High-throughput data

BBSI 2006: Lecture #(χ+4)

Takis Benos (2006)
Overview

- Transcriptomics.
  - Microarrays
  - SAGE

- Proteomics.
  - 2D, gels, 2D DIGE
  - Mass-spec
Transcriptomics: questions

Question: which genes or groups of genes are differentially expressed between two (or more) cell types/samples?

Microarray method: mRNA/cDNA is labeled and hybridizes on an array of genes (cDNAs); the intensity of the signal corresponds to the abundance of the mRNA
**cDNA arrays**

A. RNA Isolation

B. cDNA Generation

C. Labeling of Probe
   - Reverse Transcriptase
   - Fluorescent Tags

D. Hybridization to Array

E. Imaging

- Sample A > B
- Sample B > A
- Sample A = B

Source: [http://www.fao.org/DOCREP/003/X6884E/x6884e00.jpg](http://www.fao.org/DOCREP/003/X6884E/x6884e00.jpg)
Affymetrix microarrays

Source: http://www.affymetrix.com/
cDNA arrays: points of caution

Variability/noise:

• cross-hybridization variability

• *a priori* knowledge of gene structure

• fluorescence dye variability

• machine printing variability

• exposure variability
Affy chips: general comments

Variability/noise:

• cross-hybridization variability
• a priori knowledge of gene structure
• alternative spliced messages?
Experimental design

Location/spot variability:

• replicate spots

• distribute them around the array
Experimental design (cntd)

Dye variability:

• dye swap

Source:
http://www.stat.purdue.edu/research/coalesce/bioinformatics/Center_for_Bioinformatics/protein_array_analysis.html
Experimental design (cntd)

Array variability:

- replicate whole experiment! (not just technical replicas)
Data pre-processing

Data extraction:

• Identify (and exclude) “damaged” areas
• Spot identification
• Spot quality control
• Quantification

Data transformation:

• Typically, log-values are considered
Data pre-processing (cntd)

Data normalization:

- Within-slide
- Between slides
Data analysis

Supervised learning (classification):

• Main aim: to build robust classifiers
• $k$ (known) classes of genes exist
• Examples of expression levels for these genes are available
• Rules are learnt from the examples and applied in new cases (of unknown class)
• Application in disease classification, disease progression, response to treatment, etc
Data analysis (cntd)

Unsupervised learning (clustering):

- **Main aim**: to identify subsets (clusters) of genes that “behave similarly”
- No labels exist *a priori*
- The number of clusters, $k$, is usually unknown

- *Application in discovery of biological information*
Unsupervised learning

- K-means clustering algorithm:
  1. Start with a guess for the $k$ cluster centers
  2. Select $k$ centroids at random or at the maximum distance from each other (Euclidean distance)
  3. For each point, find the closest cluster centroid
  4. Replace each centroid by the coordinate-wise average of all data points that are closest to it
  5. Repeat steps #3 and #4 until no change in the cluster memberships
  6. Repeat the algorithm for different values of $k$
Unsupervised learning (cntd)

• Hierarchical clustering algorithm (divisive):
  1. Calculate all pairwise distances between data points
  2. The two closest points are joined into a cluster
  3. Calculate the centroid of the cluster and calculate the pairwise distances from this point to all other points
  4. Repeat steps #2 and #3 until no points left

• Hierarchical clustering algorithm (agglomerative):
The reverse of the divisive algorithm.
Hierarchical clustering

- Agglomerative:
  - ABCDE
  - ABC
  - AB
  - A

- Divisive:
  - ABCDE
  - ABC
  - EF
  - E

Hierarchical clustering (cntd)

Source: http://www.oncology.cam.ac.uk/images/JB2.gif
Serial Analysis of Gene Expression


**SAGE**

<table>
<thead>
<tr>
<th>SAGE Tag</th>
<th>Tag Count (in 100,000 tags)</th>
<th>Absolute abundance</th>
</tr>
</thead>
<tbody>
<tr>
<td>CATGGAGTGCTTTAAT</td>
<td>33 TAGS</td>
<td>0.033%</td>
</tr>
<tr>
<td>CATGGAGTGCTTTT</td>
<td>63 TAGS</td>
<td>0.063%</td>
</tr>
<tr>
<td>CATGGAAGAAGAAGA</td>
<td>22 TAGS</td>
<td>0.022%</td>
</tr>
<tr>
<td>CATGGAGTGAGGGTGG</td>
<td>9 TAGS</td>
<td>0.009%</td>
</tr>
</tbody>
</table>

Trends in Genetics
Microarrays vs. SAGE

Microarrays:
• hybridization variability
• *a priori* knowledge of the genes (exact or non-exact structure)

SAGE:
• time/resource consuming
• sequencing errors decrease efficiency
Proteomics technologies

- 2D gels (classical)
- 2D Difference gel electrophoresis (DIGE)
- Mass fingerprinting (e.g., MALDI-TOF)
- Antibody arrays
- Multi ligand arrays
2D gels

Source: http://www.millipore.com
2D DIGE

2D DIGE (cntd)

Cy5 and Cy3 DIGE 2D Gel

Courtesy: Massimo Trucco MD, Children’s Hospital of Pittsburgh

BBSI 2006 1-JUN-2006
Mass-spec fingerprinting

Protein sequence

Theoretical proteolytic peptides

Mass spectrum

Theoretical mass spectrum

Courtesy: Steve Ringquist PhD, RANGOS Research Center
2D gels & mass-spec

Source: http://www.amershambiosciences.com
Antibody arrays

**ChemiArray™ Protocol**

1. Incubate sample with antibody-coated array 1–2 hrs
2. Incubate array with Cytokine antibody cocktail 1–2 hrs
3. Incubate array with conjugated HRP-streptavidin 1 hr
4. Detect signal
5. Data analysis

