A Drug-Gene Network for Understanding Drug Mechanism of Action

by

NERMİN PINAR KARABULUT

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfillment of
the requirements for the degree of
Master of Science

Sabancı University

SPRING, 2012

A Drug-Gene Network for Understanding Drug Mechanism of Action

Approved by:

Asst. Prof. Murat Çokol ....
(Thesis Supervisor)

Assoc. Prof. O. Uğur Sezerman ....

Asst. Prof. Erdal Toprak ....

Asst. Prof. Gürdal Ertek ....

Asst. Prof. Cemal Yılmaz ....

Date of Approval: 27.06.2012

A Drug-Gene Network for Understanding Drug Mechanism of Action

Nermin Pınar Karabulut

Computer Science and Engineering, Master's Thesis, 2012

Thesis Supervisor: Asst. Prof. Murat Çokol

Keywords: Chemogenomics, high-throughput screening, biological networks, biological statistics, chemical structural and side effect similarity

Abstract

Chemogenomics experiments, where genetic and chemical perturbations are combined, provide data for discovering the relationships between genotype and phenotype. Here, we computationally analyzed the largest chemogenomics dataset, which combines more than 300 chemicals with virtually all gene deletion strains in the yeast *S. cerevisiae*. Traditionally, analysis of chemogenomic datasets has been done considering the sensitivity of the deletion strains to chemicals, and this has shed light into drug mechanism of action and finding drug targets. We also considered the deletion strains which are resistant to chemicals. We found a small set of genes whose deletion makes the yeast cell resistant to many chemicals. Curiously, these genes were enriched for functions related to RNA metabolism. Our approach allowed us to generate a network of drugs and genes that are connected with resistance or sensitivity relationships. As a quality assessment, we showed that the higher order motifs found in this network make biological sense. Moreover, by using this network, we constructed a biologically relevant network projection pertaining to drug similarities, and subsequently analyzed this network projection in detail. We propose the drug similarity network as a useful tool for understanding drug mechanism of action.

İlaçların Etki Mekanizmalarını Anlayabilmek için İlaç-Gen Ağı

Nermin Pınar Karabulut

Bilgisayar Bilimi ve Mühendisliği, Yüksek Lisans Tezi, 2012

Tez Danışmanı: Yrd. Doç. Dr. Murat Çokol

Anahtar Kelimeler: Kemogenomik, yüksek veri taraması, biyolojik ağlar, biyolojik istatistik, kimyasal yapı ve yan etki benzerliği

Özet

Genetik ve kimyasal karışıklıkların birleştirildiği kemogenomik deneyler, genotip ve fenotip arasındaki ilişkinin keşfedilmesi için veri sağlar. Bu araştırmada biz, 300 tane kimyasala karşı *S. cerevisiae*'daki bütün delesyon suşlarının büyüme bilgisinin içerildiği, şu ana kadar üretilmiş olan en büyük kemogenomik veriyi hesaplamalı olarak analiz ettik. Şimdiye kadarki kemogenomik veri analizlerinde hep delesyon suşlarının kimyasallara karşı 'duyarlılık' ilişkisi incelenmiştir ki bu ilaçların etki mekanizmalarının anlaşılmasına ve ilaç hedeflerini bulmaya ışık tutar. Biz ise bunun yanında bir de delesyon suşlarının kimyasallara karşı olan 'direnç' ilişkisini inceledik. Öyle bir gen kümesi bulduk ki bu genlerin hücreden silinmesinin, maya hücresini birçok ilaca karşı dirençli kıldığını farkettik. İlginç bir şekilde, bu genlerin RNA metabolizmasıyla ilgili fonksiyonlarda tesadüfen beklenmeyecek kadar işlevi olduğunu bulduk. Bu projedeki yaklaşımımız bize ilaçların ve genlerin birbiriyle 'duyarlılık' ve 'direnç' ilişkileriyle bağlandığı bir *ilaç-gen ağı* oluşturmamızı sağladı.

Bu ağa kalite kontrol olarak yaptığımız analizlerde, ağdaki yüksek dereceli motiflerin aslında biyolojik olarak anlam ifade ettiğini farkettik. Bununla beraber, bu ağı kullanarak ilaç benzerlikleriyle ilgili, biyolojik amaca uygun bir ağ yansıması oluşturduk, ve sonrasında bu ağı ayrıntılı olarak analiz ettik. Elde ettiğimiz bu *ilaç benzerlik ağı*nı, ilaçların etki mekanizmalarını daha iyi anlayabilmek için yararlı bir araç olarak sunuyoruz.

to my lovely mother

Acknowledgments

I wish to express my sincere appreciation and gratitude to my advisor Murat Çokol for his continuous support and worthwhile guidance throughout my masters studies and being always available. Moreover, he was a chance for me to strengthen my background in biology.

I am also thankful to my thesis defense committee members: Uğur Sezerman, Erdal Toprak, Cemal Yılmaz and Gürdal Ertek for their presence and invaluable review. Especially, I owe special gratitude to Uğur Sezerman for his support, suggestions and guidance in addition to being willing to help me for years.

I am grateful to The Scientific and Technological Research Council of Turkey (TUBITAK) for providing financial support (BIDEB 2210) throughout my M.Sc. education.

I specially thank to Azat Akhmetov, Murodzhon Akhmedov and Leyli Javid Khayati for the discussions and cooperation we did together.

Last but not least, I would like to present special thanks to my family, especially to my mother Nuray, for her endless and unconditional love, support and patience.

Contents

List of Figures

List of Tables

1 Introduction

Currently, the trend in drug discovery is the target identification [5] which is based on the assumption that a drug can change the activity of a defected function by binding to the protein (target) responsible for that particular function. A common target-based approach is high-throughput screening (HTS) where numerous chemicals are tested against a chosen target related to a disease, and observations are made on the inhibition ability of chemicals [6]. Another approach includes computer-aided models where the library of chemicals are attempted to dock to the target proteins *in silico*. These target-oriented approaches reveal novel compounds increasingly, but only several of them can be approved due to developmental and experimental costs. Hence, a reasonable selection of chemicals should be done beforehand to reduce these costs. On the other hand, the mechanism of action of most US Food and Drug Administration (FDA)-approved drugs is still unknown despite the use of them in curing certain diseases [7]. However, in some cases, a cell may show same response to certain drugs which leads to discovering the drugs that have similar mechanism, as a result, suggesting candidates of chemicals to be most likely effective in certain diseases. In addition, similar drugs can be used as replacement for each other in some cases, such as a drug having less side effects can be used instead of a similar drug having more side effects. Therefore, understanding drug mechanism of action is a considerable problem for drug discovery and therapeutic intervention.

In this study, we propose a drug similarity network that highlights drug mechanism of action by analyzing a previously published chemogenomic screening dataset. To the best of our knowledge, sensitivity relationships between genes and drugs have been widely used so far. We, however, also took into account the resistance relationships, and proved that resistance interactions between genes and chemicals have also biological meaning. The proposed technique involves (i) finding multi-drug resistance (MDR) and multi-drug sensitivity (MDS) genes, (ii) constructing a deletion strain-drug network by using fitness defect scores of deletion strains in the presence of a particular drug and performing quality assessments to qualify the robustness of the network, (iii) generating a drug similarity network using sensitivity and resistance relationships between drugs and deletion strains, again performing several quality assessments, and quantifying interrelationships between the similarities found in this network and certain orthogonal datasets, including chemical structural similarities and side effects similarities of drugs.

2 Background Information and Related Work

In order to better understand the data and the method we have used in this project, Section 2 gives a background information about chemogenomics, biological networks and statistics in addition to related works done so far related to these fields. Throughout this section, a gene used to represent a drug target is actually the gene that encodes the protein inhibited by the corresponding drug.

2.1 *Saccharomyces cerevisiae*

Saccharomyces cerevisiae is one of the species of yeast. In addition to its importance in industry (i.e. baking and brewing), it is one of the most studied eukaryotic organisms in cell biology owing to its suitability for genetic manipulation, accessibility of its whole genome sequence information [8] and having many homologs of proteins in human cells, spanning cell cycle proteins, signaling proteins and so on. In addition, yeast is currently the only organism to be able to detect all targets in the cell in parallel *in vivo.* As a consequence, many of the chemical genomic studies use *S. cerevisiae* as the model organism when we survey in the literature. Moreover, the largest chemogenomics dataset which was published by Hillenmeyer *et al.*, investigates the genotype-phenotype relationships within *S. cerevisiae* cells. Therefore, we exploited the genomic responses of *S. cerevisiae* cells to chemical compounds to infer drug mechanism of action throughout this project.

2.2 Chemogenomics

The emerging field of chemogenomics investigates the genomic responses to small chemical compounds. Over the past decade, the chemical-genomic screening using *S. cerevisiae* has been leading to discover drug mechanism of action, primary drug targets, or secondary drug targets, in other words *off targets* that cause unexpected side effects, and also help to reveal genes buffering drug target pathways. In summary, chemogenomics experiments facilitate to identify cellular responses against small chemical compounds *in vivo* [5]. Numerous approaches are implemented to achieve these goals:

1. Haploinsufficiency profiling (HIP)
2. Homozygous profiling (HOP)
3. Chemical genomic-genetic interaction profilings combination

4. Multi-copy suppression profiling (MSP)

5. Chemical-genomic expression profiling

Chemical genomic fitness profiling screens where two perturbants (deletion strain and chemical compound) are incorporated, measure the growth responses of distinct strains, heterozygous (HIP) or homozygous (HOP) deletion collections, in diverse chemical compounds treated cultures in order to understand drug mechanism of action in addition to gene dispensability, multi-drug resistance and gene functions within the yeast *S. cerevisiae*. Understanding drug mechanism of action contributes to find drug targets, to use combination of drugs in therapeutics or to discover the drug resistant genes such as antibiotic resistant genes. Moreover, it is noteworthy to state that comparing similarities of drug chemogenomic fitness profiles enables to discover unknown mechanism of a drug from known mechanism of a drug [9]. Each deletion collection has a deletion from start to stop codon of a single gene, which is further replaced with a unique 20-base-pair DNA tag, or oligonucleotide barcode whose abundance facilitates to measure the growth rate of the strains in an ensemble of competing cells under any condition of choice, using high-density microarrays (or arrays) [10, 1, 11].

Haploinsufficiency profiling (HIP) which is obtained by deleting one copy of each gene from the diploid cells [12], is one of the approaches used in chemogenomics screening to reveal chemical compounds that target proteins encoded by essential genes by measuring the growth responses of the distinct deletion strains in the presence of a particular drug because the growth rate (fitness) of the strain whose deleted gene is encoding the drug target, results in a further decrease by the drug inhibiting the target protein [12], and disappears from the pool over time [5]. This assumption is based on the fact that one copy of each gene is generally sufficient for the optimal growth of the diploid organisms even though there are rare exceptions. On the other hand, haploinsufficiency, abnormal phenotype caused by loss of function of one copy of gene, can be identified in the presence of a particular drug targeting protein encoded by the deleted gene which further decreases the gene function, in addition to identification of rare haploinsufficient genes under optimal growth conditions [8]. One of the advantages of the HIP approach is that any prior knowledge related to compound and its target is not necessary to detect drug target in this parallel screening [5]. Moreover, a treatment in which the gene encoding the protein targeted by the drug is absent may help to discover the secondary drug targets (off targets) [13]. However, HIP assay has some disadvantages/shortcomings, too [5]:

- Yeast cells are used in these experiments owing to having human homologs. However, certain human genes do not have yeast homologs, resulting in unidentified targets in human cells.

- HIP approach is based on the assumption that growth of the strain whose deleted gene is encoding the target protein decreases when it is exposured to the drug targeting the corresponding gene product. However, there are some genes whose deletion do not decrease the growth of the strain when inhibited by the drug. In such cases, the drug targets cannot be identified.

- In some cases, deletion of one copy of a particular gene is not enough to detect drug targets, whereas further deletion (reducing gene dosage by more than one copy) of the corresponding gene may highlight the drug target.

Likewise, the other approach is *homozygous profiling (HOP)* where both copies of each non-essential gene are deleted from either diploid or haploid cells. Over time, the growth responses of some strains decrease in the condition of choice and eventually, these strains disappear from the pool like in HIP assays. However, this time the target genes cannot be identified directly due to deleting both copies of the particular gene. On the other hand, performing the HOP approach, the complete loss of function of the deleted gene is provided, which subsequently specifies essential genes for growth in the presence of a particular drug, resulting in a hypothesis that almost all genes are required for adequate growth in at least one condition [1]. Furthermore, non-essential genes are propounded to subscribe to genetic robustness [14, 15] and also take a role in drug-target pathways indirectly that makes it required in the corresponding condition.

In order to find drug targets directly by using the HOP approach, it is combined with genetic interaction data coming from Synthetic Genetic Analysis (SGA) and this approach may be defined as *chemical genomic-genetic interaction profilings combination.* The genetic interaction profile of a non-essential (or conditionally essential) gene is compared to fitness profiles of non-essential genes in the presence of a particular drug by performing the HOP approach. The high correlation between these two profiles indicates that the conditionally essential gene is encoding the candidate target protein of the corresponding drug used in the HOP treatment [16, 5]. However, the correlation does not need to be a conclusion of inhibition of the target gene product by the drug, it may also occur owing to inhibition of gene products in the target pathways or in the cellular functions [17]. This is one of the advantages of the chemical genomic-genetic interaction profilings combination approach. Moreover, once more yeast genetic interaction data is available, chemical genomic-genetic interaction profilings combination will be more informative to understand drug mechanism of action [17].

Multi-copy suppression profiling (MSP) is obtained by increasing gene dosage instead of decreasing gene dosage to one copy like in HIP approach or deleting two copies of each gene such in HOP approach. By increasing gene dosage, the gene encoding the target protein is over-expressed, resulting in demonstrating resistance to the chemical treatment [18, 19]. MSP genome-wide assays competitively screen DNA clone libraries to detect genes showing resistance to the drug when over-expressed [10].

On the other hand, chemical-genomic expression profiling is a genome-wide approach where only one perturbant (chemical compound) is included and the genes are in their wild-types. Chemical-genomic expression profiling is performed by examining the mRNA expression profiles of the genes in the presence of chemical compounds or deletion mutants, resulting in identification of the gene functions due to the fact that shared cellular functions most probably show similar expression patterns [20]. The advantage of this approach is that a transcriptional response causing change in the mRNA expression profile is not a conclusion of decrease in yeast growth produced by the inhibiting drug [5]. Therefore, the genes encoding drug targets can be identified by performing this approach even though the gene deletions do not cause a decrease in the strain growth when inhibited by the drug.

2.2.1 Hillenmeyer *et al.* dataset

The dataset given by Hillenmeyer *et al.* [1] is the largest chemogenomics dataset which combines more than 300 unique set of small chemical compounds and diverse environmental stress conditions with virtually all deletion strains in the yeast *S. cerevisiae*. The dataset reports the growth responses of yeast cells, including whole genome heterozygous and homozygous deletion collections, in the presence of distinct chemical compounds or environmental stress conditions. The deletion collections encapsulate approximately *6000* heterozygous gene deletion strains, around *1000* of them are essential genes, and approximately *5000* viable homozygous deletion strains.

The gene deletion collections are obtained by using four oligonucleotide barcodes, or tags which have unique 20-base-pair DNAs. Two of the tags are uptag and downtag on the sense strand and the remaining tags are on the antisense strand as complementary tags. Using the abundance of these tags by measuring in microarray, the growth rates of the corresponding deletion strains are quantified.

The way this experiment is done is the following: They put all deletion strains into different pools. This means that there should be around 6000 types of yeast cells in each pool corresponding to each deletion strain. Then, they put different chemical compounds into every pool and a competitive growth occurs in each pool between deletion strains in the presence of a particular drug. They subsequently compare the growth rate of each deletion strain in the presence of a certain drug, called treatment, with no drug condition, called control.

