Analysis for Gene Expression Data of the NCI 60 Cancer Cell Lines Using MCMC on a Hierarchical Effects Model

Jae K. Lee
University of Virginia
(Formerly at Lab of Molecular Pharmacology, NCI)
November 11, 2000
IPAM Functional Genomics Workshop
OUTLINE

• NCI Large Screening Program of 60 Cancer Cell Lines
• Statistical Issues in High Throughput Array Data
• Hierarchical Effects Model Using MCMC
• Web-Based Interactive Analysis Tool (GS-HEM)
60 Cancer Cell Lines (12 reference pool lines)
- Developmental Therapeutics Programs (DTP), NCI

**CELL LINES OF THE NCI DRUG SCREEN**

<table>
<thead>
<tr>
<th>Colon</th>
<th>Lung</th>
<th>CNS</th>
</tr>
</thead>
<tbody>
<tr>
<td>COLO 205</td>
<td>A549/ATCC</td>
<td>SF-268</td>
</tr>
<tr>
<td>HCC-2998</td>
<td>EKVX</td>
<td>SF-295</td>
</tr>
<tr>
<td>HCT-116</td>
<td>HOP-62</td>
<td>SF-539</td>
</tr>
<tr>
<td>HCT-15</td>
<td>HOP-92</td>
<td>SNB-19</td>
</tr>
<tr>
<td>HT29</td>
<td>NCI-H226</td>
<td>SNB-75</td>
</tr>
<tr>
<td>KM12</td>
<td>NCI-H23</td>
<td>U251</td>
</tr>
<tr>
<td>SW-620</td>
<td>NCI-H322M</td>
<td></td>
</tr>
<tr>
<td>Ovarian</td>
<td>Ovarian</td>
<td>Leukemia</td>
</tr>
<tr>
<td>IGROV1</td>
<td></td>
<td>CCRF-CEM</td>
</tr>
<tr>
<td>Melanoma</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOX IMVI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MALME-3M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SK-MEL-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SK-MEL-28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SK-MEL-5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UACC-257</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UACC-62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Renal</td>
<td>Renal</td>
<td>Renal</td>
</tr>
<tr>
<td>786-0</td>
<td>A498</td>
<td></td>
</tr>
<tr>
<td>ACHN</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Melanoma</td>
<td>Melanoma</td>
<td>Melanoma</td>
</tr>
<tr>
<td>CAKI-1</td>
<td>RXF 393</td>
<td></td>
</tr>
<tr>
<td>RENOLI</td>
<td>SN12C</td>
<td></td>
</tr>
<tr>
<td>TK-10</td>
<td>UO-31</td>
<td></td>
</tr>
<tr>
<td>Prostate</td>
<td>Prostate</td>
<td>Prostate</td>
</tr>
<tr>
<td>PC-3</td>
<td>DU-145</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The NCI Cancer Drug Discovery - Development Pipeline

Inventory

60-Cell line screen

In vivo studies

Clinical Trials

~300,000 cmpds

> 68,000 cmpds

60 cell lines

Database A (Activity Patterns)

>300,000 cmpds

> 150 specific targets

>30K cmpds

>68K cmpds

>100 descriptors

Database S ( >100 3-D Molecular Structure)

~300K cmpds

>30,000,000 numbers

Database T (Molecular Targets)

>100 descriptors

>4,000,000 numbers

mRNA expression database
-cDNA microarray
-Oligonucleotide array

~ 68K pds

>4,000,000 numbers

>68,000 cmpds

> >300,000 cmpds

~68K numbers
60 Cancer Cell Line Screening Data

• **Activity Data (A)**
  - Drug Potency Activity; \( GI_{50}, TGI \)
    - 29,026 open compounds (Sep. '99)
    - 6,205 drugs tested more than once (Jan. '00)
    - 118 mechanism of action drug compounds

• **Target Data (T)**
  - Protein (41 protein expression data)
  - cDNA Hybridization Expression
    - Microarray (9,706 genes x 60 cells)
    - Oligonucleotide Array (6,800 genes x 60 cells)
OUTLINE

• NCI Large Screening Program of 60 Cancer Cell Lines
• Statistical Issues in High Throughput Array Data
• Hierarchical Effects Model Using MCMC
• Web-Based Interactive Analysis Tool (GS-HEM)
Analyzing Gene Expression Array Data

Need pre-processing and pre-screening to avoid (large number of) irrelevant results from various artifacts

- **quality control:**
  - optimization criteria
    - normalization (scaling or centering factors)
    - reproducibility (ratio, Chen et al. 97; ave. difference, Affymetrix)
    - sensitivity (power for identifying differential expression levels)
  - thresholding or missing values
    - maximum number of informative data points
- **subsetting:** depends on inference goals to identify (e.g., Tibshirani et al., 2000)
  - genes with distinctive patterns in specific cases
  - genes with major expression variations
Myths in Gene Chip Study

