An Overview of Functional Genomic Data

Curtis Huttenhower

Harvard School of Public Health
Department of Biostatistics

04-10-10
Greatest Biological Discoveries?
Are We There Yet?

- How much biology is out there?
- How much have we found?
- How fast are we finding it?
Are We There Yet?

Our job is to create computational microscopes:
To ask and answer specific biomedical questions using millions of experimental results.
1. Genomic data:
   Types, sources, and considerations

2. Microarrays:
   Gene expression, R/Bioconductor, MeV, and GenePattern

3. Integration:
   Scalable data integration and biological networks
All I Really Need to Know about Biology I Learned from One PowerPoint Slide
All I Really Need to Know about Biology
I Learned from One PowerPoint Slide

Cell

Stimulus

DNA

mRNA
All I Really Need to Know about Biology I Learned from One PowerPoint Slide

Cell

Stimulus

DNA

mRNA

Proteins

Transcription

Translation
Sequencing: Genomes

www.ncbi.nlm.nih.gov/sites/genome
www.ensembl.org
genome.ucsc.edu
www.sanger.ac.uk/Projects
genome.jgi
img.jgi.doe.gov
www.nmpdr.org
www.genedb.org
www.biomart.org
www.genedb.org
www.sanger.ac.uk/Projects
Sequencing: RNA and High-Throughput

- www.ncbi.nlm.nih.gov/dbEST
- www.estinformatics.org
- www.sagenet.org
- www.trace.ensembl.org
- www.ebi.ac.uk/microarray-as/ae
- www.ncbi.nlm.nih.gov/dbEST
- www.estinformatics.org
- www.sagenet.org
- www.trace.ensembl.org
- www.ebi.ac.uk/microarray-as/ae
Interactions: Physical and Genetic

- **BioGRID**: www.thebiogrid.org
- **IntAct**: www.ebi.ac.uk/intact
- **Human Protein Reference Database**: www.hprd.org
- **MINT**: mint.bio.uniroma2.it
- **DOMINE**: domine.utdallas.edu
- **dip.doe-mbi.ucla.edu**
- **mips.helmholtz-muenchen.de**
- **bond.unleashedinformatics.com**
Regulation: Transcription Factors and miRNAs

[jaspar.genereg.net](jaspar.genereg.net)

[www.biobase-international.com](www.biobase-international.com)

[redfly.ccr.buffalo.edu](redfly.ccr.buffalo.edu)

[regulondb.ccg.unam.mx](regulondb.ccg.unam.mx)

[www.oreganno.org](www.oreganno.org)

[www.cisred.org](www.cisred.org)

[www.mirbase.org](www.mirbase.org)

[www.targetscan.org](www.targetscan.org)

[pictar.mdc-berlin.de](pictar.mdc-berlin.de)
Curation:
Pathways, Functions, and Biological Roles

- the Gene Ontology: www.geneontology.org
- KEGG: www.genome.jp/kegg
- PANTHER Classification System: www.pantherdb.org
- Reactome: www.reactome.org
- Pathway Interaction Database: pid.nci.nih.gov
- Pathways for the People: www.wikipathways.org
Microarrays

www.ncbi.nlm.nih.gov/geo

www.ebi.ac.uk/microarray-as/ae

www.ebi.ac.uk/gxa

biogps.gnf.org
Outline

1. Genomic data:
   Types, sources, and considerations

2. Microarrays:
   Gene expression, R/Bioconductor, MeV, and GenePattern

3. Integration:
   Scalable data integration and biological networks
All I Really Need to Know about Biology I Learned from One PowerPoint Slide

Cell

Stimulus

DNA

mRNA

Proteins
Add red dye

Hybridize

Control mRNA sample

Add green dye

Mix

Experimental mRNA sample

Experimental condition X

Normal Cells

Spot slide with gene sequences

Gene A

Gene C

Gene B

Gene D

Hybridize

Scan

Conditions

Genes

Conditions

1.5 1 0.1 0.15 1.5 2

0.5 0.2 1 0
Microarray Data Flow

- Microarray experiment
- Image analysis
- Database
- Data selection
- Missing value estimation
- Normalization