Then, the authors give two definitions: one of them is that a gene deletion strain is considered sensitive to a condition (chemical compound) if its growth rate in the condition is slower than the control (no drug condition). The second is that some of deletion strains are sensitive to multiple conditions, hence the deleted genes of these strains are considered to be necessary for resistance to diverse perturbations. That's why they are mentioned as multidrug resistance (MDR) genes, where multiple means more than *20%* of all unique compounds. Moreover, around *50* MDR genes are reported as highly enriched for functions related to endosome transport, vacuolar degradation, and transcription [1].

In order to represent the quantification of the growth rates of the strains, z-score and log-ratio values are used. In addition, p-values of the growth rates are calculated from z-scores using t-distribution test [1].

Conditions spanning some chemical compounds, contain certain repeated experiments in which same drug in the same concentration is used on several repeats. The correlation for pairs of replicates is 0.72 which is much better than random and supports that the data has a small noise and proves the reproducibility of the experiment (Figure 1).

2.3 Biological Statistics

Biological statistics are widely used to analyze experimental data. In this project, we also used several statistics methods such as comparing p-values to extract significant results, multiple hypothesis comparison, or using t-distribution to reveal significantly different samples. Here, we indicate certain background information related to these topics. [21] was used to gain information related to biological statistics given below.

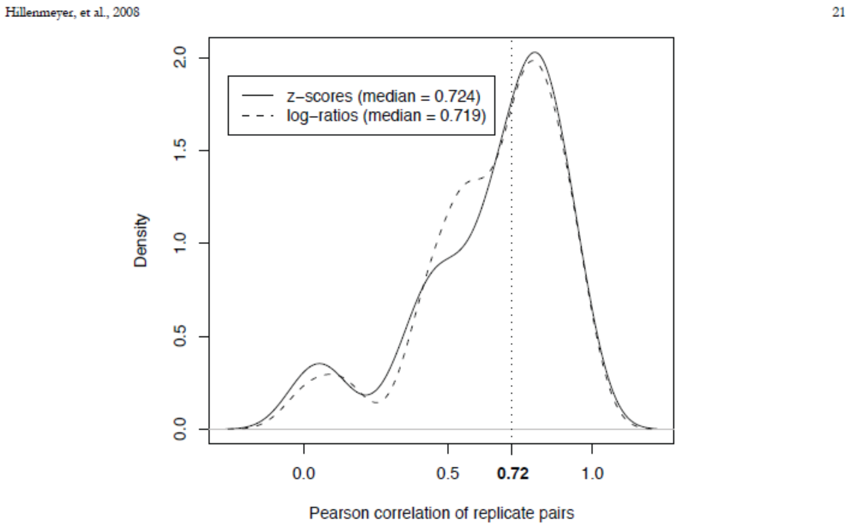


Figure 1: Distribution of correlation for pairs of replicates [1]. The correlation between conditions where same drugs are used in same concentration levels, is 0.72 which proves the reproducibility of the experiment given by Hillenmeyer *et al.* [1].

2.3.1 Types of variables

The scaling of variables can take up different forms [21] e.g. categorical, ordinal, ratio etc. Here, I characterize variables as categorical and continuous which are further used throughout the other sections of the thesis.

- Categorical variable, also known as discrete or qualitative variable, is used to classify observations into a small number of categories. If there are only *2* categories, then the variable is called *dichotomous* variable. On the other hand, *nominal* variable represents a variable that has two or more categories such as class information of protein domains etc.

- Continuous variable, also called quantitative variable, is referred to represent measurable variables that have numerical values such as temperature, mRNA expression level etc.

2.3.2 Hypothesis testing

The null hypothesis generally corresponds to the statement that things are same as each other, or there is not any statistically significant difference between two measured quantities. The alternative hypothesis, on the other hand, posits that things are different from each other. The aim of the statistical analysis is to capture whether the observed property is different from the expectation under the null hypothesis. If different, the null hypothesis can be rejected. In data analysis, the alternative hypothesis is more appealing since it highlights into exciting discoveries. Therefore, one should attempt to capture interesting patterns included in the data, suggesting the alternative hypothesis. However, it is noteworthy to state that the probability of getting a difference between two samples, just by chance, must be calculated if the null hypothesis is really true. The null hypothesis can be rejected only if this probability is lower than a theoretical p-value, or significance level, which should be decided before the analysis (See Section 2.3.3).

2.3.3 P-values and significance levels

A p-value is the probability of getting the observed or a more extreme outcome when the null hypothesis is true [21]. Two types of probabilities can be examined: one-tail probability and two-tail probability. If the p-value is calculated as the probability of getting the observed result, or either less or more than the observed result, it would be one-tailed probability. However, when the both tails (sufficiently large and sufficiently small) are taken into

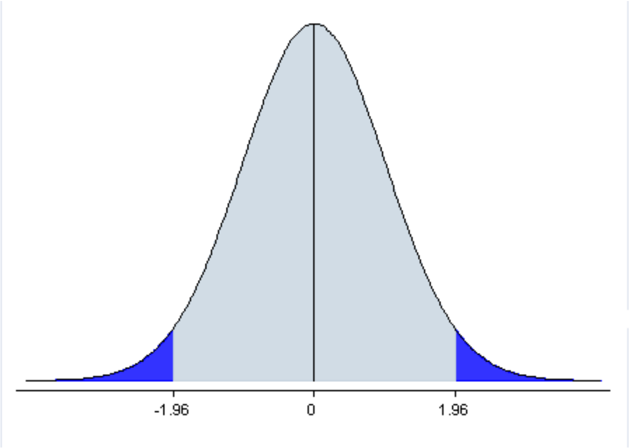


Figure 2: Probability from both sufficiently large (right-tail probability) and sufficiently small (left-tail probability) regions on the normal distribution [2].

account, the two-tailed probability is assumed as in Figure 2 [22].

The conventional significant level of p-value in biology is *0.05* which means that if the probability to observe an outcome is less than *0.05*, the null hypothesis can be rejected, as a result the alternative hypothesis is true. On the other hand, if the probability is greater than or equal to *0.05*, the null hypothesis cannot be rejected. *0.05* significance level states that even though the null hypothesis is true, there is *5%* chance to reject the null hypothesis which corresponds to false positive, or Type I, error. On the other hand, if the null hypothesis is not rejected even if the alternative hypothesis is true, false negative, or Type II, error occurs [21] which is shown in Table 1.

If the chosen significance level is higher than *0.05*, the chance of a false positive, wrong conclusion, is increased, whereas the chance of false negative is decreased. However, if the chosen significance level is lower than *0.05*, while detecting false positive is decreased, this time the chance of false negative is increased. Therefore, there is a threshold when choosing the sig-

	Alternative (real)	Null (real)
Alternative (analysis)	Null hypothesis correctly rejected	False positive, Type I error
Null (analysis)	False negative, Type II error	Null hypothesis correctly retained

Table 1: Tabulated relation between reality and statistical analysis

nificance level, which should be optimized according to the costs of false negatives and false positives inherited in the data.

2.3.4 t-Distribution test

Student t-test is one of the t-distribution that is employed to compare the means of two samples [21]. Student t-test is used when there are two types of variables: nominal variable and measurement variable. The nominal variable has two categorical values [21]. Then, using Student t-test, the means of two samples whose elements are measurement variable, are compared. As expected, the null hypothesis of the t-test is that the mean values of two samples are same.

2.3.5 Multiple comparisons

When there are multiple hypothesis that should be taken into account, the significance level should be chosen carefully. This is an open research area that there is not any universally accepted approach to decide the significance level in multiple hypothesis problem.

1. **Bonferroni correction**, is one of the approaches for multiple comparisons that uses a significance level found as dividing the conventional p-value by the number of statistical tests. As a result, the significance

level is chosen lower than conventional p-value *0.05*, which would further decrease the chance of false positives. However, in some cases, there are numerous number of statistical tests which concluded with a very small p-value (Let's say there are *1000* tests, so p-value would be equal to *0.05/1000 = 0.00005*) that increases the chance of false negatives this time. Benjamini–Hochberg procedure can be used as an alternative approach in such cases.

2. **Holm-Bonferroni method** is similar to Bonferroni method, but has slight difference. The algorithm of Holm-Bonferroni method is given in Algorithm 1.

3. **Benjamini–Hochberg procedure** controls the proportion of significant results which are actually false positives, by setting false discovery rate to a constant percentage. The algorithm of Benjamini–Hochberg procedure is given in Algorithm 2.

2.4 Biological Networks

Biological networks facilitate to characterize many complex biological systems. Categorizing the biological networks into 2 classes reveals *molecular networks* such as protein-protein interaction networks, metabolic networks, regulatory networks, RNA networks etc., and *phenotypic networks*, including co-expression networks, genetic interaction networks, chemical-genetic networks and so on [24]. The great amount of the studies to date are used *Escherichia coli* and *Saccharomyces cerevisiae* as the model organisms owing to the advantage of available set of data for these organisms. However, after

Algorithm 1 Holm-Bonferroni algorithm [23]

```
Input:
    • $k$: the number of hypothesis to be tested
    • $\alpha$: the significance level
    • $P$: the set of p-values of null hypothesis
Output: The rejected null hypothesis
    1. Order p-values in the set $P$ from smallest to largest
    2. Compare the smallest p-value, $p$ of the set $P$ to $\alpha/k$
    3. If $p$ is smaller than $\alpha/k$,
        (a) Reject null hypothesis
        (b) Swap $k$ with $k - 1$
        (c) Exclude the smallest p-value, $p$ from the set $P$
        (d) Return to step $1$
    4. Else, stop. None of the remaining null hypothesis can be rejected.
```

Algorithm 2 Benjamini–Hochberg algorithm [21]

Input:

- i: ranks of p-values
- m: total number of hypothesis to be tested
- P: the set of p-values of null hypothesis
- Q: the chosen false discovery rate

Output: the rejected null hypothesis

1. Order p-values in the set P from smallest to largest. The smallest p-value has rank of $i = 1$, the second has rank of $i = 2$, and so on
2. Compare each p-value, p in the set P to $(i/m)*Q$
3. The largest p-value that has $p < (i/m)*Q$ is significant
4. The other p-values smaller than the p-value, p found in step 3 are also significant.

completion of the Human Genome Project, the quantity of data concerning to human cells is increased, resulting in better investigation of networks pertaining to human cells [24]. Nevertheless, the emergence of advances in network theory highlights certain principles related to network topology [24], for instance, ability to identify the effect of a randomly perturbed node or ability to discover the existence of certain patterns in the network. In the following, I specify the most important principles to analyze and to reveal certain characteristics of the biological networks.

2.4.1 Degree and degree distribution

According to edge types, a network can be classified into two different categories: Directed network and undirected network. In directed networks, links have directions, and each node has an incoming degree and an outgoing degree which show the number of links coming to and leaving from the corresponding node, respectively. On the other hand, in undirected networks, the number of links that a particular node has, constitutes the degree of the corresponding node [4].

The average degree of an undirected network is defined as:

$< k > = 2 * L/N$, where:

- $<>$ denotes the average
- k is the degree
- L is the total number of links in the network
- N is the total number of nodes in the network

The degree distribution of a network, $P(k)$, corresponds to the probability of a selected node having exactly k links.

$P(k) = \frac{N(k)}{N}$, where:

- $P(k)$ is the degree distribution of nodes having $k = 1, 2, ..$
- $N(k)$ is the count of nodes having $k = 1, 2, ..$
- N is the total number of nodes in the network

$P(k)$ is used to identify the topological type of the network, which is explained in Section 2.4.3 in detail.

2.4.2 Clustering coefficient

The clustering coefficient gives information about how well the neighborhoods of a particular node are connected to each other. In order to quantify this phenomenon, the number of triangles that go through the corresponding node is determined as:

$C_D = [2 * n_D]/[k * (k - 1)]$, where:

- C_D is the number of triangles go through node D
- k is the number of links node D has
- n_D is the number of links connecting k neighbors of node D

If clustering coefficient of node D equals to or is in proximity of 1, then node D is at the center of a fully-connected network or tending to be a center node, called as *hub*. In contrast, if the clustering coefficient equals to *0* or is in proximity of *0*, then node D is either isolated or part of a slack interlinked network. The average clustering coefficient over all nodes of the network,$< C >$, characterizes the network modularity [25].

Moreover, using above information, $C(k)$ is defined as the average clustering coefficient of nodes having $k = 1, 2, ..$ links. $C(k)$ is independent of the size of the network likewise $P(k)$ [4], hence, $C(k)$ and $P(k)$ are both used to detect network characteristics (See Section 2.4.3).

2.4.3 Network types

Characterizing the degree, k, of individual nodes reveals general characteristics of the network. For instance, in certain networks, most of nodes have approximately same number of links which equal to the average degree, $< k >$, of the network, whereas scarce of nodes have links more or less than the average distribution. This type of node's degrees follow a Poisson distribution, and the type of the network is defined as *Random Network*.

However, many networks, spanning especially biological networks, contain a few nodes having many more links than the average node has [26]. Such nodes constitute the *hubs* of the network and hold the network together, suggesting that the nodes represented by these hubs must have a special and important role [27]. The hubs can be classified into *2* categories:

- *Party hubs,* take role inside modules and in coordinating certain cellular processes as a local coordinator [28, 29]
- *Date hubs*, connect distinct processes as a global coordinator [28, 29]

Nevertheless, in the networks containing hubs, the number of nodes having exactly k links which also means the degree distribution $P(k)$, demonstrates a power law:

$P(k) \sim k^{-\gamma}$, where:

- k is the degree of the node
- γ is the degree exponent
- $P(k)$ is the degree distribution of nodes having $k = 1, 2, ..$

These types of networks having a power degree distribution, are defined as *Scale-free* networks. Meanwhile,γ value is generally between *2* and *3* in most of the networks [4, 26]. Notably, it is also noteworthy to state that the smaller γ value is more important in terms of the role of the hubs because it means that the largest hub is linked with a large amount of all nodes in the network [4].

2.4.4 Network motifs

Network motifs are certain n-node subgraphs (n equals to *3*, *4* or more) being observed more (at a statistically significant level) than the randomized where the randomization is obtained by keeping the network topology same (For more detail, see Section 4). In order to detect the motifs that describe the network characteristics, all possible subgraphs are determined and counted in the real network. After the network is randomized, if the probability of observing a subgraph in the real network greater number of times than in the randomized version of the network is lower than a significance level (Let's say conventional p-value 0.05), then these subgraphs are assumed as network motifs [26, 4]. Due to the fact that most of the network types have specific motifs, such as feed-forward loop, bi-fan etc., the dynamical behavior, the type and characteristics of the network can be highlighted by detecting the network motifs [26].

2.4.5 Connected components

A connected component for an undirected network is defined as a subgraph, where any two nodes in the subgraph are connected to each other by paths but they are not connected to any other nodes outside of the subgraph. The largest connected component of a network is called giant component, which quantifies local functional clustering when comparing to the randomized versions of the network, obtained by keeping the network topology same [25].

2.4.6 A clique

A clique in an undirected graph is a subset of vertices that every vertex in the subset is connected to other vertices of the subset by an edge. A clique can also be called as a complete subgraph. The maximal clique of a network is a clique of the network which cannot be comprised by any larger clique within the network [30].

2.4.7 Small-world effect

Most of the complex networks are assumed to demonstrate a small-world effect which is defined as there are a few links between each pair of nodes in the network. Thus, this short path length between any pair of nodes implies that perturbing a particular node causes defects in the activities of neighborhoods of that node quickly and easily, in addition to the network itself [4, 24].

2.4.8 Centrality

In order to determine the importance of a node in a network, there are numerous types of measurement, including degree centrality, betweenness centrality, eigenvector centrality, closeness centrality that all of them use different algorithms to identify the centrality. In our network analysis throughout this project, we used betweenness centrality to identify importance of nodes. Once the shortest paths between each pair of nodes in the network are identified, the number of shortest paths that pass from a randomly chosen node is calculated. That number constitutes the betweenness centrality of the randomly chosen node [31].

2.4.9 PageRank

PageRank is used to identify the important nodes in large networks that its algorithm is based on webgraphs, where nodes are World Wide Web pages and directed edges are hyperlinks between pages. The basic assumption behind this algorithm is that the more web pages a particular web page is linked with, the more importance that web page would gain. PageRank can be applied to the biological networks whose edges are directed such as metabolic networks [32].

