JMLR: Workshop and Conference Proceedings 8: 136-144 Machine Learning in Systems Biology

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

Lan Umek LAN.UMEK@FRI.UNI-LJ.SI
Faculty of Computer and Information Science
University of Ljubljana, Slovenia

Petra Kaferle PETRA.KAFERLE@IJS.SI
Jožef Stefan Institute
Ljubljana, Slovenia

Mojca Mattiazzi MOJCA.MATTIAZZI@IJS.SI
Jožef Stefan Institute
Ljubljana, Slovenia

Aleš Erjavec ALES.ERJAVEC@FRI.UNI-LJ.SI
Faculty of Computer and Information Science
University of Ljubljana, Slovenia

Črtomir Gorup CRT.GORUP@GMAIL.COM
Faculty of Computer and Information Science
University of Ljubljana, Slovenia

Tomaž Curk TOMAZ.CURK@FRI.UNI-LJ.SI
Faculty of Computer and Information Science
University of Ljubljana, Slovenia

Uroš Petrovič PETRA.KAFERLE@IJS.SI
Jožef Stefan Institute
Ljubljana, Slovenia

Blaž Zupan BLAZ.ZUPAN@FRI.UNI-LJ.SI
Faculty of Computer and Information Science
University of Ljubljana, Slovenia and
Dept. of Human and Mol. Genetics,
Baylor College of Medicine, Houston, USA

Editor: Sašo Džeroski, Pierre Geurts, and Juho Rousu

Abstract

We report on development of an algorithm that can infer relations between the chemical structure and biochemical pathways from mutant-based growth fitness characterizations of small molecules. Identification of such relations is very important in drug discovery and development from the perspective of argument-based selection of candidate molecules in target-specific screenings, and early exclusion of substances with highly probable undesired side-effects. The algorithm uses a combination of unsupervised and supervised machine learning techniques, and besides experimental fitness data uses knowledge on gene subgroups (pathways), structural descriptions of chemicals, and MeSH term-based chemical and pharmacological annotations. We demonstrate the utility of the proposed approach

in the analysis of a genome-wide *S. cerevisiae* chemogenomics assay by Hillenmeyer *et al.* (Science, 2008).

Keywords: chemical genomics, structure-activity relationship, subgroup discovery, hierarchical clustering, supervised learning, MeSH term enrichment

1. Introduction

One of the promises of the post-genomics era was the identification of novel drug targets and design of more efficient and specific drugs with fewer side-effects. In reality, pipelines of pharmaceutical companies did not improve much due to genomic data alone. One of the main reasons for that is the lack of methods to combine characteristics of potential drug molecules with rich genomic data. Such data comes in many flavors: from raw genome sequence data, phenotypic profiles such as gene expression profiles, functional and physical interactions of genes and proteins, to rich annotations of genes and their products by complex ontologies. These together define phenomes (*i.e.*, genome-wide phenotypes) of a cell or an organism.

Of special interest for the identification and characterization of potential drug molecules are recently developed chemogenomic approaches. These profiles allow to measure changes in the phenome that were caused by the molecule's activity. When applied to a collection of mutants, we gain a data set with a vast potential for the generation of chemogenomics hypotheses. One such data set was recently reported by Hillenmeyer *et al.* (Hillenmeyer et al., 2008), where growth fitness in the presence of a number of chemicals was observed in a set of genome-wide single-gene deletion mutants. The authors used yeast *S. cerevisiae* as a model organism, and reported that a surprisingly large proportion (97%) of gene deletions exhibited a measurable growth phenotype.

The aim of the research reported here was to see if the data set published by Hillenmeyer *et al.* (Hillenmeyer et al., 2008) could be used further to relate genetic pathways with structural and pharmacological properties of drugs. We extended the information from fitness data by associating chemicals with their structural descriptions, and mined subsets of mutants that stem from single-deletions of genes common to a specific pathway. Our effort, in which we queried a number of data bases to complement experimental results and to allow for further analysis, could be enlisted under *integrative bioinformatics* or *integrative systems biology*. These emerging fields strive to relate a plethora of existing molecular biology data bases and experimental repositories (Hoon et al., 2008).

