A Subgroup Discovery Approach for Relating Chemical Structure and Phenotype Data in Chemical Genomics

Lan Umek[1], Petra Kaferle[2], Mojca Mattiazzi[2], Aleš Erjavec[1], Črtomir Gorup[1], Tomaž Curk[1], Uroš Petrovič[2] and Blaž Zupan[1,3]

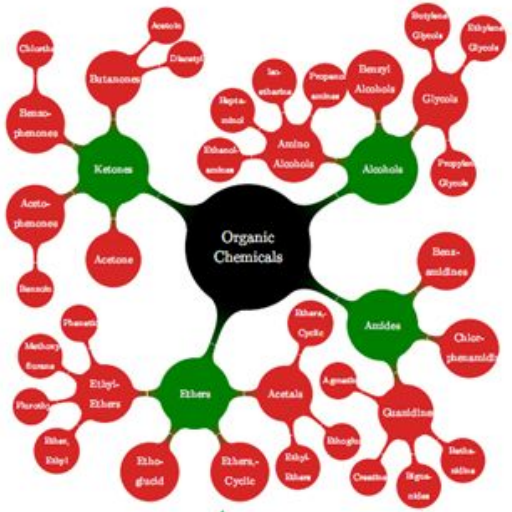
1 – University of Ljubljana, Faculty of Computer and Information Sciences
2 – Jožef Stefan Institute
3 – Department of Human and Molecular Genetics, Baylor College of Medicine, Houston, USA

Ljubljana, 6. 9. 2009

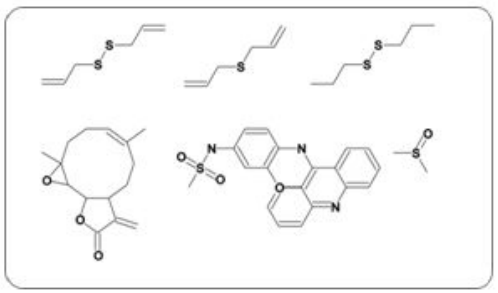
Motivation

- Relate chemical structure and *KEGG* (Kyoto Encyclopedia of Genes and Genomes) pathways

Chemical annotations :
pharmacological, structural

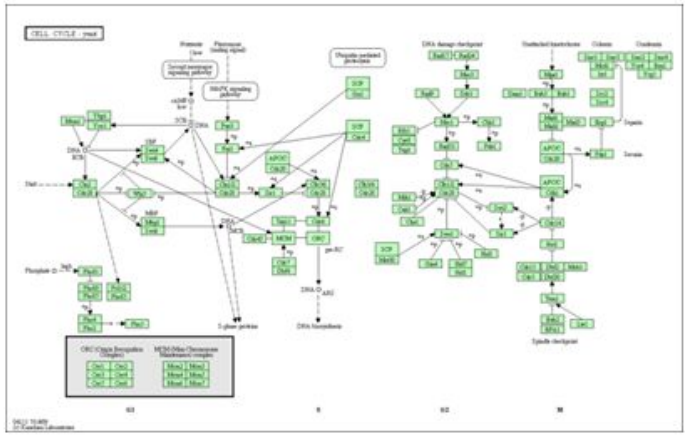


small molecules

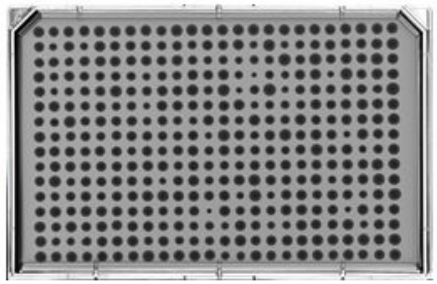


?

KEGG pathway (gene set)



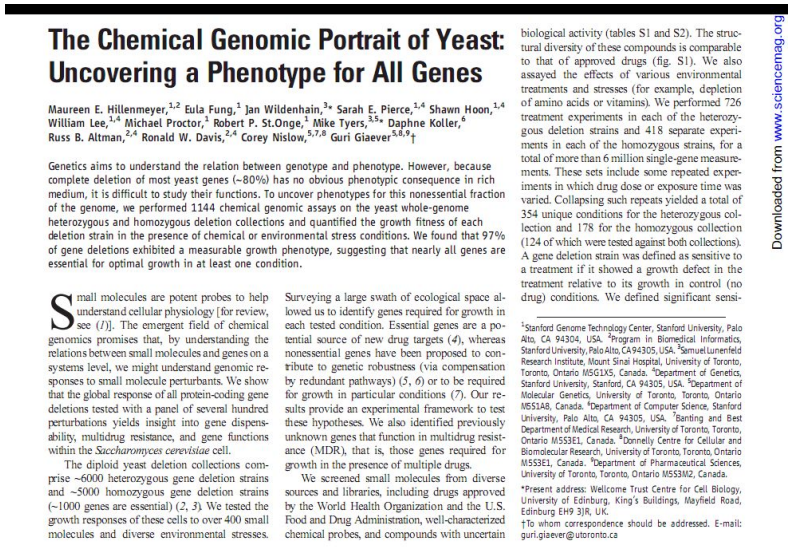
collection of single gene
knock-out strains (yeast)



chemogenomical experiment

Data set

- results of 71 chemogenomics experiments
- *phenotype profile* (4262 features): fitness (growth rate) of budding yeast *Saccharomyces cerevisiae* mutants when exposed to each of small molecules
- *drug characterization*: 126 structure-based features (obtained from *Dragon* software)