2.5 Related Work

In the study proposed by Giaever *et al.* [8], HIP approach was employed in chemogenomics analysis in order to identify haploinsufficient genes under optimal growth conditions or stressed conditions. First, they tested six heterozygous strains individually in the presence of a drug which targets the reduced gene product in the corresponding heterozygous deletion. The results found from this analysis suggested that decreasing gene dosage contributes to detect phenotypes. Therefore, they subsequently conducted a larger analysis by examining *233 S. cerevisiae* heterozygous deletion collections in parallel in the presence of tunicamycin. Haploinsuffiency was observed in 3 loci in the presence of drug tunicamycin, one of them is ALG7 which is the known target of tunicamycin, in addition to the newly discovered 2 loci, YMR007w and YMR266w. In 2004, they conducted one more HIP approach again, this time including complete collection of heterozygous deletion strains, *6000* strains, and *10* chemical compounds [33]. In this analysis, not only are putative and novel cellular interactions revealed in addition to the known ones, but they also found a chemical structure shared by three compounds that have different therapeutical effect even though all of them inhibit ERG24 heterozygous deletion strain. Similarly, Lum *et al.* [34] investigated *3500* heterozygous deletions in the presence of *78* chemical compounds, reporting many of the known drug targets.

On the other hand, Parsons *et al.* [17] performed chemical genomic-genetic interaction profiling combination by incorporating HOP screening and genetic interactions using Synthetic Genetic Array (SGA) on *4700* viable haploid deletion strains within *S. cerevisiae* in the presence of *12* different chemical compounds. The authors performed two-dimensional hierarchical clustering of HOP profiles with an ensemble of genetic interaction profiles. The results suggested several drug-target pairs to cluster together, and as a result, contributed to identify uncharacterized genes in specific roles. In

2006, they expanded their chemical data to *82* different chemical compounds, performing homozygous profiling on *4700* homozygous deletion collections, and subsequently implementing cluster analysis using two-dimensional hierarchical clustering and probabilistic sparse matrix factorization analysis [35]. They elucidated novel drug-target relations and effects of certain drugs, contributing to understand drug mechanism of action in addition to putative findings at the end of these analysis. Kapitzky *et al.* [9] performed a cross-species chemogenomics screening by combining homozygous deletion collections of *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. By combining deletion strains of both species allowed them to get more accurate findings related to drug mechanism and to observe more conservation between two species in compound-functional module relation rather than compound-gene relation.

Hillenmeyer *et al.* employed the largest chemogenomics screening to date by combining HIP and HOP approaches and including *5000* homozygous deletion strains, *6000* heterozygous deletion strains, *1144* environmental stress conditions and small chemical compounds where over *300* of them are unique (See Section 2.2.1 for more detail).

Hoon *et al.* [10] integrated three genome-wide screens, including homozygous profiling (HOP), heterozygous profiling (HIP) and multi-copy suppression profiling (MSP) in order to identify targets and to analyze cellular processes. They profiled *6000* strains against *200* chemical compounds.

Hughes *et al.* [20] studied chemical-genomic expression profiling by creating a compendium of gene expression profiles, including *300* mutant expression profiles and numerous drug treatments. They elucidated the correlation between mRNA expression profiles of mutants and known drugs, in addition to identification and experimental confirmation of eight uncharacterized open reading frames (ORF) required for certain cell processes.

In 2010, Hillenmeyer *et al.* [12] computationally analyzed the previously obtained chemogenomics data in [1]. First, they defined co-fitness metric which gives the correlation of growth rates of deletion collections in the presence of different compounds, and subsequently used this metric to quantify how well it predicts gene functions in comparison to other large scale datasets, resulting in an observation of better predictions. Moreover, they implemented a machine-learning approach to predict drug-target interactions by using combination of fitness defect scores of strains in the condition of choice, chemical structural similarity between drugs, and therapeutic class information of drugs. The results of the machine-learning approach were certain known or robust novel drug-target predictions where two of top *12* novel predictions were further verified experimentally: nocodazole with Exo84 and clozapine with Cox17. Nevertheless, *5* of top *12* predictions were validated by literature findings.

3 Motivation and Contribution of the Thesis

The prevailing picture from the previous works given in Section 2.5 is that most of the chemogenomics studies have produced the chemogenomics data and have inferred certain observations according to the analysis of the experiments. However, to the best of our knowledge, except a handful of studies as in [12], [9], [35], there are no studies applying any statistical, machine-learning or network based approaches to neither the previously produced chemogenomics data nor the data produced by themselves. A machine learning based approach was implemented only in [12], where the fitness defect scores of strains in the condition of choice, therapeutic class information of drugs, and chemical structural similarity between drugs were combined to predict novel drug-target interactions. From the literature search, we observed certain network-based studies related to drug-target interaction data [36, 25], contributing to understand the relationship between drug targets and disease-gene products, also to discover unknown drug targets using known ones. However, the drug-target data information used in these works is not purely inferred from chemogenomics experiments. The drug-target interaction information is taken from DrugBank or such repository websites where interactions may be obtained using not only chemogenomics experiments but also several types of biological or computational studies. Therefore, all of them motivated us to propose the approach given in Section 4 which analyzes chemogenomics data by using network-based analysis and statistical tools in order to better understand drug mechanism of action.

The proposed method enables to reveal unknown drug mechanism of actions, and to discover genes that have similar responses in the cellular context. In this research, we constructed a drug similarity network from the deletion strain-drug interaction network that reveals significant similarities between drugs which subsequently help to discover certain unknown effects of a drug from the most similar drug whose mechanism is known.

4 Methods & Materials

The method proposed in this work can be summarized as in the flowchart given in Figure 3, and each step is explained in detail in the following sections. Besides, it is noteworthy to remark that the Hillenmeyer *et al.* dataset is chosen as the chemogenomics dataset to analyze in the proposed method owing to being the largest chemogenomics dataset, and also being reproducible (For more detail, see Section 2.2.1).

4.1 Fitness Scores Combination

In Hillenmeyer *et al.* dataset, same drugs in different concentration levels were sometimes used. Moreover, as we mentioned in Section 2.2.1, there are also some repeated experiments where a same drug was used in same concentration levels to demonstrate the reproducibility of the experiment. However, we are interested in inferring relationships between chemical compounds (drugs) and genes in this project. Therefore, we combined the fitness scores represented with z-scores of such experiments. In statistics, the widely used methods for combination are Fisher's method which is based on p-values, and Stouffer's method to get one unique z-score from different z-score values. As we used fitness scores data represented by z-scores, we combined z-scores of such repeated experiments by using Stouffer's method whose equation is given in Equation 4.1 where k is the number of repeated experiments, Z_i is the z-score of the corresponding experiment, and Z is the final z-score for such repeated experiments.

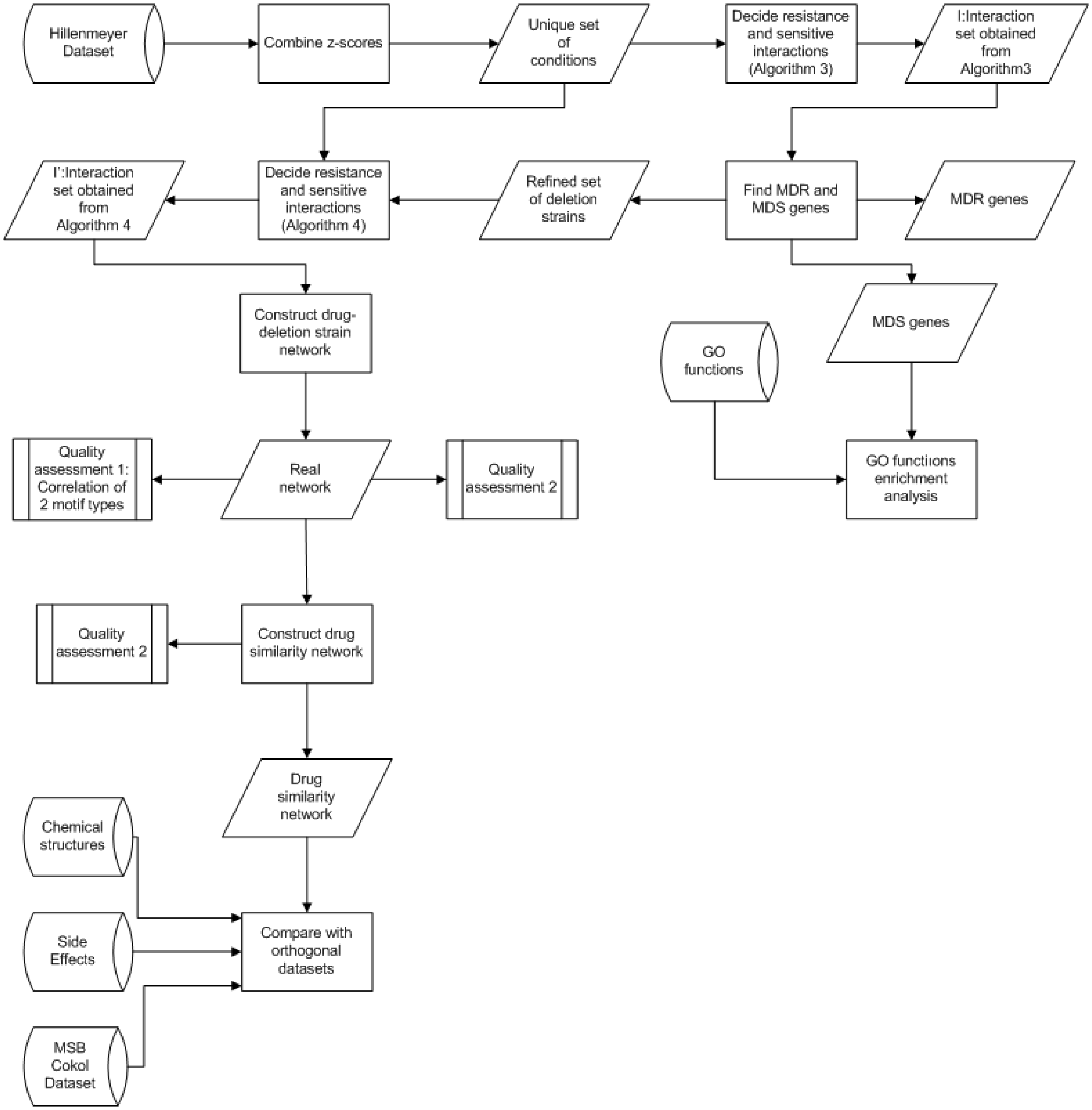

Hillenmeyer Dataset
Combine z-scores
Unique set of conditions
Decide resistance and sensitive interactions (Algorithm 3)
I:Interaction set obtained from Algorithm3
I':Interaction set obtained from Algorithm 4
Decide resistance and sensitive interactions (Algorithm 4)
Refined set of deletion strains
Find MDR and MDS genes
MDR genes
Construct drug-deletion strain network
MDS genes
GO functions
Quality assessment 1: Correlation of 2 motif types
Real network
Quality assessment 2
GO functions enrichment analysis
Quality assessment 2
Construct drug similarity network
Drug similarity network
Chemical structures
Side Effects
Compare with orthogonal datasets
MSB Cokol Dataset

Figure 3: The flowchart of the proposed methodology

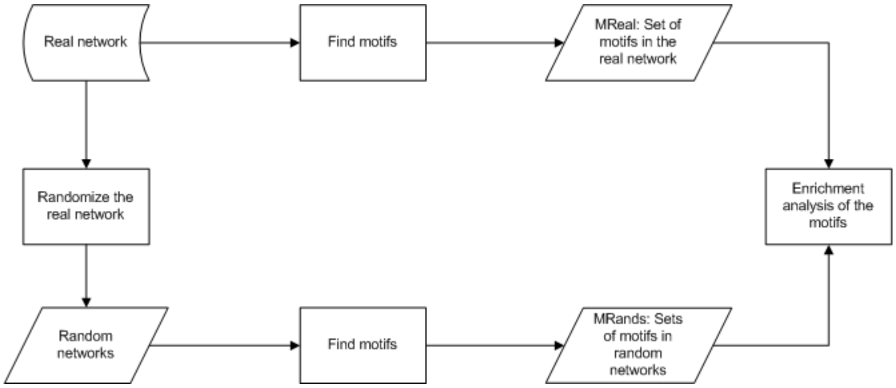

Real network
Find motifs
MReal: Set of motifs in the real network
Randomize the real network
Enrichment analysis of the motifs
Random networks
Find motifs
MRands: Sets of motifs in random networks

Figure 4: The flowchart of the subprocess: Quality assessment 2

$$Z = \frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}} \tag{4.1}$$

Once collapsing such repeated experiments using Stouffer's method, we were able to have a unique set of results for each drug and to reduce the condition number to ~300 from ~700, all of which are unique compounds.

4.2 Definition of Resistance and Sensitivity

As mentioned in Section 2.2.1, in the dataset we used, all the deletion strains were grown competitively in different pools including different chemical compounds, and then, the growth rate of each strain in the presence of a particular drug (treatment) was compared with the growth of that strain in no drug condition (control). If a deletion strain grows slower than normal (no drug condition) in the presence of a particular drug, we infer a 'sensitivity' relationship between this particular deletion strain-drug pair which is the given definition in Hillenmeyer *et al.* article [1]. On the contrary, if a deletion strain grows faster than normal in the presence of a particular drug, we infer a ‘resistance’ relationship between this particular deletion strain-drug pair.

4.2.1 Illustration for inferring resistance and sensitivity relationship

Figure 5 illustrates a basic example for how the experiment was done and how the relationships were decided. First tube (pool) represents the control condition where all yeast strains grow competitively without any chemical compound. In the second tube, when drug 1 is put, deletion strain B grows faster than normal (no drug condition). However, deletion strain A grows slower, no strain A in the second tube anymore.

It can subsequently be concluded that deletion strain B grows faster than normal in the presence of drug 1. In addition, deletion strain A grows slower than normal in presence of drug 1.

In our study, we define that deletion B is resistant to drug 1, and deletion strain A is sensitive to drug 1. And we infer a *resistance* relationship between strain B-drug 1, and a *sensitivity* relationship between strain A-drug 1.

In the third tube, deletion strain A grows faster than normal (no drug condition) in the presence of drug 2 this time. However, deletion strain C grows slower, no strain C in the third tube anymore.

Similar to before, we infer a *resistance* relationship between strain A-drug 2, and a *sensitivity* relationship between strain C-drug 2.

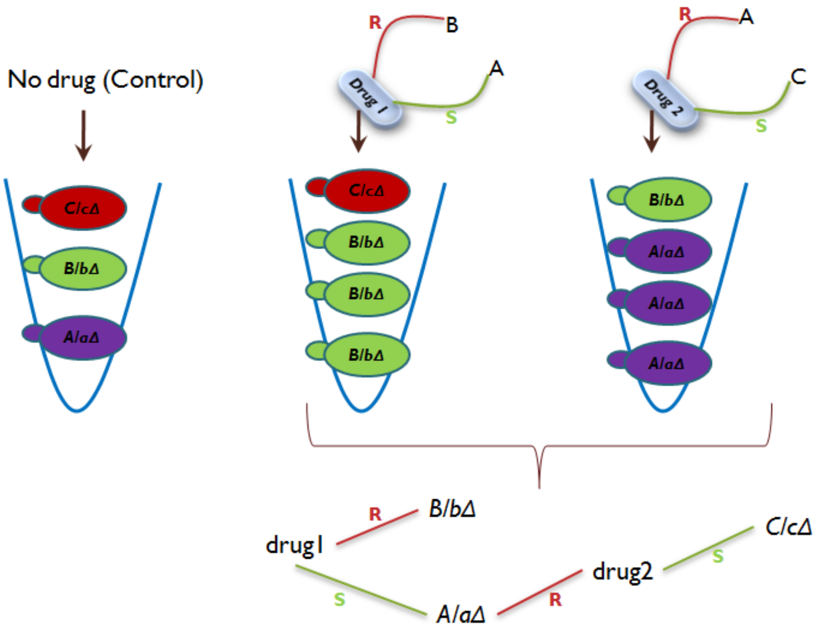


Figure 5: Example for how to infer resistance and sensitivity relationships between deletion strains and drugs. First tube (pool) represents the control condition where all yeast strains grow competitively without any chemical compound. In the second tube, when drug 1 is put, deletion strain B grows faster than normal (no drug condition). However, deletion strain A grows slower, no strain A in the second tube anymore. In the third tube, deletion strain A grows faster than normal in the presence of drug 2 this time. However, deletion strain C grows slower. In conclusion, it can be inferred from the second tube that deletion strain B is resistant to drug 1, whereas deletion strain A is sensitive. Looking at the third tube, we can say that drug 2 has resistance and sensitivity relationships with deletion strain A and deletion strain C, respectively.

