DNA Microarrays and Clustering of Gene Expression Data

Martha L. Bulyk
mlbulyk@receptor.med.harvard.edu

Biophysics 205
Spring term 2008

Traditional Method: Northern Blot

• RNA population on filter (gel); DNA probe is hybridized
• transcript length; related transcripts

Technologies to Generate Large mRNA Expression Datasets

• Serial Analysis of Gene Expression (SAGE) - sequencing (Velculescu et al. Science. 1995. 270(5235):484-7.) - digital, not analog

• Glass slide microarrays – fluorescence
  spotted: (Schena et al. Science. 1995. 270(5235):467-70.)

• Nylon membrane macroarrays (“dot blots”) – radioactivity
**DNA on the arrays:** originally PCR products, now typically just oligos (increased specificity, plus time, labor, & reagent savings)

(a)  
\[ \text{5'} \quad (10) \quad (7-15) \quad 3' \]

(b)  
\[ \text{5'} \quad (10) \quad (7-15) \quad 3' \]

preferably ~60 nt long

---

Printhead of a microarray robotic spotter printing onto a slide.

Glass slide DNA microarrays: Schena, Shalon, Davis, Brown '95, Science

- long oligos

---

**Ink-jet in situ synthesis of oligonucleotide arrays**

(Agilent technical brochure)
Labeling of Transcripts for Microarray Hybridization

- RNA purification
- Anneal oligo-dT primer or just random hexamers (5' NNNNNN 3')
- Reverse transcription including Cy3-dUTP or Cy5-dUTP
- Labeled cDNA pool

- Cy3 and Cy5 are fluorophores with minimally overlapping spectra.
- Other labeling protocols exist.

Two-color DNA microarrays

- Robotic arrays: 10,000 spots (genes/exons) per slide, ~50-100 μm diameter
- Agilent arrays: up to 244K spots per slide, ~10 μm diameter
- NimbleGen arrays: >385K spots per slide, 16x16 μm

Sample: Cy5-labeled sample 2 = Cy3-labeled
- Red: [sample 1] > [sample 2]
- Green: [sample 2] > [sample 1]
- Yellow: [sample 2] = [sample 1]

(robotarray image from Iyer et al., '99, Science, 283:83-7)

Microarray #1: expt 1 / control
Microarray #2: expt 2 / control
Microarray #3: expt 3 / control
Microarray #4: expt 4 / control

→ Can compare expt 1 vs expt 2 vs expt 3 vs expt 4
Light-Directed Synthesis of DNA Chips

Light comes through gaps in a photolithographic mask; a series of photolithographic masks is used to synthesize the oligos on the array.

Affymetrix Oligonucleotide Arrays ("GeneChips™")

- up to $1 \times 10^6$ features per array
- down to $7 \mu m \times 7 \mu m$
- single-stranded
- 25 nt in length
- PM/MM pairs

PM = "Perfect Match"
MM = "Mismatch"
Note: • Affymetrix chips are 1-color (1 sample) expts - above image is an overlay of 2 chips; • spotted microarrays are 2-color (2 samples)

NimbleGen arrays:
Maskless array synthesis

Unlike Affymetrix arrays, no photolithographic masks need to be synthesized here! Only the program that toggles the individual micromirrors on the array of micromirrors is changed, permitting flexible array design.

Expression Arrays vs. Northern Blots

• DNA "probes" are spotted on array, transcripts (or cDNAs) are hybridized • relative abundances for a given transcript

• RNA population on filter (gel); DNA "probe" is hybridized • transcript length; related transcripts
"Tiling" arrays:
Arrays of probes covering a genomic region at high density; no standard definition of probe type (oligo or PCR product) or density (overlapping or separate by some spacer)

Examples:
(a) 5' <- 3'
(b) 5' <- 3'

Experimental Noise

same cDNA pool
same RNA

(from Livesey et al. Curr Biol. '00. 10:301-310)

Experimental Noise – Duplicate Hybridizations

Importance of replicates!
Assessing how similar 2 genes’ expression patterns are over a collection of DNA microarray experiments (from different time points, cell types, etc.).

