Data Mining For Genomics

Alfred O. Hero III
The University of Michigan, Ann Arbor, MI

ISTeC Seminar, CSU
Feb. 22, 2003

1. Biotechnology Overview
2. Gene Microarray Technology
3. Mining the genomic database
4. The post-genomic era
I. Biotechnology Overview

- **Genome:** All the DNA contained in an organism. The operating system/program for gene structure/function of an organism.

- **Genomics:** Investigation of structure and function of very large numbers of genes undertaken in a simultaneous fashion.

- **Bioinformatics:** Computational extraction of information from biological data.

- **Data Mining:** Algorithms for extracting information from huge datasets using user-specified criteria.
THE STRUCTURE OF DNA

one helical turn
= 3.4 nm

Sugar-phosphate backbone

Base

Hydrogen bonds

http://www-stat.stanford.edu/~susan/courses/s166/node2.html
Central Dogma: From Gene to Protein

Source: NHGRI http://www.genome.gov/
Towards a unified theory . . .

- DNA
  - Map Databases
  - GenBank
  - EMBL
  - DDBJ

- RNA
  - Gene Expression?
  - Development?

- Proteins
  - PDB
  - SwissPROT
  - PIR

- Circuits
  - Regulatory Pathways?
  - Metabolism?

- Phenotypes
  - Clinical Data?
  - Neuroanatomy?

- Populations
  - Biodiversity?
  - Molecular Epidemiology?
  - Comparative Genomics?

Source: http://www.biotech.ucdavis.edu/powerpoint/powerpoint.htm
Hierarchy of biological questions

- **Gene sequencing**: what is the sequence of base pairs in a DNA segment, gene, or genome?
- **Gene Mapping**: what are positions (loci) of genes on a chromosome?
- **Gene expression profiling**: what is pattern gene activation/inactivation over time, tissue, therapy, etc?
- **Genetic circuits**: how do genes regulate (stimulate/inhibit) each other’s expression levels over time?
- **Genetic pathways**: what sequence of gene interactions lead to a specific metabolic/structural (dys)function?
## Sequencing Milestones

<table>
<thead>
<tr>
<th>Organism</th>
<th># of genes</th>
<th>% genes with inferred function</th>
<th>sequencing complete</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. Coli</td>
<td>4,288</td>
<td>60</td>
<td>1997</td>
</tr>
<tr>
<td>Yeast</td>
<td>6,600</td>
<td>40</td>
<td>1996</td>
</tr>
<tr>
<td>C. Elegans</td>
<td>19,000</td>
<td>40</td>
<td>1998</td>
</tr>
<tr>
<td>Drosophila</td>
<td>12,000-14,000</td>
<td>25</td>
<td>1999</td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>25,000</td>
<td>40</td>
<td>2000</td>
</tr>
<tr>
<td>Mouse</td>
<td>26,000-40,000</td>
<td>10-20</td>
<td>2002</td>
</tr>
<tr>
<td>Human</td>
<td>26,383-39,114</td>
<td>10-20</td>
<td>2001</td>
</tr>
</tbody>
</table>

Nucleic Acid Hybridization
II. Gene Microarray Technologies

- High throughput method to probe DNA in a sample
- Two principal microarray technologies:
  1) Affymetrix GeneChip
  2) cDNA spotted arrays

- Main idea behind cDNA technology:
  1) Specific complementary DNA sequences arrayed on slide
  2) Dye-labeled RNA from sample is distributed over slide
  3) RNA binds to probes (hybridization)
  4) Presence of bound RNA-DNA pairs is read out by detecting spot fluorescence via laser excitation (scanning)

- Result: 10,000–50,000 genes can be probed at once
Specialized cDNA Array: Eye-Gene

wt RNA → Label RNA with fluorescent dye (Red) → Isolate RNA → I-gene slides → Gene Expression

ko RNA → Label RNA with fluorescent dye (Green) → Isolate RNA

I-Gene Array: Probe Generation

1. Isolate RNA from tissue
2. Construct cDNA library in a plasmid vector
3. Pick clones from library and store in 96-well plate glycerol stocks
4. PCR amplification from glycerol stocks
5. Removal of excess primers and agarose gel analysis
6. Conversion to 384-well plates