To serve our aim, we developed a specific data mining approach. In particular, we used a combination of unsupervised learning (clustering) to find groups of chemicals with similar mutant-based fitness profiles, and supervised learning to check if discovered groups of chemicals can be characterized in terms of common chemical structure. The proposed search algorithm evaluates such hypotheses across a number of genetic pathways and tests a variety of plausible subgroups of chemicals. While the particular approach is new and for the first time described in this report, it in part resembles rule-based subgroup discovery techniques (Lavrač et al., 2004; Ženko and Struyf, 2005) and bi-clustering approaches (Van Mechelen et al., 2004). From the former, we borrow the idea of finding subsets of characteristic data items, and from the latter the idea that items have to be

similar in two different aspects, in our case, in structure of the chemicals and corresponding phenotype response.

The paper proceeds with the description of the data set used in our experiments, and of preprocessing (data selection) steps. We continue with a detailed description of the algorithm, experimental results and a discussion.

2. Data

In early 2008, Hillenmeyer *et al.* published a comprehensive analysis that included 1144 chemical genomic assays on the yeast whole-genome heterozygous and homozygous gene deletion collections and quantified the growth fitness of each deletion strain in the presence of chemical or environmental stress conditions (Hillenmeyer et al., 2008). This study generated the first available data set based on which systematic analysis of functional relations between biochemical pathways and chemical structure is possible.

In the analysis reported here we focussed on experiments on homozygous strains. From the initial set of 418 genome-wide screens, we removed experiments with environmental stress, irradiated drugs, inorganic compounds, platinum compounds, norcantharidin, cantharidin analog and cantharidin disodium, with the aim to focus on the chemical space of organic substances. We also discarded assays for which the molecular formula was unavailable according to the supplementary data (Hillenmeyer et al., 2008), assays that used mixtures of two chemicals and experiments where time of growth was different than 20 generations.

From the remaining 136 assays, the filtering of assays with the same small molecule at different concentrations was based on manual inspection of graphs of quantile functions of fitness values (Figure 1). From such sets of experiments, we selected the one with the sharpest transition in the related graph. In almost all cases the lowest concentration was selected. If two assays used the same chemical at the same concentration, only the first assay listed in the data was used. The resulting set included 74 assays.

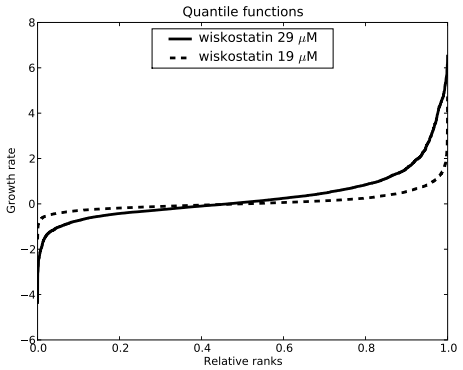

Figure 1: Quantile function of mutant growth fitness for wiskostatin is shown. Here, the experiment concentration of $19\mu M$ was selected due to sharper quantile function.

We used NCBI's PubChem (`http://pubchem.ncbi.nlm.nih.gov/`) to obtain SMILES structural descriptors (Weininger, 1988). We failed to do so for three chemicals in the collection, and thus proceeded with 71 chemicals and their related assays. SMILES descriptors were converted to an array of molecular descriptions (constitutional and topological descriptors, molecular properties, connectivity indices, atom-centred fragments, functional group counts) using the Dragon (Talete srl, 2007) software. We removed constant, near-constant, and highly correlated (correlation exceeded 0.9) descriptors and drug-like indices (like Ghose-Viswanadhan-Wendoloski index). For the analysis, we used 126 molecular descriptors.

The resulting 71 assays, each corresponding to an application of a specific small molecule, included growth fitness measurement of 4262 single-mutants. Of these, we have removed 507 mutants with missing growth fitness values. The above preprocessing thus produced a data matrix that included the growth fitness scores for 3755 single gene deletion mutants in the presence of 71 different assays.

