Networking Genes and Drugs: Understanding Gene Function and Drug Mode of Action from Large-scale Experimental Data

Diego di Bernardo
The problems we (and everybody else) are tackling:

- metabolic networks
- protein networks
- transcript networks

What is the role of my gene?

Which small molecule (drug) can modify the pathway of interest?
A ‘simple’ protein-protein interaction network (yeast S. cerevisiae)
Part I: Understanding Gene Function

The human and mouse Interactome

Belcastro V et al, UNPUBLISHED (confidential)
Reverse-engineering: from data to model

*In vivo* perturbations | Measure response | Learn | Computer model

Knockout | Drug | Stress | Overexpress | RNAi | Learning Algorithm | Cell network
Reverse engineering human and mouse gene networks:

Expression Data (human and mouse)

Reverse-engineering

MI > threshold

Gene network
Expression Data:

**HUMAN (HG-U133A)**
- 702 experiments (20255 hyb.)
- 22283 probesets (P)
- 14340 genes

**MOUSE (Mouse430_2)**
- 797 experiments (8895 hyb.)
- 45101 probesets (P)
- 28219 genes

Data structure:

```
T
P1
P2
...
Exp. 1

P1
P2
...
Exp. 2

P1
P2
...
Exp. N
```

microarrays

...
Reverse-engineering:

Bayesian Networks

Information-theoretic

Ordinary differential equations

\[ P(A/B, C, D, E) = P(A/B, C) \]

\[ \text{MI}(A, H) = 0 \]
\[ \text{MI}(A, B) > 0 \]
\[ 0 < \text{MI}(A, D) < \min(\text{MI}(A, B), \text{MI}(B, D)) \]

\[ \frac{dA}{dt} = \theta_1 A + \theta_2 B + \theta_3 C \]
or more generally:
\[ \frac{dA}{dt} = f(A, B, C, \theta) \]

BANJO
(Hartemink, A. Nature Biotechnology, 2005.)

ARACNE
(Basso et al., Nature Genetics, 2006)

NIR and TSNI (Gardner, et al, Science, 2003; Bansal et al, Bioinformatics, 2006; Della Gatta et al, Genome Research, 2008)

DYNAMIC AND STEADY-STATE (n-way)

STEADY-STATE (2-way)

DYNAMIC AND STEADY-STATE (n-way)
Reverse-engineering:

BANJO
(Hartemink, A. Nature Biotechnology, 2005.)

ARACNE
(Basso et al., Nature Genetics, 2006)

NIR and TSNI
(Gardner, et al, Science, 2003; Bansal et al, Bioinformatics, 2006; Della Gatta et al, Genome Research, 2008)

DYNAMIC AND STEADY-STATE (n-way)

STABLE-STATE (2-way)

DYNAMIC AND STEADY-STATE (n-way)
Mutual Information:

Independent genes

Genes that regulate each other

\[ I(X, Y) = \sum_{x \in \{1, ..., r\}, y \in \{1, ..., s\}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \]
Computation of Mutual Information:

\[ \begin{array}{cccccccccc}
\text{P1} & 2.3 & 2 & -1 & 0.7 & 2.6 & -0.7 & -0.7 & -0.3 & \ldots \\
\text{P2} & 5 & 8.2 & 1 & -0.7 & -0.2 & 0.3 & 2.3 & -0.7 & \ldots \\
\end{array} \]

Discretization

\[ (\text{Liu et al., 2002}) \text{ Data Mining and Knowledge Discovery, 6, 393-423.} \]

Matrix of Occurrences

\[ \begin{array}{ccc}
- & 1 & 1 \\
= & 1 & 0 \\
+ & 1 & 0 \\
\end{array} \]

(3 bins)
How to obtain one huge datasets? (Dataset merging)

Exp. #1

\[ \begin{array}{ccc}
- & 1 & 2 & 1 \\
\hline \\
+ & 1 & 0 & 1 \\
\end{array} \]

Exp. #2

\[ \begin{array}{ccc}
- & 1 & 2 & 0 \\
\hline \\
+ & 1 & 1 & 0 \\
\end{array} \]

... / #common microarrays

Exp. #N

\[ \begin{array}{ccc}
- & 0 & 2 & 1 \\
\hline \\
+ & 1 & 2 & 1 \\
\end{array} \]
Frequentist approach to MI:

<table>
<thead>
<tr>
<th>p1 \ p2</th>
<th>-</th>
<th>+</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>10/90</td>
<td>12/90</td>
</tr>
<tr>
<td>=</td>
<td>8/90</td>
<td>20/90</td>
</tr>
<tr>
<td>+</td>
<td>6/90</td>
<td>18/90</td>
</tr>
</tbody>
</table>

Mutual Information (MI) is the amount of information two random variables share.