4.3 MDR and MDS Genes Detection

In Hillenmeyer *et al.* dataset, some of deletion strains are sensitive to multiple conditions, hence the deleted genes of these strains are considered to be necessary for resistance to diverse perturbations. That's why they are mentioned as multi-drug resistance (MDR) genes, where multiple means more than 20% of all unique compounds (See Figure 6). Similar to MDR genes, we can also speak of multi-drug sensitive (MDS) genes whose deletion strains were resistant to multiple drug conditions, which mean that the deleted genes of these strains seem to be necessary for sensitivity to diverse perturbations (See Figure 7).

However, here is the issue that a threshold must be chosen to decide whether a deletion strain is sensitive or resistant to a condition. We use the same sensitivity threshold used by Hillenmeyer *et al.* in their study, $p < 0.01$, but the corresponding z-score, $z > 2.33$ was set as the threshold in our study as we used the fitness score data represented by z-scores. In order to identify resistance interactions, $z < -2.33$ which is the correspondence of upper tail p-value $p > 0.99$, was chosen.

In Algorithm 3, the detail of referring sensitivity and resistance interaction types between deletion strains and compounds are given. The output of the Algorithm 3 is matrix D', where the entity $d'(i, j)$ demonstrates resistance (1), sensitivity (-1) or no interaction (0) information between deletion strain i and compound j.

Once defining sensitivity and resistance interactions between deletion strains and compounds, it is straightforward to find deletion strains that are sensitive or resistant to more than 20% of the unique conditions.

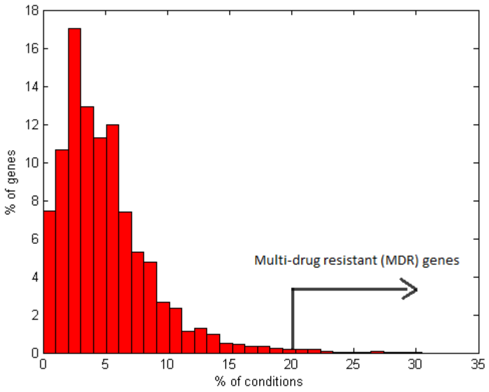


Figure 6: Multi-drug resistance (MDR) genes. x-axis shows the percentage of conditions, whereas y-axis shows the percentage of genes. Genes showing sensitivity to greater than 20% of all unique conditions at z > *2.33*, are assumed as MDR genes.

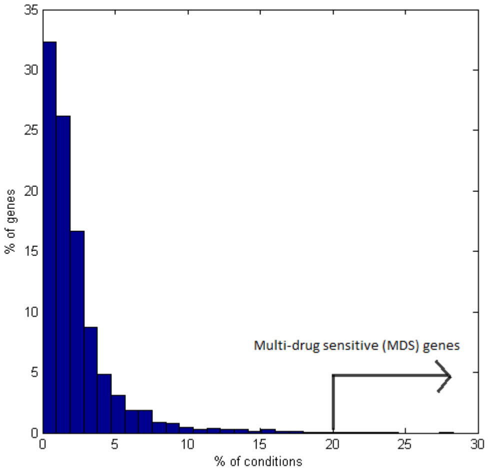


Figure 7: Multi-drug sensitive (MDS) genes. x-axis shows the percentage of conditions, whereas y-axis shows the percentage of genes. Genes showing resistance to greater than 20% of all unique conditions at z < −*2.33*, are assumed as MDS genes.

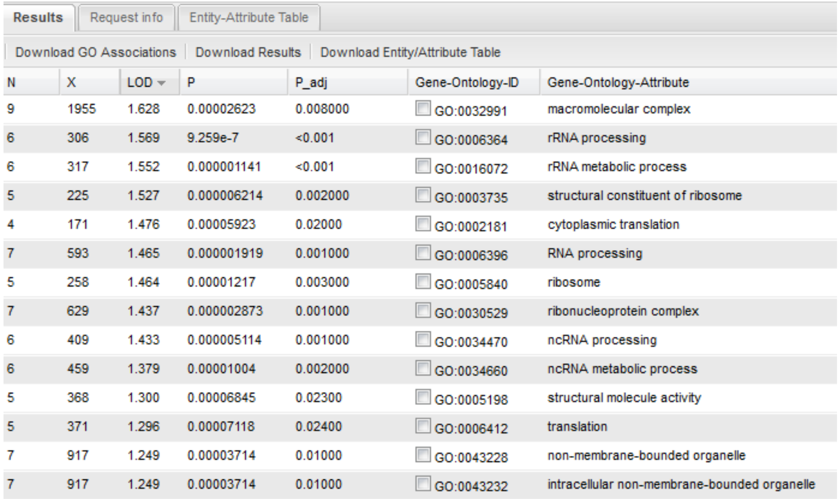
Algorithm 3 Assigning resistance and sensitivity interactions to detect multi-drug resistance (MDR) and multi-drug sensitive (MDS) genes. Fitness defect z-scores of the deletion strains are transformed into discrete numbers at given thresholds for resistance and sensitivity interactions. The output of the algorithm is matrix D', where the entity $d'(i,j)$ demonstrates resistance ($\mathit{1}$), sensitivity ($-\mathit{1}$) or no interaction ($\mathit{0}$) information between deletion strain i and compound j.

Input:

- $z - low$ score
- $z - up$ score
- Matrix D, showing continuous fitness scores (z-scores) of deletion strains in rows against chemical compounds in columns

Output: Matrix D', showing categorical fitness scores of deletion strains in rows against chemical compounds in columns

```
1. For each entity $d(i,j)$ in matrix $D$, do
    (a) if $d(i,j) < z - low$, then assign $\mathit{1}$ to $d'(i,j)$ in matrix $D'$
    (b) if $d(i,j) > z - up$, then assign $-\mathit{1}$ to $d'(i,j)$ in matrix $D'$
    (c) else, assign $\mathit{0}$ to $d'(i,j)$ in matrix $D'$
```



N	X	LOD	P	P_adj	Gene-Ontology-ID	Gene-Ontology-Attribute
9	1955	1.628	0.00002623	0.008000	□ GO:0032991	macromolecular complex
6	306	1.569	9.259e-7	<0.001	□ GO:0006364	rRNA processing
6	317	1.552	0.000001141	<0.001	□ GO:0016072	rRNA metabolic process
5	225	1.527	0.000006214	0.002000	□ GO:0003735	structural constituent of ribosome
4	171	1.476	0.00005923	0.02000	□ GO:0002181	cytoplasmic translation
7	593	1.465	0.000001919	0.001000	□ GO:0006396	RNA processing
5	258	1.464	0.00001217	0.003000	□ GO:0005840	ribosome
7	629	1.437	0.000002873	0.001000	□ GO:0030529	ribonucleoprotein complex
6	409	1.433	0.000005114	0.001000	□ GO:0034470	ncRNA processing
6	459	1.379	0.00001004	0.002000	□ GO:0034660	ncRNA metabolic process
5	368	1.300	0.00006845	0.02300	□ GO:0005198	structural molecule activity
5	371	1.296	0.00007118	0.02400	□ GO:0006412	translation
7	917	1.249	0.00003714	0.01000	□ GO:0043228	non-membrane-bounded organelle
7	917	1.249	0.00003714	0.01000	□ GO:0043232	intracellular non-membrane-bounded organelle

Figure 8: Gene Ontology (GO) enrichment test for multi-drug sensitive genes (MDS). Results are obtained by using the tool: FuncAssociate 2.0 [3]. MDS genes are found highly enriched for RNA metabolism related functions.

Around 50 MDR genes were found highly enriched for functions related to endosome transport, vacuolar degradation, and transcription in Hillenmeyer *et al.* [1] article. We also found around 10 MDS genes highly enriched for functions related to RNA metabolism (Figure 8). In our further analysis, we discarded MDR and MDS genes as we are interested in compound-specific profiles of genes which show growth phenotype to a limited number of conditions.

4.4 Strain-Drug Network Construction

As in the illustration given in Figure 5, sensitivity and resistance relationships can be thought as a network, where drugs and deletion strains constitute the nodes of the network, and resistance and sensitivity relationships between drugs and deletion strains are the edges of the network. The network itself is a bipartite graph consisting of deletion strain-drug interactions where a deletion strain and a drug are linked to each other with either a resistance or sensitivity relation, whereas there is no edge between any of two deletion strains or between any of two drugs [25]. However, there is again the same issue as in MDR-MDS detection that a threshold must be set to decide whether a growth rate shows sensitivity or resistance. As different from Algorithm 3, we did not set a general threshold to use in every growth rate. Since we would subsequently construct a drug similarity network (See 4.5) in the next step, we wanted to make all compounds equally related to the deletion strains. By doing this, the suggesting results that come from drug similarity network would have equal priors.

In Algorithm 4, the detail of referring sensitivity and resistance interaction types between deletion strains and compounds are given. Likewise the Algorithm 3, the output of the Algorithm 4 is matrix D', where the entity $d'(i,j)$ demonstrates resistance (*1*), sensitivity (*−1*) or no interaction (*0*) information between deletion strain i and compound j.

By performing Algorithm 4, a deletion strain is assigned as sensitive or resistant to a condition if its fitness defect z-score is in the top one hundred fitness scores or the lowest one hundred fitness scores, respectively of that particular condition. The top or lowest *one hundred* z-scores per drug were chosen because by choosing less z-scores such as the top or lowest *fifty* z-scores per drug, we could not obtain significantly biologically meaningful higher order motifs whose details are given in Section 4.4.1. Besides, in Figure 9, the top and lowest one hundred fitness growth z-scores per drug are shown. As seen from the figure, certain drugs have high difference between chosen z-scores, whereas the other drugs have less.

Algorithm 4 Assigning resistance and sensitivity interactions to construct the deletion strain-drug network. A deletion strain is assigned as sensitive or resistant to a condition if its fitness defect z-score is in the top one hundred fitness scores or the lowest one hundred fitness scores, respectively of that particular condition. The output of the algorithm is matrix D', where the entity $d'(i,j)$ demonstrates resistance (*1*), sensitivity ($-$*1*) or no interaction (*0*) information between deletion strain i and compound j.

Input:

- Matrix D, showing continuous fitness scores (z-scores) of deletion strains in rows against chemical compounds in columns

Output: Matrix D', showing categorical fitness scores of deletion strains in rows against chemical compounds in columns

```
1. for each column (condition) $j$ in matrix $D$, do
    (a) Add the indexes of the highest 100 z-scores in column $j$ to the set $S$
    (b) for each index $i$ in $S$
        i. Assign $-1$ to the entity$d'(i,j)$ of the matrix $D'$
    (c) Add the indexes of the lowest 100 z-scores in column $j$ to the set $S'$
    (d) for each index $i$ in $S'$
        i. Assign 1 to the entity$d'(i,j)$ of the matrix $D'$
    (e) Assign 0 to the remaining elements of $j$. column of $D'$matrix
```

4.4.1 Quality assessments

We did two quality assessments pertaining to the deletion strain-drug network in order to qualify the robustness of the network.

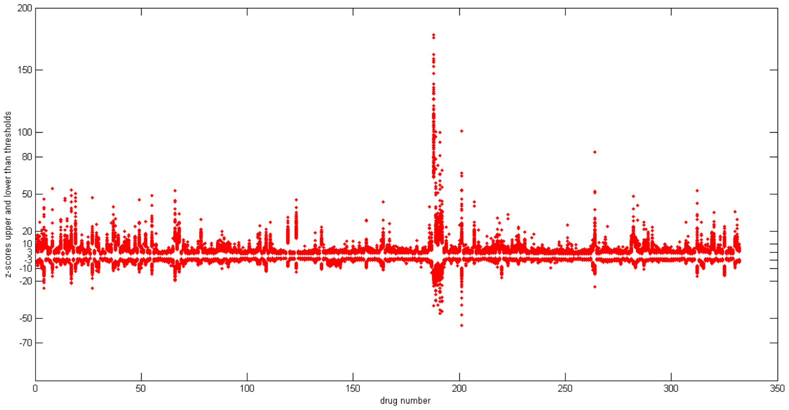


Figure 9: The top and lowest one hundred fitness z-scores per drug which are considered sensitive and resistant interactions, respectively obtained by performing Algorithm 4. Every red dot represents a z-score. x-axis shows the drug number, whereas y-axis shows the values of the top and lowest one hundred z-scores of the corresponding drug. As seen, certain drugs have high difference between chosen z-scores, whereas the other drugs have less.

First, as in Figure 10, for each drug pair, we counted the number of deletion strains that 2 drugs have resistance (R) relationships with, or sensitivity (S) relationships with, in order to investigate whether there is a relation between resistance and sensitivity interactions.

Second, we wanted to understand if any of the higher order motifs in this network is enriched or depleted. We counted the number of motifs, consisting of 2 drugs and 2 deletion strains, and R or S relationships between them in the real deletion strain-drug network. The motifs are demonstrated in Figure 11. In the first motif, drug1 and drug2 have resistance edges to both of the strains, whereas in the seventh motifs, drug1 and drug2 have sensitivity edges to both strains. In the second motif, drug2 has resistance edges to both of the strains, but drug1 has sensitivity relationship to strainA,

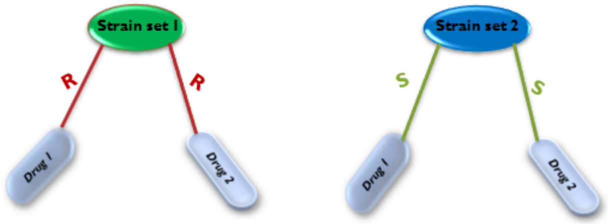


Figure 10: Motifs that are used in quality assessment 1. In this example, if the deletion strains in strain set1 are resistant to both of drug1 and drug2, the number of deletion strains in strain set1 constitutes the RR motif number for drug1-drug2 pair. SS motifs for drug1-drug2 pair can be counted similarly. Overall, this counting is done for all drug pairs.

and resistance relationship to strainB. In the third motif, both drug1 and drug2 have resistance edges to strainA, but have sensitivity edges to strainB. In motif *4*, drug1 has sensitivity relationship to strainA, and has resistance relationship to strainB, but drug2 has opposite relations to these strains. In the fifth motif, drug1 has sensitivity relationships to both of the strains, but drug2 has resistance edges to them. On the other hand, in motif *6*, drug1 has resistance edge to strainA but sensitivity edge to strainB, whereas drug2 has sensitivity edges to both of the strains.

We analyzed these motifs statistically. For example, if drug1 has an R (resistance) relationship with deletion strainA and an S (sensitivity) relationship with deletion strainB, we expect to observe that another drug which has an R relationship with deletion strainA, also has an S relationship with deletion strainB. We did such analysis by comparing the observations of motifs in the real network to those in randomized versions of the network. The processes are demonstrated with the flowchart given in Figure 4.

