Genes, Circuits and Systems

Circuits & System Workshop, 2005
Giovanni De Micheli
Centre Systèmes Intégrés
Gene Expression Analysis

DNA microarray: Disruptive technology for uncovering gene function on a global scale

What we know:
A (nearly) complete list of the genes

What we don’t know:
Gene functions and interactions
The global picture

- Measuring biological samples
  - Biosensors --> Circuits
- Analyzing biological information
  - Software tools --> Signal Processing
- Modeling and controlling interactions
  - Laboratory on a chip --> Systems

Bio-discovery uses technologies and tools similar to those used with traditional circuits and systems
Sensors and data

- Data in context
- Data analysis
- Sensors and data
DNA/RNA Hybridization

DNA is denatured by heating

DNA

B A C T

DNA is renatured on cooling

Test (red)

blueprint
The micro-array operation

a

b

c

control cells

RNA extraction

PCR amplification of complete or partial ORF's or cDNA's

Cy3

reverse transcription

Cy5

pool targets

data analysis

standard microscope glass slide (25x75 mm) polylysine coated

spotting

hybridisation

scanning
Three phases

- Micro-array design and fabrication
- Micro-array hybridization and readout
  - After sample preparation
- Micro-array data analysis

- This technologies exists since the mid-nineties
  - Ubiquitous in top-level research
  - Still virtually unused in clinical practice
  - Plenty of opportunities for refinement
Established micro-array fabrication technologies

- DNA spotting on glass substrates
  - Relatively inexpensive and reproducible
  - Pat Brown (Stanford)
- Mask-based (semiconductor-like) fabrication
  - Oligonucleotides (25mers)
  - High-density, high cost
  - Affymetrix (Mountain view)
Micro-array readout

- Labeled (Indirect)
- Non-labeled (direct)

- Electrical techniques
  - Changes in the capacitive behavior of an electrode/solution bio-sensing interface

- Integrated optical sensors
  - Molecular light absorbance
    - Visible, UV
  - Optical arrays
Electrical measurements

- Dedicated on chip circuitry for capacitance measurements

**Non Complementary**

**Complementary**

Probes

Carlotta Guiducci Lausanne 12-03-05
Optical readout

- Laser scanning:
  - Red: sample
  - Green: control

- Measure: $\log \left( \frac{\text{Intensity}_{\text{red}}}{\text{Intensity}_{\text{green}}} \right)$

- Data organized in matrices:
  - Rows: genes
  - Column: experiments
Industrial Technology roadmap

- Optimized chip layout and user-friendly slide format with integrated PCR and dual detection setup
- Integrated PCR and detection on the same chip
- Integrated on-chip sample preparation
- First prototype and proof-of-concept of buried channels
- Full sample-to-answer chip with sample preparation and electronic detection

[STM]
Data analysis

- Data in context
- Data analysis
- Sensors and data
Objectives

- Clinical analysis
  - Discover signatures of diseases
  - Discover effect of medications
- Bio-discovery
  - Gene regulatory networks
- Pharmacogenomics
  - Design drugs with direct impact on genetic features
Identifying Disease Genes

X. Chen & P.O. Brown et al
Molecular Biology of the Cell
Vol. 13, 1929-1939, June 2002
Disease Subtype Classification

Today

- Patient side:
  - More efficient treatment, i.e.
    avoid side effects of useless therapy

- Health Care System side:
  - Efficient use of information,
    time and resources

Diversity of gene expression in adenocarcinoma of the lung
Garber et Al. PNAS November 20, 2001 vol 98 n.24 13784-13789
Understanding of Bio-mechanisms

Today

- Experiments are non-pathological varying environmental condition
- i.e. how gene expression of endothelial cells changes with blood pressure variation

Endothelial cell Diversity revealed by global expression profiling, Chi, Chang, Haraldsen, PNAS, September 16 2003, vol100, n.19, 10623-10628
Data analysis

- To find common patterns in gene expression levels
  - Upregulated/downregulated genes
- To group similar patterns
The curse of dimensionality

