Introduction to Microarrays

BMI/CS 576
www.biostat.wisc.edu/bmi576.html
Mark Craven
craven@biostat.wisc.edu
November 2002

Announcements 11/7

• HW #2 due date postponed until 11/11
  – BMI file servers were screwed up yesterday
  – just electronic hand-in this time
• the plan
  – today: intro to microarrays & gene expression
  – 11/12: guest lecture by Prof. David Page on classifying
gene expression profiles to distinguish cancer subtypes
  – 11/14: clustering gene expression profiles
  – 11/19, 11/21: more guest lectures
• guest lectures will be considered material for final exam
Gene Expression & Microarrays

- so far we’ve taken a pretty static view of cells
- to really understand a genome, we need to understand
  - how genes interact with each other
  - how active various genes are under different conditions
- one way to do this is to take “snapshots” of cells
  - how much of each mRNA is there in the cell?
    - microarrays enable us to measure this for 1000’s of genes simultaneously
  - how much of each protein is there in the cell?
    - even more informative, but the technology is not as well developed

Microarrays

- microarrays provide a tool for answering a wide range of questions about the dynamics of cells
  - how active are various genes in different cell/tissue types?
  - how does the activity level of various genes change under different conditions?
    - stages of a cell cycle
    - environmental conditions
    - diseases
    - knockout experiments
  - what genes seem to be regulated together?
- can also be used to answer questions about static properties (e.g. genotyping), but we’ll focus on the former class of questions
Microarrays

- a.k.a. DNA chips, gene chips, DNA arrays etc.
- two general types that are popular
  - spotted arrays (pioneered by Pat Brown @ Stanford)
  - oligonucleotide arrays (pioneered by Affymetrix Inc.)
- both based on the same basic principles
  - anchoring pieces of DNA to glass/nylon slides
  - complementary hybridization

Complementary Hybridization

- due to Watson-Crick base pairing, complementary single-stranded DNA/RNA molecules hybridize (bond to each other)
Complementary Hybridization

- one way to do it in practice
  - put (a large part of) the actual gene sequence on array
  - convert mRNA to cDNA using reverse transcriptase

```
TCGCCAAGCTTATGG  actual gene
AGCGGTTCGAATACC  cDNA
```

reverse transcriptase

```
UCGCCAAGCUUAUGG  mRNA
```

Spotted Arrays

- robot puts little spots of DNA on glass slides
  - each spot is DNA analog of one of the mRNAs we want to measure
Spotted Arrays

- two samples (reference and test) of mRNA are reverse transcribed to cDNA, labeled with fluor dyes and allowed to hybridize to array

reference

![Reference mRNA](image)

![Labeled cDNA](image)

test

![Test mRNA](image)

![Labeled cDNA](image)

Spotted Arrays

- lasers applied to the arrays yield an emission for each fluorescent dye

![Labeled arrays](image)
Spotted Arrays

• here is an example of the resulting image

![Spotted Arrays Image]

Spotted Arrays

• we can’t detect the absolute amount of mRNA present for a given gene, but we can measure amount relative to a reference sample
• typically we have a set of measurements

\[ G_i = \log \frac{\text{red}_i}{\text{green}_i} \]

where red is the test expression level, and green is the reference level for gene \( G \) in the \( i \) th experiment
Oligonucleotide Arrays

• most common are Affymetrix’s GeneChips™

Oligonucleotide Arrays

• instead of putting entire genes on an array, put sets of DNA 25-mers (oligonucleotides)
• oligos are synthesized on the chip using a photolithography process similar to that used to make semiconductor chips
• mRNA samples are processed separately instead of in pairs
Oligonucleotide Arrays

• given a gene to be measured, select 20 different 25-mers for the gene

```
gene
   ↓  ↓  ↓  ↓  ↓
25-mers
```

• selection criteria
  – specificity
  – hybridization properties
  – ease of manufacturing

Oligonucleotide Arrays

• put each of these 25-mers on the chip
• additionally a slight variant (that differs only at the 13th base) of each is put next to it
  – this helps factor out false hybridizations
• the measurements for a gene are derived from these 40 separate measurements
  – present/absent calls
  – numerical quantity proportional to amount of mRNA present
A Gene Expression Profile

Several Computational Tasks

- clustering genes: which genes seem to be regulated together
- clustering samples: which treatments/individuals have similar profiles
- classifying genes: to which functional class does a given gene belong
- classifying samples: to which class does a given sample belong
  - e.g., does this patient have ALL or AML
  - e.g., does this chemical act like an AHR agonist, or a PCB or …
- inferring regulatory networks: what is the “circuitry” of the cell