Network randomization was done by edge swapping with the following steps:

- For *700000* iterations
  - Choose *2* edges from the bipartite network randomly
  - Swap the endpoints of the edges

700000 times edge swapping were done which are approximately *10* times of total edge number of the network. Eventually, one random network was constructed. Network topology remained same since edge distribution, the numbers of sensitivity and resistance edges per drug and per deletion strain were still same. We did such randomization *1000* times to identify the significance of the higher order motifs in the whole network. *fold enrichment* value which shows the enrichment and depletion values of the motifs in comparison to the randomized versions of the network, and empirical *p-value* for each higher order motif were calculated as below:

$$foldenrichment_i = s_{real}(Motif_i)/[(\sum_{j \in R} s_{random_j}(Motif_i))/N] \tag{4.2}$$

$$p_i = \begin{cases} length\{s_{random_j}(Motif_i) \geq s_{real}(Motif_i)\}/N & foldenrichment \geq 1, \\ length\{s_{random_j}(Motif_i) \leq s_{real}(Motif_i)\}/N & foldenrichment < 1 \end{cases} \tag{4.3}$$

where

- $i = 1, 2, 3, .., 7$ denotes the higher order motif number
- p_i is the p-value of $Motif_i$
- $foldenrichment_i$ is the fold enrichment value of $Motif_i$
- $s_{real}(Motif_i)$ is the number of occurrences of $Motif_i$ in the real network
- $s_{random_j}(Motif_i)$ is the number of occurrences of $Motif_i$ in the j^{th} random network
- R is the set of random networks
- N is the size of the set R that equals to the number of random networks
- $length\{s_{random_j}(Motif_i) \geq s_{real}(Motif_i)\}$ is the count for cases where the number of occurrences of $Motif_i$ in randomized network j is bigger than or equal to the number of occurrences of $Motif_i$ in the real network
- $length\{s_{random_j}(Motif_i) \leq s_{real}(Motif_i)\}$ is the count for cases where the number of occurrences of $Motif_i$ in randomized network j is smaller than or equal to the number of occurrences of $Motif_i$ in the real network

Comparison with the conventional p-value *0.05* is not enough here to decide significance level of the empirical p-values since multiple hypothesis were taken into account. There are *7* motifs which may either be depleted or enriched, hence, *14* hypothesis were compared in this analysis. For multiple comparison, Bonferroni Correction method given in Section 2.3.5 was used to find cutoff p-value that gives $0.05/14 = 0.0035$ rather than conventional *0.05*. Moreover, since $1/0.0035 = 285$, *285* randomizations are enough to test the significance of an obtained p value. However, we made *1000* randomizations.

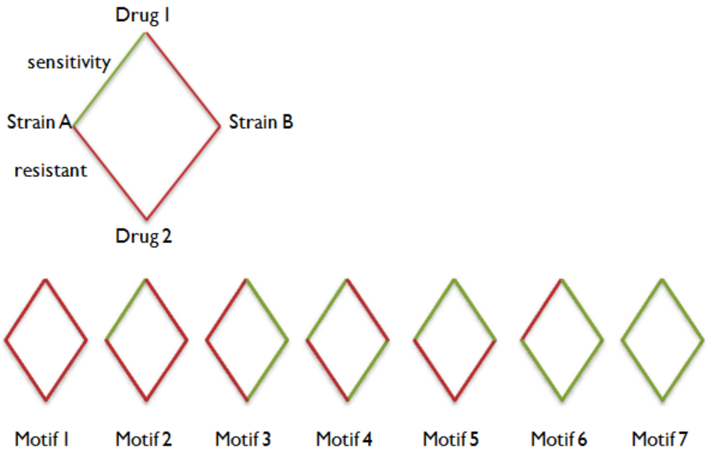


Figure 11: The higher order motifs of the deletion strain-drug network. *7* different higher order motifs can be defined by using *2* drugs, *2* deletion strains and *2* types of edges. In the first motif, drug1 and drug2 have resistance edges to both of the strains, whereas in the seventh motifs, drug1 and drug2 have sensitivity edges to both strains. In the second motif, drug2 has resistance edges to both of the strains, but drug1 has sensitivity relationship with strainA, and resistance relationship with strainB. The other motifs can also be analyzed similarly.

4.5 Drug-Drug Similarity

As the next step, we subsequently generated a drug similarity network from the bipartite deletion strain-drug graph. If a drug pair has frequently S or R relationships to the same set of deletion strains, we assume this drug pair as *similar* to each other in the drug similarity network. On the other hand, if a drug pair has inverse relationships to the same set of deletion strains, such as one of the drugs has R relations, whereas the other has S relations to the same deletion strains, we consider the drugs in this pair different to each other. By doing this, we built up the drug similarity network shown in Figure 14, which demonstrates the similarities and dissimilarities between drugs.

However, in order to be able to consider two drugs similar or different to each other, a threshold must be used to decide similarity or dissimilarity because it is not enough to say that a drug pair is similar or different to each other by only observing limited number of shared deletion strains. Therefore, we randomized the deletion strain-drug network, and calculated a p-value and a fold enrichment value for each drug pair in the network as in Equation 4.3 and Equation 4.2, respectively. However, when assessing significance for each drug pair, *1000* randomizations as in higher order motifs were not enough since there are much more hypothesis to test. As there are ~ 330 chemical compounds used in Hillenmeyer *et al.* dataset, we have ~ 55000 drug pairs to test which reduce the significant cutoff p-value substantially, leading to much more randomizations. In detail, since we have ~ 55000 hypothesis to compare, $0.05/55000 = 9.09e - 7$ must be the significant cutoff p-value. In order to be able to compare to $9.09e - 7$ significant level, at least $1/9.09e - 7 = 1100000$ randomizations must be done. On the other hand, more than *1000* randomizations of the deletion strain-drug network and counting the number of motifs for each drug pair are computationally intensive processes. However, after randomizations, almost same number of motifs were observed for each drug pair in the network due to the fact that we defined same number of sensitivity and resistance interactions for each drug pair in Algorithm 4 which also means that all drug pairs have same number of links to each other, hence, lose their specificity in the network, leading to a type of random network. Therefore, we concluded that if we randomized the real network *500* times, we would have actually $\sim$ *27000000* randomizations owing to having $\sim$ *55000* drug pairs, meaning that each drug pair is one randomization result. This intuition provides us to get much more randomization results in a computationally non-intensive way.

Then, p-value and fold enrichment value for each drug pair were calculated by comparing the motif number of similar and different edges in the real network to number of those in $\sim$ *27000000* random networks.

Eventually, we converted strain-drug network to a drug similarity network by comparing the p-value and fold enrichment value of each similar and different interactions of each drug pair to $< 9.09e-7$ and $>$ *5*, respectively.

4.5.1 Quality assessment

Once the drug similarity network was constructed, as a quality assessment, we again wanted to understand if any of the higher order motifs in this network is enriched or depleted. However, this time we counted the number of motifs, consisting of *3* drugs linked with similar or different relationships as shown in Figure 12. We did such analysis by comparing the observations of motifs in the real network to those in randomized versions of the network as shown with the flowchart given in Figure 4. However, network randomization of the drug similarity network was done with the following steps:

- For *10000* iterations
    - Choose *2* edges, both similar or both different, from the drug

similarity network randomly

- Swap the endpoints of the edges

10000 times edge swapping were done which are approximately *10* times of total edge number of the drug similarity network. Eventually, one random network was constructed. Network topology still remained same. We did such randomizations *1000* times, and calculated a p-value and a fold enrichment value for each of the higher order motif by using the equations given in Equation 4.3 and Equation 4.2, respectively. Comparison with the conventional p-value *0.05* is again not enough here to decide significance level of the empirical p-values since multiple hypothesis were taken into account. There are *4* motifs which may either be depleted or enriched, hence, *8* hypothesis were compared in this analysis. For multiple comparison, Bonferroni Correction method given in Section 2.3.5 was used to find cutoff p-value that gives $0.05/8$ = 0.006 rather than conventional *0.05*. Moreover, as $1/0.006 = 158$, *158* randomizations are enough to calculate the empirical p-value. However, we again made *1000* randomizations as in randomization of deletion strain-drug network.

Besides, the randomized versions of the real network which are used to calculate the significance of subgraphs with *3* nodes (motifs), also keep the same number of occurrences of all subgraphs with 2 nodes of the real network [26].

4.5.2 Chemical structural similarity

Once we found similarities and dissimilarities between drugs by inferring from the drug similarity network, we wanted to compare these similarities with the

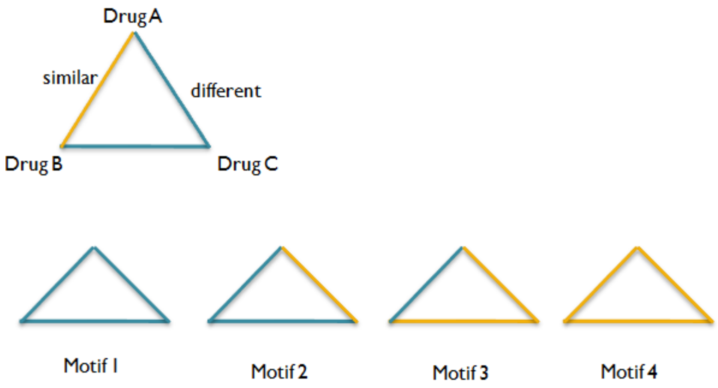


Figure 12: The higher order motifs of the drug similarity network. *4* different higher order motifs can be defined by using *3* drugs and *2* different types of edges. In the first motif all drugs are different from each other, whereas in the fourth, all of them are similar. The other motifs can also be analyzed similarly.

chemical structural similarities of the chemical compounds.

Molecular fingerprints are one of the properties to encode the structure of a molecule. A series of binary digits (bits) are the widely used type of fingerprints format that represent the presence or absence information of certain substructures in the molecule. The similarity between two molecules can be identified by comparing their fingerprints.

We represented each chemical compound in SMILES (Simplified Molecular Input Line Entry System) strings [37] in order to analyze their chemical structure. We used *Pybel* which is a Python module to access OpenBabel toolkit [38], to calculate the fingerprints of chemical compounds represented by SMILES. Each chemical compound was defined with *3* different binary vectors, corresponding to *3* types of fingerprints formats provided by Python *Pybel* module:

- FP2 is a path-based fingerprint that investigates linear segments of size *1* to *7* atoms
- FP3 uses the SMART strings in patterns.txt
- FP4 uses the SMARTS strings in SMARTS_InteLigand.txt

Even though one can add its own queries to the above stated files (patterns.txt etc.), we did not add any additional substructure to these files.

After calculating fingerprints binary data for each of the chemical compound, we used Tanimoto coefficient to calculate structural similarity between drug pairs. However, another metrics such as Hamming distance, Dice coefficient [12] and so on, are able to be used.

4.5.3 Side effects similarity

Then, we wanted to compare similarities found in the drug similarity network to side effect similarities of chemical compounds.

In order to obtain the side effect information pertaining to chemical compounds, we used meddra adverse effects data from Side Effect Resource, SIDER 2 [39]. The number of common chemical compounds between Hillenmeyer *et al.* compounds and SIDER compounds are *53*. SIDER provides *4199* different side effects related to chemical compounds.

We formed a vector for each of *53* chemical compounds that consists of binary values, *0* or *1* where *1* represents the presence of a particular side effect, and *0* represents the absence of it. The distance for each drug pair was subsequently calculated by simply counting the shared side effects between two drugs. Then, we compared side effect similarities of drug pairs to the number of similar or different edges between drug pairs.

5 Results & Discussions

5.1 Verification of Deletion Strain-Drug Network

In the deletion strain-drug network, there are *6013* nodes which comprise *5681* deletion strains and *332* conditions, consisting of *326* chemical compounds and *6* environmental stress conditions. The edge number of the network is *66400* where half of them are resistance edges and the other half are sensitivity edges.

As a result of quality assessment *1*, if *2* drugs have an S relationship with the same set of deletion strains, we observed that these *2* drugs will also have an R relationship with another set of deletion strains, compared to the random. We called these types of relations as SS and RR motifs (See Figure 10) that the correlation result and plot for RR and SS motifs of drug pairs are given in Table 2 and Figure 13, respectively. This observation proves the biological meaning of resistance interactions.

The results for quality assessment *2* are given in Table 3, where 'fold enrichment' value shows the enrichment and depletion values of the motifs in comparison to the randomized versions of the network. We found all p-values equal to *0*. We, however, are able to only guarantee that p-values are lower than 10^{-3}due to making *1000* randomizations. As all of p-values are lower than the cutoff p-value *0.0035* calculated in Section 4.4.1, we can say that all of the higher order motifs are significant.

In the left half of *Motif1* and *Motif7*, *2* drugs have same response to the same strain as shown in Figure 11. Thus, in the right half part, we expect to see the same response, too. As both of the drugs have also resistance edges and sensitivity edges to the same strain in the right half of *Motif1* and *Motif7*, respectively, we expect these motifs to be enriched. However, in the left half of *Motif2* and *Motif6*, drugs show different responses to the same deletion, hence, we expect to see these different responses in the right half, too. However, in the right half parts of these motifs, both drugs show same responses. Therefore, we expect these motifs to be depleted. In the left half of *Motif3*, both drugs have same relationships to the same strain, hence, in the right half part, we expect to see this same response. Since both drugs have sensitivity edges to the same strain in the right half part of *Motif3*, as we expected, it is enriched. On the other hand, in the left half parts of *Motif4* and *Motif5*, drugs have different responses. Therefore, we expect to observe these different responses in the right half of these motifs, too. As both drugs have also inverse relationships in the right, these motifs are enriched.

In conclusion, all of the values pertaining to the higher order motifs prove that this deletion strain-drug network is robust, and make biological sense about drug mechanism of action.

correlation	**p-value**
0.371	$< 10^{-4}$

Table 2: Correlation between SS and RR motifs of drug pairs as a result of quality assessment 1

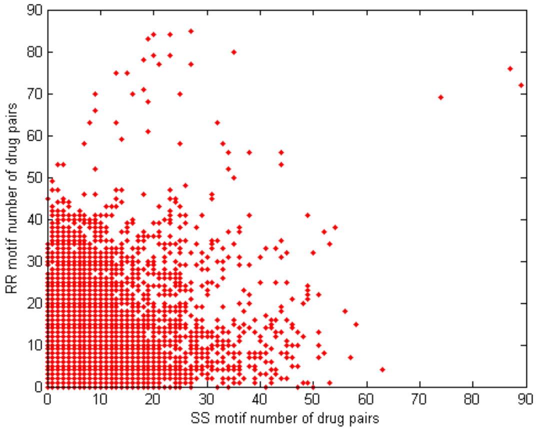


Figure 13: Correlation plot for SS and RR motifs of drug pairs. Every dot in the plot represents a single drug pair. x-axis demonstrates the SS motif number of the corresponding drug pair (the number of deletion strains that both drugs in the pair have sensitivity relationships with), whereas y-axis shows the RR motif number of that drug pair (the number of deletion strains that both drugs in the pair have resistance relationships with).

	Motif *1*	**Motif** *2*	**Motif** *3*	**Motif** *4*	**Motif** *5*	**Motif** *6*	**Motif** *7*
p-value	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$
fold enrichment	*2.270*	*0.951*	*1.777*	*1.084*	*1.491*	*0.898*	*1.991*

Table 3: Quality assessment 2 result for higher order motifs in the deletion strain-drug network. The occurrences of higher order motifs in the real network are compared to the occurrences of those in random networks. The fold enrichment value for each motif which represents how many times the corresponding motif is enriched or depleted in the real network, indicates that the higher order motifs are biologically meaningful.

5.2 Drug Similarity Network

There are totally *162* nodes (drugs), *504* similar edges and *317* different edges in the drug similarity network. Yellow edges represent similar relationship between drug pairs, whereas blue edges represent different relationship. In Figure 14, the blue and purple circular nodes show the same drug in different concentration levels. As expected, they were found similar to each other in our network. The green diamond nodes in the right upper connected component form a clique, and show drugs found similar to each other even though their chemical structures based on FP2 fingerprint are different ($\textit{Tanimoto coefficient} < 0.2$) except 2 drug pairs. However, it is noteworthy to state that their chemical structures based on FP3 and FP4 fingerprints are similar for all drug pairs in that clique ($\textit{Tanimoto coefficient} \geq \textit{0.2}$). 3-vertex, 4-vertex and 5-vertex cliques found in connected components except the giant component, are shown with different colored and shaped nodes in this figure. The correlation result and plot for similar and different edges between drug pairs are given in Table 4 and Figure 15, respectively.