Farjo, R & Yu, J. Vision Research 42 (2002)
I-Gene Array: Printing and Processing

384-well plate

cDNAs printed on glass slides

Slide processing
1. Target labeling
2. Hybridization
3. Scanning
4. Data Analysis

Farjo, R & Yu, J. Vision Research 42 (2002)
• Treated sample labeled red (Cy5)
• Control data labeled green (Cy3)
Single-Chip Raw Data Analysis

Problem: Experimental Variability

- **Population** – too wide genetic diversity
- **Cell lines** - poor sample preparation
- **Slide Manufacture** – slide surface quality, dust deposition
- **Hybridization** – sample concentration, wash conditions
- **Cross hybridization** – similar but different genes bind to same probe
- **Image Formation** – scanner saturation, lens aberrations, gain settings
- **Imaging and Extraction** – misaligned spot grid, segmentation

Microarray data is intrinsically Statistical!
III. Mining Statistical Genomic Data

Questions:

- How to estimate true Cy5 and Cy3 from raw data?
- How to compensate for experimental variability?
- How to extract expression profile ratios from a set of up to 50,000 probe responses?
- How to specify gene profile selection criteria for mining in this data?
- How to discover complex genetic pathways to disease, aging, etc?
Mining Statistical Genomic Data.

Answers:

- Spot Extraction: Estimate Cy3 and Cy5 concentrations
  - Image processing, image segmentation, anova models
- Comparing between microarray experiments
  - Statistical invariance, equalizing transformations, normalization
- Gene filtering and screening
  - Simultaneous statistical inference, T tests, FDR
- Discovery of genetic pathways
  - Clustering, dependency graphs, HMM’s
Spot Extraction Issues

- Technical noise and variability
- Laser gain and calibration
- Cy3/cy5 channel bleedthrough
- Image formation gain
- Spot gridding algorithm
- Spot segmentation algorithm
Technical Noise and Variability

Good Signal

Weak Signal

Irregular Spots

Comet Tails

Source: http://stress-genomics.org/
Gain Effects

Weak

Normal

Saturated

Optimal gain can be studied by information theory
Rate Distortion Lower Bound

MSE

Gain

Hero Springer-03
Standard Spot Segmentation Method

- **Addressing** – Locate “center of description” for each spot
- **Spot Segmentation** – Classification of pixels either as signal or background.
- **Spot Quantification** – Estimation of hybridization level/ratio of spot

Source: C. Ball, Stanford Microarray Database
Segmentation via Morphological Operators

Original Image

Alternate-Sequential Filtered

Watershed Transformed

Final Segmented Image

Siddiqui, Hero and Siddiqui, Asilomar-02

The University of Michigan Dept. of EECS
Spot EigenAnalysis

- Gray level covariance matrix over each spot boundary is calculated.
- Eigen analysis of each covariance matrix is performed.
- Trends in direction of eigenvectors indicate systematic bias in spot printing.

Siddiqui, Hero and Siddiqui, Asilomar-02
Add Dimension: Expression Profiles

Cy5/Cy3 hybridization profiles
Problem: Intrinsic Profile Variability

Across gene variability

Within gene variability
Solution: Experimental Replication

Issues:
- Control by experimental replication is expensive
- Surplus real estate allows replication in layout
- Batch and spatial correlations may be a problem
Comparing Across Microarray Experiments

Question: How to combine or compare experiments A and B?
Un-Normalized Data Sets

Within experiment intensity variations mask A B differences:

Experiment A (Wildtype)  Experiment B (Knockout)

Hero & Fleury, ISSP-03
Two Approaches

- If quantitative gene profile comparisons are required:
  - must find normalization function to align all data sets within an experiment to a common reference.