MI can be used to measure how dependent two probes are.

\[
I(X,Y) = \sum_{x \in \{1..r\}, \ y \in \{1..s\}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}
\]

\[
\hat{I}(X,Y) = \sum_{x \in \{1..r\}, \ y \in \{1..s\}} f(x, y) \log \frac{f(x, y)}{f(x)f(y)}
\]

where

\[
f(z) = \frac{n_z}{n}
\]

\[
f(x, y) = \frac{n_{xy}}{n}
\]

dibernardo.tigem.it
Network statistics and properties

- **HUMAN**
  - 20,255 experiments
  - 22,283 probes
  - Threshold 0.04
  - 4,817,629 edges

- **MOUSE (Mouse430_2)**
  - 8,895 experiments
  - 45,101 probes
  - Threshold 0.025
  - 14,461,095 edges

- **Mouse+Human**
  - 10,415 genes
  - Threshold **H 0.04, M 0.025**
  - 3,283,347 edges

- 20,123 of the human genes belong to the same component.
Interactome validation on experimentally verified interactions

• The Golden standard is a collection of experimentally validate edges for a total of **105,688** edges from a wide range of publicly available databases:
Results -- Validation of gene-to-gene interactions

Validation (PPV vs Sensitivity)

- ~76% of the predicted edges were true.
- This doesn’t mean that the remaining ~24% are not correct.

- COXPRESdb: a database of coexpressed gene networks in mammals
Results -- Validation of gene-to-gene interactions

~76% of the edges predicted are true,
• This doesn’t mean that the remaining ~24% are not correct.

Genes involved into the spindle check point. We are currently experimentally validating these interactions via Y2H:
**Netview: Online visualization tool**

**Query the database with a Gene Symbol**

<table>
<thead>
<tr>
<th>Specie</th>
<th>human</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identifier</td>
<td>gene_symbol</td>
</tr>
<tr>
<td>Gene Symbol</td>
<td>FOS</td>
</tr>
<tr>
<td>Tissue</td>
<td>ALL</td>
</tr>
<tr>
<td>Neighbors</td>
<td>10</td>
</tr>
<tr>
<td>Depth</td>
<td>2</td>
</tr>
</tbody>
</table>

**Neighbors**: # of nodes directly connected to the queried node.
**Depth**: # of network levels to explore (root is the queried node).
Netview: Online visualization tool

- **jSquid**: a Java applet for graphical on-line network exploration
  Bioinformatics 2008 24(12):1467-1468.
Using the network to understand gene function

1. Neighbors selection.

2. Neighbors Enrichment analysis via hypergeometric distribution.

3. Gene function prediction.

Cell Cycle gene ontology

other gene ontology

\[ p\text{-value: } 6.6614\times10^{-29} \]
Using the network to understand gene function: validation

- 58% of the genes were correctly assigned to a gene function.

- This doesn’t mean that the remaining 42% are not properly assigned to a gene function.

- We are now validating experimentally gene function for 8 genes predicted to be localised in mitochondria.
Conclusion of Part I:

• Using expression data from a wide variety of tissues and cell lines enables the identification of functional modules within the cell regulatory network

• It is possible to predict functional and physical interactors of a gene using co-expression networks

• It is possible to predict the function of a gene from its interactors (i.e. co-expressed genes)

• We are now looking at how this global interactome network can be useful in interpreting gene expression data and to understand the global organisation of the cell regulatory network
Part II: Understanding Drug Mode of Action

The Drug Network

Iorio F et al, UNPUBLISHED (confidential)
Drug Discovery Problem

We want to investigate the mode of action of a novel drug...

Therapeutic Target

Off Target

Drug Molecule
The Connectivity Map DataSet (microarrays):

**small molecules:** 1309 perturbagens tested
(FDA approved and nondrug bioactive compounds)

**cell lines:**
MCF7  (human epithelial breast cancer)
PC3    (human epithelial prostate cancer)
HL60   (human leukemia)
SKMEL5 (human melanoma)
ssMCF7 (MCF7 grown in a different vehicle)

**Concentration and treatment**
10mM  (when the optimal concentration is unknown) x 6h

**Negative Control**
cells in the same plate and treated with vehicle alone (medium, DMSO...)

---

[Lamb et al, Science 2006]
General Cellular Response to a Drug:

Using a novel rank aggregation method (next slide)

Prototype Ranked List (PRL) for Drug A

[Iorio et Al, Journal of Computational Biology 2009]