A Typical Genomic Study

Cases (10’s - 100’s)

Variables (10,000’s - 100,000’s)

A Typical Clinical Study

Cases (1,000’s - 1,000,000’s)

Variables (10’s - 100’s)

- Underdetermined system

Kohane et al., *Microarrays for Integrative Genomics*, 2003

G. De Micheli -- ISCAS 2005
Clustering Analysis

- Considers both rows and columns
- Focuses on subsets
- Biclustering
  - Two-dimensional clustering
  - Hierarchical Clustering
  - Grid-based Clustering
  - Vector Quantization
  - Principal Component Analysis
  - K-medoids
  - Self-Organizing Maps
- Subspace Clustering
  - Nonlinear Component Analysis
  - K-means
  - On-line Clustering
- Multidimensional Scaling
- Density-based Partitioning
- Independent Component Analysis
Bi-clustering algorithms

- Several approaches
  - $\delta$-biclustering [Cheng & Church 2000]
    - Heuristic
  - P-Clustering [Wang et al. 2002]
    - Heuristic
  - $\delta$-Pclustering [Yoon et al. 2004]
    - Exact

- Advantages:
  - Capture relevant subspaces
  - Capture coherence and fluctuation

- Disadvantages:
  - Computationally hard
Computational Challenge

- Find all relevant maximal biclusters
  - Rank them according to some metric
    - E.g. Mean score residue (MSR)
- Exploit methods for large scale data handling
  - BDDs and ZDDs
- Use implicit representation to avoid full set representation and to support operations
Zero-suppressed BDDs
for set representations

[Minato 86]
Set operations with ZDDS

- ZDDs support well recursive operations
  - Like union and intersection
- Recursive algorithms are w.c. exponential
  - But linear with representation size
- Representation size grows mildly with problem size
Gene Expression Data

Yeast cell cycle data (Tavazoie et al., 1999)

G. De Micheli -- ISCAS 2005

[Yoon 2004]
Keypoints

- Clustering is a key difficult problem to solve
- Methods based on implicit set representations (e.g. with ZDDs) can cope with large data set
- For all practical purposes, micro-array data can be efficiently processed today
- Outstanding issues
  - Biological validation of data
  - Statistical analysis of results
Data and its context ontologies

- Data in context
- Data analysis
- Sensors and data
GO – The 3 Ontologies

- Every gene is classified from three points of view:
  - Biological process
  - Molecular function
  - Cellular component
- Directed acyclic graphs
  - Each node is uniquely labeled
    - Ex: GO:000xxxx
  - Flat description files
GO - Flat Files

GO:0000067$Gene_Ontology ; GO:0003673
@part_of:biological_process ; GO:0008150
@is_a:cellular process ; **GO:0009987**
@is_a:cell growth and/or maintenance ; GO:0008151 ; synonym:cell physiology @ is_a: physiological process ; GO:0007582
@is_a:cell proliferation ; GO:0008283
@part_of:cell cycle ; GO:0007049 ; synonym:cell-division cycle
@part_of:DNA replication and chromosome cycle ; **GO:0000067**
@is_a:physiological process ; **GO:0007582**
@is_a:cell growth and/or maintenance ; GO:0008151 ; synonym:cell physiology @ is_a: cellular process ; GO:0009987
@is_a:cell proliferation ; GO:0008283
@part_of:cell cycle ; GO:0007049 ; synonym:cell-division cycle
@part_of:DNA replication and chromosome cycle ; **GO:0000067**
@part_of:cellular_component ; GO:0005575
@part_of:molecular_function ; GO:0003674
GoMiner - Positioning in Ontology
Cluster validation by GO

- A cluster with genes in a single category
  - Clear biochemical meaning
  - Precise positioning in GO
Data and its context
clinical traits

- Data in context
- Data analysis
- Sensors and data
Goal

- Correlate clinical traits with genetic data
- Example:
  - Biopsy genetic analysis and imaging data
- Objective:
  - Better diagnosis of diseases and response to medication