Unsupervised: Clustering
Unsupervised: Decomposition
Supervised: Decomposition
Networks and Data Integration
Spot Identification

Individual spots are recognized
Size and shape adjusted per spot
Additional flagging of bad (X) or non-present (NA) spots

Different Spot identification methods: Fixed circles, circles with variable size, arbitrary spot shape (morphological opening)
Spatial Defects
Data Normalization: Definition

- Normalization is an attempt to compensate for systematic bias in data.
- Normalization attempts to remove the impact of non-biological influences on biological data:
  - Balance fluorescent intensities of the two dyes
  - Adjust for differences in experimental conditions (b/w replicate gene expression experiments)
  - Probe-specific intensities for Affymetrix data (b/w arrays)
- Normalization allows you to compare data from one experiment to another (after removing experiment-specific biases)
Normalization: Effects on Intensity

Non-normalized

Normalized

Same mRNA hybridized in both channels
Normalization: LOESS and Quantile

\[ A = \frac{(\log \text{Green} + \log \text{Red})}{2} \]

\[ M = \log \text{Red} - \log \text{Green} \]

Local estimate

Use the estimate to bend the banana straight
8.4 Estrogen Data: A 2x2 Factorial Experiment with Affymetrix Arrays

This data is from the estrogen package on Bioconductor. A subset of the data is also analysed in the factDesign package vignette. To repeat this case study you will need to have the R packages affy, estrogen and hgu95av2cdf installed.

The data gives results from a 2x2 factorial experiment on MCF7 breast cancer cells using Affymetrix HGU95av2 arrays. The factors in this experiment were estrogen (present or absent) and length of exposure (10 or 48 hours). The aim of the study is the identity genes which respond to estrogen and to classify these into early and late responders. Genes which respond early are putative direct-target genes while those which respond late are probably downstream targets in the molecular pathway.

First load the required packages:

```r
source("http://bioconductor.org/biocLite.R")
biocLite()
biocLite("hgu95av2cdf")
biocLite("estrogen")
```

The data files are contained in the `data` directory of the estrogen package:

```r
> datadir <- file.path(find.package("estrogen"), "data")
> dir(datadir)
```

The target file is called phenoData.txt. We see there are two arrays for each experimental condition, giving a total of 8 arrays.

```r
> targets <- readTargets("phenoData.txt", path=datadir, sep="", row.names="filename")
> ab <- ReadAffy(filenames=dir.path(datadir, targets$filename))
> eset <- zma(ab)
```

Background correcting

Normalizing

Calculating Expression

There are many ways to construct a design matrix for this experiment. Given that we are interested in the early and late estrogen responders, we can choose a parameterization which includes these two contrasts.

```r
> treatments <- factor(c(1,2,3,4), labels=c("E10", "E10", "E48", "E48"))
> contrasts(treatments) <- cbind(Time=c(0,0,1,1), E10=c(0,1,0,0), E48=c(0,0,0,1))
> design <- model.matrix(~treatments)
> colnames(design) <- c("Intercept", "Time", "E10", "E48")
```

The second coefficient picks up the effect of time in the absence of estrogen. The third and fourth coefficients estimate the log2-fold change for estrogen at 10 hours and 48 hours respectively.
R/Bioconductor: An Example

```r
> fit <- lmFit(cistr, design)
> cont.matrix <- contrasts(cistr$design, c(0, 1, 0), c(1, 0, 0))
> fit2 <- contrasts.fit(fit, cont.matrix)
> fit2 <- eBayes(fit2)

We can examine which genes respond to estrogen at either time by computing moderated F-statistics on 2 degrees of freedom:

```r
d stat <- Fstat(fit2)
p.value <- pf(F.stat, df1=length(f.stat[,2]), df2=length(f.stat[,2]), lower.tail=FALSE)
```

Now we consider those genes with moderated F-statistics with p-values below 0.0001, and classify these according to whether they are significantly up or down regulated at the early or late times:

