Strategies for Mining the NCI’s Screening Databases:

Data (NCI60, Xenograft)
Informatics (Bio and Chemo)

Laboratory of Computational Technologies
Anders Wallqvist, Ruili Huang, Narmada Thanki,
Xiang-Jun Lu, Alfred Rabow

NIH/NCI/DCTD/DTP/STB
Drs. Doroshow, Collins, Shoemaker

spheroid.ncifcrf.gov
Information → Knowledge → Hypothesis Mining → Tools → Visualization

- Successes
- Pitfalls
- Strategies
- Recommendations
Data Generation / Data Analysis

Compounds

Functional screen

Phenotypic readout

Database

Decision

Gene Function

Drug Function

Statistics Mathematics
Data Fusion

- Mechanistic
- Specific
- General

- Single
- Descriptive
- Scale
- Detail

- Context

Diagram shows the relationship between data fusion and dimensions such as specific vs. general, mechanistic vs. descriptive, single vs. group, and context vs. scale.
Data Fusion

“Interactive WEB”
NCI_{60}: lung, renal, colorectal, ovary, breast, prostate, central nervous system, melanoma, leukemia

\sim 100,000 \text{ compounds}

GI_{50}: 50\% \text{ cancer cell growth inhibition concentration}

Wallqvist et al. JCIM (2006)
GI<sub>50</sub> SOM
Mechanism of Action Categories:
[M] Anti-mitotic...............large and functional
[S] DNA synthesis.............low lipophilicity
[P] Phosphatase/kinase.........most diverse signal
[R] Membrane active..........high lipophilicity
[Q] Xenobiotic metabolism...reactive groups
Chemistries: Modeling GI50

Selecting potent compounds

- large (many previous studies)
- complex: many features
  (Oprea, Blake, Veber, Veith)
- local effects (SOM regionalization)
  - potency scales with selectivity
    (Huang et al.)
Gene Expression vs GI50

6 NSCs selected from highest + correlations
Hypothemycin, LF, PD98059
Rosen et al.
Sellers et al.
Pitfall

GI50: Gene Expression Correlations

N=8835

N=4907

~90k unique GI50:gene expression profiles

Pearson Correlation (p<0.005)
### Linking Pathway Gene Expression to GI$_{50}$

<table>
<thead>
<tr>
<th>cells</th>
<th>cells</th>
</tr>
</thead>
<tbody>
<tr>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>↓</td>
<td>↓</td>
</tr>
</tbody>
</table>

**coherent pathway**

**non-coherent pathway**

Linking Pathway Gene Expression to GI$_{50}$

For pathway $P$:

**Genes in $P$**

- $gene_1$
- $gene_2$
  - $r_{in,1}$
  - $r_{in,2}$
  - $r_{in,i}$
  - $r_{in,n}$
- ... ...
- $gene_n$

**Genes not in $P$**

- $gene_1$
- $gene_2$
  - $r_{out,1}$
  - $r_{out,2}$
  - $r_{out,j}$
  - $r_{out,m}$
- ... ...
- $gene_m$

Pearson Correlation: $r$

$R_{in} = \{ r_{in,1}, r_{in,2}, \ldots, r_{in,i}, \ldots, r_{in,n} \}$

$R_{out} = \{ r_{out,1}, r_{out,2}, \ldots, r_{out,j}, \ldots, r_{out,m} \}$

Kruskal-Wallis $H$, $p$

$R_{in} > R_{out} \iff H > 0$

$R_{in} < R_{out} \iff H < 0$

Drug is significantly associated with $P$ if: $H > 0$ and $p < 0.05$

$H$ defines a Fitness Score for pathways against GI$_{50}$
Relating Fitness Scores to Drug Response

Nucleotide sugar metabolism
Kegg (hsa00520)
24 member pathway

Pearson Correlation
mRNA_{pathway} : GI50_{node}

Kruskall Wallis Statistic

Pathway Fitness (coherence)

Inhibitors and structural analogs
Kegg Pathways

MAPK pathway fitness

Candidate Agents

New drugs?

New targets/MOA?
Xenograft data

Experimental Design
- 1363 NSC tested
- 31 formulations
- 187 treatment schedules
- 50 tumor models
- 6 implantation sites
- 15 mice strains
- >5,000 combinations of experiments

Measurements
- Tumor weight reduction (TW50)
- Survival time (ST150)
- Toxicity (survival control vs treatment)
- Therapeutic index (TW50, ST150/Tox)
survival time (ST)

T/C, %

mg/kg/injection

r²=0.82

modelled log(TW50)

measured log(TW50)

tumor weight (TW)

toxicity (TX)

Sensitivity: TP/(TP+FN)