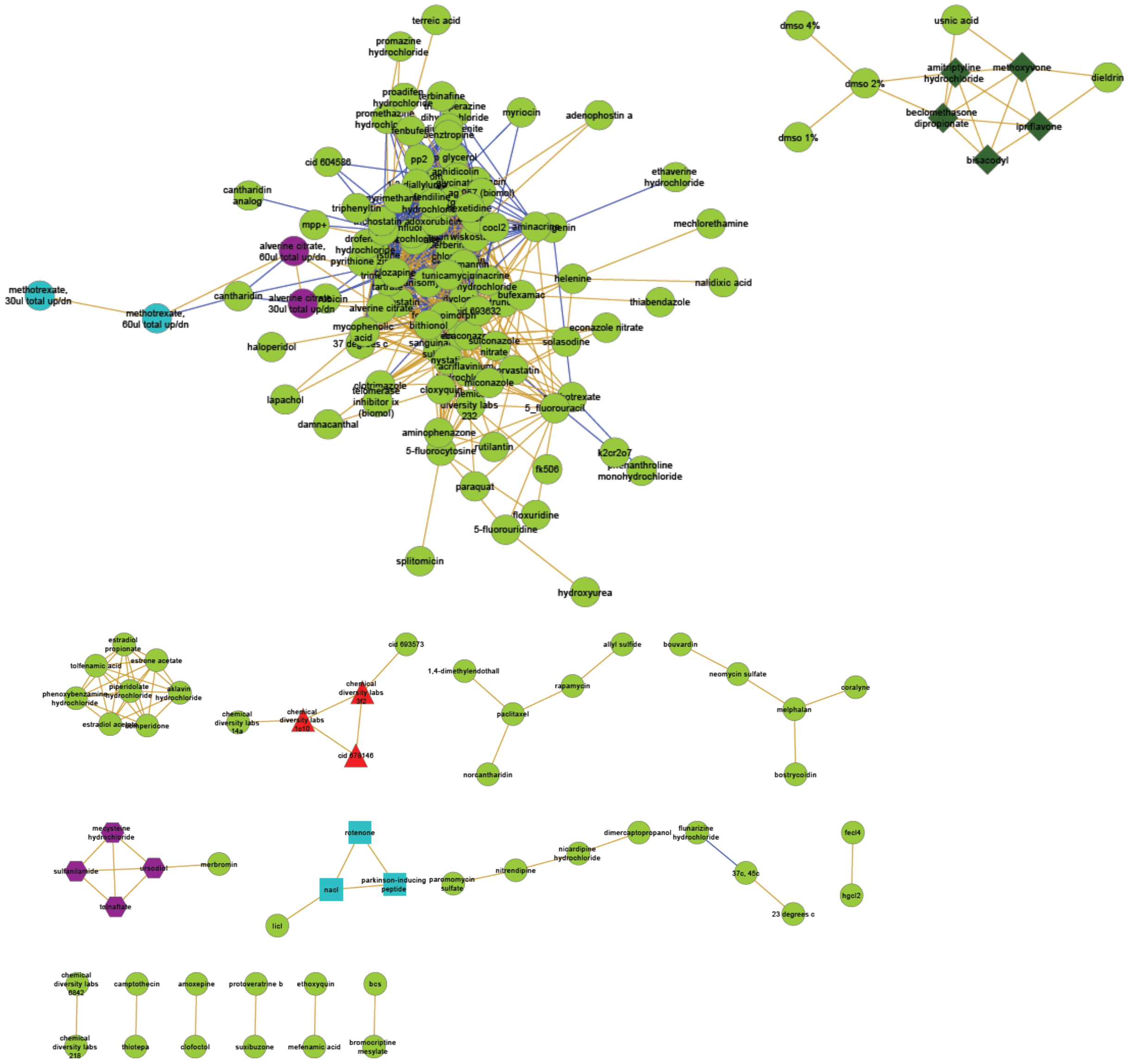


Figure 14: Drug similarity network. Force directed layout is used to draw the network. Nodes are compounds. Yellow edges represent similar relationship between drug-pairs, whereas blue edges represent different relationship. The blue and purple nodes with circle shape show the same drug in different concentration levels. The drugs found in different cliques of connected components except the giant component, are represented with different colors and shapes, spanning red triangles, blue squares, purple hexagons and dark green diamonds.

correlation	**p-value**
-0.110	$< 10^{-4}$

Table 4: Correlation between similar and different edges between drug pairs in the drug similarity network

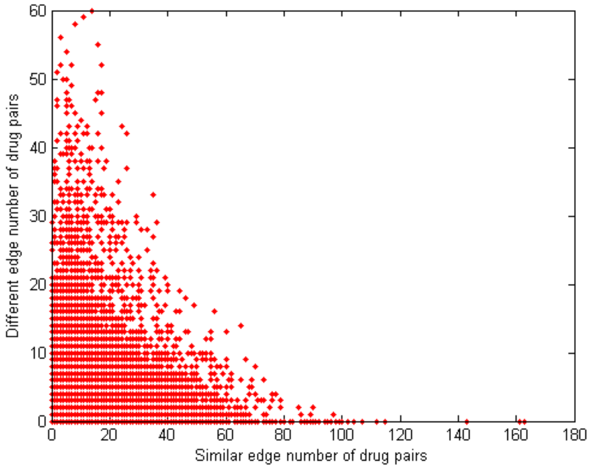


Figure 15: Correlation plot for similar and different edges between drug pairs in the drug similarity network. Every red dot in the plot represents a drug pair. x-axis shows the similar edge number between drugs in the corresponding drug pair, whereas y-axis shows the different edge number between them.

The results for quality assessment *2* are given in Table 5 where ‘fold enrichment’ value shows the enrichment and depletion values of the motifs in comparison to the randomized versions of the network. Likewise the deletion strain-drug network, we again found all p-values equal to *0*. However, we are able to only guarantee that p-values are lower than 10^{-3}due to making *1000* randomizations. As all of p-values are lower than the cutoff p-value *0.006* calculated in Section 4.5, we can say that all of the higher order motifs are significant.

In order to interpret the significance of fold enrichment values, first, we can simply think that drugA and drugC in Figure 12 are similar to each other. If drugA is different from drugB, we expect drugC to be different from drugB. As we expect, *Motif*2 is enriched. However, we do not expect drugC to be similar to drugB. Therefore, as we expected, it is depleted in *Motif*3. Moreover, we expect *Motif1* and *Motif4* to be enriched because all of the edges are in the same type. As *Motif1* could not be observed in the real network even though it was encountered in randomized versions of the real network, fold enrichment value for *Motif1* was found as *0*.

	Motif *1*	**Motif** *2*	**Motif** *3*	**Motif** *4*
p-value	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$
fold enrichment	*0.0*	*6.926*	*0.059*	*4.650*

Table 5: Quality assessment result for higher order motifs of the drug similarity network. The occurrences of higher order motifs in the real network are compared to the occurrences of those in random networks. The fold enrichment value for each motif which represents how many times the corresponding motif is enriched or depleted in the real network, indicates that the higher order motifs are biologically meaningful. The fold enrichment value for Motif *1* is *0* because it is not observed in the real network even though it is encountered in random networks.

In conclusion, all of these observations pertaining to the drug similarity network prove the robustness of the network. Network visualization of drug similarity network supplies an important source to understand drug-drug relations and drug mechanism of action [25].

5.2.1 Network properties

The plots of degree distribution function, $P(k)$ (See Section 2.4.1) shown in Figure 16 and the clustering coefficient function, $C(k)$ (See Section 2.4.2) shown in Figure 17 demonstrate neither random network's properties nor scale-free network's properties exactly. However, as we mentioned in Section 2.4.3, the majority of drugs may have a few interactions, whereas the minority of them are highly connected, representing network hubs as shown in Figure 16. Moreover, the degree distribution $P(k)$ approximates a power law, $P(k) \sim k^{-\gamma}$ where γ equals to *2.64*. Since $2 < \gamma < 3$, the hubs are relevant and the largest hub is linked with a small number of nodes [4]. Therefore, we are able to strongly conclude that the drug-similarity network is a scale-free network.

Maximum degree is *47* which means the hub that has the largest link number, has connections to *47* drugs. It proves the above statement that the largest hub is connected to small fraction of all nodes. The largest hub of the network is *berberine chloride.* Properties of the largest*10* hubs are given in Table 6. We could not observe any relation between molar mass of a compound and being a hub. However, their chemical structures, especially based on FP3 fingerprints format, are highly similar where Tanimoto coefficients between drug pairs are at least *0.2*, most of the times $>$ *0.4*. Moreover, if we sort the drugs according to the PageRank, the top *10* PageRank belong to the largest *10* hubs because unlike in the directed networks, in undirected networks, PageRank is proportional to degree of nodes, as a result, does not give any information related to considerable nodes [32]. On the other hand, the projection of these drugs within the network is demonstrated in Figure

21. In the below network of Figure 21, the complete subgraph of these drugs are able to be seen clearly that means all nodes representing the largest *10* hubs of the drug similarity network, have also connections between each others.

On the other hand, minimum degree of the network is *1*, whereas the average degree $< k >$ is $(\mathit{2} * L)/N = (2 * 821)/162 = 10.14$ where L and N denote the edge number and the node number, respectively (For more detail, see Section 2.4.1).

There are *99* drugs in the largest connected component of the network which is also called giant component [25]. This giant component size of the real network is significantly smaller than those of randomized versions with *fold enrichment*$= 0.62$ and $p < \mathit{10}^{-\mathit{3}}$. This result suggests that certain drugs form local clusters within the network.

As seen in Figure 18, maximum clustering coefficient of the network is *1*, whereas the minimum is *0*. On the other hand, the average clustering coefficient, $< C >$ of the network is *0.46* that identifies the overall aptitude of nodes to generate clusters [4]. The average clustering coefficient, $< C >$ of the real network is significantly larger than those of randomized versions with *fold enrichment*$= 3.20$ and $p < \mathit{10}^{-\mathit{3}}$. Therefore, the drug similarity network can be considered as *small-world* that there are a few links between each pair of nodes in the network (See Section 2.4.7).

Betweenness centrality and PageRank scores of the drug similarity network are given in Figure 19 and Figure 20, respectively that maximum betweenness centrality and PageRank score were found as *687.84* and *2.89*, respectively which belong to the largest hub of the network, *berberine chlo-*

ride, as expected.

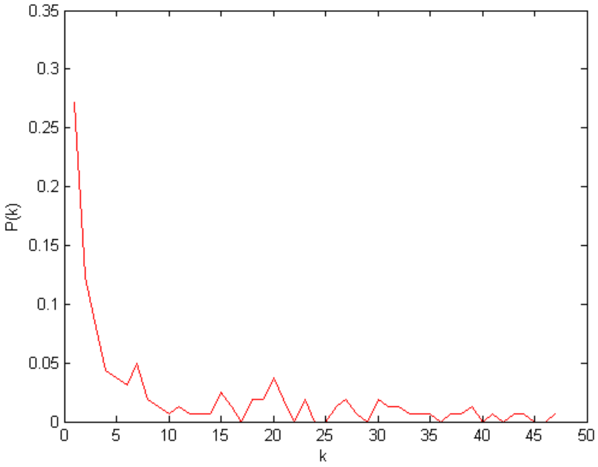


Figure 16: P(k): Degree distribution of a selected node with k links in the drug similarity network. The majority of drugs have a few interactions, whereas the minority of them are highly connected, representing network hubs. The degree distribution $P(k)$ approximates a power law, $P(k) \sim k^{-\gamma}$ where γ equals to $\mathit{2.64}$. Since $\mathit{2} < \gamma < \mathit{3}$, the hubs are relevant and the largest hub is linked with a small number of nodes [4]. Therefore, we are able to strongly conclude that the drug-similarity network is a scale-free network.

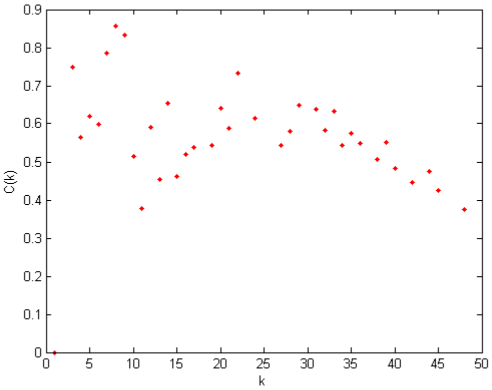


Figure 17: C(k): Average clustering coefficient of the nodes with k links in the drug similarity network

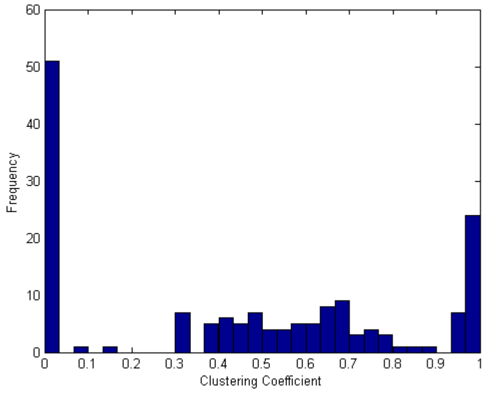
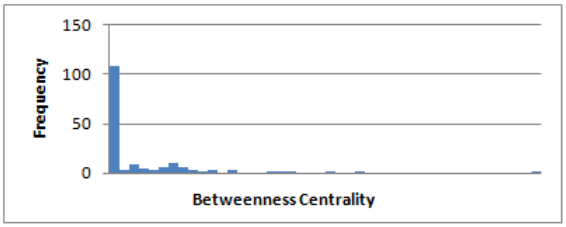
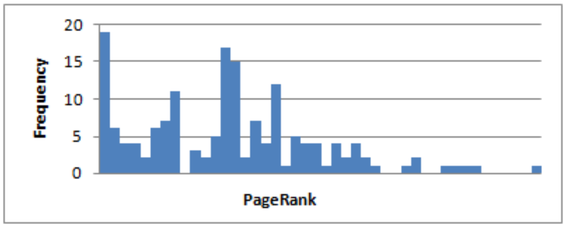


Figure 18: Frequency of clustering coefficient of each node in the drug similarity network. Maximum clustering coefficient of the network is *1*, whereas the minimum is *0*. On the other hand, the average clustering coefficient, $< C >$ of the network is *0.46* that identifies the overall aptitude of nodes to generate clusters [4]. The average clustering coefficient, $< C >$ of the real network is significantly larger than those of randomized versions with *fold enrichment*= 3.20 and $p < 10^{-3}$. Therefore, the drug similarity network can be considered as *small-world* that there are a few links between each pair of nodes in the network.



