

Informatics Methods for Correlating Molecular Structure with Genomic Data

**Paul Blower, Michael Fligner, Joseph Verducci,
and Chihae Yang**

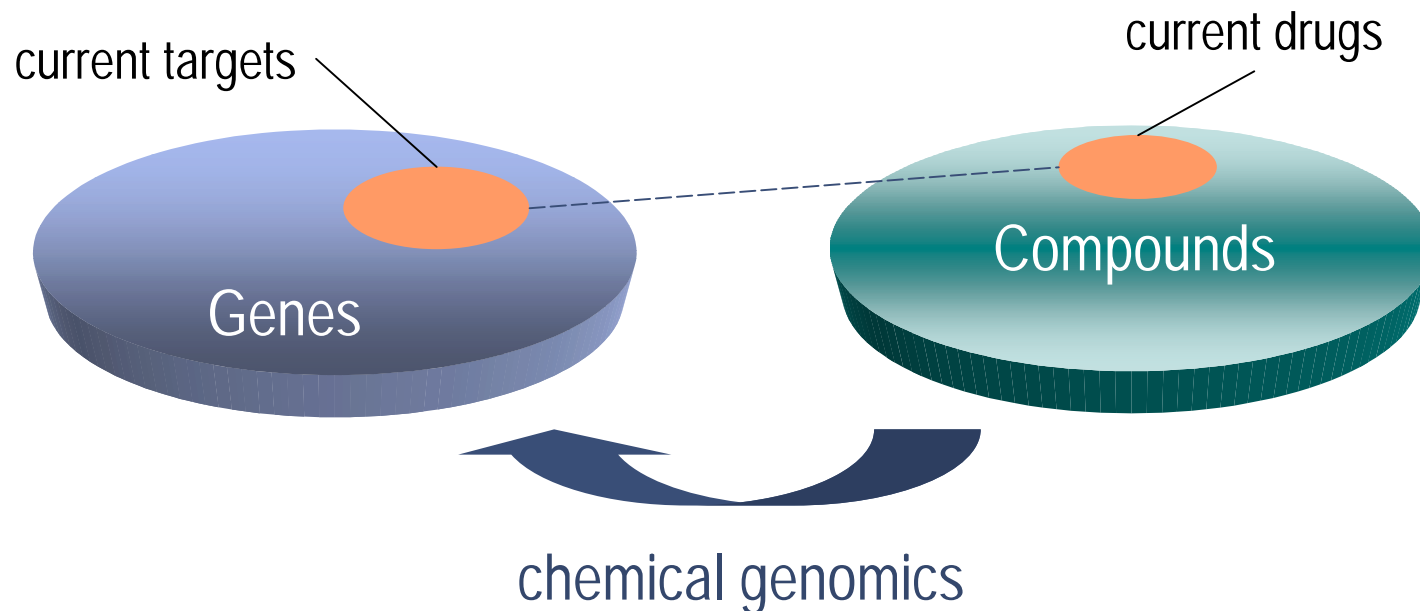


Challenges in the Post Genomics Era

- Large volume of data
- Missing data, Noisy data
- Experimental errors
- Biologically meaningful interpretation
- Too many genes that look “promising” or “interesting” as drugable targets...
- Viable methodologies to connect to chemistry

Chemical Genomics

- A new paradigm of drug discovery where genomic or proteomic responses of whole cells or tissues are linked to chemical compound classes in an iterative process.
- Overall process is faster and results provide better clinical candidates with fewer side effects and lower toxicity.