- If only ranking of gene profile differences is required:
  - No need to normalize: can apply rank order transformation to measured hybridization intensities
A vs B Microarray Normalization Method

Exp A

Housekeeping
Gene Selector

Inverse
Mean
Unif Tran

Exp B

Inverse
Mean
Unif Tran

Normalized A

Normalized B
Un-Normalized Data Set (Wildtype)

Unnormalized, data=wild vs knockout

Hero & Fleury, ISSP-03
Normalized Data Set (Wildtype)
Rank Order Statistical Transformation

- **Rank order algorithm**: at each time point replace each gene intensity with its relative rank among all genes
  - The relative ranking is preserved by (invariant to) arbitrary monotonic intensity transformations.
Mining Gene Expression Data

- Issues
  - Feature space
  - Feature selection criteria
  - Statistical robustification
  - Cross-validation
  - Experimental Validation
Y/O Human Retina Study

(2001H Retina Gene Study)

16 individuals in 2 groups of 8 subjects

Selection criteria:

\[ \xi_1(g) = \frac{|O(g) - Y(g)|}{\sigma_0^2(g) + \sigma_Y^2(g))^{1/2}} \]

\[ \xi_2(g) = \frac{1}{\sqrt{\sigma_0^2(g) + \sigma_Y^2(g))^{1/2}}} \]
Fred Wright’s Human Fibroblast Data

Lemon & et al. 2001

18 individuals in
3 groups of 6 subjects

Selection criteria:

\[ \xi_1(g) = (\mu_{100}(g) - \mu_{50}(g))(\mu_{50}(g) - \mu_0(g)) \]
\[ \xi_2(g) = (\sigma_{100}^2(g) + \sigma_{50}^2(g) + \sigma_0^2(g))/3 \]
Mouse Retinal Aging Data

Yosida et al: 2003

Selection criteria:

\[
\xi_1(g) = \Delta_{M21,M2}(g) = (\mu_{M21}(g) - \mu_{M2}(g))^2
\]

\[
\xi_2(g) = \max_{t=3,\ldots,6} \{ \text{var}(\Delta_{t+1,t})(g) \}
\]

24 mice in
6 groups of 4 subjects
NRL Knockout vs Wildtype Retina Study

12 knockout/wildtype mice in 3 groups of 4 subjects

Selection criteria:
\[ \xi_1(g) = \Delta_{K,W}^2(g) = \|\mu_K(g) - \mu_W(g)\|^2 \]
\[ \xi_2(g) = \max\{\text{var}_K(g), \text{var}_W(g)\} \]
Data Mining with a Single Criterion

- Paired t-test with False Discovery Rate:

\[ T(g) = \frac{\xi_1(g)}{\xi_2(g)} > T_{2(m-1)}^{-1}(1 - \alpha/2) \]

- For Y/O Human study:

\[ T(g) = \frac{|\overline{O}(g) - \overline{Y}(g)|}{\sqrt{(\sigma^2_O(g) + \sigma^2_Y(g))/2}} \]
Multicriterion scattergram: Paired t-test

8226 Y/O retina genes plotted in multicriteria plane

Fleury et al. ICASSP-02
Multicriterion Selection Criteria

- Seek to find Pareto-optimal genes which strike a compromise between two criteria.

A, B, D are Pareto optimal.

Pareto Fronts
Multicriterion scattergram: Pareto Fronts

Pareto fronts
- first
- second
- third

Buried gene

Fleury & et al. ICASSP-02
Cross-Validation Approach: Resampling

# replicates = m = 4
# time points = t = 6
# profiles = 4^6 = 4096
Bayesian approach: Posterior Analysis

\[ P(i|Y) = P(\text{gene } i \text{ on PF } | \text{ data } Y) \]
### Pareto Front Likelihood Table

