Functional Genomics Lecture

ECE Genomic Signal Processing (Dr. Rosen)
Tuesday, Apr. 21, 2009, 6-8 pm, 616 Bossone

Bahrad A. Sokhansanj, PhD
School of Biomedical Engineering, Science & Health Systems
DNA the molecule of life

Trillions of cells
Each cell:
- 46 human chromosomes
- 2 meters of DNA
- 3 billion DNA subunits (the bases: A, T, C, G)
- Approximately 30,000 genes code for proteins that perform most life functions

The Central Dogma of Molecular Biology

**Replication**
- DNA duplicates

**Transcription**
- RNA synthesis
  - RNA polymerase
  - Nucleus
  - mRNA
  - Nuclear membrane
  - Cytoplasm

**Translation**
- Protein synthesis
  - ribosome
  - mRNA
  - protein

ANDY VIERSTRATE 1999
A GENE REGULATORY NETWORK

INPUT

signal A

receptor proteins

cascade of interacting kinase proteins or other molecules

inactive transcription factor A

active transcription factor A

DNA

cis-regulatory DNA sequence elements

RNA polymerase

target gene

OUTPUT

mRNA

protein

OUTPUT

cell functions

INPUT

signal B

receptor proteins

inactive transcription factor B

active transcription factor B

inhibitory factor

feedbacks
Prepare cDNA Probe

"Normal" Tumor

RT / PCR
Label with Fluorescent Dyes

Combine Equal Amounts

Hybridize probe to microarray

Microarray Technology

http://www.accessexcellence.org/AB/GG/microArray.html
The Reality of Biological Data

Can we use these data in a numerical model?
Microarrays Produce Ugly Data

- Microarray fabrication is inconsistent: experiments have poor repeatability
  - size, shape, and alignment of spots varies from array to array
  - defects in the slide and different washing protocols result in intra- & inter-slide variations in background

- Results are highly sensitive to image analysis
  - spot recognition, intensity quantification and normalization, background subtraction

- DNA chips are more consistent, but still not perfect - and they cost much more
More Ugliness – Gels for Proteomics

Monocyte cell line exposed to Y.pestis and Y.pseudotuberculosis

2DE of 4 hr exposures, MOI 5:1, control (no pathogen). Protein identification in progress. ▲ (unchanged), ⚫ (upregulated), ⬤ (downregulated). Two separate areas of gels

DIGE and MS are improvements, but still, at best, semi-quantitative.
Prepare cDNA Probe

"Normal" Tumor

RT / PCR Label with Fluorescent Dyes

Combine Equal Amounts

Hybridize probe to microarray

Microarray Technology

Prepare Microarray

SCAN

http://www.accessexcellence.org/AB/GG/microArray.html
<table>
<thead>
<tr>
<th>Gene</th>
<th>HR 1</th>
<th>HR 4</th>
<th>HR 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>YPO3718/pgl</td>
<td>-1.609</td>
<td>-1.114</td>
<td>1.089</td>
</tr>
<tr>
<td>YPO2686/pgm</td>
<td>-1.157</td>
<td>-1.260</td>
<td>-1.172</td>
</tr>
<tr>
<td>YPO2995/ct</td>
<td>-1.861</td>
<td>1.094</td>
<td>-1.083</td>
</tr>
<tr>
<td>YPO0078/pfkA</td>
<td>-2.191</td>
<td>-1.212</td>
<td>-1.134</td>
</tr>
<tr>
<td>YPO0920/fbaA</td>
<td>-1.536</td>
<td>1.063</td>
<td>1.124</td>
</tr>
<tr>
<td>YPO1254/bglA</td>
<td>1.304</td>
<td>1.147</td>
<td>1.094</td>
</tr>
<tr>
<td>YPO0166/none</td>
<td>1.666</td>
<td>1.075</td>
<td>1.215</td>
</tr>
<tr>
<td>YPO2977/glk</td>
<td>-1.174</td>
<td>-1.308</td>
<td>-1.111</td>
</tr>
<tr>
<td>YPO1135/galM</td>
<td>2.292</td>
<td>-1.020</td>
<td>1.079</td>
</tr>
</tbody>
</table>
By observing chip data, one can infer which genes are highly expressed or not expressed, or in general the relative expression levels of all genes.

Inference: genes x, y, z are highly expressed under conditions W while genes a, b, c are not expressed.
Information Derivable from Chip Data

- By comparing gene expression levels under two conditions, one can infer which genes’ expression levels are affected.