3. Methods

We will assume that our data is a random sample $e_1 = (X_1, Y_1), \ldots, e_n = (X_n, Y_n)$, $(n = 71)$ where each pair $e_i = (X_i, Y_i)$ represents a chemogenomical experiment. Each experiment consists of two vectors: X_i is a set of DRAGON-based structural descriptors of the i-th chemical used in the assay, and Y_i is a resulting vector of phenotype responses, consisting of growth fitness scores for 3755 single gene deletion budding yeast mutant.

We here propose a method that aims to relate chemical structures of the small molecules involved in experiments with their characteristic phenotypic profile. In particular, we are looking for subgroups of experiments (chemicals) where:

- experiments in the subgroup have similar phenotypic profile in some specific subsets of mutants,
- the set of chemicals in the subgroup can be reliably discriminated from other chemicals in the data set using DRAGON-based structural descriptions.

The subsets of mutants were identified based on the annotation of a gene to a specific KEGG pathway (Kanehisa et al., 2007). We have only used pathways that include more than two mutated genes. As of April 2009, there were 98 such pathways, covering 760 genes in total.

We have applied a specific search algorithm that uses unsupervised learning to find subgroups of chemicals with similar gene set-based phenotypic profile, and supervised learning to identify those subgroups which can be successfully characterized by the set of chemical structure descriptors. The final step of the analysis is a MeSH term enrichment-based characterization of resulting subsets of chemicals. Both steps, the search algorithm and chemical characterization of the subsets are described below.

3.1 Search Algorithm

The algorithm searches for characterizable sets of chemicals that resulted in similar phenotypic profiles for a subset of mutants. The algorithm is executed all gene sets (one KEGG pathway represents one gene set), and includes the following steps:

1. Choose a subset of phenotypic features (genes from a specific KEGG pathway) GS and define the dissimilarity measure δ_{GS} between two experiments e_i, e_j (phenotypic profiles) using a weighted Manhattan metric:

$$\delta_{GS}(e_i, e_j) = \sum_{k \in GS} \frac{|Y_{ik} - Y_{jk}|}{\max_l Y_{lk} - \min_l Y_{lk}} \tag{1}$$

   where Y_{ik} represents k–th component of random vector Y_i.

2. Perform hierarchical clustering (Kaufman and Rousseeuw, 1990) of the experiments with δ_{GS} using Ward's minimum-variance linkage (Ward, 1963).

3. Traverse the resulting dendrogram to identify various candidates for subgroups. Consider only subgroups consisting of at least $min_{size} = 4$ chemicals and no bigger than $max_{size} = 10$.

4. For all subgroups (of chemicals) identified in the previous step, estimate the degree of separability from the rest of the chemicals in the data set. For each subgroup, we first classify chemicals based on their membership in the subgroup. We then perform leave-one-out to estimate the accuracy of support vector machine (SVM)-based class-prediction. SVM is presented with DRAGON-based chemical structure descriptors and the classification in the current subgroup. Area under ROC (AUC) is used to measure the predictive accuracy. SVM with linear kernel as implemented in SVMlight library (Linear Learner with default parameters) (Fan et al., 2008) was used in our experiments. Subgroups with AUC equal to 0.75 or above are retained and reported to the user.

3.2 Characterization of Subgroups

The discovered subgroups include a set of chemicals which share a similar phenotype response in a KEGG pathway-specific subset of mutated genes. Each reported subset of mutated genes is therefore characterized by the name of their respective KEGG pathway. We also need a simple, readable characterization of chemicals in the subgroup. For this, we have used terms from the *chemical classification* and *pharmacological classification* part of Medical Subject Headings (MeSH) ontology. Annotations of chemicals with MeSH terms were retrieved from NCBI's PubChem (`http://pubchem.ncbi.nlm.nih.gov/`). We then used enrichment analysis to find terms characteristic for a given subset of chemicals. Given all chemicals annotated to term t and chemicals in subgroup G we test if there exists a relationship between membership of the subgroup G and term t using *Fisher's exact test*. The p–values from Fisher's test are then used for ranking annotated terms. We report terms with the associated p–value less than 0.05.

