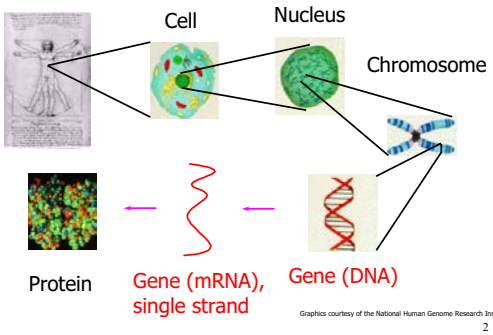
Microarrays

- Technology behind microarrays
- Data analysis approaches
- Example applications

1

Molecular biology overview



Graphics courtesy of the National Human Genome Research Institute

2

Gene expression

- Cells are different because of **differential gene expression.**
- About 40% of human genes are expressed at any one time.
- Gene is expressed by **transcribing** DNA into single-stranded mRNA
- mRNA is later **translated** into a protein
- Microarrays measure the level of mRNA expression

3

Basic idea

- mRNA expression represents dynamic aspects of cell
- mRNA expression can be measured with latest technology
- mRNA is isolated and labeled with fluorescent protein
- mRNA is hybridized to the target; level of hybridization corresponds to light emission which is measured with a laser
- Higher concentration ➡ more hybridization ➡ more mRNA

4

A demonstration

- DNA microarray animation by A. Malcolm Campbell.

5

Experimental conditions

- Different tissues
- Different developmental stages
- Different disease states
- Different treatments

6

Background papers

- Background paper 1
- Background paper 2
- Background paper 3

7

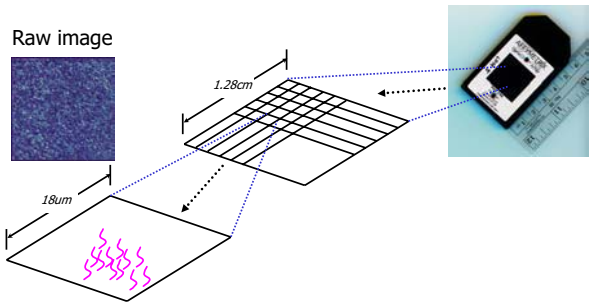
Microarray types

The main types of gene expression microarrays:

- Short oligonucleotide arrays (Affymetrix)
- cDNA or spotted arrays (Brown lab)
- Long oligonucleotide arrays (Agilent Inkjet)
- Fiber-optic arrays
- ...