The Chemical Genomic Portrait of Yeast: Uncovering a Phenotype for All Genes

Maureen E. Hillenmeyer,[1,2] Eula Fung,[1] Jan Wildenhain,[3]* Sarah E. Pierce,[1,4] Shawn Hoon,[1,4] William Lee,[1,4] Michael Proctor,[1] Robert P. St.Onge,[1] Mike Tyers,[3,5]* Daphne Koller,[6] Russ B. Altman,[2,4] Ronald W. Davis,[2,4] Corey Nislow,[5,7,8] Guri Giaever[5,8,9]†

Genetics aims to understand the relation between genotype and phenotype. However, because complete deletion of most yeast genes (~80%) has no obvious phenotypic consequence in rich medium, it is difficult to study their functions. To uncover phenotypes for this nonessential fraction of the genome, we performed 1144 chemical genomic assays on the yeast whole-genome heterozygous and homozygous deletion collections and quantified the growth fitness of each deletion strain in the presence of chemical or environmental stress conditions. We found that 97% of gene deletions exhibited a measurable growth phenotype, suggesting that nearly all genes are essential for optimal growth in at least one condition.

Small molecules are potent probes to help understand cellular physiology [for review, see (*1*)]. The emergent field of chemical genomics promises that, by understanding the relations between small molecules and genes on a systems level, we might understand genomic responses to small molecule perturbants. We show that the global response of all protein-coding gene deletions tested with a panel of several hundred perturbations yields insight into gene dispensability, multidrug resistance, and gene functions within the *Saccharomyces cerevisiae* cell.

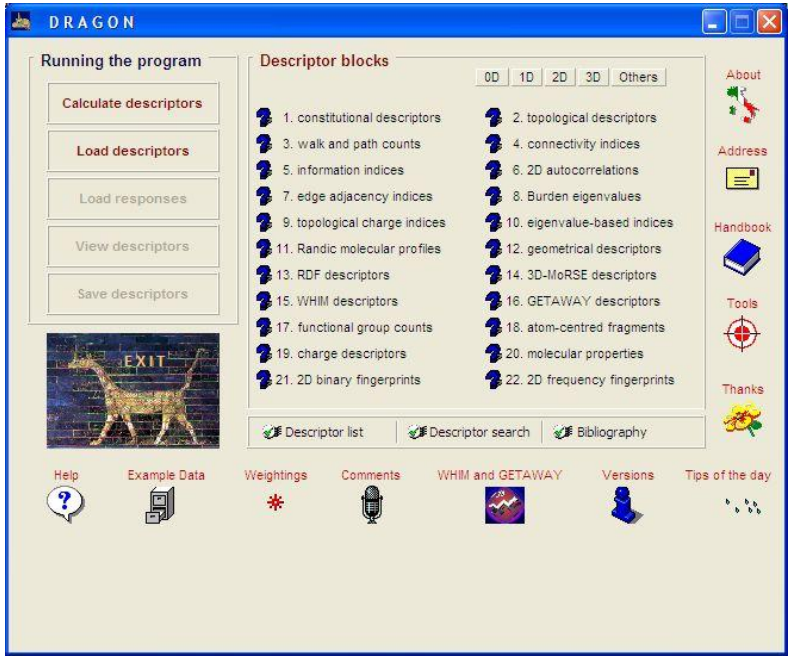
The diploid yeast deletion collections comprise ~6000 heterozygous gene deletion strains and ~5000 homozygous gene deletion strains (~1000 genes are essential) (*2*, *3*). We tested the growth responses of these cells to over 400 small molecules and diverse environmental stresses. Surveying a large swath of ecological space allowed us to identify genes required for growth in each tested condition. Essential genes are a potential source of new drug targets (*4*), whereas nonessential genes have been proposed to contribute to genetic robustness (via compensation by redundant pathways) (*5*, *6*) or to be required for growth in particular conditions (*7*). Our results provide an experimental framework to test these hypotheses. We also identified previously unknown genes that function in multidrug resistance (MDR), that is, those genes required for growth in the presence of multiple drugs.