3.3 Implementation

The proposed method was developed in Python within the Orange data mining framework (Demšar et al., 2004), which implements unsupervised and supervised techniques, leave-one-out evaluation and ROC analysis. Orange Bioinformatics toolbox (Curk et al., 2005) was used to access KEGG pathways, obtain MeSH terms and chemical annotations, and perform enrichment analysis of MeSH terms.

4. Results

Our algorithm discovered 25 subgroups. Eleven of them resulted in at least one enriched classification term (pharmacological or chemical), eight of them (for which all terms were annotated to at least 2 small molecules) are presented in Table 1. They include 40 small molecules (56.5% of the experiments). The highest AUC score was 0.876 for a subgroup not shown in the Table (no associated enriched terms).

Table 1: A selection of subgroups (chemicals and their associated phentoypic profiles) as discovered by the proposed algorithm. Reported are the number of small molecules in a subgroup, AUC scores, associated KEGG pathway, and enriched chemical and pharmacological MeSH terms.

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

5. Discussion

The algorithm presented in this paper enables inference relations between chemical structure and biochemical pathways. Identification of such relations is very important for drug discovery, since it allows for an argument-based selection of candidate molecules in target-specific screenings, and early exclusion of substances with highly probable undesired side-effects.

The experimental analysis we report in the paper uses the first, and currently the only publicly available data set that observes chemically-induced phenotypes in a genome-wide set of single-gene mutations. With availability of single-mutant collections for a range of model organisms, and promises of corresponding RNA-interference platforms that could also be applied for genome-wide phenotype screening of human samples, we expect the emergence of similar data sets in the near future. The presented computational approach should therefore not be regarded as a single-application attempt, but rather as an enabling technology that could help us in data analysis and hypothesis formation from the soon-to-emerge experimental data.

The comprehensive evaluation of the results in Table 1 is beyond the scope of this paper. An ultimate test would require a number of wet-lab experiments to either confirm or discard the proposed hypotheses. We have, though at the scanning stage, found some of the proposed hypotheses very interesting. One of the identified subgroups, consisting of six molecules (Figure 2), is related to the cellular process "cell cycle". Disturbances in cell cycle regulation are the hallmark of cancer. Enrichment analysis of the chemicals revealed that the subgroup contains both anticarcinogenic agents from the data set, parthenolide and amsacrine. Moreover, the most characteristic phenotypic marker by which the six substances were clustered together was the relative growth fitness of mutants in three genes (*LTE1*, *DBF2* and *CDH1*), which are all involved in mitotic exit. Parthenolide, the more thoroughly studied of the two identified anticarcinogenic substances, is indeed thought to affect this phase of the cell cycle (Fonrose et al., 2007). This example thus illustrates the biological relevance of the proposed method and illustrates usefulness of such methods. For example, identification of anticarcinogenic activity of parthenolide was identified in an experimental screen (Jonathan J. Ross and Birnboim, 1999) of the type which is notoriously error-prone. The method presented here demonstrates that such computational analysis prior to experimental screens could importantly increase the likelihood of a positive outcome of the screens.

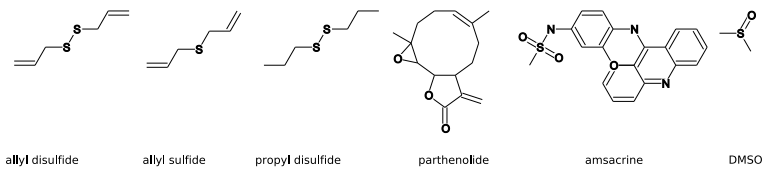

Figure 2: Subgroup of six small molecules having similar impact on cell cycle genes. Their enriched chemical terms are disulfides and allyl compounds ($p = 0.0048$), their enriched pharmacological terms are anticarcinogenic agents ($p = 0.0055$) and protective agents ($p = 0.0161$).