Similarity measures

- Euclidean distance
- non-Euclidean metrics:
  - Manhattan (Hamming) distance
  - Pearson correlation coefficient

Euclidean Distance

\[ d(x,y) = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2} \]

- \( N \) = \# of experimental conditions (ex. 9 time points)
- \( x_i \) = measured expression level of gene \( x \) in experiment \( i \)
- \( y_i \) = measured expression level of gene \( y \) in experiment \( i \)
Pearson Correlation Coefficient

\[ r(x,y) = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{N} (y_i - \bar{y})^2}} \]

- \( N \) = number of experimental conditions (e.g., 8 time points)
- \( x_i \) = measured expression level of gene \( x \) in experiment \( i \)
- \( y_i \) = measured expression level of gene \( y \) in experiment \( i \)
- \( \bar{x} \) = mean expression level of gene \( x \) over all experiments
- \( \bar{y} \) = mean expression level of gene \( y \) over all experiments
- \( r = [-1..1] \)
  - 0 = no correlation
  - 1 = perfectly correlated
  - -1 = perfectly anti-correlated

Genes are Clustered According to How Similar Their Expression Profiles Are

One commonly used similarity measure:
Pearson Correlation Coefficient

\( r = 1 \) perfectly correlated
\( r = -1 \) perfectly anti-correlated
\( r = 0 \) no correlation

Representing Microarray data in Expression Space

- Normalized expression data from microarrays
- Gene 1
- Gene 2
- Experiment 1
- Experiment 2
- Experiment 3

8
Grouping Genes with Similar Patterns in Expression Space

Power of Clustering mRNA Expression Data

- Identify tissue-specific genes
- Predict gene function based on co-expressed genes
- Distinguish cancers that can’t readily be distinguished otherwise
  - acute myeloid (AML) vs acute lymphoblastic leukemia (ALL)
  - distinguish subclasses of diffuse large B-cell lymphoma (DLBCL)
  - distinguish subclasses of breast cancers
- Impact decisions regarding how to treat patients
  → “Personalized medicine”

What might be interesting to examine?

- wildtype versus mutant cells (ex. knockout)
- effect of environment (growth conditions, stress, etc.)
- natural changes in cell during growth (cell cycle)
- normal versus disease cells (ex. breast cancers)
- different tissues (ex. heart, thymus, etc.)
- stages of development
- response to a drug
Some Common Clustering Algorithms

Unsupervised:
• k-means
• self-organizing maps (SOMs)
★ • hierarchical
★ • principal component analysis (PCA)

Supervised:
• support vector machines (SVMs)

k-Means Clustering
• user-specified # of clusters
• all objects initially assigned to one of k clusters
• compute distances between clusters (average expression vector)
• objects moved between clusters, and
remain in new cluster if more similar to average expression vector
• iterate until minimized intra-cluster distances &
maximized inter-cluster distances

(Tavazoie et al., Nat Genet, ’99, 22(3):281-5)

EXAMPLE:
Hierarchical Clustering (8600 Genes) of Time Course of
Serum Stimulation of Primary Human Fibroblasts

Background:
1) normal human fibroblasts: + growth factors → growth
   − growth factors → G0 (nondividing)
2) Growth factors are provided by fetal bovine serum in culture.
3) Serum is normally encountered by cells in vivo in the context
   of a wound.

Experimental setup:
1) cultured primary human fibroblasts
2) deprived cells of serum for 48 hrs
3) stimulated cells with serum

(Iyer et al., ’99, Science, 283:63-71)
Samples and Microarrays:

- ~10,000 spots: ~9,800 cDNAs representing ~8,600 genes
- 12 timepoints: 0, 15 min, 30 min, 1, 2, 4, 6, 8, 12, 16, 20, 24 hrs
- no replicates

Hierarchical Clustering (8600 Genes) of Time Course of Serum Stimulation of Primary Human Fibroblasts

Cluster & TreeView programs
http://rana.lbl.gov/EisenSoftware.htm

Along x-axis: timepoints
Along y-axis: genes

red = increased (here, relative to t=0)
green = decreased

Hierarchical Clustering
Data suggest that exposure to serum is not interpreted as a general mitogenic stimulus, but rather as a specific physiological signal specifying that a wound has occurred, as serum is normally encountered by cells in vivo in the context of a wound.