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

Lan Umek¹, Petra Kaferle², Mojca Mattiazzi², Aleš Erjavec¹, Črtomir Gorup¹, Tomaž Curk¹, Uroš Petrovič² and Blaž Zupan¹,³

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

KEGG pathway (gene set)

small molecules

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)
Sample from the data

<table>
<thead>
<tr>
<th>small molecule</th>
<th>Input features</th>
<th>Output features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Structure of small molecule</td>
<td>Mutant-based fitness</td>
</tr>
<tr>
<td></td>
<td># of N atoms</td>
<td># of S atoms</td>
</tr>
<tr>
<td>mycophenolic acid</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>BCNU</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>rotenone</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>papuamide B</td>
<td>13</td>
<td>0</td>
</tr>
</tbody>
</table>
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. Klösgen: Applications and Research Problems of Subgroup Mining (1999)
  - X. Su: Subgroup Analysis via Recursive Partitioning (2009)

- multilabel prediction (prediction of several discrete variables)
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, ...)

Lan Umek et al. Subgroup Discovery for Relating Chemical Structure and Phenotype Data
Example
Example
Example
Example
Example

Input space

Output space
Search algorithm

- **Input space**
  - **space of structural descriptors**
  - **Output space**
  - **KEGG pathway**
  - **Hierarchical clustering**

- **traverse the clustering tree**
- **analyse subgroups of sufficient size**

- **Input space**
  - **separate two classes**
  - **using a selected data mining approach**
  - **estimate AUC using leave-one-out evaluation technique**

- **AUC over predefined threshold**
- **interesting subgroup**
- **label instances**
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

MeSH ontology via PubChem

Physiological Effects of Drugs
Protective Agents
Anticarcinogenic Agents

members of subgroup

set of small molecules
● subgroup

Anticarcinogenic Agents

<table>
<thead>
<tr>
<th>subgroup</th>
<th>yes</th>
<th>no / ?</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>complement</td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>65</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>69</td>
<td>71</td>
</tr>
</tbody>
</table>

Fisher's exact test

p-value = 0.006 (enriched term)
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

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