```r
> results <- classifyTestsF(fit2, p.value=0.0001)
> vennDiagram(results)
```

```
> topTable gives a detailed look at individual genes:
```r
> topTable(fit2, coef="E10", n=20)

```

```
> topTable(fit2, coef="E48", n=20)
```

We see that 111 genes were up regulated at 10 hours, 58 of these still up at 48 hours. Also, 38 genes were down-regulated at 10 hours, 19 of these still down at 48 hours.

limma: Linear Models for Microarray Data
User's Guide
Gordon Smyth, Natalie Thorne and James Wittenhall
The Walter and Eliza Hall Institute of Medical Research
14 May 2004
MeV: An Example
MeV: Unsupervised Clustering

www.tm4.org/mev
MeV: Principle Component Analysis

www.tm4.org/mev
GenePattern: An Example
**GenePattern: Gene Set Enrichment Analysis**

### GSEA Report for Dataset dilution_rate_01_human

**Enrichment in phenotype:** Glucose (6 samples)
- 612 / 1228 gene sets are upregulated in phenotype Glucose
- 189 gene sets are significant at FDR < 25%
- 134 gene sets are significantly enriched at nominal p-value < 5%

**Enrichment in phenotype:** Other (30 samples)
- 616 / 1228 gene sets are upregulated in phenotype Other
- 70 gene sets are significantly enriched at FDR < 25%
- 55 gene sets are significantly enriched at nominal p-value < 5%

**Dataset details**
- The dataset has 4063 features (genes)
- No probe set => gene symbol collapsing was requested, so all 4063 features were used

**Gene set details**
- Gene set size filters (min=5, max=500) resulted in filtering out 664 x 1892 gene sets
- The remaining 1229 gene sets were used in the analysis
- List of gene sets used and their sizes (restricted to features in the specified dataset)

**Gene markers for the Glucose versus Other comparison**
- The dataset has 4063 features (genes)
- # of markers for phenotype Glucose: 1638 (40.3%) with correlation area 40.2%
1. Genomic data: Types, sources, and considerations

2. Microarrays: Gene expression, R/Bioconductor, MeV, and GenePattern

3. Integration: Scalable data integration and biological networks
Machine Learning for Data Integration

High Similarity
Low Similarity
High Correlation
Low Correlation

G1 + G2 + G4 + G9 + G3 - G7 - G8 - G2 - G5

0.9 0.7 0.1 0.2 0.8 0.5 0.05 0.1 0.6

P(G2-G5|Data) = 0.85

100Ms gene pairs → 1Ks datasets

MIPS
KEGG
the Gene Ontology

Low Correlation
High Correlation

Low Similarity
High Similarity

Frequency
Low Correlation
High Correlation
Not coloc.
Coloc.

Dissim.
Similar
Meta-Analysis for Data Integration

Simple regression: All datasets are equally accurate

\[ y_{e,i} = \beta_e + \varepsilon \]

Random effects: Variation within and among datasets and interactions

\[ y_{e,i} = \beta_e + \eta_e + \varepsilon_{e,i} \]

\[ \beta_e \propto \sum_i w_{e,i} y_{e,i} \]

\[ w_{e,i} = \frac{1}{s_{e,i}^2 + \hat{\Delta}_e^2} \]

\[ \rho' = \frac{1}{2} \log \left( \frac{1+\rho}{1-\rho} \right) \]

\[ z = \frac{\rho' - \mu_{\rho'}}{\sigma_{\rho'}} \]

Z-score

Pearson correlation
Meta-Analysis for Data Integration

Evangelou 2007

Hibbs 2007

\[ \rho = \frac{1}{2} \log \left( \frac{1+\rho}{1-\rho} \right) \]

\[ z = \frac{\rho - \mu_\rho}{\sigma_\rho} \]

Recall (log scale)

log$_2$(Precision / Random)

Supervised Bayesian integration
Unsupervised meta-analytic integration
Random effects meta-analysis
Quantile normalization