6. Conclusions

This report presents the first attempt to analyze chemogenomic data by relating chemical structures to biochemical pathways. An example is given to demonstrate the biological relevance of the proposed method. It should be noted, however, that for the full extent

of the usefulness of the proposed method, more comprehensive data sets are required. As a lesson from the study, screens uniformly covering the chemical space in the selection of tested molecules are likely to provide the best predictive power. Further impact of the combined experimental and computational methods described here for drug discovery and development will be achieved when technical limitations for conducting genome-wide screens in mammalian cells will be overcome; importantly, the method presented here for yeast data is, with only slight modifications, useful also for mammalian systems.

Acknowledgment

The study was supported by grants from Slovenian Research Agency (J2-9699, L2-1112).

References

T. Curk, J. Demšar, Q. Xu, G. Leban, U. Petrovič, I. Bratko, G. Shaulsky, and B. Zupan. Microarray data mining with visual programming. *Bioinformatics*, 21(3):396–8, 2005.

J. Demšar, B. Zupan, and G. Leban. Orange: From Experimental Machine Learning to Interactive Data Mining, White Paper, 2004.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

Xavier Fonrose, Frederic Ausseil, Emmanuelle Soleilhac, Veronique Masson, Bruno David, Isabelle Pouny, Jean-Christophe Cintrat, Bernard Rousseau, Caroline Barette, Georges Massiot, and Laurence Lafanechere. Parthenolide Inhibits Tubulin Carboxypeptidase Activity. *Cancer Res*, 67(7):3371–3378, 2007.

Maureen E. Hillenmeyer, Eula Fung, Jan Wildenhain, Sarah E. Pierce, Shawn Hoon, William Lee, Michael Proctor, St, Mike Tyers, Daphne Koller, Russ B. Altman, Ronald W. Davis, Corey Nislow, and Guri Giaever. The chemical genomic portrait of yeast: Uncovering a phenotype for all genes. *Science*, 320(5874):362–365, 2008.

Shawn Hoon, Smith, Iain M. Wallace, Sundari Suresh, Molly Miranda, Eula Fung, Michael Proctor, Kevan M. Shokat, Chao Zhang, Ronald W. Davis, Guri Giaever, Robert P. St Onge, and Corey Nislow. An integrated platform of genomic assays reveals small-molecule bioactivities. *Nat Chem Biol*, 4(8):498–506, 2008. URL `http://dx.doi.org/10.1038/nchembio.100`.

J. Thor Arnason Jonathan J. Ross and H. Chaim Birnboim. Low Concentrations of the Feverfew Component Parthenolide Inhibit In Vitro Growth of Tumor Lines in a Cytostatic Fashion. *Planta Med*, 65(2):126–129, 1999.

Minoru Kanehisa, Michihiro Araki, Susumu Goto, Masahiro Hattori, Mika Hirakawa, Masumi Itoh, Toshiaki Katayama, Shuichi Kawashima, Shujiro Okuda, Toshiaki Tokimatsu, and Yoshihiro Yamanishi. Kegg for linking genomes to life and the environment. *Nucleic Acids Research*, 36(Database issue):D480–D484, December 2007.

L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley, 1990.

Nada Lavrač, Branko Kavšek, Peter Flach, and Ljupčo Todorovski. Subgroup discovery with CN2-SD. *Journal of Machine Learning Research*, 5:153–188, 2004.

Talete srl. Dragon for Windows (Software for Molecular Description Calculations), Version 5.5, 2007. URL `http://www.talete.mi.it/`.

Iven Van Mechelen, Hans-Hermann Bock, and Paul De Boeck. Two-mode clustering methods: a structured overview. *Statistical Methods in Medical Research*, 13(5):363–394, 2004.

Bernard Ženko and Jan Struyf. Learning predictive clustering rules. In *4th Int'l Workshop on Knowledge Discovery in Inductive Databases: Revised Selected and Invited Papers, volume 3933 of LNCS*, pages 234–250. Springer, 2005.

Joe H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

David Weininger. SMILES, a chemical language and information system. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988.