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.
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?
<table>
<thead>
<tr>
<th>Clone ID</th>
<th>GenBank</th>
<th>GeneName</th>
<th>Symbol</th>
<th>UniGene</th>
<th>LocusLink</th>
<th>Chr.</th>
<th>Molecular Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRA-0298</td>
<td>BC013125</td>
<td>Similar to Rhodopsin</td>
<td>LOC212541</td>
<td>Mm.2965</td>
<td>212541</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>MRA-0299</td>
<td>unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRA-0300</td>
<td>NM_008938</td>
<td>Peripherin 2</td>
<td>Prph2</td>
<td>Mm.5032</td>
<td>19133</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>MRA-0301</td>
<td>NM_008831</td>
<td>Prohibitin</td>
<td>Phb</td>
<td>Mm.2355</td>
<td>18673</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>MRA-0302</td>
<td>bad seq</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRA-0303</td>
<td>bad seq</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRA-0304</td>
<td>unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRA-0305</td>
<td>M19381</td>
<td>Calmodulin 1</td>
<td>Calm1</td>
<td>Mm.34246</td>
<td>12313</td>
<td>7</td>
<td>calcium ion binding::</td>
</tr>
<tr>
<td>MRA-0306</td>
<td>M28727</td>
<td>Tubulin, alpha 2</td>
<td>Tuba2</td>
<td>Mm.197515</td>
<td>22143</td>
<td>2</td>
<td>GTP binding::</td>
</tr>
<tr>
<td>MRA-0307</td>
<td>BF469955</td>
<td>RIKEN cDNA 1110018F16 gene</td>
<td>1110018F16Rik</td>
<td>Mm.40490</td>
<td>68594</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>MRA-0308</td>
<td>J00376</td>
<td>Crystallin, alpha A</td>
<td>Cryaa</td>
<td>Mm.1228</td>
<td>12954</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>MRA-0309</td>
<td>BB284055</td>
<td>Expressed sequence AIS97479</td>
<td>AIS97479</td>
<td>Mm.28817</td>
<td>98404</td>
<td>1</td>
<td>ATP binding::phospholipid transporter::ATP-binding cassette (ABC) transporter::</td>
</tr>
<tr>
<td>MRA-0310</td>
<td>NM_007378</td>
<td>ATP-binding cassette, subfamily A (ABC1), member 4</td>
<td>Abca4</td>
<td>Mm.3918</td>
<td>11304</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Clone ID</th>
<th>Biological Process</th>
<th>Cellular Components</th>
<th>Tissue Expressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRA-0298</td>
<td>eye, adult-retina, eyeball, retina, spinal ganglion, embryonic body between diaphragm and</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRA-0299</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRA-0300</td>
<td>integral membrane protein:</td>
<td>eye, adult-retina, nervous system, retina, eyeball</td>
<td></td>
</tr>
<tr>
<td>MRA-0301</td>
<td>embryo, whole embryo, mammary, kidney, colon, nervous system, skin, melanoma, tonsil,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRA-0302</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRA-0303</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRA-0304</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRA-0305</td>
<td>cell cycle:</td>
<td>embryo, whole embryo, hippocampus, testis, gonad, forelimb, branchial arches, mammary</td>
<td></td>
</tr>
<tr>
<td>MRA-0306</td>
<td>microtubule-based process:</td>
<td>mammary, embryo, whole embryo, brain, spinal cord, spinal ganglion, head, neural retina</td>
<td></td>
</tr>
<tr>
<td>MRA-0307</td>
<td>microtubule-based movement:</td>
<td>nervous system, spleen, cortex, muscle, tail cell, head, hippocampus, spinal cord, basal ganglia</td>
<td></td>
</tr>
<tr>
<td>MRA-0308</td>
<td>sensory organ development:</td>
<td>eyeball, head, embryo, whole embryo, neural retina, eye, adult, spleen</td>
<td></td>
</tr>
<tr>
<td>MRA-0309</td>
<td>cytoplasm:</td>
<td>heart, liver, head, amygdala, mammary gland, pancreas, mammary, urinary bladder, embryonic body</td>
<td></td>
</tr>
<tr>
<td>MRA-0310</td>
<td>vision: transport: phospholipid transfer to membrane:</td>
<td>heart, brain, head, adult-retina, pineal-glands, embryonic body between diaphragm and</td>
<td></td>
</tr>
</tbody>
</table>