Requirements

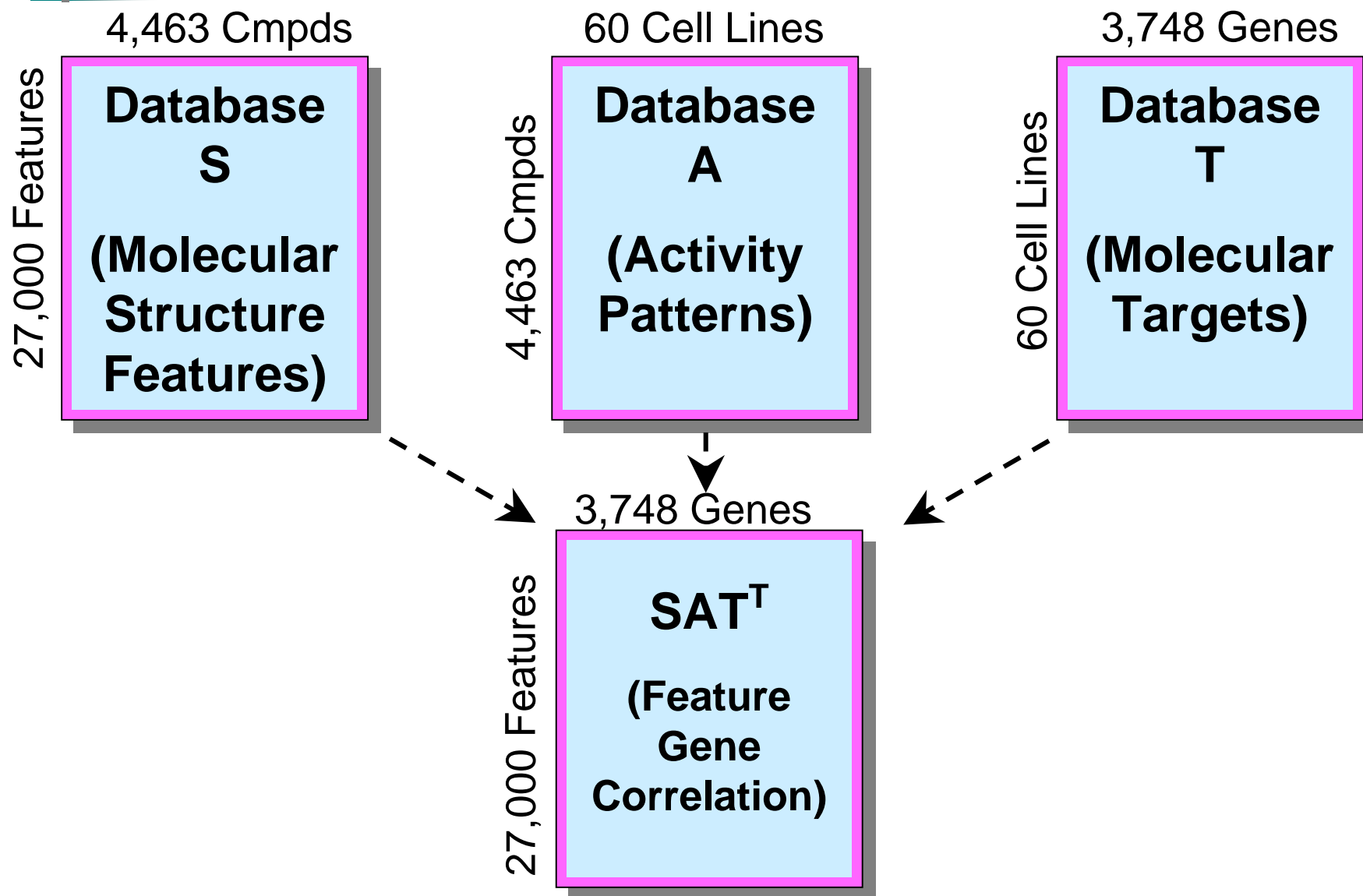
- Gene or protein family classifications
- Chemical compound scaffolding method
- Chemically and biologically intuitive informatics methods
- Experimental databases
 - Compound biological activities
 - Genomics or proteomics data (e.g., gene expression patterns)



Overview

- NCI Datasets
- Compound Gene Correlations
- Structure-based Data Mining
- Compound Testing

Conceptual Framework





NCI Gene Expression Dataset

- Microarrays spotted with 9703 cDNA elements

- mRNA isolated from NCI 60 cancer cell lines

- Leukemia (6)

- Melanoma (7)

- Breast (8)

- Ovarian (6)

- CNS (6)

- Lung (9)

- Prostate (2)

- Colon (7)

- Kidney (8)

- 12 cell lines used for reference pool

- Fluorescence tagged during hybridization

- DNA elements are from Washington Univ. Merck
IMAGE

- ~3700 named genes

- ~ 1,900 human homologues

- 4104 EST

* Source: <http://discover.nci.nih.gov>; U. Scherf, et. al., *Nature Genet.*, **2000**, **24**, 236–44.

