Biology Implications of Microarray Data Analysis

- human health
  tamoxifen & cancer (Hilsenbeck et al.)
  aging & calories (Lee et al. 1999)
  obesity & diabetes (Nadler et al. 2000; Soukas et al. 2000)

- DNA–RNA–protein basic biology
  E. coli operon discovery (Craven et al. 2000)

- cell systems biology
  cyclic regulation (Spellman et al. 1998)
  function/pathway discovery (Hughes et al. 2000)
  signaling in MAPK pathways (Roberts et al. 2000)

Microarray Image Analysis
oligonucleotide arrays

Affymetrix chip technology
Lockhart et al. (1996); *Nat Gen* 1999 Supplement;
Schadt Li Su Wong (2000); Li Wong (2000)

probe array of 20 sequences from gene
typically 25 nucleotides: unique to gene
uniform (!) hybridization characteristics
synthesized in 24-50μm region
10^6 to 10^7 copies of probe

PM = perfect match: correct gene sequence
MM = mismatch: center nucleotide changed
MM should have random & cross hybridization

nonlinear relationship between PM & MM
ideally high PM and low MM
depends on hybridization kinetics

using gene chips

prepare RNA sample with label
hybridize RNA sample to array
stain/wash process
scan array with confocal laser scanner ($$)
– excite features with laser
– detect photon emission from labeled RNA
– convert into 16-bit intensity value

statistics on PM & MM
difference or ratio? how to average over 20 pairs?

software to process chip data

GeneChip (Affymetrix); Schadt Wong (R: public?)
– image segmentation (identify spot)
– background correction
– scaling/normaliztion to compare arrays
– ascertain presence/absence in sample
– assess if gene shows differential expression

background/gradient correction

why correct for background?
measured intensity includes “noise”
hybridization to reading

large scale background variation
– affects hundreds-thousands of spots on chip
– scars and blotches on parts of chip
– intensity gradients across chip
– processing irregularities
– block correction or local averaging?
– mask out artifacts (manual/automatic)

small scale background variation
– diffusion of spot construction
– diffusion of spot reading

array scaling for comparison

confounding of conditions & arrays
overall intensity can vary
– amount of RNA
– hybridization process
– scanning process

assumption: intensity difference among arrays are linearly related with zero intercept
use average across chip to normalize

$$\beta = \frac{\sum_{ij}(PM_{ij} - MM_{ij})_{chip1}}{\sum_{ij}(PM_{ij} - MM_{ij})_{chip2}}$$

$j = 1, \cdots, n$ genes; $i = 1, \cdots, N$ features per gene
rescale chip by 1000 / mean intensity on chip

problem: large dynamic range in intensities (1000)
low vs. high intensity? (change-point)

assess gene presence

Affymetrix assessment:
positive: $PM - MM \geq SDT$ & $PM/MM \geq SRT$
negative: $PM - MM \leq SDT$ & $PM/MM \leq SRT$
– SD(R)T = statistical difference (ratio) threshold
– log average: $10*$mean of $\log(PM/MM)$
– positive/negative fractions