FDA approved

1-Specificity: 1-TN/(TN+FP)

random
Activity Class

Fitness Scores

Xenograft Efficacy
399 Anticancer Medicines in Development (283 nonbiologicals) 123 (45%) have structural analogs in NCI screening set

Ind’s success
Recommendations

Statistics:
- beyond sorting
- clustering
  - SOM
- decision trees
  - random forests
- curse of dimensionality
  - false positives
  - positive predicitive value

Data Sharing
- chemistry
- gene expression
- mutation
- SNP
- ‘cancer genes’
- negative results

Reverse mining
- retrospective testing
  - clinical trials
  - preclincinal data
Molecular properties

Toxicology

Cellular growth inhibition

Gene expression

SNP copy number

Clinical trials

Proteomics

Xenografts

Karyotype

Methylation status

microRNA

“"It is not enough to know the principles, one needs to know how to manipulate”

- Dictionnaire de Trevoux, quoted by Michael Faraday in the first edition of Chemical Manipulation (1827)
Chemistries: Modeling GI50

$\text{GI}_{50} = F_1(\text{properties}) = c_1x_1 + c_2x_2 + \ldots + c_Nx_N$

Training: $r^2 = 0.77$

Testing: $r^2 = 0.67$
Xenograft data

\[ \text{outcome} = B \times [ \text{(treatment)} \times \text{(chemistry)} \times \text{(cellular growth inhibition)} ] \]

exptl design \hspace{1cm} \text{properties} \hspace{1cm} \text{GI50}

Treatment variations alone account for a log order of difference in efficacy outcome.
Molecular Classes

- Antineoplastic Antibiotics
- Direct Membrane
- Antimitotic
- Intercalating
- DNA Polymerase
- Chelating
- Kinase/Phosphatase
- CDK
- Ion Channel
- Golgi
- Purine Antimetab.
- Pyrimidine Antimetab.
- Topo I
- Topo II
- Alkylating

Golgi Kinase/Phosphatase

CDK

Ion Channel

Purine Antimetab.

Pyrimidine Antimetab.

Topo I

Topo II

Alkylating
<table>
<thead>
<tr>
<th></th>
<th>Act 1</th>
<th>Act 2</th>
<th>Act 3</th>
<th>Interm.</th>
<th>Act 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>GI50</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chemistry</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mRNA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pathways</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xenograft</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Act 1</td>
<td>Act 2</td>
<td>Act 3</td>
<td>Interm.</td>
<td>Act 4</td>
</tr>
<tr>
<td>------------------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>GI50</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Chemistry</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>mRNA</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Pathways</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Xenograft</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Act 1</td>
<td>Act 2</td>
<td>Act 3</td>
<td>Interm.</td>
<td>Act 4</td>
</tr>
<tr>
<td>----------------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>GI50</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Chemistry</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mRNA</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Pathways</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xenograft</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Chemistry Meets Biology

<table>
<thead>
<tr>
<th></th>
<th>Act 1</th>
<th>Act 2</th>
<th>Act 3</th>
<th>Interm.</th>
<th>Act 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>GI50</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Chemistry</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mRNA</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Pathways</td>
<td></td>
<td></td>
<td></td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Xenograft</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Pathway Fitness - Cohesiveness

- Relationship between the number of genes in a pathway that are shared with other pathways and the cohesiveness of the pathway

- Genetic Information Processing
  - highest percentage of cohesive pathways
  - least number of shared genes

- Environmental Information Processing
  - lowest percentage of cohesive pathways
  - largest number of shared genes

More cohesive: protein biosynthesis, mitosis, energy transfer
Less cohesive: apoptosis, chromatin remodeling, transport

Gene --- Pathway --- Drug

Connectivity Maps
Lamb et al., 2006
Rapamycin Family

h_vegf
h_ires
h_ran
GO:74 cell cycle
GO:1525 angiogenesis
GO:6099 tca cycle

h_ctc
GO:3724 RNA helicase
GO:6631 fatty acid metabolism

- Structurally similar to temsirolimus (N=24)
  179 gene expressions are correlated with these 24 NSCs
## Chemistry Meets Biology

<table>
<thead>
<tr>
<th></th>
<th>Act 1</th>
<th>Act 2</th>
<th>Act 3</th>
<th>Interm.</th>
<th>Act 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>GI50</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Chemistry</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>mRNA</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Pathways</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Xenograft</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>
Look. We know *that* it works ---- *that* is no longer the question. What we now want to know is *how*… *How* now brown cow?”
insensitive  GI50  sensitive
ABCB1 mRNA expression
GI50 correlations
Thiosemicarbazone NSC73306
MDR substrates
Gene Expression vs GI50
H-acceptor path 3
+ charge
POS
NEG
Thiosemicarbazone NSC73306
Phosphatase_kinase agents produce near maximal tumor weight reduction for modest values in GI50 and Therapeutic Index.