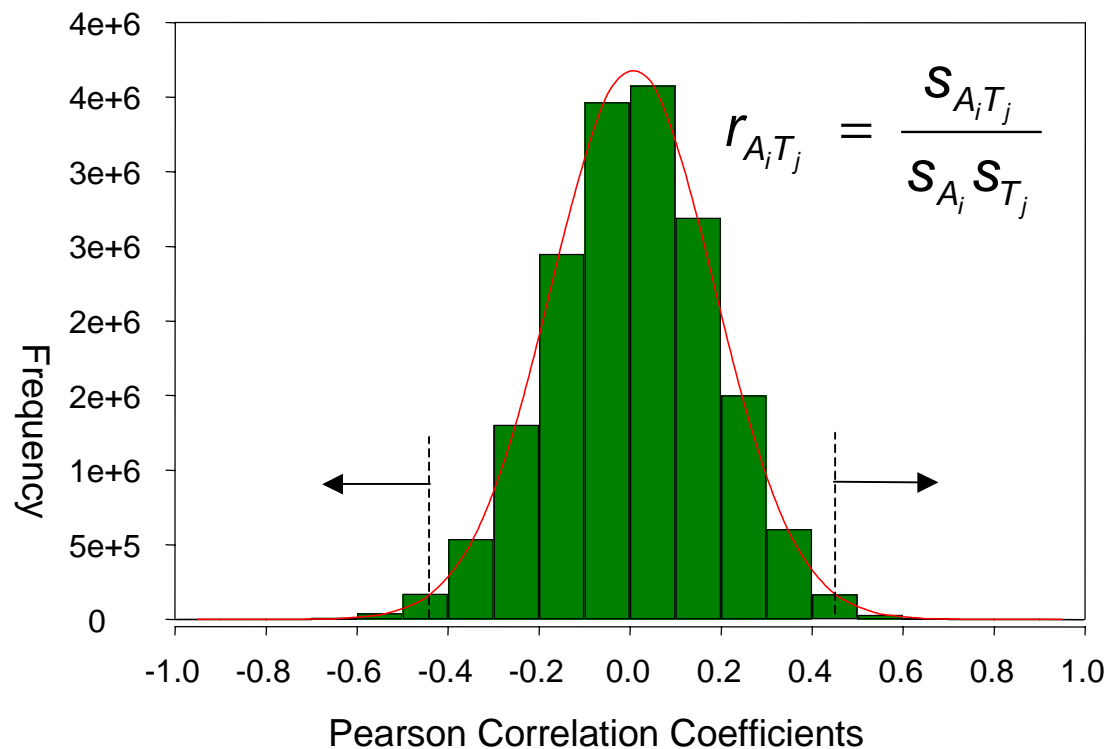


Compounds Used in Study

- NCI 4,463 compounds tested 2 or more times
- Each compound tested at 5 concentrations, usually 10^{-4}M - 10^{-8}M
- Used growth inhibition (GI_{50}) of compounds over NCI60 cell lines

Gene-Compound Correlations

- Across NCI60 cell lines
- At 5 % error rate, correlations higher than .45 or lower than -0.45 were considered significant





Selection of Genes

- Based on statistical techniques
 - Genes with high variance over 60 cell lines
 - Genes correlated with compound activity from AT matrix
 - Genes correlated with cell origins
- Based on gene hierarchy
- Based on results of compound exposure



