System Pharmacology: Application of Network Theory in Predicting Potential Adverse Drug Reaction Based on Gene Expression Data

Duy Pham[1], Bao-Khanh Le [1], Tu-Bao Ho[2,3], Ly Le[1]
[1] Department of Biotechnology, International University, HCMC, Vietnam
[2] School of Knowledge Science, Japan Advanced Institute of Science and Technology, Japan
[3]John von Neumann Institute, Vietnam National University, HCMC, Vietnam
bao@jaist.ac.jp, ly.le@hcmiu.edu.vn

Abstract- **In drug development process, adverse drug reaction (ADR) is one of the biggest challenges to evaluate the drug safety for passing to the market. Genomic expression data following in vitro drug treatments and thus have become widely used in ADR identification and prediction. In this research, we develop the prediction method by using system pharmacology-based study. We performed the proteomic, small molecular compounds - protein interaction and ADR data based on Connectivity Map database. A major protein-drug-side effect (PDS) network and a protein-drug (PD) network were obtained and analyzed by followed network centrality study, which allows for selection of side effects that are defined as central nodes. From the result, the top ranking of novel side effects was identified. In a case study, we established prediction models for 2, 3, 7, 8-tetrachlorodibenzo-p-dioxin (TCDD) in breast cancer treatment adverse events. In conclusion, the network-based approach provided the relationship between protein targets network and side effects based on the gene expression profiles and can predict the potential side effects for new a combinatory drug in the drug development process.**

***Index term*:** Adverse drug reactions, Side effects, Biological network, Protein–protein interaction, Systems biology, Systems pharmacology, Network centrality, Connectivity Map, CMap, Dioxin, Breast cancer, 2, 3, 7, 8-tetrachlorodi- benzo-p-dioxin, TCDD, dioxin, Network theory, Gene expression profile.

I. INTRODUCTION

Adverse drug reaction (ADR) is a 30 years old term which defined by WHO as "a response to a drug that is noxious and unintended and occurs at doses normally used in man for the prophylaxis, diagnosis or therapy of disease, or for modification of physiological function" [1]. In all disease treatment provided, ADRs always can be occurred by many causes. On drug development, during the lead optimization phase, many of compounds fail to be a drug; only 11% of remaining compounds can be accepted at the first stage of a clinical trial [2]. ADR is the one major reason for failure in drug clinical trial, so detect them as soon as possible by prediction is the currently trending and powerful way to decrease negative effects of drugs in disease treatment.

Omics data provide the opportunity to investigate ADRs clearly because it provides a huge amount of data at the molecular level [3]. A large number of methods were used to identify ADR are proposed. SIDER (Side Effect Resource), one of the most important ADR databases which predict and investigate adverse drug reaction [4]. SIDER was widely used to predict ADR as well as the side effects in several others studies [5-9]. Moreover, network analysis [10], machine learning [6], mathematical matrix construction [8], statistical test, and data mining [5, 7] are several computational methods which were using to predict ADRs based on the association of drugs, protein targets and adverse effects. Specially, M. Campillos *et al.* applied similarity approach to 746 marketed drugs and a network of 1018 side effect-driven drug-drug to find the relationship between a side effect and drug target [11]. Recently, the computational method to predict Adverse Drug Reaction using gene expression profile data – Connectivity Map [12], was publish by K. Wang *et al.* [13]. CMap has been widely used to predict new drug indications (i.e., drug repositioning) [14-16]. Connectivity map can significantly enrich true positive drug-indication pairs. The gene expression association suggested that not only therapeutic effects, but also side effects can be characterized by transcriptomic profiles [13]. From these researches, the ADR can be completely predicted by the method based on the association between gene expression of drug and drug targets (on and off).

For case study, *2, 3, 7, 8-tetrachlorodibenzo-p-dioxin (dioxin)* - TCDD is one of the highest potency carcinogens compound [17]. However, several experimental and epidemiological evidence [18-20] were showed that TCDD may have a protective effect for breast cancer [21]. Hsu *et al.* were publishing the microarray data of TCDD treatment in MCF-7 (breast cancer cell line) to Gene Expression Omnibus (GEO) database. It could potentially become the novel pathways associated for the

inhibitory effects of TCDD on breast cancer [21]. Although, TCDD is very toxicity compound [17], so there are many ADRs may occur after treatment. Thus, the mission to identify or predict the ADRs is highly important in the TCDD drug development.