Genome Sequencing Status

- Whole genome has been sequenced for over 1000 viruses and over 100 microbes
- Plant and animal genomes sequenced
  - Oat, soybean, barley, rice, wheat, corn
  - Mouse, zebrafish, human
- Plant and animal genomes in progress
  - Cotton, tomato, potato…
  - Rabbit, dog, chicken…

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>

Source: http://www.biotech.ucdavis.edu/powerpoint/powerpoint.htm
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
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**

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)
I-Gene Array: Image Formation

Target cDNA → Microarray → Laser scanner → PMT → CRT

Probe cDNA

overlay images

analysis
• 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: Cy5/Cy3 or Cy5-Cy3?
  - Image processing, image segmentation, non-linear 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

Streaks

Comet Tails

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

Weak  Normal  Saturated

Optimal gain can be studied by information theory
Spot Segmentation Failure Modes

Grid misalignment

Laser Misalignment

Source: C. Ball, Stanford Microarray Database
Steps in Conventional Segmentation

- **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

Mathematical morphology unifies these steps
Segmentation via Morphological Operators

Original Image

Alternate-Sequential Filtered

Watershed Transformed

Final Segmented Image

Siddiqui, Hero and Siddiqui, Asilomar-02
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
Circularity Coefficient: $\text{mean}(r^2)/(\text{mean}(r))^2$

- Plot for Radius vs. Sphericity Coefficients (measure of circularity) of spots
- Spots with lower sphericity coefficients appear in lower half of plane
- Closer the sphericity coeff. is to 1, the better it is
- Deviation from circularity may give cause to discard data

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

Cy5/Cy3 hybridization profiles
Problem: Intrinsic Profile Variability

Across-gene variability

Within-gene variability
Solution: Experimental Replication

Exps 1, 2, M

M replicates

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

Experiment A

Experiment B

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

Exp B

- Unif Tran

Inverse

Mean

Normalized A

Normalized B

Unif Tran

Unif Tran

Exp A

Exp B
Un-Normalized Data Set (Wildtype)
Normalized Data Set (Wildtype)
Profile Rank Order Statistics

- 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

(Yosida & et al: 2002)

16 individuals in 2 groups of 8 subjects

Selection criteria:

\[ \xi_1(g) = \overline{O}(g) - \overline{Y}(g) \]
\[ \xi_2(g) = \frac{\sigma^2_O(g) + \sigma^2_Y(g)}{2} \]
Fred Wright’s Human Fibroblast Data

18 individuals in 3 groups of 6 subjects

Selection criteria:

\[ \xi_1(g) = (\mu_{100}(g) - \mu_{50}(g)) (\mu_{50}(g) - \mu_{0}(g)) \]

\[ \xi_2(g) = \frac{(\sigma_{100}^2(g) + \sigma_{50}^2(g) + \sigma_{0}^2(g))}{3} \]
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:

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

- For Y/O Human study:

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

![Diagram showing Pareto-optimal front and points](attachment:image.png)
Multicriterion scattergram: Pareto Fronts