• **Myth 1: Can do with each single hybridization**
  • **Question:** how can we do array study for 3-4 biological factors using 4 chips?
  • **Answer:** don’t do it!
    - Enormous false positive findings, say $10^{-2}$ from single chip $\rightarrow$ from duplicates $10^{-2} \times 10^{-2}$
    - Waste of time for a week for statisticians, for several months for biologists

• **Myth 2: Can do without a statistical design**
  - statistical factors of variability
    • gene
    • variety: types of sample, treatment, time
    • individual sample
    • array
    • dye (microarray)
Reciprocally Labeled Pairs of Microarray data
Analyzing Array Data (continued)

- replication and experimental design (blocking)
  - replicates of genes on a chip and/or of treatments on several chips, especially for interaction
  - blocking errors from individual, array, and dye (not interested in identifying them separately, but need to have replicates to “factor them out” together)
Example: Experimental design on an array study

• **Microarray study on** comparing a treatment effect at two different time points with two individual replicates

<table>
<thead>
<tr>
<th>Chip 1</th>
<th>Chip 2</th>
<th>Chip 3</th>
<th>Chip 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cy3</td>
<td>Cy3</td>
<td>Cy3</td>
<td>Cy3</td>
</tr>
<tr>
<td>Cy5</td>
<td>Cy5</td>
<td>Cy5</td>
<td>Cy5</td>
</tr>
<tr>
<td>I1-T1</td>
<td>Ref</td>
<td>I1-T2</td>
<td>Ref</td>
</tr>
<tr>
<td>Ref</td>
<td>I2-T1</td>
<td>I2-T2</td>
<td>Ref</td>
</tr>
</tbody>
</table>

• Replicates for arrays, dyes, individuals are shared.

• Treatment and time point factors are separately replicated from individual, array, and dye factors.
OUTLINE

• NCI Large Screening Program of 60 Cancer Cell Lines
• Statistical Issues in High Throughput Array Data
• Hierarchical Effects Model Using MCMC
• Web-Based Interactive Analysis Tool (GS-HEM)
Estimating Interaction Effects

- Need to identify interaction expression levels
- Linear Model for interaction (Kerr and Churchill, 2000)
  \[ Y_{ijk} = G_i + C_j + (GC)_{ij} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \sim N(0, \sigma^2) \]
- Problem in large screening:
  - astronomical number of interaction parameters
    - e.g., Two-way interaction model with 9 x 10,000 levels requires estimation of more than 100,000 parameters
- Multiple layers and correlated structure of random variation
- Unbalanced and missing data structure (GLM, EM)
- Small sample bias in variance estimation (REML)
Hierarchical Effects Model

• Model
  - Layer 1: Array experimental variability:
    \[ Y_{ijkl} = r_{ijk} + \varepsilon_{ijkl}, \quad \varepsilon_{ijkl} \sim \mathcal{N}(0, \sigma^2), \text{ given } G_{ijk} \]
  - Layer 2: Biological variability:
    \[ r_{ijk} = G_i + C_j + \delta_{ij} + \alpha_{ijk}, \quad \alpha_{ijk} \sim \mathcal{N}(0, \nu_{ij}^2) \]
  - Priors for parameters
\( r_{ijk} \sim N(\mu + g_i + c_j + \delta_{ij}, \nu_{ij} \cdot \sigma_{ij} \cdot \sigma_{ij} \cdot \sigma_{ij} \cdot \sigma_{ij}) \)

\( Y_{ijkl} \sim N(r_{ijk}, \sigma_{ij} \cdot \sigma_{ij} \cdot \sigma_{ij} \cdot \sigma_{ij}) \)

\( i=1,\ldots,G; j=1,\ldots,C \)
\( k=1,\ldots,m_{ij}; l=1,\ldots,n_{ijk} \)

DAG (Directed Acyclic Graph) for Bayesian Hierarchical Effects Model (HEM)
• Why hierarchical?
  - Experimental reasons:
    • Chronological experimental procedure
    • Several different hierarchical layers of errors
  - Statistical reasons:
    • Want to decompose error variation into several components, while utilizing all variation information among replicates
    • Estimate interaction effects of over-parameterized models, especially for large data sets, taking into account unbalanced, missing data structure
    • Predict cases when no experimental data available