To understanding the characteristics of networks comprehensively, the global and local topological analysis methods of are always utilized [22]. From a global perspective, centrality is a measurement in detecting the importance of vertex in the whole network [23, 24]. Network centralities give us the weighty nodes that have a meaningful position in the overall network architecture [24]. In this sense, many network centralities have been defined to manifest the significance of a node for a biological network by several features, such as node degree, betweenness, eigenvector measures, etc. [24]. In this study, we identify the centralities measurement of protein-drug-side effect (PDS) network to evaluate and rank the side effects as well as ADRs which based on associated with gene expression levels of TCDD in breast cancer cell line.

In this study, we supposed two major hypotheses are: (1) the similar gene expression level may lead to the same protein-protein interactions tendency; (2) potential side effect should be contributed by the high interaction of proteins-proteins interaction network. The network centrality network will be calculated to show the potential side effects for the breast cancer treated compound.

II. MATERIAL AND METHOD

A. DATA MINING

Microarray data were extracted from the Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/). We used web-based software GEO2R which available on http://www.ncbi.nlm.nih.gov/geo/geo2r/ and post-process in R environment to identify the gene expression signature profiles from GEO sample datasets. In this case study, we analyze microarray data from the dioxin against breast cancer research [21]. They provided two sets of microarray experiment with three biological replicates (GSM188013, GSM188016, GSM188020 are control samples (DMSO) and GSM188014, GSM188018, GSM188022 are 100 nM TCDD treated samples for 16 hours) that are based on Affymetrix Human Genome U133A Platform in MCF-7 cell line (Breast cancer cell line). For each probe, t-test with multiple test correction of Benjamini and Hochberg [25] is applied by GEO2R which automatically calculated its corresponding t-statistic and adjusted P value [26].

The Connectivity Map (also known as CMap) Version 2 is a collection of 7,000 expression profiles representing 1,309 compounds from cultured human cells treated microarray experiment and simple pattern-matching algorithms - Kolmogorov–Smirnov test [27, 28].

Let n and t be the number of probe sets in the feature set (22, 283) and the number of probe sets in the tag list (input) respectively. Instruct all n probe sets by the level of their differential expression for the current instance i. Building a vector V of the position (1..n) of each probe set in the signature list to list of all probe sets and screen these components in ascending order such that $V(j)$ is the position of tag j, where $j = 1, 2,..., t$. Then, the KS test [27] is applied by computing the two values:

$$a = max_{j=1}^{t}\left[\frac{j}{t} - \frac{V(j)}{n}\right]$$

$$b = max_{j=1}^{t}\left[\frac{V(j)}{n} - \frac{j-1}{t}\right]$$

Set $ks_i = a, if\ a > b$. Set $ks_i = -b, if\ b > a$. The up scores and down scores are ks_i up and ks_i down, respectively. The connectivity score Si is set to zero where ks_i up and ks_i down have the same sign. Otherwise, set $s_i = ks_i\ up - ks_i\ down$ s_i, p to be $max(s_i)$ and q to be $\min(s_i)$ across all instances in the collection C. The connectivity score S_i for the remaining instances is defined as ${s_i}/{p}$ if $s_i > 0$ or $-{s_i}/{q}$ if $s_i < 0$ [18]

Based on CMap, we query the gene expression signatures to find the drug connection to small molecule drugs database. The permuted result is obtained and extracting the compound based on p-value < 0.05 and positive enrichment score (connectivity score) condition for instance of each compounds (using mean scores).

For chemical-protein target annotation, we use STITCH (Search Tool for Interactions of Chemicals) database (http://stitch.embl.de/) [29]. STITCH contains interactions for between 300,000 small molecules and 2.6 million proteins from 1133 organisms. By evidence extracted from experiments, databases and the literature, chemicals are linked to other chemicals and proteins [29, 30]. We derived chemical – proteins interaction following parameters: active prediction methods all enabled except text mining; no more than 50 interactions; high confidence score (0.700); and network depth equal to 2.