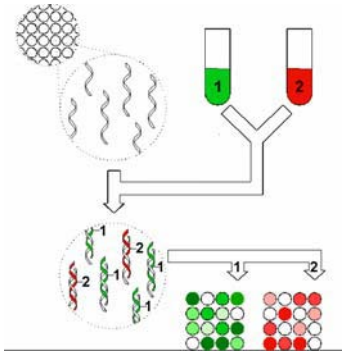
8

Affymetrix chips



9

Competitive hybridization



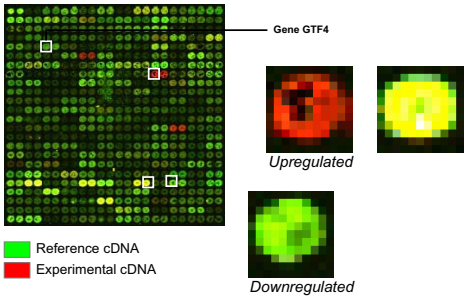
10

Microarray image data

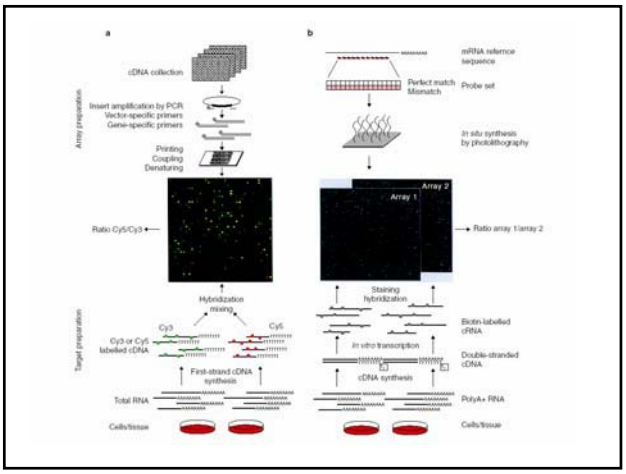
mouse heart versus liver hybridization

11

More images



12



Characteristics of microarray data

- Extremely high dimensionality
  - Experiment = ($gene_1$, $gene_2$, …, $gene_N$)
  - Gene = ($experiment_1$, $experiment_2$, …, $experiment_M$)
  - N is often on the order of 10^4
  - M is often on the order of 10^1
- Noisy data
  - Normalization and thresholding are important
- Missing data
  - For some experiments a given gene may have failed to hybridize

14

Data mining challenges

- Too few experiments (samples), usually < 100
- Too many columns (genes), usually > 1,000
- Too many columns lead to false positives
- For exploration, a large set of all relevant genes is desired
- For diagnostics or identification of therapeutic targets, the smallest set of genes is needed
- Model needs to be explainable to biologists

15

Data processing

- Gridding
  - Identifying spot locations
- Segmentation
  - Identifying foreground and background
- Removal of outliers
- Absolute measurements
  - cDNA microarray
    - Intensity level of red and green channels
  - Affymetrix chips
    - Average difference of PM and MM spots

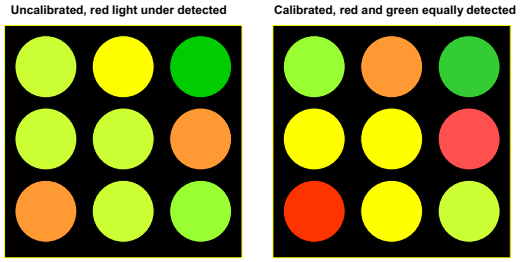
16

Data normalization

- Normalize data to correct for variances
  - Dye bias
  - Location bias
  - Intensity bias
  - Pin bias
  - Slide bias
- Control vs. non-control spots
  - Maintenance genes

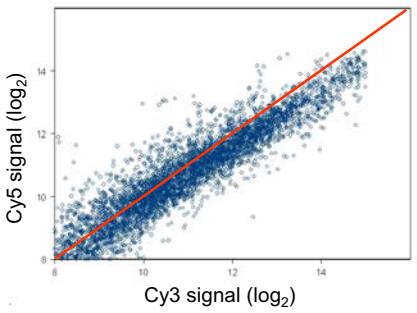
17

Data normalization



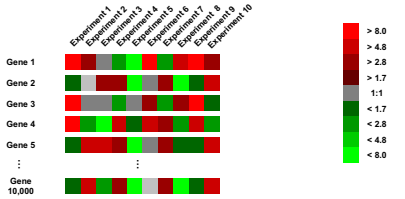
18

Normalization



19

Result of data normalization



20

Data analysis

- What kinds of questions do we want to ask?
  - Clustering
    - What genes have similar function?
    - What regulatory pathways exist?
    - Can we subdivide experiments or genes into meaningful classes?
  - Classification
    - Can we correctly classify an unknown experiment or gene into a known class?
    - Can we make better treatment decisions for a cancer patient based on gene expression profile?

21

Clustering goals

- Find natural classes in the data
- Identify new classes / gene correlations
- Refine existing taxonomies
- Support biological analysis / discovery
- Different Methods
  - Hierarchical clustering, SOM's, k-means, etc

22

Clustering techniques

- Distance measures
  - Euclidean: $\sqrt{\Sigma (x_i - y_i)^2}$
  - Vector angle: cosine of angle = $x.y / \sqrt{(x.x)}\ \sqrt{(y.y)}$
  - Pearson correlation
    - Subtract mean values and then compute vector angle
    - $(x-\underline{x}).(y-\underline{y}) / \sqrt{((x-\underline{x}).(x-\underline{x}))}\ \sqrt{((y-\underline{y}).(y-\underline{y}))}$