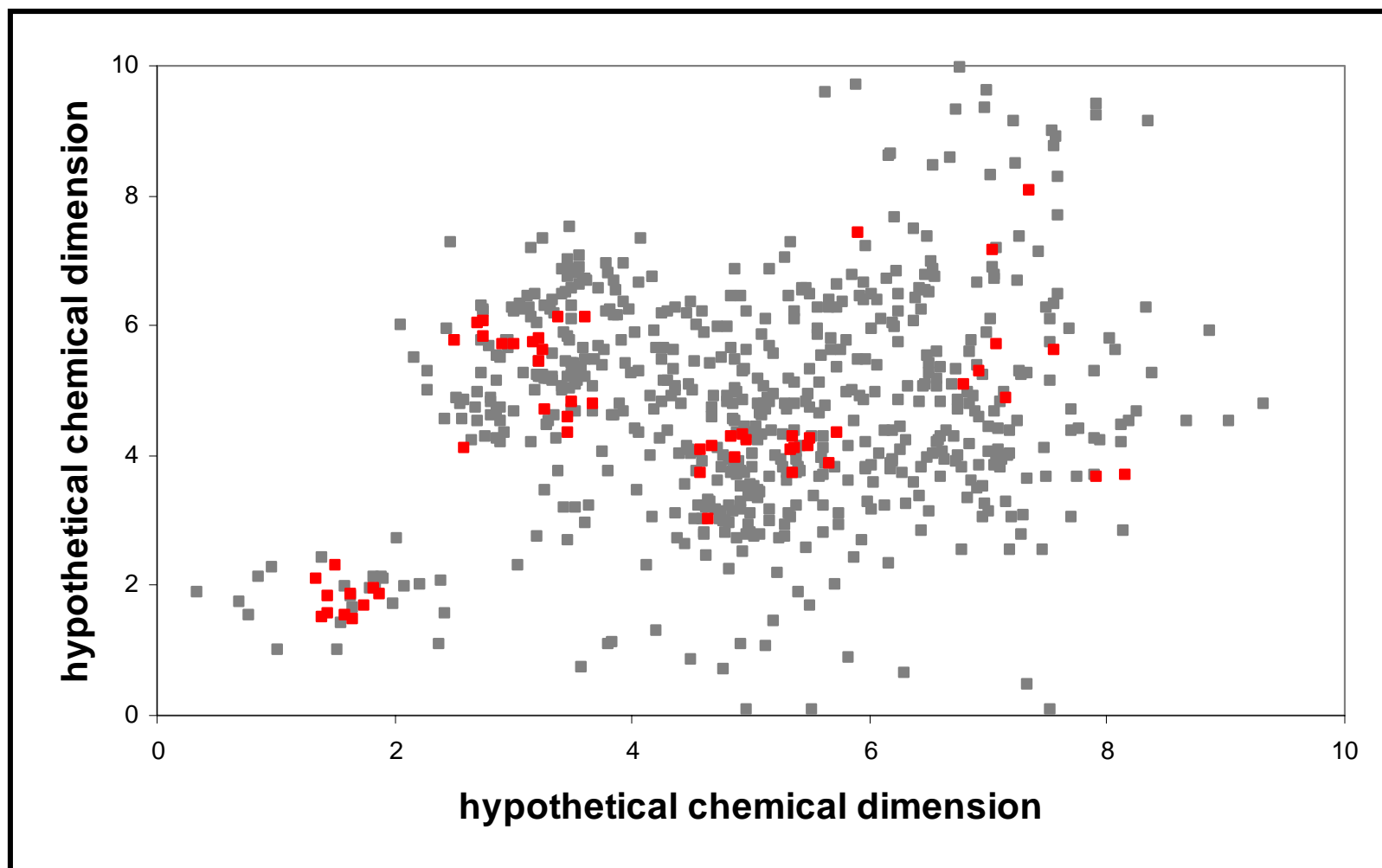
Selection of Genes from T Matrix

- Studentized range test with 7 cell clusters*
 - 476 genes with significant difference between 2 clusters
 - 391 genes (82%) involved melanoma or leukemia
- Calculate sub-[AT] matrix
- Select genes with high compound correlations

* Blower, et. al., *Pharmacogen. J.* **2002**, 2, 259–271

Structure-based Data Mining

Conceptual Activity Distribution





Structure-based Data Mining

Molecular Descriptors

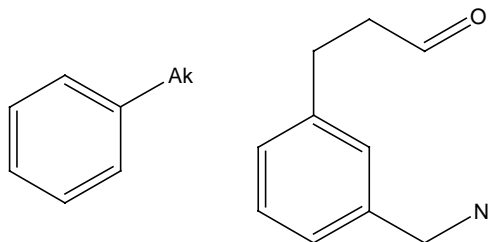
- Chemical data analyzed by common structural features
- Hierarchy comprises >27,000 structural features
- Major classes:

Amino acids
Bases, nucleosides
Benzenes
Carbocycles
Carbohydrates
Elements
Functional Groups

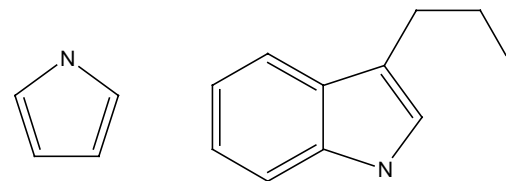
Heterocycles
Naphthalenes
Natural products
Peptidomimetics
Pharmacophores
Protective Groups
Spacer groups