With 1430 drug and 5868 side-effects, SIDER (Side Effect Resource) [31] is the best-known public source to annotate the side-effect as well as adverse drug reaction for small molecule drugs.

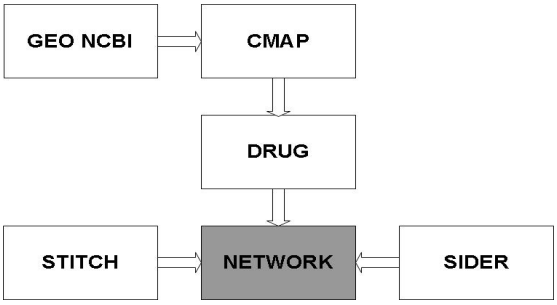


Figure 1: Data mining workflow

B. CONSTRUCTION OF PROTEIN-DRUG-SIDE EFFECT NETWORK

The screening and network design of the drugs from CMap result with corresponding proteins interaction and side effects were performed using Cytoscape software [32], version 3.3.0. The interactome protein-protein, small compounds-protein and small compounds-side effects data available for Homo sapiens were used to map to small molecular CMap drugs result. Then, Merge Networks core plugin of Cytoscape is applied (in union mode) to combine every single drug networks to accomplish the big PDS network. Besides, we merge the small compounds to protein-protein interactome to PD network for evaluating the relation between drug and protein which have an impact on side effects appearance.

C. CENTRALITY TOPOLOGICAL NETWORK ANALYSIS

PDS major network centralities were computed by using CytoScape software application: CentiScaPe [33] which available at http://apps.cytoscape.org/apps/CentiScaPe. We suppose a major undirection proteins-drug-side effects network is $G = (V, E)$. Vertices (V) are nodes which make the edges (E) between two vertices in the graph (G). We adopt to compute three centrality indexes: degree, betweenness, and stress.

Calculating for each vertex in the network, degree *deg(V)* is the number of directly connected vertices. The degree has a local attribute.

Stress is calculated by measuring the number of shortest paths between two vertices passing through a node. This centrality can indicate the relevance of a protein as functionally capable of holding together communicating side-effect nodes. Two nodes are connected by means of other shortest paths may not passing through the node *v*. To calculate the stress of a node *v*, all shortest paths in a graph *G* are calculated and then the number of shortest paths passing through v is counted. A stressed node is a node traversed by a high number of shortest paths. We computed stress centrality by the formula

$$C_{str}(V) = \sum_{s \neq V \in V} \sum_{t \neq V \in V} \sigma_{st}(V)$$

where σ_{st} is the number of shortest paths between *s* and *t* and $\sigma_{st}(v)$ is the number of shortest paths between *s* and *t* passing through the vertex *V*.

The betweenness is similar to the stress but provides a more elaborated and informative centrality index. It is calculated considering couples of nodes (V_1, V_2) and counting the number of shortest paths linking V_1 and V_2 and passing through a node *N*. Then, the value is related to the total number of shortest paths linking V_1 and V_2. The formula of betweenness computing is:

$$C_{spb}(V) = \sum_{s \neq V \in V} \sum_{t \neq V \in V} \delta_{st}(V)$$

where $$\delta_{st}(V) = \frac{\sigma_{st}(V)}{\sigma_{st}}$$

The betweenness mean is calculated by the following formula:

$$C_{avespb} = \frac{\sum_{i=1}^{n} C_{spb} i}{n}$$

where *n* is number of calculated betweenness node.