The Kru-Bor Merging Method

- $D$: The set of all the possible permutations of microarray probes;
- $X$: A set of ranked lists of probes computed by sorting, in decreasing order, the genome-wide differential expression profiles (GEP) obtained by treating cell lines with the same drug;
- $\delta: D^2 \rightarrow N$: The Spearman’s Foot-Rule distance associating to each pair of ranked lists in $X$ a natural number quantifying the similarity between them;
- $B: D^2 \rightarrow D$: The Borda Merging Function, associating to each pair of ranked lists in $X$, a new ranked lists obtained by merging them with the Borda Merging Method;

Prototype
Ranked List
For drug A

1. $n = |X|$
2. while $n > 1$
3. find $i, j : \delta(x_i, x_j) = \min_{p,q=1,...,n, p\neq q} \delta(x_p, x_q)$
4. $y = B(x_i, x_j)$
5. $X = (X \setminus \{x_i, x_j\}) \cup \{y\}$
6. $n = |X|$
7. end
The Drug Distance Matrix

For each drug

Check how many changed genes are in common

Compute similarity by using Enrichment Scores
Computation of the drug distance:

• Given a set of \( NH \) probes in \( S \) and a ranked list of \( N \) probes:

• The Enrichment Score of \( S \) on the list is defined as:
  
  \[
  \text{max}_i |P_{\text{hit}} - P_{\text{miss}}| 
  \]

  where:

  \[
  P_{\text{hit}}(S, i) = \sum_{\substack{g_j \in S \\ j \leq i}} \frac{1}{NH} 
  \]

  \[
  P_{\text{miss}}(S, i) = \sum_{\substack{g_j \notin S \\ j \leq i}} \frac{1}{(N - NH)} .
  \]
Computation of the distance:

*Total Enrichment Score*

Given two set of probe identifiers \( p = \{p_1, \ldots, p_h\} \) and \( q = \{q_1, \ldots, q_w\} \) we define the Total Enrichment Score, TES, of the *signature* \( \{p, q\} \) respect to the GEP \( x_i \), as follows:

\[
TES_i^{\{p, q\}} = \frac{ES_i^p - ES_i^q}{2}.
\] (1)

We then define as a distance between two compounds \( i \) and \( j \) the following quantity:

\[
\frac{1}{2} \left( TES_i^{\{p_j, q_j\}} + TES_j^{\{p_i, q_i\}} \right)
\]
The Drug Network is obtained by setting a threshold:

Distance
Threshold = 0.2

<table>
<thead>
<tr>
<th>Compound</th>
<th>Informations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cephaeline</td>
<td>Ipecac (a plant) Alkaoids - Protein Synthesis inhibition</td>
</tr>
<tr>
<td>Emetine</td>
<td>Steroids found in some species of digitalis (purpurea or lanata), a plants.</td>
</tr>
<tr>
<td>Digoxigenin</td>
<td>Used to treat Cardiac Diseases.</td>
</tr>
<tr>
<td>Digoxin</td>
<td>Cardiac Glycoside</td>
</tr>
<tr>
<td>Digitoxigenin</td>
<td>Endogenous hormone found in the ripe seeds of the african plant Strophanthus.</td>
</tr>
<tr>
<td>Ouabain</td>
<td>It blocks the sodium pump and it is used to cure human heart failure, angina</td>
</tr>
<tr>
<td>Proscillaridin</td>
<td>pectoris and Myocardial infarction.</td>
</tr>
<tr>
<td>Lanatoside C</td>
<td>Cardiotonic Glycoside isolated from Scilla maritima var. Alba (a plant)</td>
</tr>
<tr>
<td>Compound</td>
<td>Informations</td>
</tr>
<tr>
<td>-------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Cicloheximide</td>
<td>Antibiotic substance isolated from streptomycin-producing strains of Streptomyces griseus. It acts by inhibiting elongation during protein synthesis.</td>
</tr>
<tr>
<td>Geldanamycin</td>
<td>a benzoquinone ansamycin antibiotic that binds to Hsp90 (Heat Shock Protein 90) and alters its function.</td>
</tr>
<tr>
<td>Alvespimycin</td>
<td>Hsp90 inhibitor that has demonstrated the potential to disrupt the activity of multiple oncogenes and cell signaling pathways implicated in tumor growth, including HER2, a key signaling pathway in breast cancer.</td>
</tr>
<tr>
<td>Vorinostat</td>
<td>or suberoylanilide hydroxamic acid (SAHA) is a member of a larger class of compounds that inhibit histone deacetylases (HDAC).</td>
</tr>
<tr>
<td>Scriptaid</td>
<td>A novel histone deacetylase inhibitor</td>
</tr>
<tr>
<td>HC Toxin</td>
<td>Inhibition of Maize Histone Deacetylases by HC Toxin, the Host-Selective Toxin of Cochliobolus carbonum</td>
</tr>
<tr>
<td>Rifabutin</td>
<td>Rifabutin is a bactericidal antibiotic drug primarily used in the treatment of tuberculosis. The drug is a semi-synthetic derivative of rifamycin S. Its effect is based on blocking the DNA-dependent RNA-polymerase of the bacteria.</td>
</tr>
<tr>
<td>Compound</td>
<td>Informations</td>
</tr>
<tr>
<td>------------</td>
<td>-------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>anisomycin</td>
<td>also known as flagecidin is an antibiotic produced by Streptomyces griseolus which inhibits protein synthesis.</td>
</tr>
<tr>
<td>captothecin</td>
<td>a cytotoxic quinoline alkaloid which inhibits the DNA enzyme topoisomerase I (topo 1).</td>
</tr>
<tr>
<td>irinotecan</td>
<td>a chemotherapy agent that is a topoisomerase 1 inhibitor. Chemically, it is a semisynthetic analogue of the natural alkaloid camptothecin.</td>
</tr>
<tr>
<td>astemizole</td>
<td>a second generation antihistamine drug which has a long duration of action.</td>
</tr>
<tr>
<td>terfenadine</td>
<td>an antihistamine formerly used for the treatment of allergic conditions.</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Lanatoside C, anisomycin, caspase-3 inhibitors