**Inference**: gene X is significantly more highly expressed in diseased cell than in normal cell; hence gene X could potentially serve as a marker of the disease – differentially expressed genes.
Information Derivable from Chip Data

- By observing gene expression levels collected at different time points after a particular stimulus, one can infer how a gene’s expression level changes with time.

**Inference:** as the disease progresses, gene X’s expression decreases; hence ….
By observing expression levels of two genes collected at different time after a particular stimulus, one can infer they have similar or different expression patterns.

**Inference:** genes with similar expression patterns might be functionally related, e.g., working in the same pathway – **co-expressed genes -> co-regulated**
By comparing expression patterns of gene A collected when gene B is functioning and not functioning (e.g., knockout or mutation), one could possibly derive gene B’s effect on gene A.

Inference: genes A and B may interact directly or indirectly, or even B is the cause of A’s altered expression patterns – interaction or causality relationship.
Molecular Signatures – Genome, Proteome, Metabolome, Glycome, ...

- New technologies allow large scale, parallel measurement of cell state:
  - transcription (mRNA expression; gene chips, RT-PCR)
  - translation (protein expression; gels, MS, protein chips)
  - protein modification (gels, MS)
  - protein-protein interaction (2-hybrid, protein chips, MS)
  - metabolites (MS, NMR)
  - carbohydrates and glycosylation (MS, ?)
  - (also large scale phenotypic changes)

- At first order, we can either/both
  - identify groups of cells/tissues/populations that share common patterns
  - detect patterns of gene/protein/metabolites/etc. that correlate with previously identified phenotypic groups
Many Ways to Identify Signatures

• Identifying major “components” of variation (potentially something that has to be removed from data, such as a fundamental difference between sampled groups)
  – singular value decomposition (SVD), principle component analysis (PCA), etc.
Singular Value Decomposition

\[ X = USV^T \]

Complex Regulation Drives Yeast Cell Cycle


http://cyberia.cfdrc.com/datab/Applications/cell_tissue_bio/cellcycle/cellcycle.html
Singular Value Decomposition
(Yeast Cell Cycle Microarray Data)

(based on 38 genes, elu expression dataset)
Many Ways to Identify Signatures

• Identifying major “components” of variation (potentially something that has to be removed from data, such as a fundamental difference between sampled groups)
  – singular value decomposition (SVD), principle component analysis (PCA), etc.

• Finding groups within data
  – clustering
  – self-organizing maps
  – support vector machines
Clustering (k-means)

Figure 3

Many Ways to Identify Signatures

- Identifying major “components” of variation (potentially something that has to be removed from data, such as a fundamental difference between sampled groups)
  - singular value decomposition (SVD), principle component analysis (PCA), etc.
- Finding groups within data
  - clustering
  - self-organizing maps
  - support vector machines
- Separating known groups
  - univariate methods (i.e. B-Tests, T-Tests on each gene, ANOVA on each gene)
  - horrible “capitalization on chance” problems
  - linear discriminant analysis / canonical variate analysis
    - these methods can be generalized for undetermined data, though the relative magnitudes of variables becomes significant in that case (but that filters out potentially noisy data) OR you get capitalization by chance by using stepwise methods
Linear Separation – Group Classification

Canonical Variate Analysis (Linear Discriminant Analysis)

Find optimal linear combinations of variables that maximize inter-group differences while minimizing intra-group differences.

Nonlinear Classification? (Kernel methods)
What Do We Get From Signatures

• Pattern for discrimination between groups (responders, non-responders, different genetic populations, etc.)
  – therapeutic design
  – diagnostics

• Lists of Genes
  – what genes appear to be the most significant in determining the difference between groups or cause the formation of distinct patterns within data?
    • you get long lists when you “capitalize on chance” using univariate methods
    • shorter lists from multivariate methods or when you use “honest” statistics modified for variable selection
Figure from: Hanahan and Weinberg, 2000. Cell 100, p. 57-70
A network-based analysis of systemic inflammation in humans