23

K-means clustering

- Randomly assign points to k clusters
- Iterate
  - Assign each point to its nearest cluster (use centroid of clusters to compute distance)

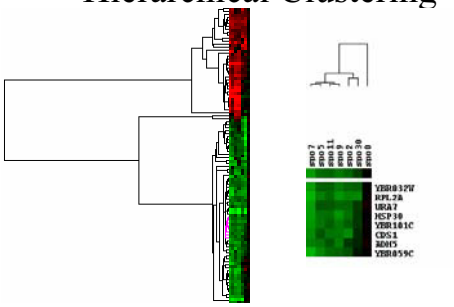
24

Hierarchical clustering

- Single linkage
  - Min distance
- Complete linkage
  - Max distance
- Average linkage
  - Similar to UPGMA

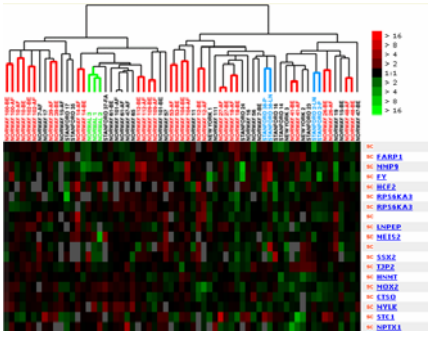
25

Hierarchical Clustering



26

Hierarchical Clustering



Perou, Charles M., et al. *Nature*, 406, 747-752 , 2000.

27

Self organizing maps (SOM)

- Artificial neural networks (Kohonen)
- Mapping of inputs to an output grid structure

28

Classification

- Challenges
  - Large number of genes and small number of experiments
  - Noise
    - Biological
    - Technical
  - Large number of irrelevant attributes

29

Classification methods

- NN
- Artificial Neural Networks
- Decision tree
  - Information gain
  - ID3, C4.5
- Bayesian Classifier
- Support Vector Machine (SVM)

30

Bayes Theorem

- **Goal:** To determine the most probable hypothesis, given the data *D* plus any initial knowledge about the prior probabilities of the various hypotheses in *H*.
- ***Prior probability of h, P(h):*** it reflects any background knowledge we have about the chance that *h* is a correct hypothesis (before having observed the data).
- ***Prior probability of D, P(D):*** it reflects the probability that training data *D* will be observed given no knowledge about which hypothesis *h* holds.
- ***Conditional Probability of observation D, P(D|h):*** it denotes the probability of observing data *D* given some world in which hypothesis *h* holds.

31

Bayes Theorem (Cont'd)

- ***Posterior probability of h, P(h|D):*** it represents the probability that *h* holds given the observed training data *D*. It reflects our confidence that *h* holds after we have seen the training data *D* and it is the quantity that Machine Learning researchers are interested in.
- ***Bayes Theorem*** allows us to compute ***P(h|D):***

$$P(h|D)=P(D|h)P(h)/P(D)$$

32

Maximum A Posteriori (MAP) Hypothesis and Maximum Likelihood

- **Goal:** To find the most probable hypothesis *h* from a set of candidate hypotheses *H* given the observed data *D*.
- ***MAP Hypothesis,*** $h_{MAP} = \underset{h \in H}{argmax}\, P(h|D)$
  $= \underset{h \in H}{argmax}\, P(D|h)P(h)/P(D)$
  $= \underset{h \in H}{argmax}\, P(D|h)P(h)$
- If every hypothesis in *H* is equally probable a priori, we only need to consider the likelihood of the data *D* given *h*, ***P(D|h).*** Then, h_{MAP} becomes the ***Maximum Likelihood***,
  $h_{ML} = \underset{h \in H}{argmax}\, P(D|h)P(h)$

33

Example

- P(cancer) = .008, P(¬cancer) = .992
- P(pos|cancer) = .98, P(neg|cancer) = .08
- P(pos|¬cancer) = .03, P(pos|¬cancer) = .97
- P(cancer|pos) = (.98*.008 = .0078) / P(pos)
- P(¬cancer|pos) = (.03*.992 = .0298) / P(pos)
- MAP implies that ¬cancer is the better hypothesis

34