Source: [http://www.chemicon.com](http://www.chemicon.com)
Multi ligand arrays

Technologies for protein-DNA interactions
Overview

• *In vivo* target identification.
  ▪ ChIP, ChIP-on-chip
  ▪ STAGE

• *In vitro* target identification.
  ▪ SELEX
  ▪ Phage display
  ▪ Protein-DNA interaction chips
  ▪ Band-shifts, QuMFRA
Chromatin immunoprecipitation (ChIP)
Chromatin immunoprecipitation (ChIP)
Chromatin immunoprecipitation (ChIP)

+ Ab
Chromatin immunoprecipitation (ChIP)

LM-PCR

Proteinase K
ChIP-on-chip

Source: Lee et al. Science 2002 298:799-804
ChIP-on-chip: yeast study

Fig. 1. Systematic genome-wide location analysis for yeast transcription regulators. (A) Methodology. Yeast transcriptional regulators were tagged by introducing the coding sequence for a c-myc epitope tag into the normal genomic locus for each regulator. Of the yeast strains constructed in this fashion, 106 contained a single epitope-tagged regulator whose expression could be detected in rich growth conditions. Chromatin immunoprecipitation (ChIP) was performed on each of these 106 strains. Promoter regions enriched through the ChIP procedure were identified by hybridization to microarrays containing a genome-wide set of yeast promoter regions. (B) Effect of P value threshold. The sum of all regulator-promoter region interactions is displayed as a function of varying P value thresholds applied to the entire location data set for the 106 regulators. More stringent P values reduce the number of interactions reported but decrease the likelihood of false-positive results.

Source: Lee et al. Science 2002 298:799-804
**Homologous recombination**

ChIP-on-chip: yeast study (cntd)

Source: Lee et al. Science 2002 298:799-804
ChIP-on-chip: yeast study (cntd)

Source: Lee et al. Science 2002 298:799-804
ChIP-on-chip: spot normalization

Single Array Error Model.

- Intensity spot ratio (X):

\[ X = \frac{a_2 - a_1}{(s_1 + s_2 + f(a_1 + a_2))/2} \]

- X is Normally distributed.
- \( f, s_1 \) and \( s_2 \) are chosen so that \( \text{Var}(X) = 1 \).

Significance of X.

\[ P(X = x) = 2 \cdot (1 - \text{erf}(|x|)) \]

\[ \text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_{0}^{z} e^{-t^2} \, dt. \]
ChIP-on-chip: multiple measurements

Uncertainty.

$$\sigma = \frac{\log_{10}(a_2 / a_1)}{X}$$

Weights ($w_i$).

- **Method**: minimum variance weighted average

Average measurement ratio.

$$\bar{x} = \frac{\sum_{i=1}^{3} w_i \cdot x_i}{\sum_{i=1}^{3} w_i}$$
SELEX
Systematic Evolution of Ligands by EXponential enrichment
Tuerk and Gold, Science (1990)

primer-1 ——— GCGNNNGCG ——— primer-2

re-amplification

PCR amplification

selection

sequencing
Phage display

Figure Selection of ganglioside-binding peptides from phage peptide library and the analysis of carbohydrate-peptide interaction

Source: http://www.glycoforum.gr.jp/science/word/glycotechnology/GT-C08E.html
## Selection data

<table>
<thead>
<tr>
<th>SELEX</th>
<th>Phage display</th>
</tr>
</thead>
<tbody>
<tr>
<td>s r s d h l t t h i r</td>
<td>5' g c g g a a g c g</td>
</tr>
<tr>
<td>5' g c g g g g g c g</td>
<td>s Q G G N l V R h L r</td>
</tr>
<tr>
<td>5' g c g g g g g a g</td>
<td>s N G G N l G R h M k</td>
</tr>
<tr>
<td>5' g c g g g t g c g</td>
<td>s A R S N l L R h T r</td>
</tr>
<tr>
<td>5' g c g t g g g c g</td>
<td>s L Q S N l V R h Q r</td>
</tr>
<tr>
<td>5' g a g g g g c g</td>
<td>s I A S N l L R h Q r</td>
</tr>
</tbody>
</table>

| s r s d E l t R h i r | 5' g c g c a g g c g          |
| 5' g c g g g g c g   | s R G D H l K D h I k        |
| 5' g c g t g g g c g | s R S D H l T T h I r        |
**Protein-DNA chips**

Protein-DNA chips (cntd)

Source: Bulyk et al., Proc Natl Acad Sci USA (2001)

Quantitative data: band-shifts

Source: http://www.piercenet.com/media/super_oct1.jpg

Source: http://www.biomedcentral.com/content/figures/1471-2091-3-13-3.jpg
Quantitative data: QuMFRA

Quantitative Multiple Fluorescence Relative Affinity
Quantitative data: QuMFRA (cntd)

\[
\frac{K_A (D_1)}{K_A (D_2)} = \frac{[P \cdot D_1]/[P] \cdot [D_1]}{[P \cdot D_2]/[P] \cdot [D_2]} = \frac{[P \cdot D_1]/[P \cdot D_2]}{[D_1]/[D_2]}
\]


BBSI 2006 1-JUN-2006 © 2006 P. Benos 49
Acknowledgements

- *Special Thanks to...*
  - Massimo Trucco MD & Steve Ringquist PhD, Children’s Hospital and RANGOS Diabetes Research Center, University of Pittsburgh *(for proteomics pictures)*
  - Eleanor Feingold PhD, Human Genetics, GSPH, University of Pittsburgh *(for slides)*