III. RESULTS

A. CONNECTIVITY MAP ANALYSIS

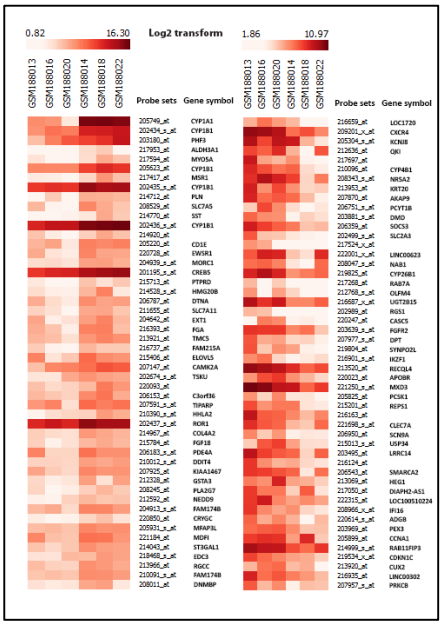


Figure 2: Top 50 up-regulated and top 50 down-regulated probe sets in MCF-7 cell line treated with TCDD. GSM188013, GSM188016, GSM188020 are control wells (DMSO) and GSM188014, GSM188018, GSM188022 are treated wells.

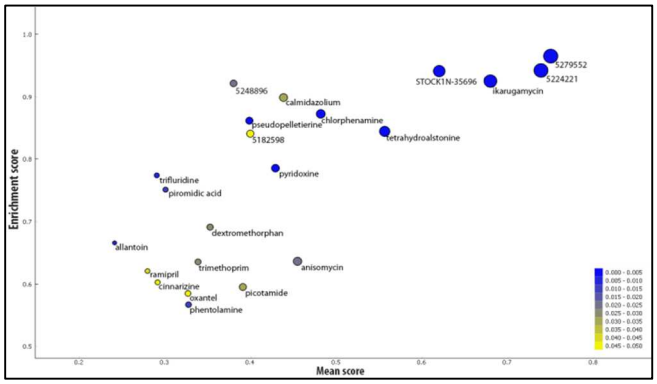


Figure 3: Connectivity map analysis: top 22 drugs have more than 50% similar gene expression to TCDD in MCF-7 cell line (X-axis: Enrichment score, Y-axis: Mean score and Color band: p-value).

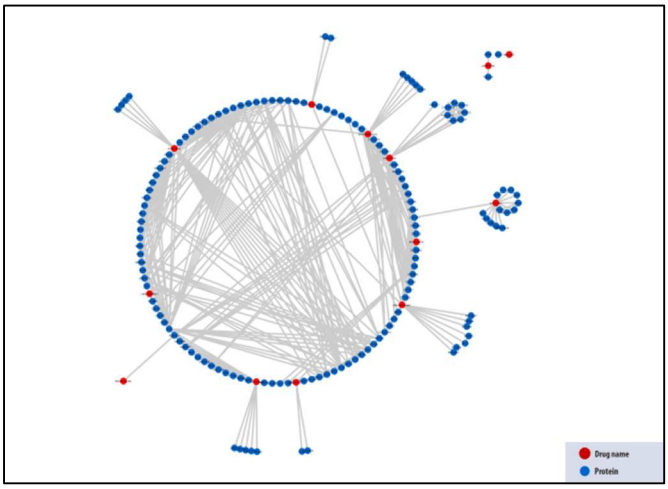


Figure 4: The drug-protein interactions network.

We obtained total 192 up-regulated probe sets and 110 down-regulated probe sets from microarray experiment [21] result data by multiple testing with 1.3 (threshold) fold-tests. The top 50 up- and down-regulated probe sets are shown in Figure 2. We were using these probe sets to query with CMap software (https://www.broadinstitute. org/cmap/). The results showed that 22 small molecule drugs (Figure 3) have more than 50% similar to TCDD in the gene expression level. Only 13 drugs have the chemical-protein interaction networks and 8 drugs have the chemical-side effect links. Besides, the others drug mechanism also can cause the unintended effect to the patients. Based on CMap result, we can have the initial connection of drugs at the molecular level.

B. PROTEIN-DRUG- SIDE EFFECT NETWORK

From 22 CMap drug results, we constructed 13/22 small drugs-proteins (9 drugs lack of protein information) interactions to make a network (Figure 4) with 162 (13 drugs and 149 proteins) nodes and 499 edges. In addition, we provided a network of drugs-side effects included 8/22 drugs and 474 corresponding side effects. After merging, we obtained a big network (Figure 5): 470 nodes (16 drugs, 149 proteins and 305 side effects) and 961 edges. In addition, 7 drugs have both protein – protein interactions and side effects that were visualized in the graph. The rest have only protein – protein interactions.