We screened small molecules from diverse sources and libraries, including drugs approved by the World Health Organization and the U.S. Food and Drug Administration, well-characterized chemical probes, and compounds with uncertain biological activity (tables S1 and S2). The structural diversity of these compounds is comparable to that of approved drugs (fig. S1). We also assayed the effects of various environmental treatments and stresses (for example, depletion of amino acids or vitamins). We performed 726 treatment experiments in each of the heterozygous deletion strains and 418 separate experiments in each of the homozygous strains, for a total of more than 6 million single-gene measurements. These sets include some repeated experiments in which drug dose or exposure time was varied. Collapsing such repeats yielded a total of 354 unique conditions for the heterozygous collection and 178 for the homozygous collection (124 of which were tested against both collections). A gene deletion strain was defined as sensitive to a treatment if it showed a growth defect in the treatment relative to its growth in control (no drug) conditions. We defined significant sensi-

[1]Stanford Genome Technology Center, Stanford University, Palo Alto, CA 94304, USA. [2]Program in Biomedical Informatics, Stanford University, Palo Alto, CA 94305, USA. [3]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, University of Toronto, Toronto, Ontario M5G1X5, Canada. [4]Department of Genetics, Stanford University, Stanford, CA 94305, USA. [5]Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5S1A8, Canada. [6]Department of Computer Science, Stanford University, Palo Alto, CA 94305, USA. [7]Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario M5S3E1, Canada. [8]Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario M5S3E1, Canada. [9]Department of Pharmaceutical Sciences, University of Toronto, Toronto, Ontario M5S3M2, Canada.

*Present address: Wellcome Trust Centre for Cell Biology, University of Edinburg, King's Buildings, Mayfield Road, Edinburg EH9 3JR, UK.

†To whom correspondence should be addressed. E-mail: guri.giaever@utoronto.ca



Sample from the data

small molecule	Input features Structure of small molecule			Output features Mutant-based fitness		
	# of N atoms	# of S atoms	Partition coefficient	YGL234W	YCL010C	YPL212C
mycophenolic acid	0	0	2.60	?	?	-0.97
BCNU	3	0	0.80	0.23	0.26	-0.23
rotenone	0	0	1.90	1.40	0.16	-0.35
papuamide B	13	0	-6.10	-0.45	0.15	-0.21

Problem

- find interesting subgroups of experiments=(small molecules, phenotype profiles) where
  - experiments in the subgroup have similar phenotype profile in some specific subset of mutants (*KEGG* pathway),
  - the small molecules in the subgroup can be reliably discriminated from other small molecules in the data set using *structural descriptors*
- relate chemical structure and phenotype profile