• Need computational statistical tools on complex hierarchical models --> MCMC
Posterior Distributions of Parameters and Missing Data

\[ \pi(\mu | \text{rest}) = \text{Normal}(\sum_{i,j} \frac{\sum_{k} \omega_{i,k} \omega_{j,k} (\mu - \mu_{i,j})^2}{\omega_{i,k} \omega_{j,k} + \sigma_{\mu,\mu}^2} \cdot (\frac{\mu_{i,j}}{\sigma_{\mu,\mu}} + \cdots + \frac{\mu_{i,j}}{\sigma_{\mu,\mu}})^{-1}) \]

\[ \pi(g_i | \text{rest}) = \text{Normal}(\sum_{i,j} \frac{\sum_{k} \omega_{i,k} \omega_{j,k} (g_i - g_{i,j})^2}{\omega_{i,k} \omega_{j,k} + \sigma_{g_i,g_i}^2} \cdot (\frac{g_{i,j}}{\sigma_{g_i,g_i}} + \cdots + \frac{g_{i,j}}{\sigma_{g_i,g_i}})^{-1}) \]

\[ \pi(c_j | \text{rest}) = \text{Normal}(\sum_{i,j} \frac{\sum_{k} \omega_{i,k} \omega_{j,k} (c_j - c_{i,j})^2}{\omega_{i,k} \omega_{j,k} + \sigma_{c_j,c_j}^2} \cdot (\frac{c_{i,j}}{\sigma_{c_j,c_j}} + \cdots + \frac{c_{i,j}}{\sigma_{c_j,c_j}} + \frac{1}{\sigma_{\delta_{i,j}}})^{-1}) \]

\[ \pi(\delta_{i,j} | \text{rest}) = \text{Normal}(\frac{\sigma_{\delta_{i,j}}^2}{\sigma_{\phi_{i,j},\phi_{i,j}}^2} \cdot \sum_k \frac{\omega_{i,k} \omega_{j,k} (\phi_{i,j} - \phi_{i,j})^2}{\omega_{i,k} \omega_{j,k} + \sigma_{\phi_{i,j},\phi_{i,j}}^2} \cdot (\frac{1}{\sigma_{\phi_{i,j}}^2} + \frac{1}{\sigma_{\delta_{i,j}}^2})^{-1}) \]

\[ \pi(\epsilon_{i,j} | \text{rest}) = \begin{cases} \text{Normal}(\mu + g_i + c_j + \delta_{i,j} - \sigma_{\epsilon_{i,j}}^2), & \text{if missing} \\ \text{Normal}(\frac{\sigma_{\epsilon_{i,j}}^2}{\sigma_{\phi_{i,j},\phi_{i,j}}^2} \cdot \sum_k \frac{\omega_{i,k} \omega_{j,k} (\mu + g_i + c_j + \delta_{i,j})^2}{\omega_{i,k} \omega_{j,k} + \sigma_{\phi_{i,j},\phi_{i,j}}^2} \cdot (\frac{1}{\sigma_{\phi_{i,j}}^2} + \frac{1}{\sigma_{\delta_{i,j}}^2})^{-1}), & \text{otherwise} \end{cases} \]

\[ \pi(\sigma_{\epsilon_{i,j}}^2 | \text{rest}) = \text{Gamma}(\frac{1}{2} + \alpha_{\epsilon_{i,j}} \cdot \sum_k \frac{1}{\omega_{i,k} \omega_{j,k} (\mu + g_i + c_j + \delta_{i,j})^2 + \beta_{\epsilon_{i,j}}}) \]

\[ \pi(\sigma_{\phi_{i,j}}^2 | \text{rest}) = \text{Gamma}(\frac{1}{2} + \alpha_{\phi_{i,j}} \cdot \sum_k \frac{1}{\omega_{i,k} \omega_{j,k} (\mu + g_i + c_j + \delta_{i,j})^2 + \beta_{\phi_{i,j}}}) \]
OUTLINE

• NCI Large Screening Program of 60 Cancer Cell Lines
• Statistical Issues in High Throughput Array Data
• Hierarchical Effects Model Using MCMC
• Web-Based Interactive Analysis Tool (GS-HEM)
GS-HEM: Interactive Web-Based Tool

- Innovations in our web-based interactive tools
  - Effective interaction & collaboration
  - Efficient investigation on various combinations of interest
  - Accessible independently from platforms and locations
  - Fully utilize a statistical package, S-PLUS in both statistical and graphic (visualization) methods
    - minimize redundant costs for statistical development
    - incorporate other programs and software
World Wide Web and On-line Analysis
Interactive Analysis Pages

### Interactive web tool I:

<table>
<thead>
<tr>
<th>Data mining, exploration, and discovery</th>
</tr>
</thead>
<tbody>
<tr>
<td>1204 genecard links (Stanford)</td>
</tr>
<tr>
<td>genecard pathways (Stanford)</td>
</tr>
<tr>
<td>1203 genecard links (Whitehead)</td>
</tr>
<tr>
<td>genecard pathways (Whitehead)</td>
</tr>
<tr>
<td>generate a cluster image</td>
</tr>
<tr>
<td>result an axis of a cluster page</td>
</tr>
<tr>
<td>pattern search</td>
</tr>
<tr>
<td>experiment data for 6215 drugs</td>
</tr>
</tbody>
</table>