Examples of Molecular Descriptors

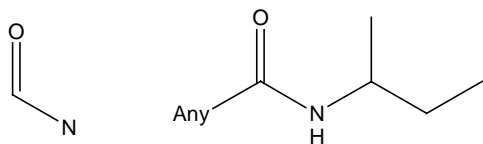
Benzenes



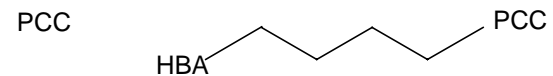
Heterocycles



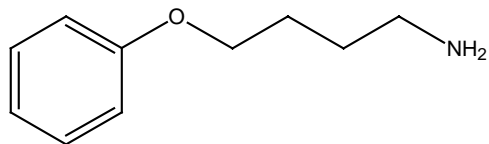
Functional Groups



Pharmacophores

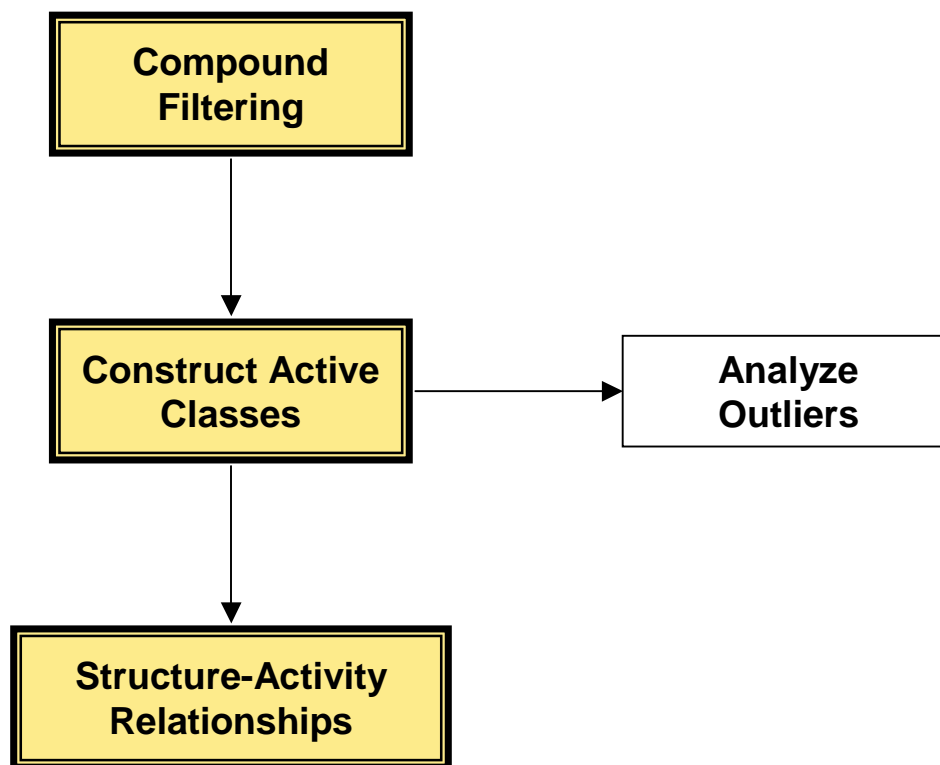


Spacer groups



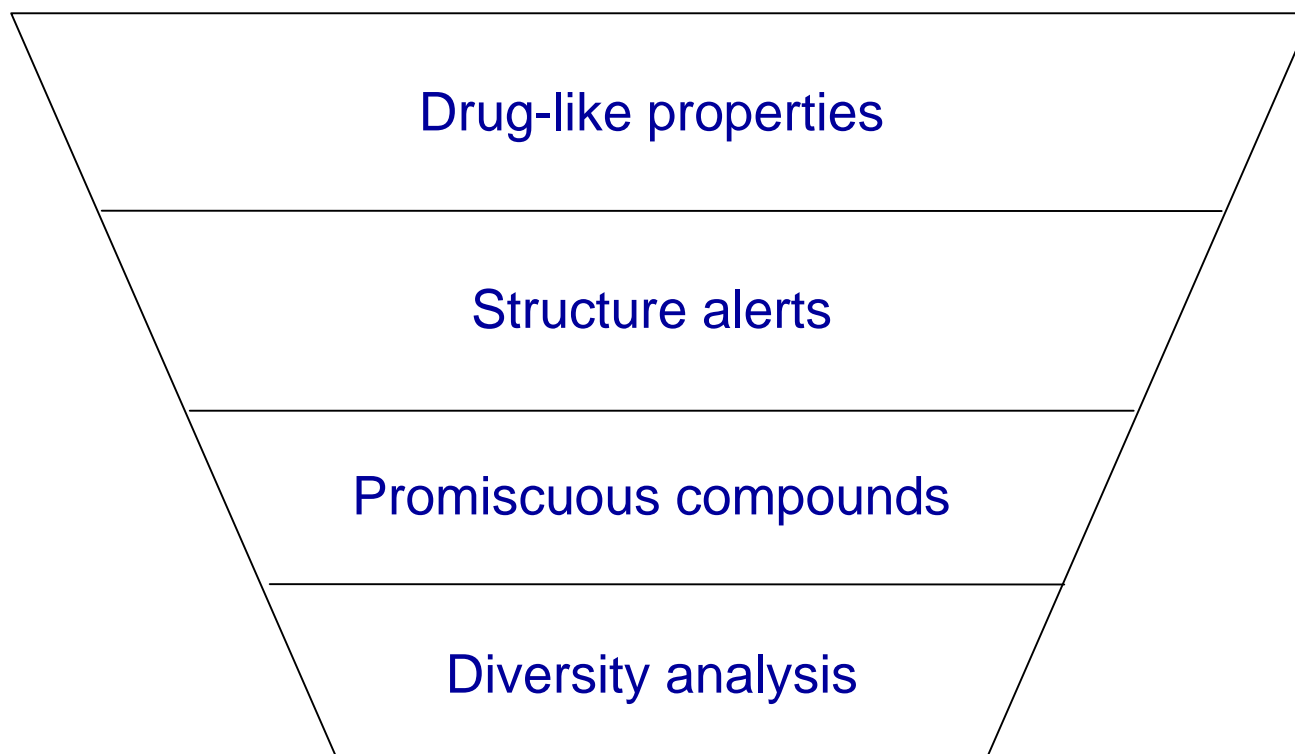
Structure-based Data Mining

General Strategy

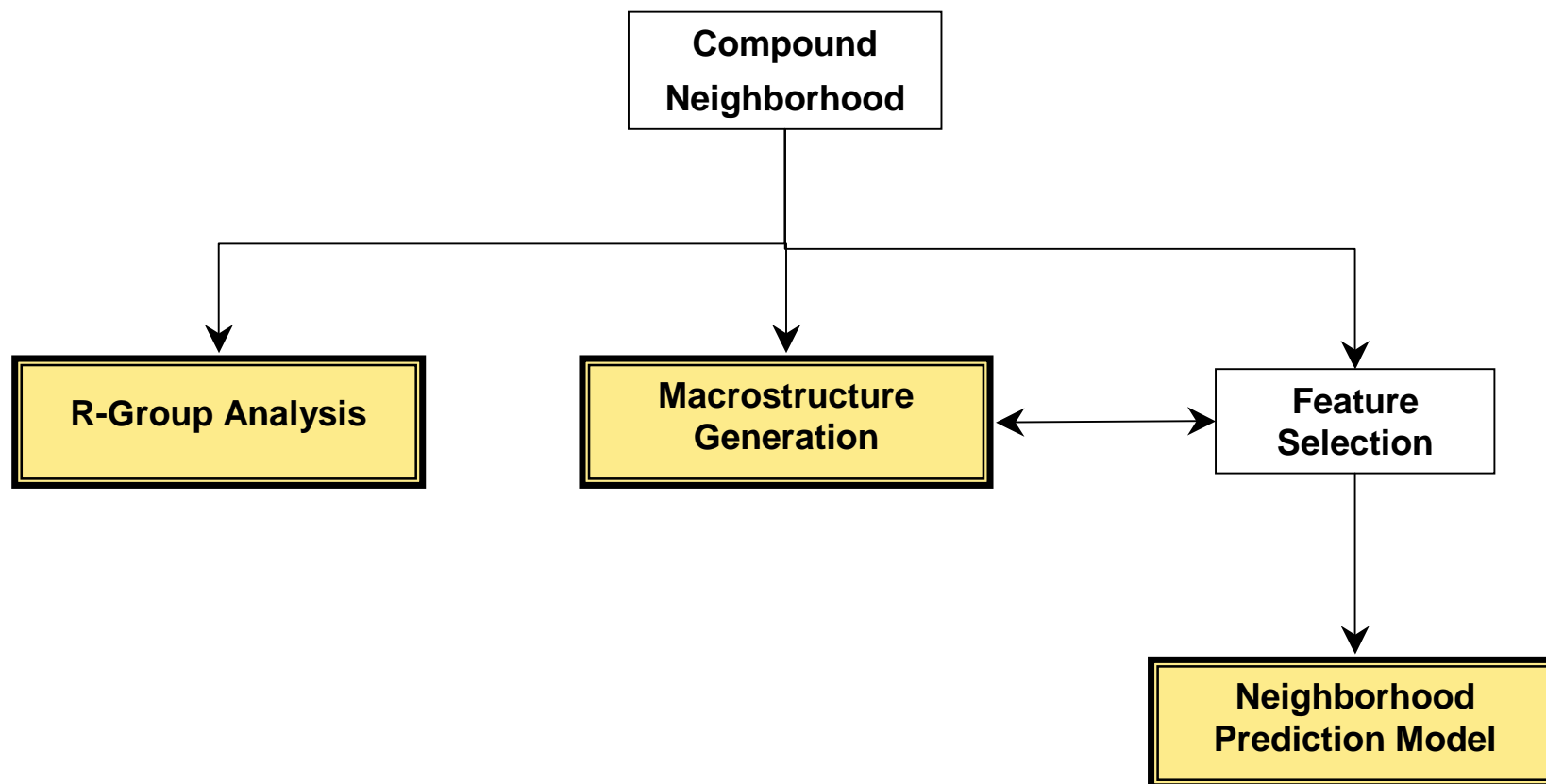




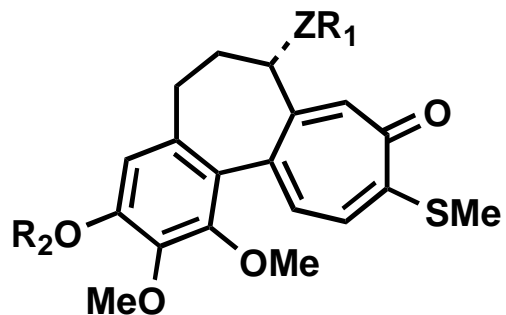
Compound Filtering



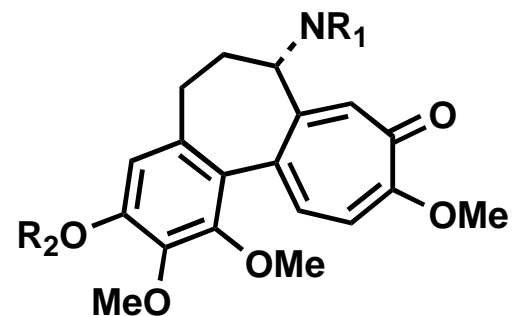
SAR Analysis



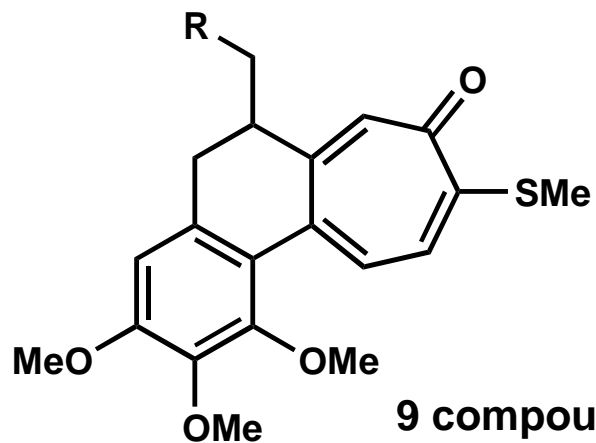
Colchicine Class



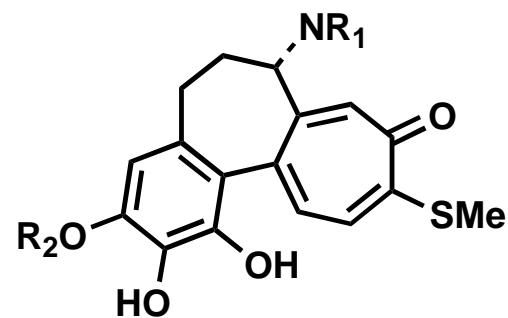
38 compounds
Ave pGI₅₀ = 7.74



23 compounds
Ave pGI₅₀ = 6.94

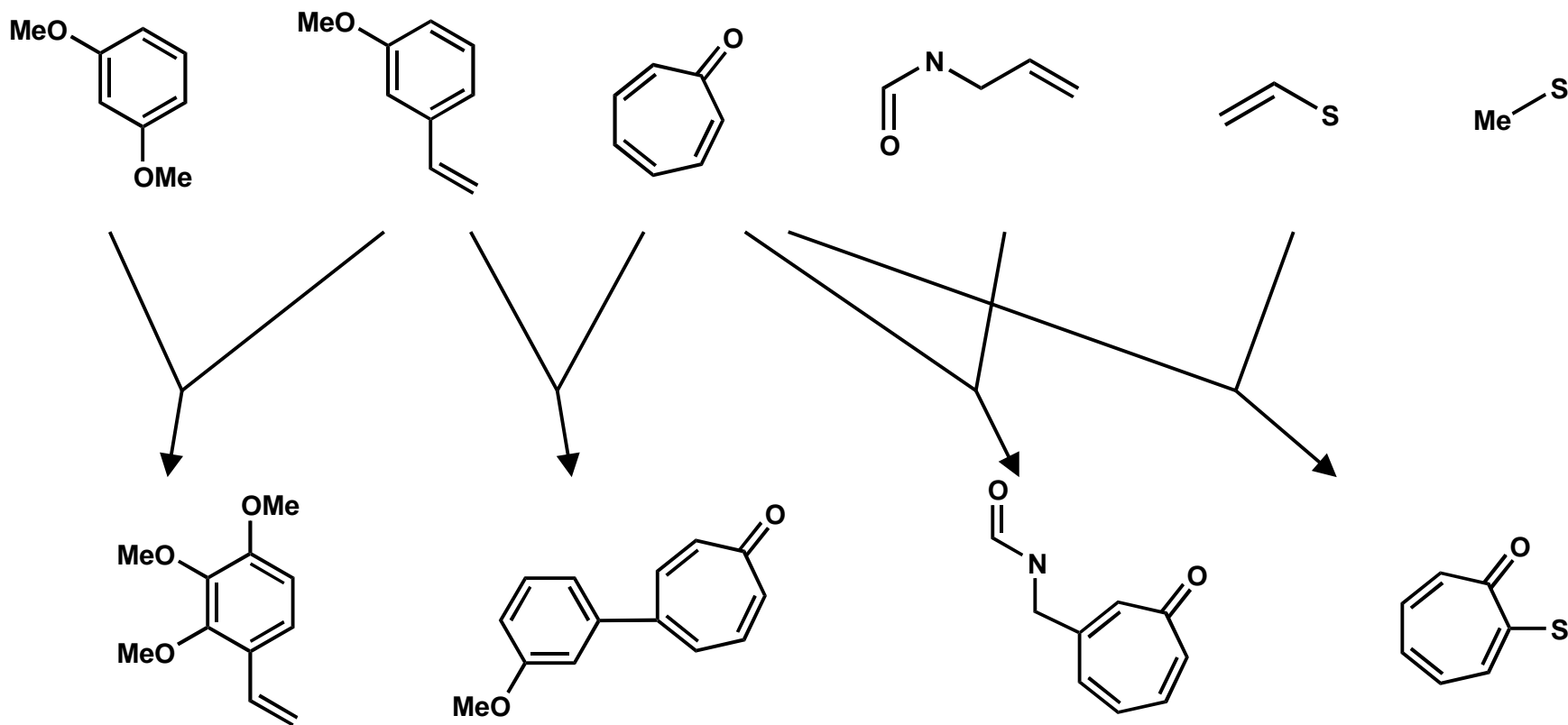


9 compounds
Ave pGI₅₀ = 6.96