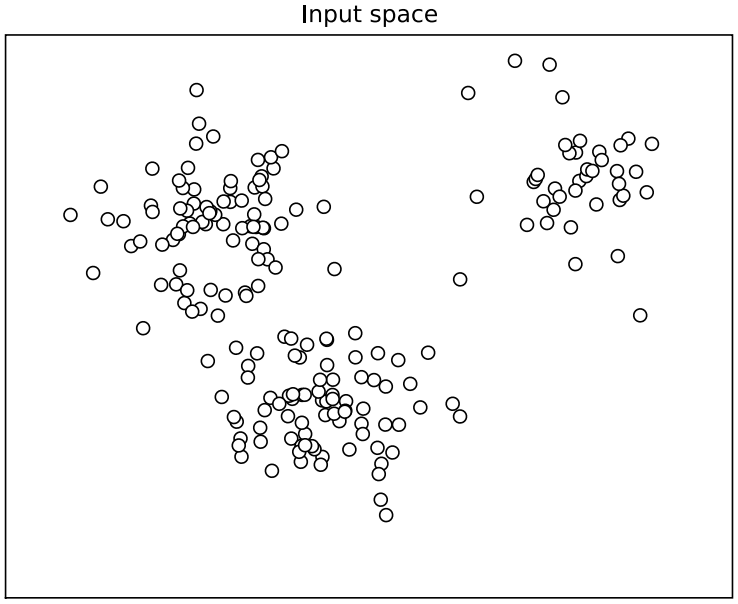
Related work

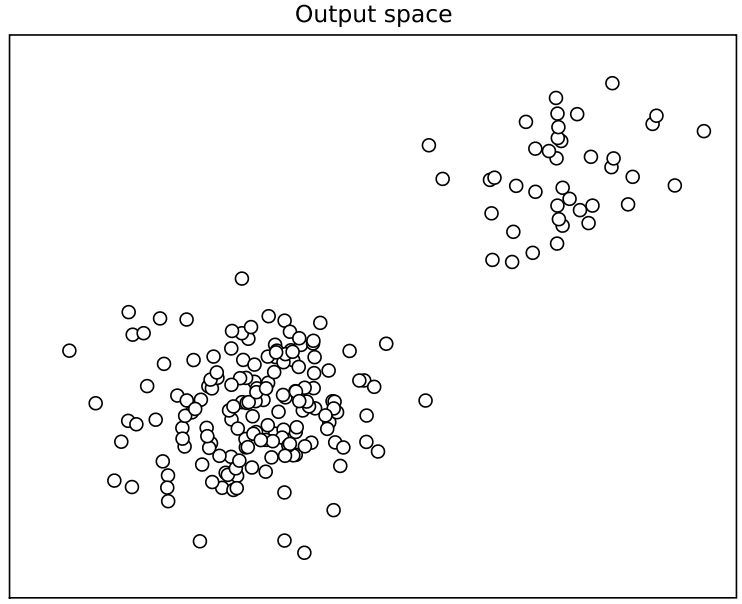
- *subgroup discovery* (one discrete outcome variable)
  - W. Klősgen: Applications and Research Problems of Subgroup Mining (1999)
  - D. Gamberger, N. Lavrač: Expert–Guided Subgroup Discovery (2002)
  - E. Suzuki: Undirected Exception Rule Discovery as Local Pattern Detection (2004)
  - B. Kavšsek: APRIORI-SD: Adapting Association Rule Learning to Subgroup Discovery (2006)
  - A. Knobbe: Exceptional Model Mining (2008)
  - X. Su: Subgroup Analysis via Recursive Partitioning (2009)
- *multilabel prediction* (prediction of several discrete variables)
  - H. Wold: Partial Least Squares Regression (1985)
  - H. Blockeel: Top-down Induction of First Order Logical Decision Trees (1998)
  - B. Ženko: Learning Predictive Clustering Rules (2007)

Our approach

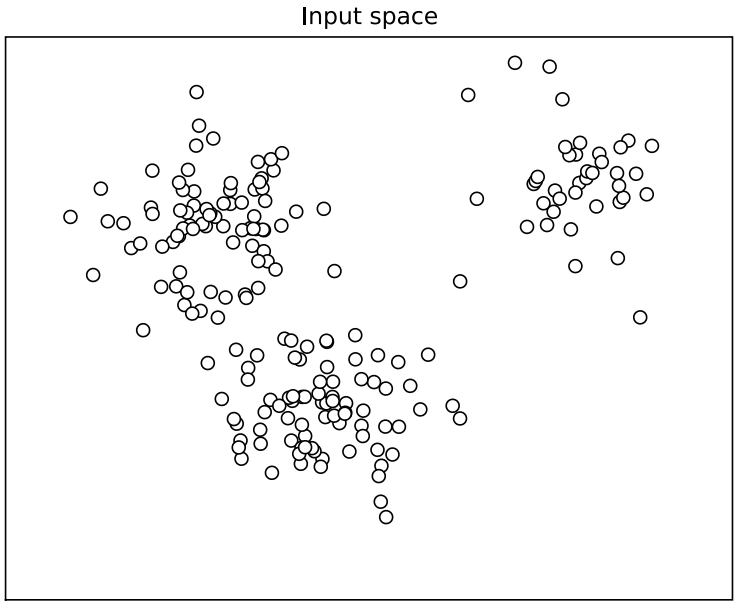
- an extension of the original subgroup discovery task
- assumes several outcome variables (of mixed types)
- does not seek for general prediction classification model (multilabel prediction)
- integrates information from several data-bases (*KEGG*, *MeSH*,. . .)