Pearson correlation

Z-score
Functional Networks

Global interaction network

Metabolism network

Conserved network

Kidney network

STRING
string-db.org

FuncNet
funcnet.eu

Endeavour
homes.esat.kuleuven.be/~bioiuser/endeavour
Biological Networks: Manipulation and Analysis

www.cytoscape.org

hub.iis.sinica.edu.tw/cytoHubba

bioputer.mimuw.edu.pl/modevo

baderlab.org/Software/MCODE
Biological Networks:
Clusters, Hubs, Bottlenecks, and Flow

Community structure in social and biological networks
M. Girvan*1 and M. E. J. Newman*5

Evidence for dynamically organized modularity in the yeast protein–protein interaction network
Jing-Dong J. Han 1, Nicolas Berlin 1, Tong Hao 1, Debra S. Goldberg 1, Gabriel F. Berriz 2, Lan Y. Zhang 2, Denis Dupuy 2, Albert R. J. M. Waltho 3, Michael E. Cusick 4, Frederick P. Roth 5 & Marc Vidal 6

Still Stratus Not Altocumulus: Further Evidence against the Date/Party Hub Distinction
Nizar N. Batada 7, Teresa Reguly, Ashton Breitkreutz, Lorrie Boucher, Bobby-Joe Breitkreutz, Laurence D. Hurst 8, Mike Tyers 9

The Importance of Bottlenecks in Protein Networks: Correlation with Gene Essentiality and Expression Dynamics
Haiyan Yu 1,2,3,4, Philip M. Kim 1,2,4, Emmett Sprecher 1,2, Valery Trifonov 2, Mark Gerstein 1,4,5

Information Flow Analysis of Interactome Networks
Patrycja Vasilyev Missiuro1,2, Kesheng Liu1, Lihua Zou3, Brian C. Ross1, Guoyan Zhao4, Jun S. Liu5, Hui Ge1,6
Biological Networks: Network Motifs

Bi-fan

Positive auto-regulation

Negative auto-regulation

delay

WGD and evolvability

speed + stability

Feedback

Coherent feed-forward

Incoherent feed-forward

www.weizmann.ac.il/mcb/UriAlon/groupNetworkMotifSW.html

mfinder
Network motifs detection tool

mavisto.ipk-gatersleben.de

FANMOD a tool for fast network motif detection

theinf1.informatik.uni-jena.de/~wernicke/motifs

\[ \text{X} \rightarrow \text{Y} \text{ represents} \]

transcription network

gene x

gene y

neuron synaptic connection network

ecological food web

\[ \text{X} \rightarrow \text{Y} \]

filter

pulse

Milo 2002

Alon 2007
The strength of these relationships indicates how cohesive a process is.
Functional Mapping: Mining Integrated Networks

Predicted relationships between genes

Low Confidence

High Confidence

Chemotaxis
Functional Mapping: Mining Integrated Networks

The strength of these relationships indicates how associated two processes are.

Chemotaxis

Flagellar assembly
Functional Mapping: Associations Between Gene Sets

- Hydrogen Transport
- Electron Transport
- Cellular Respiration
- Cell Redox Homeostasis
- Protein Processing
- Peptide Metabolism
- Protein Catabolism
- Energy Reserve Metabolism
- Vacuolar Protein Catabolism
- Positive Regulation of Protein Metabolism
- Negative Regulation of Protein Metabolism
- Organelle Depolymerization
- Organelle Fusion
- Organelle Inheritance

Edges: Associations between processes
- Light green: Moderately Strong
- Dark green: Strong
- Red: Very Strong
Functional Mapping: Associations Between Gene Sets

Edges
Associations between processes

Moderately Strong
Very Strong

Borders
Data coverage of processes

Sparsely Covered
Well Covered

- Hydrogen Transport
- Electron Transport
- Cellular Respiration
- Aldehyde Metabolism
- Cell Redox Homeostasis
- Energy Reserve Metabolism
- Vacuolar Protein Catabolism
- Peptide Metabolism
- Protein Processing
- Negative Regulation of Protein Metabolism
- Protein Depolymerization
- Organelle Fusion
- Organelle Inheritance