### Interactive web tool II:

<table>
<thead>
<tr>
<th>Identification of entities with high association</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene-drug correlations</td>
</tr>
<tr>
<td>correlation search analysis</td>
</tr>
<tr>
<td>probabilistic correlation combination search</td>
</tr>
<tr>
<td>krt threshold correlations</td>
</tr>
</tbody>
</table>

### Interactive web tool III:

<table>
<thead>
<tr>
<th>Inference and literature search</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene/gene pubmed - keyword (Stanford)</td>
</tr>
<tr>
<td>gene/gene pubmed - keyword (Whitehead)</td>
</tr>
<tr>
<td>gene/gene pubmed - cloned or accession</td>
</tr>
<tr>
<td>gene/drug pubmed - keyword (Stanford)</td>
</tr>
<tr>
<td>gene/drug pubmed - keyword (Whitehead)</td>
</tr>
<tr>
<td>gene/drug pubmed - NGC, cloned or accession</td>
</tr>
<tr>
<td>spatial correlation</td>
</tr>
<tr>
<td>multi correlation analysis</td>
</tr>
<tr>
<td>cross correlation analysis</td>
</tr>
<tr>
<td>estimation of gene &amp; drug effects by IKM</td>
</tr>
<tr>
<td>critical levels of correlation</td>
</tr>
<tr>
<td>demo applet</td>
</tr>
</tbody>
</table>
Cluster-image analysis result
Estimation of Gene & Drug Effects by GS-HEM

Markov Chain Monte Carlo on Hierarchical Effects Model (HEM)

Identify and estimate the effects of elements ordered by the magnitude of significance. Estimate the effects of each element based on the hierarchical effects model (HEM) within each group and return the estimates with highest normal scores (estimated standard dev). The grouping can be chosen as one of preset options or by user choice.

Main Data:
- Median Activity
- Median Activity 115
- Excluded Target
- Activity
- Standard (confused) genes
- Standard (3,204) genes
- Estimation genes
- Bipsides (18) genes
- Bipsides (1,230) genes
- Whitened genes
- Target
- Log Target

MCMC Implementation & Output Options
1. Direction of search for normal score: positive, negative, both
2. Number of output nodes within the normal score range
3. Break-in iteration for Markov chain convergence
4. Size of Markov chain sample (length of MCMC run)
5. Save MCMC sample for parameters of main elements: no, yes
6. Save MCMC sample for parameters of interaction effects: no, yes
7. Normal variance for main parameters
8. Gamma parameters for variance parameters
9. Gamma parameters for variance parameters
10. Save list of selected hits: yes, no

National Cancer Institute
University of Virginia
HEM Result: Estimation of Gene-Cell Interaction Effects
GS-HEM Results on Microarray data for 9 Cancer Cell Line Types
Cancer tissue specific genes

- **Group 4: Colon Cell Type**
  - 264347, Transforming growth factor beta
  - 489884, Human insulin-like growth factor binding protein 5 (IGFBP5)
  - 469842, Homo sapiens mRNA for fatty acid binding protein, complete cds

- **Group 5: Ovarian Cell Type**
  - 183950:THY-1 MEMBRANE GLYCOPROTEIN PRECURSOR
  - 489235:HADHB Hydroxyacyl-Coenzyme A dehydrogenase/3-ketoacyl-Coenzyme A thiolase/enoyl-Coenzyme A hydratase
  - 285784, ESTs, Highly similar to TUBULIN BETA CHAIN [Schizosaccharomyces pombe]

- **Group 6: Leukemia Cell Type**
  - 510301, GLUTAMINYL-TRNA SYNTHETASE
  - 361247:TISSUE FACTOR PATHWAY INHIBITOR 2 PRECURSOR Chr.7
  - 470385, Homo sapiens placental bikunin mRNA, complete cds

- **Group 8: Breast Cell Type**
  - 248955, Human mitochondrial 1,25-dihydroxyvitamin D3 24-hydroxylase mRNA
  - 236338:TP53 Tumor protein p53 (Li-Fraumeni syndrome) Chr.17
  - 429540:ELONGATION FACTOR TU, MITOCHONDRIAL PRECURSOR Chr.16
COLLABORATORS

• LMP, NCI
  - Larry Smith
  - Lorrie Tanabe
  - Uwe Scherf
  - William Reinhold
  - Yi Zhou
  - John Weinstein

• Stanford
  - Douglas Ross
  - Michael Eisen
  - Patrick Brown
  - David Botstein

• Whitehead
  - Donna Slonim
  - Jane Staunton
  - Todd Golub
  - Pablo Tamayo
  - Erik Lander
Reproducibility: Microarray MDA-MB-435 vs. MDA-N