C. CENTRALITY TOPOLOGICAL ANALYSIS OF PDS NETWORK

Centrality measures are able to show the importance of nodes in our proteins-drugs-side effects network. After calculated the PDS network, in our analysis, we suppose that the side effects with high degree nodes are very highly potential to become the side effects of TCDD (Figure 6).

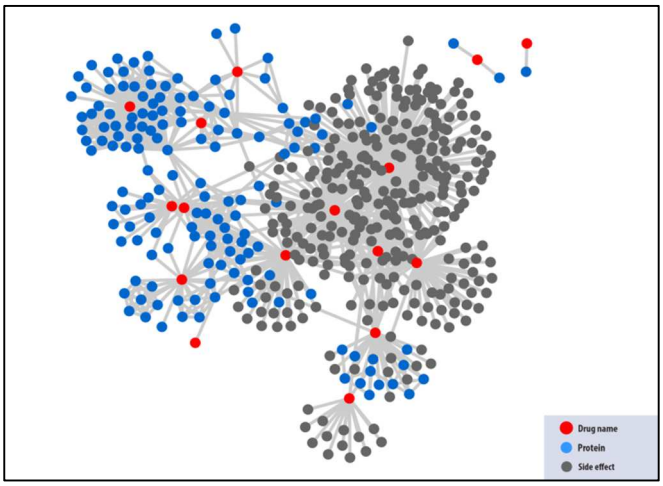


Figure 5: The connectivity network of TCDD to proteins and side-effects

We pick up top 10 side effects with the degree index greater or equal 4. Nausea, vomiting, headache, diarrhea, dizziness, pruritus, paraesthesia, anorexia, convulsion and hypersensitivity are 10 potential side effects. Moreover, we calculated the betweenness and stress centrality which have formula related to the connections with protein interactions network.

In the following work, we annotated top 10 side effects-drugs and calculated the Average Betweeness of all nearest nodes of side effect node to perform the highly connection in PDS network (Table 1). Cause that almost side effects come from same drugs so we attained a group of generally similar pharmacological actions. Angiotensin-Converting Enzyme inhibitor, Antibiotic, Alpha-adrenergic blockers, Antiviral agent, Vitamin, Antihistamine and Cough suppressant have a possibility to cause the side effects of TCDD.

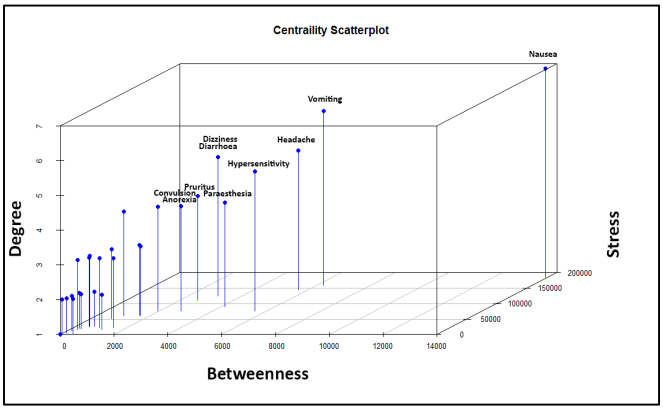


Figure 6: The 3D plot of centrality network

IV. CONCLUSION AND DISCUSSION

Systems pharmacology and its associated application can improve and accelerate the drug development process of a new drug for breast cancer as well as cancer disease treatments. Furthermore, systems pharmacology also strongly stimulates the development of single compounds that can target important in a specific network associated with breast cancer or with another type of tumor.

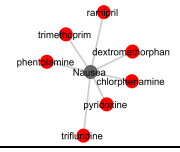
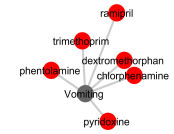
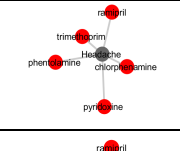
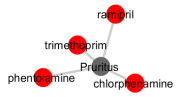
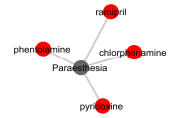
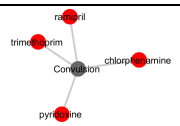
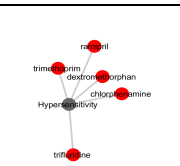
In this paper, we have investigated the computational method based on mining drug gene expression of drug data to predict related ADR. The association between gene expression level of drug and ADR are performed by drug protein-protein interaction network. The system pharmacological model of ADR prediction is completely constructed to give the detail information to decrease the negative impacts of drug in disease treatment, especially breast cancer treatment. Besides, we are just in initial steps of computing this model, but the calculating procedure is clearly provided. Our system pharmacological finding has a particular meaning and has a potential idea to support the ADR prediction method in the applying of computational biology to drug development and drug long-term treatment.