Minimum Betweenness Centrality	*0.00*
Maximum Betweenness Centrality	*687.84*
Average Betweenness Centrality	*42.37*

Figure 19: Betweenness centrality of the drug similarity network



Minimum PageRank	*0.20*
Maximum PageRank	*2.89*
Average PageRank	*1.00*

Figure 20: PageRank frequency of the drug similarity network

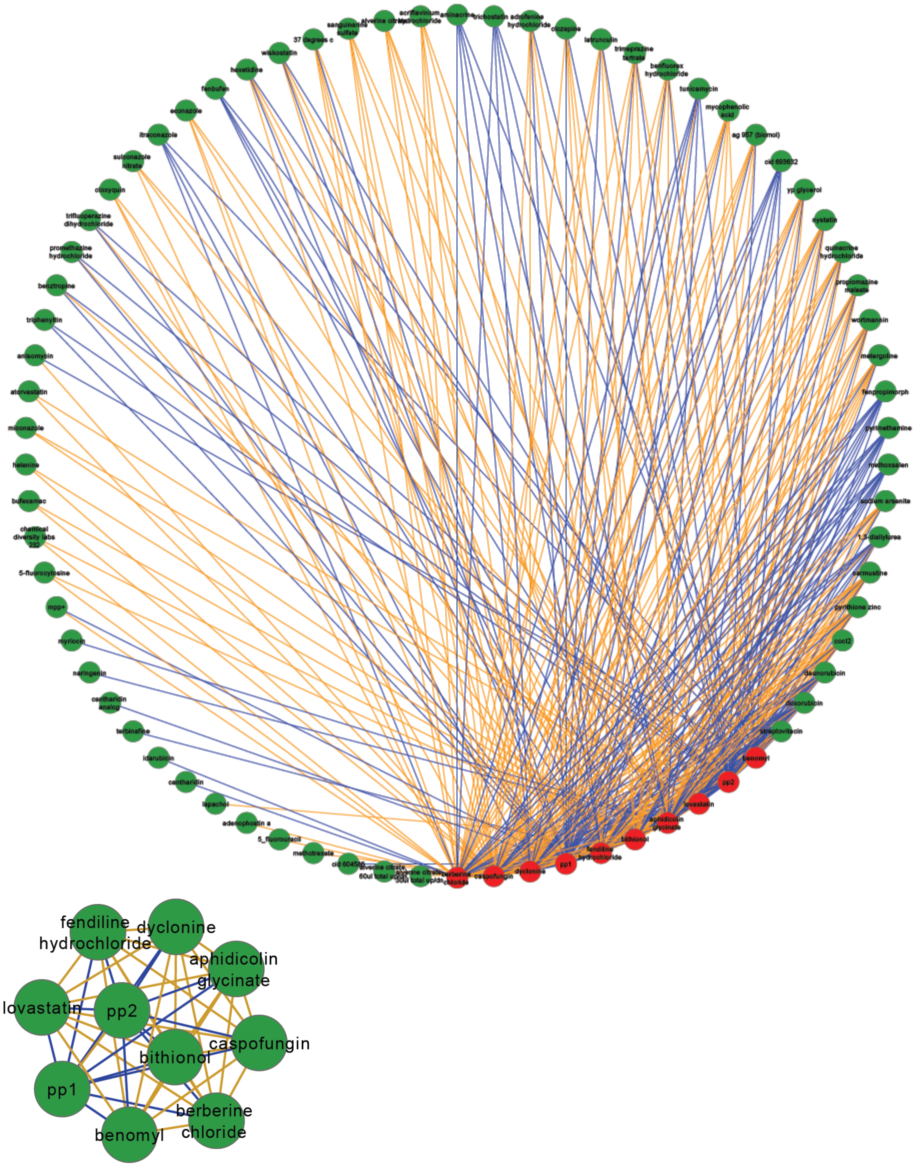


Figure 21: Subgraph of the largest **10** hubs in the drug similarity network. Degree sorted circle layout is used to draw the network. Nodes are compounds. Yellow edges represent similar relationship between drug-pairs, whereas blue edges represent different relationship. In the above network, the drugs representing the largest *10* hubs are colored with red, and all the connections of these drugs are given. In the below network, a complete subgraph of these drugs is given where all drugs are connected to each other.

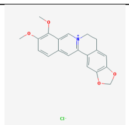
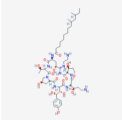
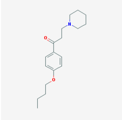
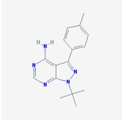
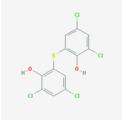
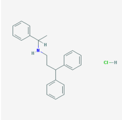
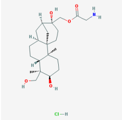
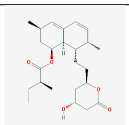
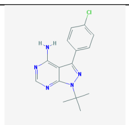
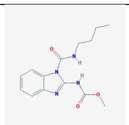
Drug name	PubChem (CID)	Degree	2D Structure	Molecular formula	Molar mass	PageRank
berberine chloride	*12456*	47		$C_{20}H_{18}ClNO_4$	*371.81g/mol*	*2.89*
caspofungin	*151068*	44		$C_{52}H_{88}N_{10}O_{15}$	*1093.31g/mol*	*2.57*
dyclonine	*3180*	43		$C_{18}H_{27}NO_2$	*289.41g/mol*	*2.46*
pp1	*1400*	41		$C_{16}H_{19}N_5$	*281.36g/mol*	*2.45*
bithionol	*2406*	39		$C_{12}H_6Cl_4O_2S$	*356.05g/mol*	*2.38*
fendiline hydrochloride	*5702162*	39		$C_{23}H_{26}ClN$	*351.91g/mol*	*2.16*
aphidicolin glycinate	*130315*	38		$C_{22}H_{38}ClNO_5$	*431.99g/mol*	*2.17*
lovastatin	*53232*	37		$C_{24}H_{36}O_5$	*404.54g/mol*	*2.09*
pp2	*4878*	35		$C_{15}H_{16}ClN_5$	*301.77g/mol*	*1.93*
benomyl	28780	34		$C_{14}H_{18}N_4O_3$	*290.32g/mol*	*1.87*

Table 6: Properties of the largest *10* hubs of the drug similarity network

5.2.2 An 8-vertex well connected component

As seen in Figure 14, there is a well connected component with 8 nodes in the drug similarity network. Here, we expanded this component by showing the interactions between the drugs in the component and deletion strains as shown in Figure 22, which is a subgraph of deletion strain-drug network. Drugs and deletion strains are nodes of the network. Green edges represent the sensitivity interactions between drugs and deletion strains, whereas red edges represent the resistance edges. The 8 drugs are shown with blue squares. The nodes shown with black circles are deletion strains. However, the shared deletion strains that all 8 compounds have resistance relationships, are shown with pink circles. As the deletion strain-drug network is bipartite, there is not any interaction between deletion strains or drugs itself. The properties of drugs in the component are given in Table 7. The betweenness centrality scores of these drugs are extremely low as expected because these drugs form a component where any additional drug has no access to these drugs.

Aclavine hydrochloride is one of them whose therapeutic indication is infection, and therapeutic uses is as an agent for anti-infection [40]. Estrone is one type of estrogens including estradiol. As an hormone replacement therapy, *estradiol acetate* is used to avoid the indications of menopause in women [41, 42, 43, 44]. On the other hand, *Estrone acetate* is also used in hormonal therapeutics. *Dopamine* takes a role as an antagonist of protein dopamine receptor D2 (one of five G protein-coupled receptors) in homo sapiens [40]. While it is used in treatment of dopamine receptor's physiological effects, it has also been showed that dopamine takes a role in pain processing in central nervous system [45]. Therapeutic indication and therapeutic use of *piperidolate hydrochloride* are pain and analgesic, respectively. In therapeutic indications such as pain and inflammation, *tolfenamic acid* is also used as an analgesic and anti-inflammatory agent. *Phenoxybenzamine hydrochloride* takes a role as an antagonist of proteins with alpha-adrenergic receptor activity. It is used as an anti-hypertensive agent in hypertension indication [40].

Consequently, we are able to infer that the drugs in this well connected component are somehow share certain therapeutics indications and uses which leads us to have intuition that yeast cells show similar response to drugs having similar therapeutics effects (See Section 6).

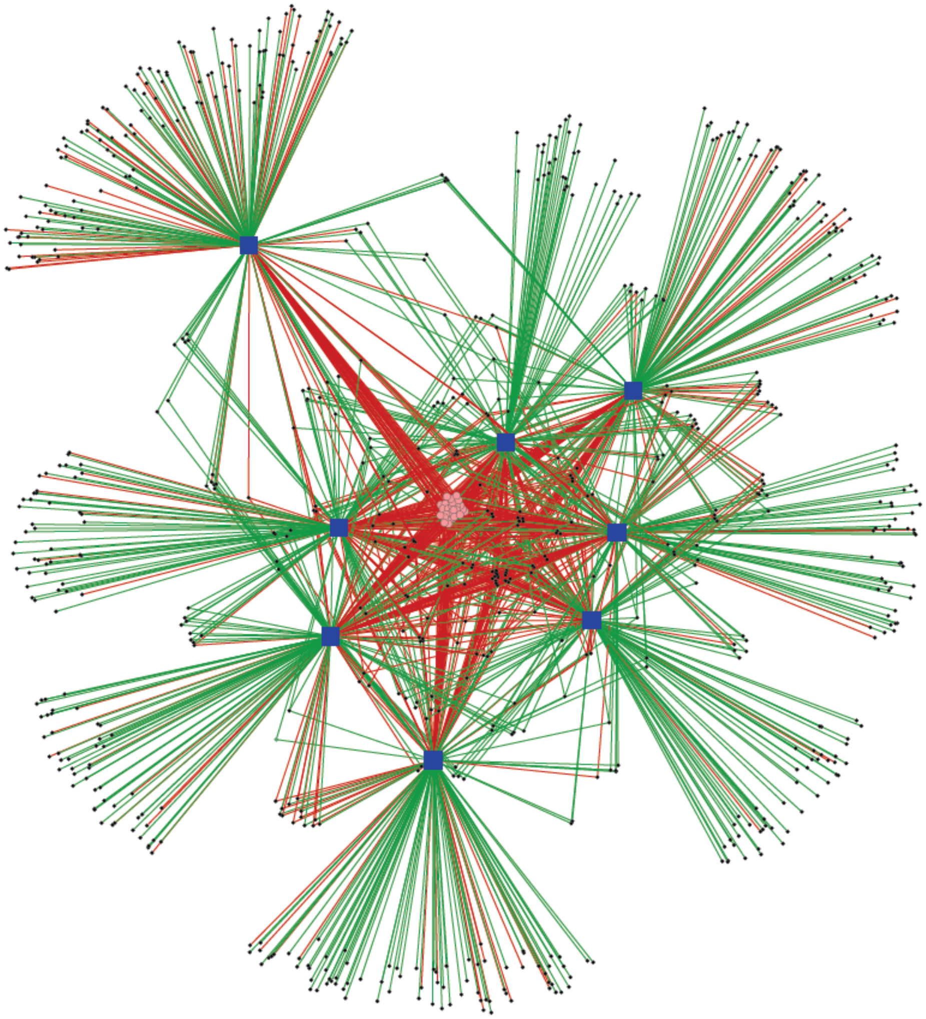


Figure 22: A well connected component with 8 nodes in the drug similarity network. Nodes are drugs and deletion strains of the drug similarity network. Green edges represent the sensitivity interactions between drugs and deletion strains, whereas red edges represent the resistance edges. The 8 drugs in the well connected component are shown with blue squares. The nodes shown with black circles are deletion strains. However, the shared deletion strains that all 8 compounds have resistance relationships, are shown with pink circles.

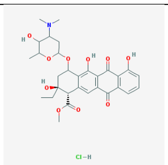
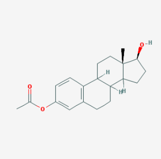
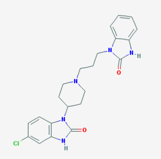
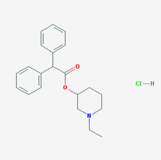
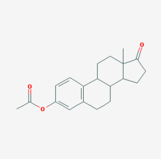
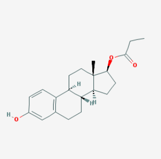
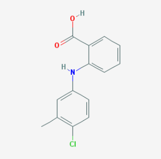
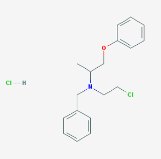
Drug name	**PubChem (CID)**	**2D Structure**	**Molecular formula**	**Molar mass**	**Betweenness centrality**
aklavin hydrochloride	*264889*		$C_{30}H_{36}ClNO_{10}$	*606.06g/mol*	*0.00*
estradiol acetate	*157050*		$C_{20}H_{26}O_3$	*314.42g/mol*	*0.17*
domperidone	*3151*		$C_{22}H_{24}ClN_5O_2$	*425.91g/mol*	*0.17*
piperidolate hydrochloride	*8520*		$C_{21}H_{26}ClNO_2$	*359.89g/mol*	*0.17*
estrone acetate	*3273*		$C_{20}H_{24}O_3$	*312.40g/mol*	*0.17*
estradiol propionate	*19571*		$C_{21}H_{28}O_3$	*328.45g/mol*	*0.17*
tolfenamic acid	*5507*		$C_{14}H_{12}ClNO_2$	*261.70g/mol*	*0.17*
phenoxybenzamine hydrochloride	*5284441*		$C_{18}H_{23}Cl_2NO$	*340.29g/mol*	*0.00*

Table 7: Properties of drugs in the well connected component with 8 nodes of the drug similarity network

5.2.3 Comparison with chemical structural similarities

Once we proved robustness of the drug similarity network, we wanted to discover reasons of the observed similarities. First, we thought that chemical structural similarities of the compounds may shed light into finding a relation with the drug similarity network.

We calculated the correlation between chemical structural similarities of drug pairs and the number of similar edges or the number of different edges that drug pairs are linked with (also means the number of shared deletion strains. For more detail, see Section 4.5).

165 edges out of *504* similar edges and *71* edges out of *317* different edges between drug pairs were found as sharing chemical structural similarity with *Tanimoto coefficient*$\geq$ *0.2*. The results given in Table 8 and Table 9 demonstrate that there is a significant correlation between chemical structural similarity and similar edge numbers of drug pairs. However, we were not able to observe any correlation between chemical structural similarity and different edge numbers of drug pairs.

Then, we tried to create a regression model to predict the chemical structural similarity score using similar and different edge numbers of each drug pair which is a multiple linear regression problem. The results are given in Table 10. Even though the coefficient of similar edge number is substantially low, its p-value is highly significant.

Therefore, the chemical structural similarity may be used to explain similar relationships between drug pairs.

	FP2	FP3	FP4
correlation	*0.35*	*0.19*	*0.31*
p-value	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$

Table 8: Correlation between chemical structural similarities and similar edge numbers between drug pairs of the drug similarity network

	FP2	FP3	FP4
correlation	-0.04	-0.04	-0.02
p-value	< 0.55	< 0.52	< 0.78

Table 9: Correlation between chemical structural similarities and different edge numbers between drug pairs of the drug similarity network

	Coefficient	Standard deviation	p-value
similar edge number	*0.001*	$5.83e-005$	$< 1.51e-070$
different edge number	$-2.33e-004$	$1.19e-004$	< 0.05

Table 10: A regression model for chemical structural similarity based on FP2, by using similar and different edge numbers between drug pairs of the drug similarity network

5.2.4 Comparison with side effect similarities

We also had an intuition that side effect similarities of drugs may cause same response on the cell, hence it may contribute to be able to gain insight into observed similarities and dissimilarities in the drug similarity network.

As we have side effect information only for *53* drugs of Hillenmeyer *et al.* dataset, we were only able to examine relationships of *1378* drug pairs for the side effects analysis. *33* similar relationships and *11* different relationships out of *1378* relationships of drug pairs were found significant ($p-value < 9.09e-7$ and *fold enrichment*$>$ *5*), hence these interactions were involved in

	similar edge numbers	**different edge numbers**
correlation	*0.14*	-0.46
p-value	< 0.46	< 0.16

Table 11: Correlation between side effect similarities and similar or different edge numbers between drug pairs of the drug similarity network

	Coefficient	**Standard deviation**	**p-value**
similar edge number	-0.02	*0.05*	< 0.73
different edge number	*0.41*	*0.15*	< 0.006

Table 12: A regression model for side effect similarity by using similar and different edge numbers between drug pairs of the drug similarity network

the drug similarity network. Likewise the chemical structural similarity, we calculated the correlation between side effects similarities of drug pairs and the number of similar edges or the number of different edges of drug pairs. The results given in Table 11 demonstrate that we were not able to observe any correlation between side effect similarities and similar edge numbers of drug pairs. On the other hand, there is a correlation between side effect similarities and different edge numbers of drug pairs. Its p-value, however, is not significant.

Likewise in chemical structural similarity, we evaluated multiple linear regression to predict the side effect similarity score using similar and different edge numbers of each drug pair. The results are given in Table 12. As seen in the table, the coefficient for similar edge numbers is not significant. However, the coefficient for different edge numbers is significant and large enough to be able to infer that side effect similarity increases when the different edge number increases.