| PPF linear contrast | $P(i|Y)$ | RPF linear contrast | $P(i|Y)$ | RPF non-parametric | $P(i|Y)$ |
|--------------------|---------|--------------------|---------|-------------------|---------|
| AFFX-ThrX-5-at     | 0.999   | AFFX-DapX-5-at     | 0.944   |                   |         |
| HG3342-HT3519-s-at | 0.998   | AFFX-ThrX-5-at     | 0.694   |                   |         |
| AFFX-DapX-5-at     | 0.998   | AFFX-ThrX-5-at     | 0.685   |                   |         |
| HG831-HT831-at     | 0.996   | HG3342-HT3519-s-at | 0.662   |                   |         |
| AFFX-ThrX-M-at     | 0.986   | HG831-HT831-at     | 0.648   |                   |         |
| X69111-at          | 0.984   | U14394-at          | 0.352   |                   |         |
| U14394-at          | 0.974   | V00594-at          | 0.301   |                   |         |
| AFFX-LysX-3-at     | 0.962   | X69111-at          | 0.287   |                   |         |
| V00594-at          | 0.955   | U45285-at          | 0.245   |                   |         |
| U45285-at          | 0.932   | AFFX-LysX-3-at     | 0.176   |                   |         |
| AB000115-at        | 0.899   | AFFX-HSAC07/X00351-5-at | 0.111 |                   |         |
| AFFX-HSAC07/X00351-5-at | 0.866 | AB000115-at        | 0.417   | D29992-at          | 0.083   |
| U73379-at          | 0.837   | U73379-at          | 0.042   |                   |         |
| AFFX-DapX-M-at     | 0.678   | V00594-s-at        | 0.042   |                   |         |
| Y09912-ma1-at      | 0.67    | U75362-at          | 0.042   |                   |         |
| U75362-at          | 0.56    | AFFX-PheX-5-at     | 0.032   |                   |         |
| AFFX-DapX-3-at     | 0.555   | U03399-at          | 0.032   |                   |         |
| V00594-s-at        | 0.554   | U75362-at          | 0.028   |                   |         |
| HG1980-HT2023-at   | 0.483   | S70585-ma1-at      | 0.014   |                   |         |
| HG3044-HT3742-s-at | 0.441   | L02320-at          | 0.009   |                   |         |
| D43636-at          | 0.389   | L05515-at          | 0.009   |                   |         |
| L27624-s-at        | 0.387   | V00594-at          | 0.009   |                   |         |
| U03399-at          | 0.378   | X69111-at          | 0.009   |                   |         |
| S69370-s-at        | 0.321   | AFFX-PheX-5-at     | 0.005   |                   |         |
| AFFX-PheX-5-at     | 0.315   | HG174-HT174-at     | 0.005   |                   |         |

*Hero & Fleury: VLSI03*
Robustification and Validation Issues

- Cross-validation recomputes Pareto fronts over resampled virtual profiles (Fleury et al: 2002).
- Bayesian Pareto front also robustifies to prior uncertainty in data (Hero & Fleury: 2002).

Computational issues:

- Cross-validated fronts: completely data-driven but computation is exponential in # replicates (m) and # time points (t).
- Bayesian Pareto fronts: requires joint density of criteria and marginalization. Computation is linear.
The Post-Genomic Era

- Whole genomes of species will be mapped
- Genetic pathways to structure, metabolism, disease, will remain as open questions
- Pathway analysis: what are the important gene interactions?
  - Requires performing many more experiments than zero-interaction analysis
  - Computational load is exponentially increasing in number of genes in pathway
  - New algorithms and models are needed
Draft Pathways for Photoreceptor Function

Wnt/Ca – calmodulin pathway

Bmp pathway

Retinoid acid pathway

Each Link: Gene Co-regulation Study

Bmp genes

RA genes

Calmodulin family
Conclusions

- Signal processing, math, computer science, statistics: ever-increasing role in genomics
- New frontiers:
  - Protein arrays
  - Mass Spect
  - Molecular Imaging
- Bottleneck will remain: computational and statistical inadequacies!
Dawning of Post-Genomic Era
Post-Post-Genomic Era?