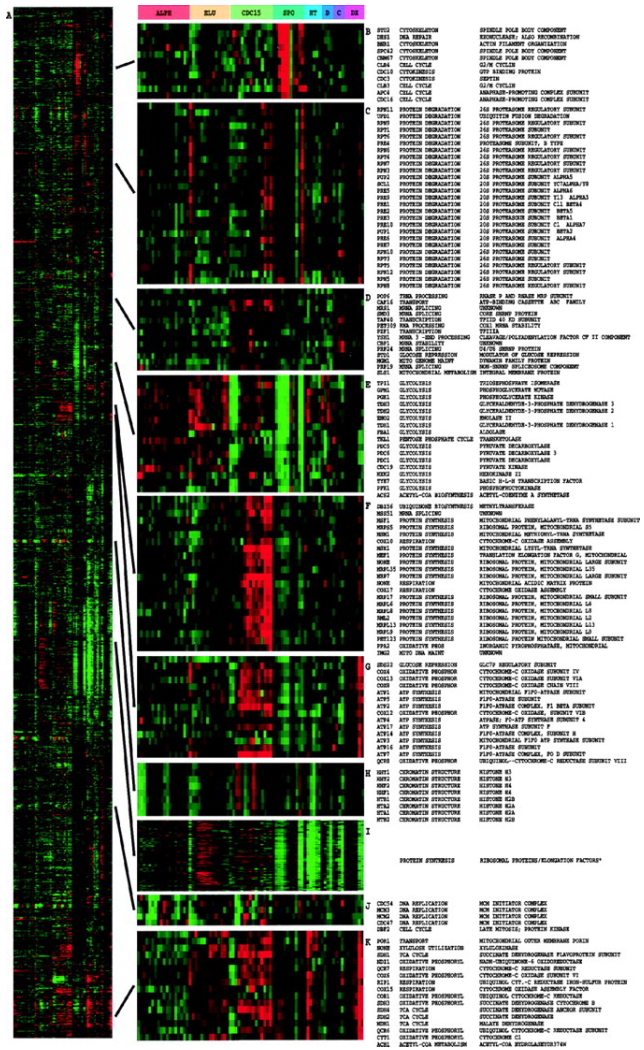
(B) the cell cycle

(C) the immediate-early response

(D) signaling and angiogenesis

(E) wound healing and tissue remodeling

Clustering yeast genes by co-expression across many conditions



(*B*) spindle pole body assembly and function

(*C*) the proteasome

(*D*) mRNA splicing

(*E*) Glycolysis

(*F*) the mitochondrial ribosome

(*G*) ATP synthesis

(*H*) chromatin structure

(I) the ribosome and translation

(*J*) DNA replication

(*K*) the TCA cycle and respiration

Clustering tumor samples with B- and T-cell types based on expression profiles

Patients with “germinal center type” expression profiles generally had higher five-year survival rates

Please See

Alizadeh, AA, MB Eisen, RE Davis, C Ma, IS Lossos, A Rosenwald, JC Boldrick, H Sabet, T Tran, X Yu, JI Powell, L Yang, GE Marti, T Moore, J Hudson Jr, L Lu, DB Lewis, R Tibshirani, G Sherlock, WC Chan, TC Greiner, DD Weisenburger, JO Armitage, R Warnke, R Levy, W Wilson, MR Grever, JC Byrd, D Botstein, PO Brown, and LM Staudt. "Distinct Types of Diffuse Large B-cell Lymphoma Identified by Gene Expression Profiling." *Nature* 403, no. 6769 (3 February 2000): 503-11.

Microarray analysis of alternative splicing with exon junction probes I

Tissue-specific splicing of OCRL1 gene

Please see figure 1 of

Johnson, JM, J Castle, P Garrett-Engele, Z Kan, PM Loerch, CD Armour, R Santos, EE Schadt, R Stoughton, and DD Shoemaker. "Genome-wide Survey of Human Alternative Pre-mRNA Splicing with Exon Junction Microarrays." *Science* 302, no. 5653 (19 December 2003): 2141-4.

Microarray analysis of alternative splicing with exon junction probes II

Brain-specific spliced isoforms of APP gene

Please see figure 2 of

Johnson, JM, J Castle, P Garrett-Engele, Z Kan, PM Loerch, CD Armour, R Santos, EE Schadt, R Stoughton, and DD Shoemaker. "Genome-wide Survey of Human Alternative Pre-mRNA Splicing with Exon Junction Microarrays." *Science* 302, no. 5653 (19 December 2003): 2141-4.

Papers for Thursday

#1 Segal, E, M Shapira, A Regev, D Pe'er, D Botstein, D Koller, and N Friedman. "Module Networks: Identifying Regulatory Modules and Their Condition-specific Regulators from Gene Expression Data." *Nature Genetics* 34, no. 2 (June 2003): 166-76.

#2 Beer, Michael A., and Saeed Tavazoie. "Predicting Gene Expression from Sequence." *Cell* 117 (16 April 2004): 185-198.

Background reading:

#3 Friedman, N. "Inferring Cellular Networks using Probabilistic Graphical Models." *Science* 303, no. 5659 (6 February 2004): 799-805.

Appendix of Probability & Statistics Primer