Oligonucleotide and complementary DNA microarrays are being used to subclassify histologically similar tumours, monitor disease progress, and individualize treatment regimens. However, extracting new biological insight from high-throughput genomic studies of human diseases is a challenge, limited by difficulties in recognizing and evaluating relevant biological processes from huge quantities of experimental data. Here we present a structured network knowledge-base approach to analyse genome-wide transcriptional responses in the context of known functional inter-relationships among proteins, small molecules and phenotypes. This approach was used to analyse changes in blood leukocyte gene expression patterns in human subjects receiving an inflammatory stimulus (bacterial endotoxin). We explore the known genome-wide interaction network to identify significant functional modules perturbed in response to this stimulus. Our analysis reveals that the human blood leukocyte response to acute systemic inflammation includes the transient dysregulation of leukocyte bioenergetics and modulation of translational machinery. These findings provide insight into the regulation of global leukocyte activities as they relate to innate immune system...
Problems:

1. Correlation of genomic and proteomic data does not provide functional information

Solution: Computational modeling as a tool

2. Complexity: Thousands of interactions that cannot be studied via single gene/protein approach

Solution: Systems Biology approach
Text-mining, protein-protein interaction databases, Gene Ontology...

*identify*

potential relationships
between variables

Dynamic Measurements, Functional Genomics, Functional Proteomics...

*constrain*

relationship models

Differences between models

*design*

Hypothesis driven
lab experiments & clinical trials

How does

change?
Fuzzy Simulation Procedure

Normalized Data for Biological Variables (e.g. mRNA ratio)

Example Rule 1 = [3 2 1]
If X is LO then Y is HI
If X is MED then Y is MED
If X is HI then Y is LO

X1 = [0 0.2 0.8]
X2 = [0 0.4 0.6]

Y = Y high - Y low / Yh + Ym + YL

Predicted Values & Error
Quantitative Measurement
Modeling methodology

Bottom Up Model

Mechanisms
Artificial Data
Pathways
Reactions

Top Down Model

Regression
Parameters
Fit Experimental Data

Modeling methodology

Bottom Up Model

Mechanisms
Artificial Data
Pathways
Reactions

Top Down Model

Regression
Parameters
Fit Experimental Data
Developing Models based on Data

Two independent "training" & "testing" time series

Least squares parameter estimation was performed

Best prediction was selected on the robustness of the model
Bayesian Network Inference

• *Science*, Volume 308, pp 523-529, April 22, 2005
  “Causal protein-signaling networks derived from multiparameter single-cell data”
K. Sachs, O. Perez, D. Pe’er, D. A.
Lauffenburger, G. P. Nolan
Biomolecular Network
Application of Bayes Nets

• “Bayesian Network”, “Bayes Net”... a generalization of “Bayesian Inference”
• Directed, acyclic graphs representing relationships between variables
• NO FEEDBACK LOOPS
  – (these have to be handled using “Dynamic Bayesian Networks”)

Example of Bayesian Net Inference

http://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html
Simplify Joint Probability Distributions

• By the chain rule of probability, the joint probability of all the nodes in the graph above is
  \[ P(C, S, R, W) = P(C) \times P(S|C) \times P(R|C,S) \times P(W|C,S,R) \]

• By using conditional independence relationships, we can rewrite this as
  \[ P(C, S, R, W) = P(C) \times P(S|C) \times P(R|C) \times P(W|S,R) \]

  where we were allowed to simplify the third term because R is independent of S given its parent C, and the last term because W is independent of C given its parents S and R.

• In general, if we had n binary nodes, the full joint would require \( O(2^n) \) space to represent, but the factored form would require \( O(n \times 2^k) \) space to represent, where k is the maximum fan-in of a node. And fewer parameters makes learning easier.
Identifying structure of Bayesian networks

• You can use MCMC (Markov Chain Monte Carlo) to “learn” parameter values based on data, or you can generate models and see how well they can predict correlations that you measure under different perturbations (the approach taken in the paper)
Network Inference from Experiments
Looking at correlations = Big mess

52 out of 55 possible(!!) correlations are found in data with reasonable p value
X and Y are correlated... now, if we inhibit X, we see an effect on Y (yellow dots) But if we inhibit Y, there is no effect on X (purple dots). So, X influences Y, but not vice versa.
Identifying a complex network

• Simulated annealing was used, with operators adding or deleting arcs in the graph. In most cases, a change that led to a higher scoring (better fit) was accepted, but there was a small (and diminishing) probability of accepting a worse scoring solution (to avoid falling in a local minimum).

• The search starts with a random graph; it was repeated 500 times to explore the whole search space – all the highest scoring networks were combined to form a “consensus graph”
Consensus results

Final result for confidence > 85% graphs
Limitation of acyclicity

Some loops were excluded by confidence > 85% but could be biological significant (big limitation of Bayes nets – but what about Dynamic Bayesian Networks?)