- Energy Reserve Metabolism
- Cell Redox Homeostasis
- Protein Depolymerization
- Organelle Inheritance

- Hydrogen Transport
- Electron Transport
- Cellular Respiration
- Aldehyde Metabolism
- Cell Redox Homeostasis
- Energy Reserve Metabolism
- Vacuolar Protein Catabolism
- Peptide Metabolism
- Protein Processing
- Negative Regulation of Protein Metabolism
- Protein Depolymerization
- Organelle Fusion
- Organelle Inheritance

- Energy Reserve Metabolism
- Cell Redox Homeostasis
- Protein Depolymerization
- Organelle Inheritance

- Hydrogen Transport
- Electron Transport
- Cellular Respiration
- Aldehyde Metabolism
- Cell Redox Homeostasis
- Energy Reserve Metabolism
- Vacuolar Protein Catabolism
- Peptide Metabolism
- Protein Processing
- Negative Regulation of Protein Metabolism
- Protein Depolymerization
- Organelle Fusion
- Organelle Inheritance

- Energy Reserve Metabolism
- Cell Redox Homeostasis
- Protein Depolymerization
- Organelle Inheritance
Functional Mapping: Associations Between Gene Sets

Edges
Associations between processes
Moderately Strong
Very Strong

Nodes
Cohesiveness of processes
Below Baseline
Baseline (genomic background)
Very Cohesive

Borders
Data coverage of processes
Sparsely Covered
Well Covered

Hydrogen Transport
Electron Transport
Cell Redox Homeostasis
Cellular Respiration
Peptide Metabolism
Protein Processing
Vacuolar Protein Catabolism
Negative Regulation of Protein Metabolism
Protein Depolymerization
Organelle Fusion
Organelle Inheritance

Aldehyde Metabolism
Energy Reserve Metabolism
Data integration summarizes an impossibly huge amount of experimental data into an impossibly huge number of predictions; *what next?*
How can a biologist take advantage of all this data to study his/her favorite gene/pathway/disease without losing information?

Functional mapping
- Very large collections of genomic data
- Specific predicted molecular interactions
- Pathway, process, or disease associations
- Underlying experimental results and functional activities in data
Functional Maps for Functional Metagenomics

Mapping organisms into phyla

Mapping genes into pathways

Mapping pathways into organisms

MG-RAST Meta Genome Rapid Annotation using Subsystem Technology
GOS 4441599.3
Hypersaline Lagoon, Ecuador

KEGG Pathways

Integrated functional interaction networks in 27 species
Efficient Computation For Biological Discovery

Massive datasets and genomes require efficient algorithms and implementations.

- Sleipnir C++ library for computational functional genomics
- Data types for biological entities
  - Microarray data, interaction data, genes and gene sets, functional catalogs, etc. etc.
  - Network communication, parallelization
- Efficient machine learning algorithms
  - Generative (Bayesian) and discriminative (SVM)

It’s also speedy: microbial data integration computation takes <3hrs.

And it’s fully documented!

huttenhower.sph.harvard.edu/sleipnir
1. Genomic data:
   Types, sources, and considerations

   - Lots of data is publicly available, of lots of different types!
   - Sequencing, interactions, expression, regulation, pathways
   - NB: most repositories also provide programmatic access

2. Microarrays:
   Gene expression, R/Bioconductor, MeV, and GenePattern

   - Gene expression microarrays measure transcript abundance
   - R/Bioconductor is a powerful tool for statistical manipulation
   - MeV and GenePattern are quick-and-easy graphical tools

3. Integration:
   Scalable data integration and biological networks

   - Genomic data is most powerful when many types are integrated
   - Biological networks are a common model for integration
   - Functional mapping is one way to summarize large data collections
Thanks!

Curtis Huttenhower

Interested? Summer, rotation, and graduate students are welcome!

http://huttenhower.sph.harvard.edu