17 compounds
Ave pGI₅₀ = 5.05

Assembling Macrostructures



R-Group Analysis

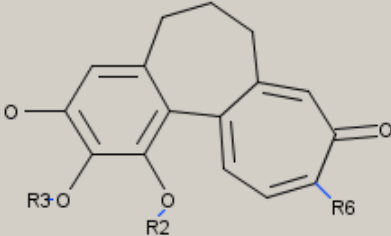
Project Browser

File Edit View Tools Window Help

Compare Structures Feature Combinations Clusters Scaffolds R-Groups Add to Hierarchy

R Group Analysis - R-Groups (Run #3)

Filters: colchicin... Choose...



Name	R2	R3	R6	Mean: colchicine 91	Z-Score: colchicine 91	Total
Result 2	Me	Me	SMe	7.92	5.6	41
Result 1	Me	Me	OMe	7.17	0.4	17
Result 11	H	H	SMe	4.72	-6.6	14
Result 4	Me	H	OMe	6.24	-0.8	2
Result 3	Me	Me	OH	4.97	-1.5	1
Result 5	Me	Me	H	7.22	0.1	1
Result 6	H	H	OMe	5.30	-1.2	1
Result 7	Me	H	SMe	7.78	0.5	1
Result 8	Ac	Ac	SMe	5.72	-0.9	1
Result 9	Me	Ac	SMe	6.41	-0.4	1
Result 10	Me	Ac	OMe	7.70	0.5	1

Change Substitution Points... Add Columns...