5.2.5 Comparison with MSB Cokol 2012 dataset

Cokol *et al.* [46] examined *175* drug pair (*25* more self-self drug pairs were also examined for control) combinations from *33* different chemical compounds in *S. cerevisiae* in order to assess synergistic (S), antagonistic (A) and independent (I) interactions between drugs. *108* interactions out of *175* interactions which are between *25* drugs out of *33* drugs, were also examined in Hillenmeyer *et al.* dataset.

There is not any significant similar interactions out of *108* interactions that will subsequently constitute the drug similarity network. However, we observed *5* significant different interactions in the Cokol *et al.* dataset, hence, also appear in the drug similarity network. *3* out of *5* of these different interactions were also found as antagonistic in the Cokol *et al.*, whereas *2* out of *5* interactions were referred as independent. Independent interactions are between 5 fluorouracil - benomyl and 5 fluorouracil - fk506. Antagonistic interactions are between benomyl - staurosporine, calyculin a - latrunculin and rapamycin - fk506.

6 Conclusions and Future Work

In this thesis, we proposed a method in order to understand drug mechanism of action. The results of the proposed method suggest similarities and dissimilarities between certain drug pairs. Once we made several quality assessments on the network and proved its robustness, we compared the observed similarities with various orthogonal datasets.

First of all, as a conclusion of comparing the results to chemical structural similarities of compounds, we concluded that drugs that have similar chemical structures, may have similar effect on the yeast cell which supports the results and hypothesis given by Giaever *et al.* [33] and Hillenmeyer *et al.* [12] that cells may show similar response to drugs having similar structures.

On the other hand, we, in some cases, observed that yeast genes show different response to drugs that have similar side effects. Therefore, we are not able to generalize the drug similarities by only comparing their side effects to each other. Side effect similarities of drugs cannot explain by itself how the drug mechanism works within the cell.

Consequently, the suggested drug similarities can be used as a valuable tool that provides a reasonable selection for further development and test of unapproved compounds, for instance, compounds that are considered to have high probability on taking role in treatment of cancer disease but not FDA-approved, can be chosen from the set of most similar drugs given by our study to those compounds that drugs in the set are FDA-approved and widely used in cancer treatment. Therefore, the best candidates can be pinpointed which further prevent the useless cost of development and test of compounds unrelated to cancer treatment. This approach also contributes us to discover unknown mechanism of a drug from known ones which are revealed as very similar to the corresponding drug at the end of our study. Moreover, if a drug has important side effects on the cell even though it is used in treatment of cancer disease, it can be replaced with another drug very similar to the corresponding drug in our study, but which has less adverse effects.

The drug similarity network revealed *caspofungin*, *berberine chloride* and *bithionol* as more similar to the drugs in the network with having similar edge degrees > *30* which suggests that these drugs may have same effect on the yeast cells with most of other drugs. On the other hand, *pp1* and *pp2* are the drugs that were found as the most different ones within the network. Therefore, it may be inferred that they may have a specific role in the mechanism of cell functions that many of other drugs do not have this specificity.

Additionally, we found a set of genes called MDS genes whose deletions make yeast cells resistant to multiple drug conditions, hence the corresponding deleted genes are required for sensitivity to diverse perturbations. MDS genes were found as highly enriched for RNA metabolism related functions which contributes us to understand the working mechanism of cell functions in response to chemical perturbations.

As a future work, we are planning to cluster the drugs in the drug similarity network according to Anatomical Therapeutic Chemical (ATC) classification, and then compare the ATC clusters to the observed similarities. Since we found drugs of the 8-vertex well connected component as sharing certain therapeutic indications, our intuition is that drugs which are in same

ATC cluster, are most likely similar to each other. In other words, yeast cells show same response to the drugs which have same therapeutic effect on the cell.

On the other hand, as the drug similarity network, a gene similarity network can also be constructed from deletion strain-drug network that may contribute to highlight gene functions within the cell. Comparing the similarities found in the gene similarity network to genetic interactions between gene pairs or protein interactions may also help to explain how the genetic or protein interactions occur.

Finally, we want to test similarities of drugs found in cliques of connected components of the drug similarity network except those of the giant component, experimentally since it would strengthen our observations and prove the suggested hypothesis.

APPENDIX - Abbreviations

- **MDR** - Multi-drug resistance
- **MDS** - Multi-drug sensitive
- **HOP** - Haploinsufficiency Profiling
- **HIP** - Homozygous Profiling
- **MSP** - Multi-copy Suppression Profiling
- **SGA** - Synthetic Genetic Analysis
- **ORF** - Open Reading Frames
- **ATC** - Anatomical Therapeutic Chemical
- **SIDER** - Side Effect Resource
- **SMILES** - Simplified Molecular Input Line Entry System
- **MSB** - Molecular Systems Biology
- **GO** - Gene Ontology
- **HTC** - High Throughput Screening
- **FDA** - Food and Drug Administration
- **FP** - Fingerprints

References

[1] 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*, 320:362–365, 2008.

[2] J. Bruin. newtest: command to compute new test @ONLINE, 2011.

[3] Gabriel F. Berriz, John E. Beaver, Can Cenik, Murat Tasan, and Frederick P. Roth. Next generation software for functional trend analysis. *Bioinformatics*, 25:3043–3044, 2009.

[4] Albert-Laszlo Barabasi and Zoltan N. Oltvai. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5:101–113, 2004.

[5] Andrew M. Smith, Ron Ammar, Corey Nislow, and Guri Giaever. A survey of yeast genomic assays for drug and target discovery. *Pharmacology & therapeutics*, 127:156–164, 2010.

[6] R. T. Jacob, M. J. Larsen, S. D. Larsen, Kirchhoff P. D., D. H. Sherman, and R. R. Neubig. MScreen: An Integrated Compound Management and High-Throughput Screening Data Storage and Analysis System. *Journal of Biomolecular Screening*, published online on June 15, 2012.

[7] Jürgen Drews. Strategic trends in the drug industry. *Drug Discovery Today*, 8:411–420, 2003.

[8] G. Giaever, D. D. Shoemaker, T. W. Jones, H. Liang, E. A. Winzeler, A. Astromoff, and R. W. Davis. Genomic profiling of drug sensitivities via induced haploinsufficiency. *Nat Genet*, 21:278–283, 1999.

[9] Laura Kapitzky, Pedro Beltrao, Theresa J. Berens, Nadine Gassner, Chunshui Zhou, Arthur Wuster, Julie Wu, Madan M. Babu, Stephen J. Elledge, David Toczyski, R. Scott Lokey, and Nevan J. Krogan. Cross-species chemogenomic profiling reveals evolutionarily conserved drug mode of action. *Molecular Systems Biology*, 6:451, 2010.

[10] 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*, 4:498–506, 2008.

[11] Hon Nian N. Chua and Frederick P. Roth. Discovering the targets of drugs via computational systems biology. *The Journal of biological chemistry*, 286:23653–23658, 2011.

[12] Maureen E. Hillenmeyer, Elke Ericson, Ronald W. Davis, Corey Nislow, Daphne Koller, and Guri Giaever. Systematic analysis of genome-wide fitness data in yeast reveals novel gene function and drug action. *Genome biology*, 11:R30+, 2010.

[13] Elke Ericson, Marinella Gebbia, Lawrence E. Heisler, Jan Wildenhain, Mike Tyers, Guri Giaever, and Corey Nislow. Off-Target Effects of

Psychoactive Drugs Revealed by Genome-Wide Assays in Yeast. *PLoS Genet*, 4:e1000151+, 2008.

[14] Zhenglong Gu, Lars M. Steinmetz, Xun Gu, Curt Scharfe, Ronald W. Davis, and Wen-Hsiung H. Li. Role of duplicate genes in genetic robustness against null mutations. *Nature*, 421:63–66, 2003.

[15] David Deutscher, Isaac Meilijson, Martin Kupiec, and Eytan Ruppin. Multiple knockout analysis of genetic robustness in the yeast metabolic network. *Nature Genetics*, 38:993–998, 2006.

[16] Michael Costanzo, Anastasia Baryshnikova, Jeremy Bellay, Yungil Kim, Eric D. Spear, Carolyn S. Sevier, Huiming Ding, Judice L. Y. Koh, Kiana Toufighi, Sara Mostafavi, Jeany Prinz, Robert P. St. Onge, Benjamin VanderSluis, Taras Makhnevych, Franco J. Vizeacoumar, Solmaz Alizadeh, Sondra Bahr, Renee L. Brost, Yiqun Chen, Murat Cokol, Raamesh Deshpande, Zhijian Li, Zhen-Yuan Lin, Wendy Liang, Michaela Marback, Jadine Paw, Bryan-Joseph San Luis, Ermira Shuteriqi, Amy Hin Yan Tong, Nydia van Dyk, Iain M. Wallace, Joseph A. Whitney, Matthew T. Weirauch, Guoqing Zhong, Hongwei Zhu, Walid A. Houry, Michael Brudno, Sasan Ragibizadeh, Balázs Papp, Csaba Pál, Frederick P. Roth, Guri Giaever, Corey Nislow, Olga G. Troyanskaya, Howard Bussey, Gary D. Bader, Anne-Claude Gingras, Quaid D. Morris, Philip M. Kim, Chris A. Kaiser, Chad L. Myers, Brenda J. Andrews, and Charles Boone. The Genetic Landscape of a Cell. *Science*, 327:425–431, 2010.

[17] Ainslie B. Parsons, Renee L. Brost, Huiming Ding, Zhijian Li, Chaoying Zhang, Bilal Sheikh, Grant W. Brown, Patricia M. Kane, Timothy R. Hughes, and Charles Boone. Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. *Nat Biotech*, 22:62–69, 2004.

[18] Hendrik Luesch, Tom Y.H. Wu, Pingda Ren, Nathanael S. Gray, Peter G. Schultz, and Frantisek Supek. A Genome-Wide Overexpression Screen in Yeast for Small-Molecule Target Identification. *Chemistry and Biology*, 12:55–63, 2005.

[19] Rebecca A. Butcher, Bhupinder S. Bhullar, Ethan O. Perlstein, Gerald Marsischky, Joshua LaBaer, and Stuart L. Schreiber. Microarray-based method for monitoring yeast overexpression strains reveals small-molecule targets in TOR pathway. *Nature chemical biology*, 2:103–109, 2006.

[20] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102:109–126, 2000.

[21] J.H. McDonald. Handbook of Biological Statistics (2nd ed.). Sparky House Publishing, Baltimore, Maryland. 2009.

[22] John E. Freund. Modern Elementary Statistics (6th ed.). Prentice hall. 1984.

[23] Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

[24] Albert-László L. Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature reviews. Genetics*, 12:56–68, 2011.

[25] Muhammed A. Yildirim, Kwang-Il Goh, Michael E. Cusick, Albert-Laszlo Barabasi, and Marc Vidal. Drug-target network. *Nat Biotech*, 25:1119–1126, 2007.

[26] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network Motifs: Simple Building Blocks of Complex Networks. *Science*, 298:824–827, 2002.

[27] H. Jeong, S. P. Mason, A. L. Barabasi, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411:41–42, 2001.

[28] Jing-Dong J. Han, Nicolas Bertin, Tong Hao, Debra S. Goldberg, Gabriel F. Berriz, Lan V. Zhang, Denis Dupuy, Albertha J. M. Walhout, Michael E. Cusick, Frederick P. Roth, and Marc Vidal. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430:88–93, 2004.

[29] Sumeet Agarwal, Charlotte M. Deane, Mason A. Porter, and Nick S. Jones. Revisiting Date and Party Hubs: Novel Approaches to Role

Assignment in Protein Interaction Networks. *PLoS Comput Biol 6(6): e1000817l*, 2009.

[30] Qi Ouyang, Peter D. Kaplan, Shumao Liu, and Albert Libchaber. DNA Solution of the Maximal Clique Problem. *Science*, 278:446–449, 1997.

[31] Tore Opsahl, Filip Agneessens, and John Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32:245–251, 2010.

[32] Gábor Iván and Vince Grolmusz. When the Web meets the cell: using personalized PageRank for analyzing protein interaction networks. *Bioinformatics*, 27:405–407, 2011.

[33] Guri Giaever, Patrick Flaherty, Jochen Kumm, Michael Proctor, Corey Nislow, Daniel F. Jaramillo, Angela M. Chu, Michael I. Jordan, Adam P. Arkin, and Ronald W. Davis. Chemogenomic profiling: Identifying the functional interactions of small molecules in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 101:793–798, 2004.

[34] Pek Y. Lum, Christopher D. Armour, Sergey B. Stepaniants, Guy Cavet, Maria K. Wolf, Scott J. Butler, Jerald C. Hinshaw, Philippe Garnier, Glenn D. Prestwich, and Amy Leonardson. Discovering Modes of Action for Therapeutic Compounds Using a Genome-Wide Screen of Yeast Heterozygotes. *Cell*, 116:121–137, 2004.

[35] Ainslie B. Parsons, Andres Lopez, Inmar E. Givoni, David E. Williams, Christopher A. Gray, Justin Porter, Gordon Chua, Richelle Sopko, Re-

nee L. Brost, Cheuk-Hei H. Ho, Jiyi Wang, Troy Ketela, Charles Brenner, Julie A. Brill, G. Esteban Fernandez, Todd C. Lorenz, Gregory S. Payne, Satoru Ishihara, Yoshikazu Ohya, Brenda Andrews, Timothy R. Hughes, Brendan J. Frey, Todd R. Graham, Raymond J. Andersen, and Charles Boone. Exploring the mode-of-action of bioactive compounds by chemical-genetic profiling in yeast. *Cell*, 126:611–625, 2006.

[36] Yoshihiro Yamanishi, Michihiro Araki, Alex Gutteridge, Wataru Honda, and Minoru Kanehisa. Prediction of drug target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, 24:i232–i240, 2008.

[37] Smiles [http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html].

[38] Noel O'Boyle, Michael Banck, Craig James, Chris Morley, Tim Vandermeersch, and Geoffrey Hutchison. Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3:33+, 2011.

[39] Michael Kuhn, Monica Campillos, Ivica Letunic, Lars J. Jensen, and Peer Bork. A side effect resource to capture phenotypic effects of drugs. *Molecular Systems Biology*, 6:343, 2010.

[40] Kathleen P. Seiler, Gregory A. George, Mary P. Happ, Nicole E. Bodycombe, Hyman A. Carrinski, Stephanie Norton, Steve Brudz, John P. Sullivan, Jeremy Muhlich, Martin Serrano, Paul Ferraiolo, Nicola J. Tolliday, Stuart L. Schreiber, and Paul A. Clemons. ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Research*, 36:D351–D359, 2008.

[41] L. Speroff. Efficacy and tolerability of a novel estradiol vaginal ring for relief of menopausal symptoms. *Obstetrics and Gynecology*, 102 (4):823–34, 2003.

[42] F Al-Azzawi, B Lees, J Thompson, and JC Stevenson. Bone mineral density in postmenopausal women treated with a vaginal ring delivering systemic doses of estradiol acetate. *Menopause (New York, N.Y.)*, 12 (3):331–9, 2005.

[43] WH Utian, L Speroff, H Ellman, and C Dart. Comparative controlled trial of a novel oral estrogen therapy, estradiol acetate, for relief of menopause symptoms. *Menopause (New York, N.Y.)*, 12 (6):708–15, 2005.

[44] L Speroff, AF Haney, RD Gilbert, and H Ellman. Efficacy of a new, oral estradiol acetate formulation for relief of menopause symptoms. *Menopause (New York, N.Y.)*, 13 (3):442–50, 2006.

[45] TS Jensen and TL Yaksh. Effects of an intrathecal dopamine agonist, apomorphine, on thermal and chemical evoked noxious responses in rats. *Brain Res*, 296 (2):285–93, 1984.

[46] Murat Cokol, Hon N. Chua, Murat Tasan, Beste Mutlu, Zohar B. Weinstein, Yo Suzuki, Mehmet E. Nergiz, Michael Costanzo, Anastasia Baryshnikova, Guri Giaever, Corey Nislow, Chad L. Myers, Brenda J. Andrews, Charles Boone, and Frederick P. Roth. Systematic exploration of synergistic drug pairs. *Molecular Systems Biology*, 7:544, 2011.