In the future, we will combine more mathematical method in biological network analysis process to give a high accuracy and precision for the side effects results. We expected to do a lot of evaluation in protein-protein network to identify the core mechanism of action to lead the side effects. Moreover, we are going to process the mining data from Library of Integrated Cellular Signatures (LINCS) which known as CMap 3.0 to expand the prediction adverse drug reaction research for breast cancer treatment.

ACKNOWLEDGMENTS

This work is funded by Vietnam National University at Ho Chi Minh City under the grant number B2015-42-02.

Table 1. Side-effect sub-network annotation.

Sub-network	Average Betweenness (PD network)	Pharmacological action
	2459.318	Angiotensin-Converting Enzyme inhibitor, Antibiotic, Alpha-adrenergic blockers, Antiviral agent, Vitamin, Antihistamine, Cough suppressant
	2869.204	Angiotensin-Converting Enzyme inhibitor, Antibiotic, Alpha-adrenergic blockers, Vitamin, Antihistamine, Cough suppressant
	2617.353	Angiotensin-Converting Enzyme inhibitor, Antibiotic, Alpha-adrenergic blockers, Vitamin, Antihistamine
	2674.045	Angiotensin-Converting Enzyme inhibitor, Antibiotic, Alpha-adrenergic blockers, Antihistamine, Cough suppressant
	2310.441	Angiotensin-Converting Enzyme inhibitor, Antibiotic, Alpha-adrenergic blockers, Antihistamine
	2310.441	Angiotensin-Converting Enzyme inhibitor, Antibiotic, Alpha-adrenergic blockers, Antihistamine
	3014.693	Angiotensin-Converting Enzyme inhibitor, Alpha-adrenergic blockers, Vitamin, Antihistamine
	2300.481	Angiotensin-Converting Enzyme inhibitor, Antibiotic, Vitamin, Antihistamine
	2300.481	Angiotensin-Converting Enzyme inhibitor, Antibiotic, , Vitamin, Antihistamine
	1897.077	Angiotensin-Converting Enzyme inhibitor, Antibiotic, Antiviral agent, Antihistamine