Color by: colchicine91 - L... << -3.0 -2.5 -2.0 2.0 2.5 3.0 >>

Ready Total:81 Mean:7.04 Std. Dev.:1.43 Filtered:81 Selected:0



Compound Dosing Experiments

- Gene expression experiments focused on breast cancer – 4 compounds x 3 cell lines
- Compounds: estradiol, 4-hydroxytamoxifen, isoflavanoid, doxorubicin
- Cell lines: MCF7 (breast, ER+), MDA mB231 (breast, ER-), HT29 (colon)
- NCI 23 cell lines for data mining
 - Removed leukemia and melanoma cell lines
 - Selected breast, ovarian, prostate, colon

Genes Regulated By Compound Dosing

Compound	MCF7		MDA mB231		MCF7 - MDA mB231	
	+	-	+	-	+	-
Estradiol	CDK5R1 DNMT3A	CTSS	CDC5L CYP3A4 ER2	GABRB3 CYP19	GABRE CDC25A CYP3A5	CTSS CDC5L
4-hydroxy tamoxifen	CDC2L5 ABCA2 GABRD	CYP2E1	CCND1 CDK10 CYP2E1	CDKL1	ABCA2 CYP3A5 CYP2E1	EGF GABRA5
Isoflavonoid	THBS2 CYP2E1	ABCC2	CDK10 BRCA2 CDC52	CDK4	ABCA2 DNMT3A CYP2E1	CDK10
Doxorubicin	ABCA2 CDK10 GABRE	PPP1R3A	TOP3A SIAT8C	COX5B	THBS2 GATA1	BRAC2



Summary

- Unique combinations of gene hierarchy, correlations, and informatics methods
- Rapidly identified compound classes with relevant genes using NCI dataset
- In-silico technique aids experimental design



Next Steps

- Experimental validation using Affymetrix GeneChip® (U133)
- Repeat the *in silico* analysis from reverse engineering perspective
- Develop SAR models for gene expression and compound structures



Acknowledgements

Ohio State Univ.

Michael Fligner
Joseph Verducci
Robert Brueggemeier
Jeanette Richardson

NCI

John Weinstein

LeadScope, Inc.

Chihae Yang
Kevin Cross
Glenn Myatt