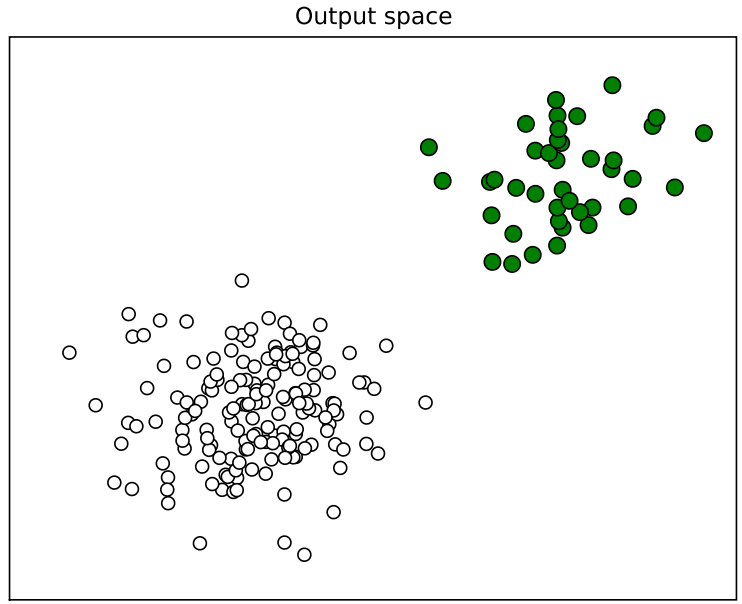
Example



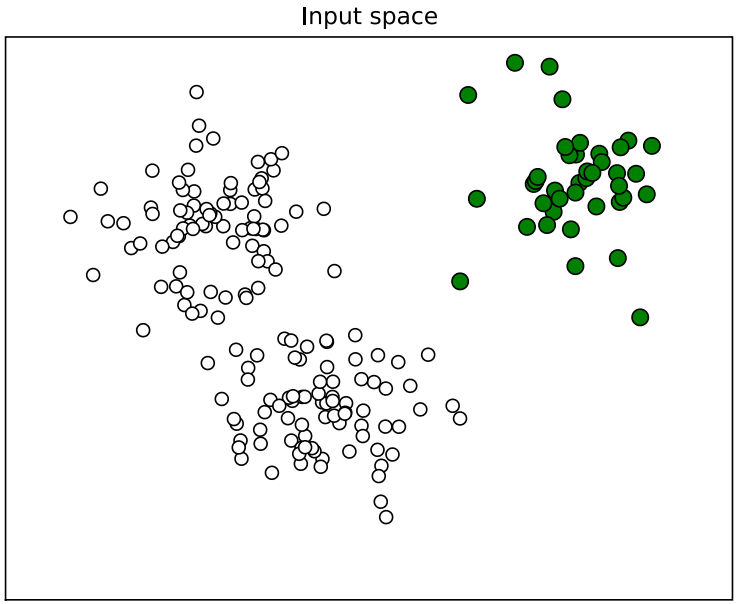


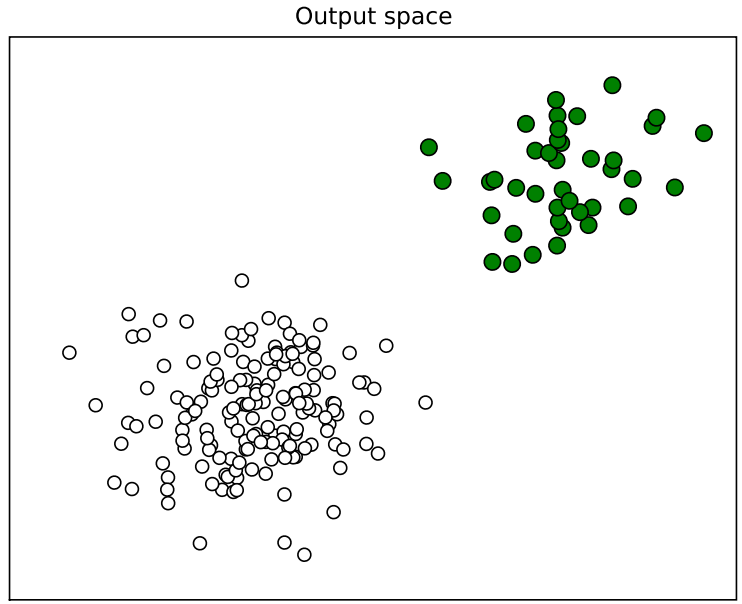
Example



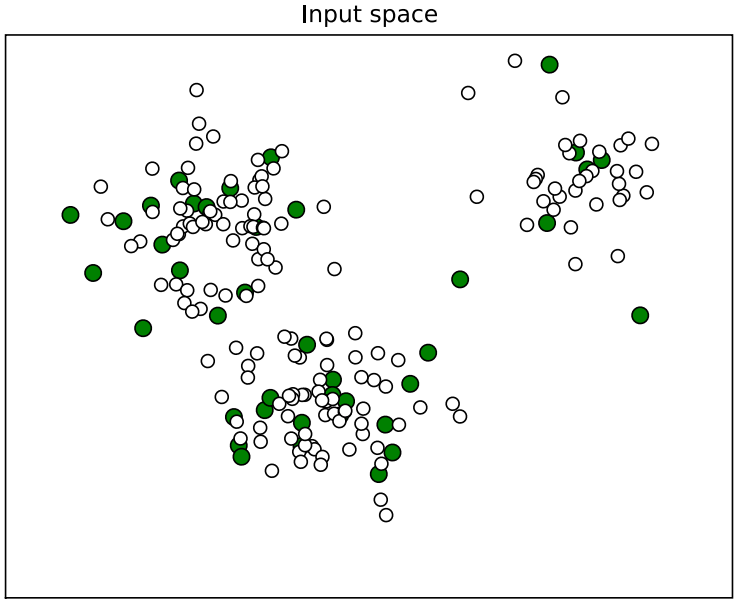


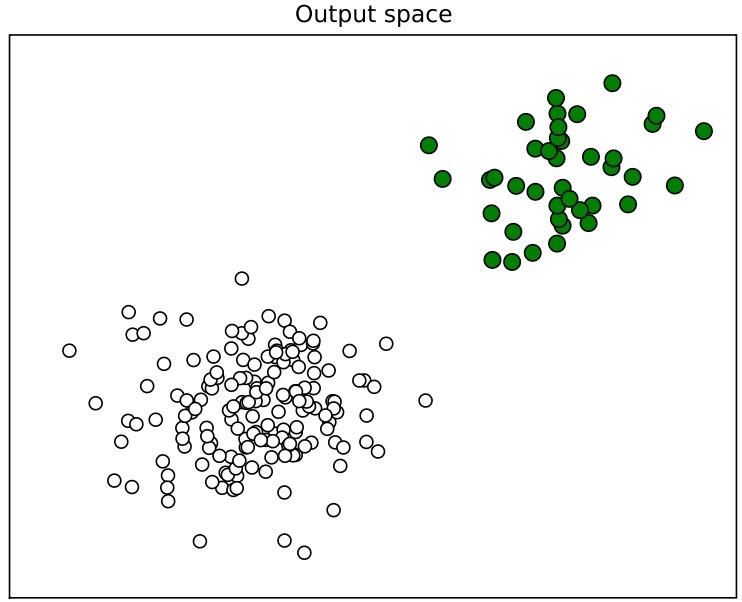
Example



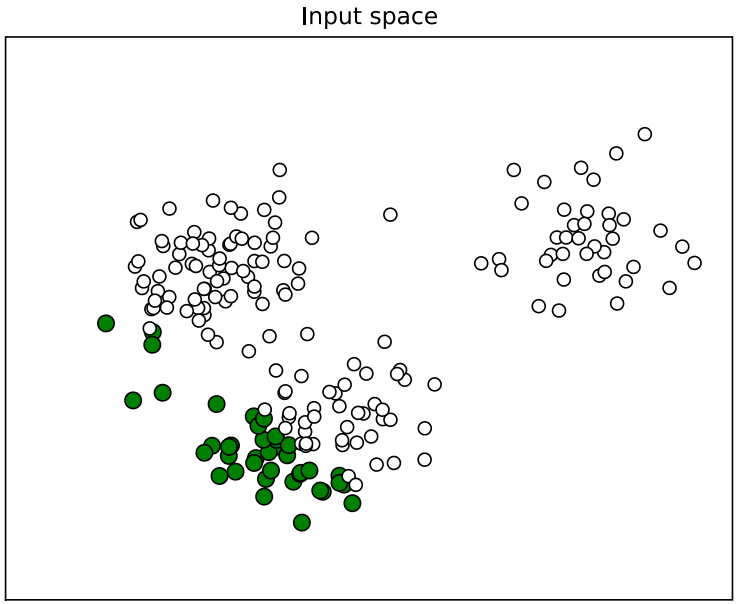


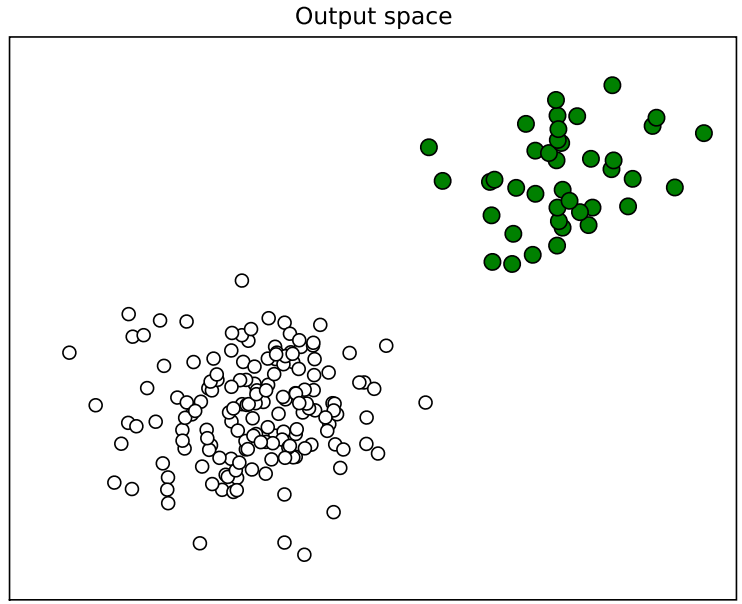
Example



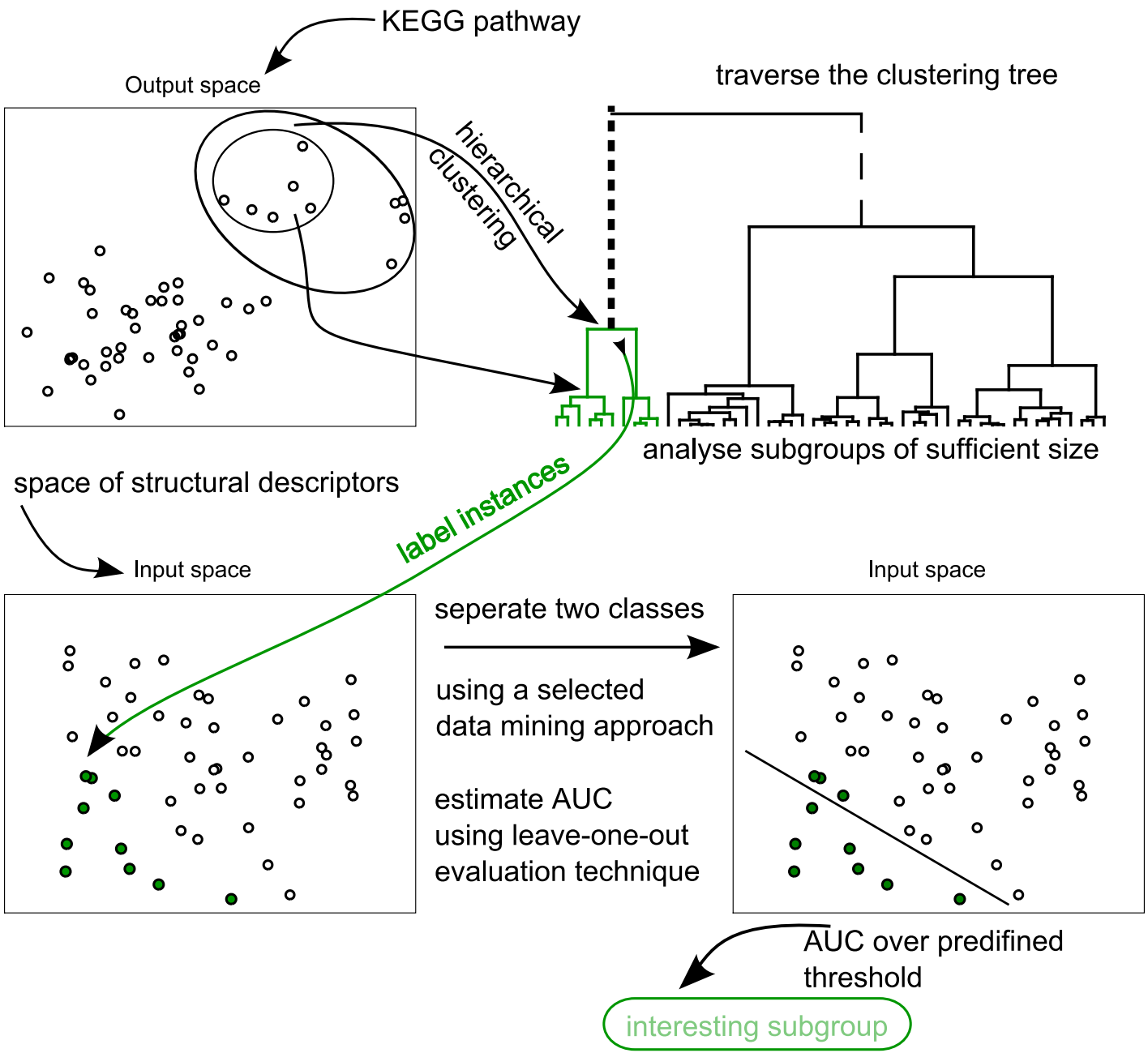


Example





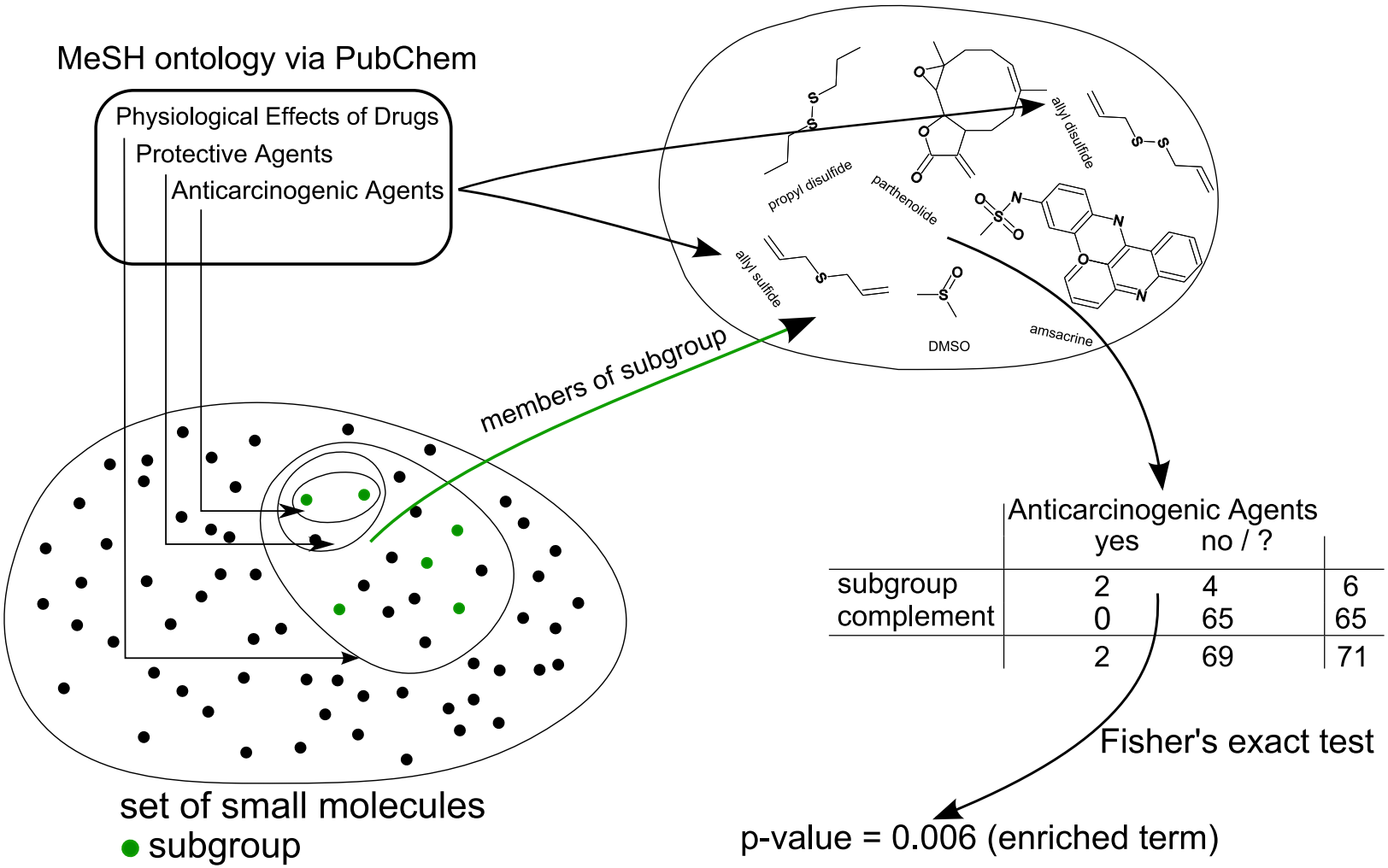
Search algorithm



Dissimilarity measure and classifiers

- clustering
  - weighted *Manhattan metric* for clustering in the space of a selected *KEGG* pathway
  - 98 different (*KEGG* pathways) were used (covering 760 genes in total)