<table>
<thead>
<tr>
<th>Compound</th>
<th>Informations</th>
</tr>
</thead>
<tbody>
<tr>
<td>monorden</td>
<td>(radicicol) antifugal metabolites. It inhibits the Hsp90 Chaperone</td>
</tr>
<tr>
<td>alsterpaullone</td>
<td>CDKs inhibitor</td>
</tr>
<tr>
<td>doxorubicin</td>
<td>a drug widely used in cancer chemotherapy. It is an anthracycline antibiotic and structurally closely related to daunomycin. Used in combination with CDKs inhibitors</td>
</tr>
</tbody>
</table>
The Drug network

There is an edge connecting two drugs if their distance is below a fixed threshold.

Distance Threshold = 0.8049
Statistics

number of connected vertices = 1302

number of edges = 41047 (~ 5% of a fully connected network with the same number of nodes)

Avg. Shortest Path length = 2.5

Avg. Local Clustering Coefficient = 0.44

Maximum Shortest Path = 7

Node Degree Empirical cdf
Community Identification by Hierarchical Clustering by Message Passing

Frey et al., Science 2007
Community Validation

<table>
<thead>
<tr>
<th>Id.</th>
<th>Most Enriched Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>Antipsychotics</td>
</tr>
<tr>
<td>65</td>
<td>COX2 Modulators</td>
</tr>
<tr>
<td>44</td>
<td>Dopaminergic Agents</td>
</tr>
<tr>
<td>73</td>
<td>Alpha and Beta Adrenergic Modulators</td>
</tr>
<tr>
<td>81</td>
<td>Serotonin Receptor Modulators, Antiparkinsonians</td>
</tr>
<tr>
<td>53</td>
<td>Protein Synthesis Inhibitors</td>
</tr>
<tr>
<td>63</td>
<td>Na+/K+ - ATPase membrane pump inhibitors</td>
</tr>
<tr>
<td>75</td>
<td>Hepatic Enzymes Inducers</td>
</tr>
<tr>
<td>16</td>
<td>Histone Deacetylase Inhibitors</td>
</tr>
</tbody>
</table>
Community Enrichment Analysis

Identified Communities: 24
Enriched Communities: 61
ATC-Code Enriched: 43
Direct Target Enriched: 16
Functional Enriched: 15

Total: 106
Neighborhoods have community structure and a large number of community is functional enriched.

Experimental validation with three compounds:

This is helpful to "recover" the mode-of-action of a novel drugs.
Mapping gene changes due to drug treatment:
Part II: Conclusions

Drug Network = A novel, efficient tool to study drugs and their mode of action by gene expression profiling

- Performance assessment showed that 91% of tested compounds were correctly classified

- The modular structure of the sub-network that surrounds a new drug elucidates the MOA of the drug
Maria Aurelia Ricci
Velia Siciliano
Alda Graziano
Giulia Cuccato
Lucia Marucci
Filippo Menolascina
Diego di Bernardo

Systems, Synthetic and Computational Biology Lab

Gennaro Gambardella
Vincenzo Belcastro

Computer Science

Mathematics

Biology

Computational Systems

Electrical Engineering

Nerviano Medical Sciences (Milano) - Italy