REFERRENCES

1. Edwards, I.R. and J.K. Aronson, Adverse drug reactions: definitions, diagnosis, and management. Lancet, 2000. **356**(9237): p. 1255-9.
2. Bender, A., et al., Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. ChemMedChem, 2007. **2**(6): p. 861-73.
3. Ho, T.B., et al., Data-driven Approach to Detect and Predict Adverse Drug Reactions. Curr Pharm Des, 2016.
4. Kuhn, M., et al., The SIDER database of drugs and side effects. Nucleic Acids Res, 2016. **44**(D1): p. D1075-9.
5. LaBute, M.X., et al., Adverse drug reaction prediction using scores produced by large-scale drug-protein target docking on high-performance computing machines. PLoS One, 2014. **9**(9): p. e106298.
6. Liu, M., et al., Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. J Am Med Inform Assoc, 2012. **19**(e1): p. e28-35.
7. Pauwels, E., V. Stoven, and Y. Yamanishi, Predicting drug side-effect profiles: a chemical fragment-based approach. BMC Bioinformatics, 2011. **12**: p. 169.
8. Vilar, S., N.P. Tatonetti, and G. Hripcsak, 3D pharmacophoric similarity improves multi adverse drug event identification in pharmacovigilance. Sci Rep, 2015. **5**: p. 8809.
9. Wallach, I., N. Jaitly, and R. Lilien, A structure-based approach for mapping adverse drug reactions to the perturbation of underlying biological pathways. PLoS One, 2010. **5**(8): p. e12063.
10. Cami, A., et al., Predicting adverse drug events using pharmacological network models. Sci Transl Med, 2011. **3**(114): p. 114ra127.
11. Campillos, M., et al., Drug target identification using side-effect similarity. Science, 2008. **321**(5886): p. 263-6.
12. Lamb, J., et al., The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science, 2006. **313**(5795): p. 1929-35.
13. Wang, K., et al., Systematic drug safety evaluation based on public genomic expression (Connectivity Map) data: myocardial and infectious adverse reactions as application cases. Biochem Biophys Res Commun, 2015. **457**(3): p. 249-55.
14. Kunkel, S.D., et al., mRNA expression signatures of human skeletal muscle atrophy identify a natural compound that increases muscle mass. Cell Metab, 2011. **13**(6): p. 627-38.
15. Ishimatsu-Tsuji, Y., T. Soma, and J. Kishimoto, Identification of novel hair-growth inducers by means of connectivity mapping. FASEB J, 2010. **24**(5): p. 1489-96.
16. Iorio, F., et al., Discovery of drug mode of action and drug repositioning from transcriptional responses. Proc Natl Acad Sci U S A, 2010. **107**(33): p. 14621-6.
17. Steenland, K., et al., Dioxin revisited: developments since the 1997 IARC classification of dioxin as a human carcinogen. Environ Health Perspect, 2004. **112**(13): p. 1265-8.
18. Bertazzi, P.A., et al., Dioxin exposure and cancer risk: a 15-year mortality study after the "Seveso accident". Epidemiology, 1997. **8**(6): p. 646-52.
19. Brown, N.M., et al., Prenatal TCDD and predisposition to mammary cancer in the rat. Carcinogenesis, 1998. **19**(9): p. 1623-9.
20. Warner, M., et al., Serum dioxin concentrations and breast cancer risk in the Seveso Women's Health Study. Environ Health Perspect, 2002. **110**(7): p. 625-8.
21. Hsu, E.L., et al., A proposed mechanism for the protective effect of dioxin against breast cancer. Toxicol Sci, 2007. **98**(2): p. 436-44.
22. Jeong, H., et al., Lethality and centrality in protein networks. Nature, 2001. **411**(6833): p. 41-2.
23. Hu, S., et al., Systems biology analysis of Sjogren's syndrome and mucosa-associated lymphoid tissue lymphoma in parotid glands. Arthritis Rheum, 2009. **60**(1): p. 81-92.
24. Borgatti, S.P., Centrality and network flow. Social networks, 2005. **1**(27): p. 55-71.
25. Benjamini, Y., and Yosef Hochberg. , Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society, 1995. **Series B** (Methodological): p. 289-300.
26. Barrett, T., et al., NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res, 2013. **41**(Database issue): p. D991-5.
27. Grover, N.B., Two-sample Kolmogorov-Smirnov test for truncated data. Comput Programs Biomed, 1977. **7**(4): p. 247-50.
28. Albano, A.M., P.E. Rapp, and A. Passamante, Kolmogorov-Smirnov test distinguishes attractors with similar dimensions. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics, 1995. **52**(1): p. 196-206.
29. Kuhn, M., et al., STITCH: interaction networks of chemicals and proteins. Nucleic Acids Res, 2008. **36**(Database issue): p. D684-8.
30. Kuhn, M., et al., STITCH 4: integration of protein-chemical interactions with user data. Nucleic Acids Res, 2014. **42**(Database issue): p. D401-7.
31. Kuhn, M., et al., A side effect resource to capture phenotypic effects of drugs. Mol Syst Biol, 2010. **6**: p. 343.
32. Shannon, P., et al., Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res, 2003. **13**(11): p. 2498-504.
33. Scardoni, G., M. Petterlini, and C. Laudanna, Analyzing biological network parameters with CentiScaPe. Bioinformatics, 2009. **25**(21): p. 2857-9.
34. Gao, J., et al., Metscape: a Cytoscape plug-in for visualizing and interpreting metabolomic data in the context of human metabolic networks. Bioinformatics, 2010. **26**(7): p. 971-3.