  - Ward's linkage
- supervised data mining approach
  - support vector machines with linear kernel

Enriched MeSH (Medical Subject Headings) terms



Example of the result

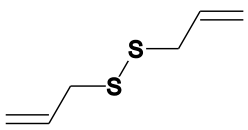
- subgroup of six small molecules ($AUC = 0.76$)
- KEGG pathway: cell-cycle

Enriched chemical terms:

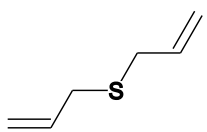
- disulfides and allyl compounds ($p = 0.0048$)

Enriched pharmacological action terms:

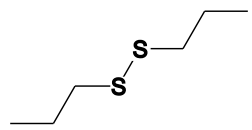
- anticarcinogenic agents ($p = 0.0055$)
- protective agents ($p = 0.0161$)



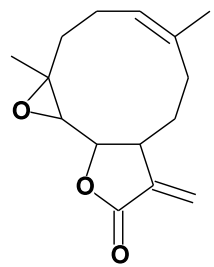
allyl disulfide



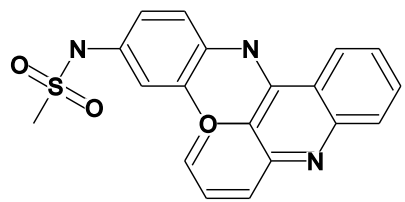
allyl sulfide