Pareto fronts

- first
- second
- third

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

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

\[ P(i|Y) = P(\text{gene } i \text{ on PF} | \text{ data } Y) \]
| 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           | 0.998 | AFFX-ThrX-5-at     | 1 U14394-at | 0.944 |
| HG3342-HT3519-s-at        | 0.998 | AFFX-ThrX-5-at           | 0.998 | AFFX-ThrX-M-at     | 1 U23435-s-at | 0.694 |
| AFFX-DapX-5-at            | 0.998 | AFFX-PhexX-M-at          | 0.996 | HG3342-HT3519-s-at | 1 AFFX-LysX-3-at | 0.662 |
| HG831-HT831-at            | 0.996 | AFFX-LysX-M-at          | 0.986 | HG831-HT831-at     | 1 AFFX-LysX-M-at | 0.648 |
| X69111-at                 | 0.984 | U14394-at                | 1 U14394-at | 0.352 |
| U14394-at                 | 0.974 | V00594-at                | 0.962 | X69111-at         | 1 AB000115-at | 0.301 |
| AFFX-LysX-3-at            | 0.962 | X69111-at                | 0.955 | U45285-at         | 1 AB000115-at | 0.287 |
| V00594-at                 | 0.955 | U45285-at                | 0.932 | AFFX-LysX-3-at     | 0.917 | U53003-at | 0.176 |
| U45285-at                 | 0.932 | AFFX-LysX-3-at          | 0.899 | AFFX-HSAC07/X00351-5-at | 0.806 | M92934-at | 0.111 |
| AB000115-at               | 0.899 | AFFX-HSAC07/X00351-5-at | 0.866 | AB000115-at       | 0.417 | D29992-at | 0.083 |
| AFFX-HSAC07/X00351-5-at   | 0.866 | AB000115-at             | 0.837 | U73379-at         | 0.13 | HG831-HT831-at | 0.069 |
| U73379-at                 | 0.837 | HG831-HT831-at          | 0.678 | V00594-s-at       | 0.074 | S79522-at | 0.042 |
| AFFX-DapX-M-at            | 0.678 | V00594-s-at             | 0.67  | U75362-at         | 0.037 | V00594-s-at | 0.042 |
| Y09912-ma1-at             | 0.67  | U75362-at               | 0.56  | AFFX-PheX-5-at    | 0.028 | D43636-at | 0.032 |
| U75362-at                 | 0.56  | AFFX-PheX-5-at          | 0.555 | U03399-at         | 0.009 | U22377-at | 0.032 |
| AFFX-DapX-3-at            | 0.555 | U03399-at               | 0.554 | V00594-s-at       | 0.028 | U75362-at | 0.028 |
| V00594-s-at               | 0.554 | V00594-s-at             | 0.483 | HG1980-HT2023-at  | 0.032 | U43636-at | 0.014 |
| HG1980-HT2023-at          | 0.483 | HG1980-HT2023-at        | 0.441 | HG3044-HT3742-s-at | 0.009 | HG1980-HT2023-at | 0.009 |
| HG3044-HT3742-s-at        | 0.441 | HG3044-HT3742-s-at      | 0.389 | D43636-at         | 0.009 | HG3044-HT3742-s-at | 0.009 |
| D43636-at                 | 0.389 | D43636-at               | 0.387 | L27624-s-at       | 0.009 | D43636-at | 0.009 |
| L27624-s-at               | 0.387 | L27624-s-at             | 0.378 | S69370-s-at       | 0.009 | S69370-s-at | 0.009 |
| S69370-s-at               | 0.378 | S69370-s-at             | 0.378 | AFFX-PheX-5-at    | 0.005 | S69370-s-at | 0.005 |
| AFFX-PheX-5-at            | 0.378 | AFFX-PheX-5-at          | 0.315 | HG174-HT174-at    | 0.005 | HG174-HT174-at | 0.005 |

Hero&Fleury:VLSI03
Robustification and Validation Issues

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

**Computational issues:**

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

ISTeC Seminar, Colorado State University, 2/03

The University of Michigan Dept. of EECS
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

GENETIC DAWN

SERGUEI
FRANCE

© 2002 CARTOONISTS & WRITERS SYNDICATE. http://CartoonWa.com
Post-Post-Genomic Era?

CORRIGAN
TORONTO STAR
Toronto
CANADA

HAPPY MOTHER’S DAY...MA!

...ANOTHER REASON TO PAN CLONING

CARTOONISTS & WRITERS SYNDICATE: http://CartoonWeb.com
Or....

Well, people succeeded in reducing biodiversity down to one species, and they always thought it would be them.
Oligonucleotide GeneChip Microarray