CAD -- Computer Aided Diagnosis
Radiological Data

MRI

Coding into a categorical imaging scoring schema

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>49A34</td>
<td>45A9</td>
<td>52A28</td>
<td>4A34</td>
<td>61A31</td>
<td>99A6</td>
<td>46A15</td>
<td>41A31</td>
<td>70P</td>
<td>69A29</td>
</tr>
<tr>
<td>Imaginog score</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Genomic Data

Microarray

Coding into an expression value scoring schema

<table>
<thead>
<tr>
<th>CLID</th>
<th>PATIENT ID</th>
<th>shx013: 49A34</th>
<th>shv060: 45A9</th>
<th>shq077: 52A28</th>
<th>shx009: 4A34</th>
<th>shx014: 61A31</th>
<th>shq082: 99A6</th>
<th>shq083: 46A15</th>
<th>shx008: 41A31</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMAGE:74</td>
<td>ISG20</td>
<td>-1.02</td>
<td>-2.34</td>
<td>1.44</td>
<td>0.57</td>
<td>-0.13</td>
<td>0.12</td>
<td>0.34</td>
<td>-0.51</td>
</tr>
<tr>
<td>IMAGE:76</td>
<td>TNFSF13</td>
<td>-0.52</td>
<td>-4.06</td>
<td>-0.29</td>
<td>0.71</td>
<td>1.03</td>
<td>-0.67</td>
<td>0.22</td>
<td>-0.09</td>
</tr>
<tr>
<td>IMAGE:36</td>
<td>LOC93343</td>
<td>-0.25</td>
<td>-4.08</td>
<td>0.06</td>
<td>0.13</td>
<td>0.08</td>
<td>0.06</td>
<td>-0.08</td>
<td>-0.05</td>
</tr>
<tr>
<td>IMAGE:23</td>
<td>ITGA4</td>
<td>-1.375</td>
<td>-1.605</td>
<td>0.155</td>
<td>-0.015</td>
<td>0.035</td>
<td>-0.035</td>
<td>0.505</td>
<td>-0.865</td>
</tr>
</tbody>
</table>
**Radiogenomic Data**

**Patients (20)**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Trait 1</td>
<td>0 1 0 1 1 1 1 1 1</td>
<td>1 0 1 0 0 1 0 1 0</td>
<td>1 1 1 0 1 1 1 1 1</td>
<td>1 1 1 1 0 0 1 1 0</td>
<td>1 1 1 0 0 1 1 0 1</td>
<td>0 1 1 1 0 0 1 1 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trait 2</td>
<td>1 -0.12 -2.34 1.44 0.57 -0.13 0.12 0.34 -0.51 0.47</td>
<td>-0.52 -4.06 -0.29 0.71 1.03 -0.67 0.22 -0.09 0.1</td>
<td>-0.25 -4.08 0.06 0.13 0.08 0.06 -0.08 -0.05 -6.56E-09</td>
<td>-1.375 -1.605 0.155 -0.015 0.035 -0.035 0.505 -0.865 1.325</td>
<td>-0.4 -3.13 0.62 1.91E-08 0.56 -0.42 -0.91 0.12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trait 3</td>
<td>1 1 1 0 1 1 1 1 1</td>
<td>1 1 1 1 0 0 1 1 0</td>
<td>0 0 1 1 1 1 1 1 1</td>
<td>1 1 1 0 0 1 1 0 1</td>
<td>1 1 1 1 0 0 1 1 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trait 4</td>
<td>1 1 1 1 1 1 0 0 1</td>
<td>0 0 1 1 1 1 1 1 1</td>
<td>1 1 1 0 0 1 1 0 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trait 5</td>
<td>1 1 1 0 1 1 1 1 1</td>
<td>0 0 1 1 1 1 1 1 1</td>
<td>1 1 1 0 0 1 1 0 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Clinical Data**
- Trait 1: 0 1 0 1 1 1 1 1 1
- Trait 2: 1 0 1 0 0 1 0 1 0
- Trait 3: 1 1 1 0 1 1 1 1 1
- Trait 4: 1 1 1 1 0 0 1 1 0
- Trait 5: 1 1 1 0 0 1 1 0 1