propyl disulfide



parthenolide



amsacrine

DMSO

Overview of the results

size	AUC	pathway	chemical classification	pharmacological classification
5	0.855	nitrogen metabolism	sulfur compounds	myeloablative agonists
				toxic actions
5	0.855	ubiquinone biosynthesis	hydrocarbons, halogenated,	antineoplastic agents
			nitrogen mustard compounds	alkylating
7	0.819	biosynthesis of steroids	disulfides	none
5	0.782	drug metabolism		
		other enzymes	urea	none
5	0.779	alanine and aspartate		
		metabolism	disulfides	none
6	0.756	cell cycle - yeast	disulfides, allyl compounds	protective,
				anticarcinogenic agents
6	0.756	folate biosynthesis	azirines, sulfur compounds	antineoplastic,
				alkylating agents
8	0.752	one carbon pool	allyl compounds	protective, antineoplastic,
		by folate		anticarcinogenic agents

Conclusions

- subgroup discovery method
  - requires data instances with two-sets of descriptors
  - suitable for applications with data–rich domain
- demonstration of utility on a problem from chemical genomics
  - identification of subgroups of small molecules with similar effects on known gene sets (mutant-based phenotypes)

Ongoing work

Problems:

- small data sets
- selection of the small molecules

Ongoing work:

- comparison of the results with different approaches
- automated rating of hypothesis interestingness (PubMed)