**Imaging data**

**Genomic Data**

- **IMAGE:** ISG20 || interferon stimulated gene 20kDa
  - Trait 1: -1.02 -2.34 1.44 0.57 -0.13 0.12 0.34 -0.51 0.47
- **IMAGE:** TNFSF13 || tumor necrosis factor (ligand) superfamily, member 13
  - Trait 1: -0.52 -4.06 -0.29 0.71 1.03 -0.67 0.22 -0.09 0.1
- **IMAGE:** LOC93343 || hypothetical protein BC011840
  - Trait 1: -0.25 -4.08 0.06 0.13 0.08 0.06 -0.08 -0.05 -6.56E-09
- **IMAGE:** ITGA4 || integrin, alpha 4 (antigen CD49D, alpha 4 subunit of VLA-4 receptor)
  - Trait 1: -1.375 -1.605 0.155 -0.015 0.035 -0.035 0.505 -0.865 1.325
- **IMAGE:** SGCE || sarcoglycan, epsilon
  - Trait 1: -0.4 -3.13 0.62 1.91E-08 0.56 -0.42 -0.91 0.12
- **IMAGE:** RPS4Y || ribosomal protein S4, Y-linked
  - Trait 1: -0.83 -1.3 0.65 -2.07 0.31 -2.16 0.88 -2.37 0.36
- **IMAGE:** SMCY || Smcy homolog, Y chromosome (mouse)
  - Trait 1: -1.99 -2 0.21 -2.15 1.25 -2.7 0.38 -2.71 0.1
- **IMAGE:** DBY || DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide, Y chromosome
  - Trait 1: -1.29 -0.99 0.34 -1.35 -0.03 -0.95 0.41 -1.01 0.42
- **IMAGE:** EIF1AY || eukaryotic translation initiation factor 1A, Y chromosome
  - Trait 1: -1.06 -0.72 0.78 0.01 0.24 -0.91 0.48 -0.6 0.54

G. De Micheli -- ISCAS 2005
Co-clustering

- Unsupervised cluster search method in higher dimensional spaces
- Correlate data and provide qualitative information about pathology
- Results:
  - Method was applied to renal carcinoma with good correlation
- Outstanding issue:
  - Scoring of radiological data
  - Can this be automated?
Immunohistochemical Automated Quantification

- Acquiring quantitative and qualitative information from immuno-stains

↓

- Automated image processing methods to standardize IHC analysis
Example of IHC images

EGFR/erb-B receptors positivity (in carcinoma cells) as brown stain
Negative carcinoma cells and other cells in the sample as blue stain
System view

- A closed loop system
  - Identification
  - Optimum control

Data Analysis

- Genetic data
- Clinical data

Treatment
Microarray data and genetic pathways

- Data in context
- Data analysis
- Sensors and data
Gene regulatory networks

Example: Phage $\lambda$

[Myers 2003]
Gene regulatory networks
Example: Phage λ decision circuit

[MYERS 2003]
Models of gene evolution

- Finite-state behavior
  - FSMs
- Transients
  - ODEs
  - Hybrid systems
- Stochasticity
  - Markov models
- Asynchrony
Gene regulatory networks
Example: Phage λ asynchronous model
Gene regulatory networks

Databases

- Pathways
  - HS, animals and plants
- Genomic
  - BLAST searches
- Chemical
  - Compound structures
- BRITE
  - Drug taxonomy
Gene regulatory networks
Circadian cycle in homo sapiens

[Diagram showing gene regulatory network]
Summary and conclusions

- Genomic information has been exploding
  - Its effective use is still limited
- Application areas include:
  - Clinical medicine, pharmaceutics, biology
- Technologies include:
  - Modeling, abstraction, circuit design, data analysis, dedicated languages, optimal control

CAS researchers master these technologies and can widen their horizons to these application areas