Network-based Inferring Drug-Disease Associations from Chemical, Genomic and Phenotype Data

Yu-Fen Huang
Institute of Information Systems and Applications
National Tsing Hua University
HsinChu 300, Taiwan
giselleyfh@gmail.com

Hsiang-Yuan Yeh
Department of Computer Science
National Tsing Hua University
HsinChu 300, Taiwan
d926708@oz.nthu.edu.tw

Von-Wun Soo
Institute of Information Systems and Applications
National Tsing Hua University
HsinChu 300, Taiwan
soo@cs.nthu.edu.tw

***Abstract*—With the information of drug, disease phenotype and protein interactions accumulating rapidly, to investigate the relationships between drugs and diseases is a critical importance issue. Until recently, few studies attempt to discover drug-disease associations on a network basis. We integrate drug and phenotype information and protein interaction network together and apply a network propagation approach to infer and evaluate the likelihood of the probability between drug and disease based on gene expression profile. In the experiments, we adopt prostate cancer as our test data. We validate our results to the manually curated associations in Comparative Toxicogenomics Database. Our experimental studies show that our proposed method obtains high specificity and sensitivity (AUC=0.98) and clearly outperforms previous existing methods. Our proposed method discovers potential drug-disease associations that drew the attention of biologists and provides a new perspective for toxicogenomics and drug reposition evaluation.**

Keywords-Drug-disease association, network propagation, phenotype network, protein network, chemical structure

I. INTRODUCTION

Disease, as an intricate phenotype, is often caused by the accumulation of multi-factor-driven alterations and is involved in dysfunctions of dozens of abnormal genes [1]. Drugs achieve their therapeutic functions by targeting biological processes in which the products of those abnormal genes participate. However, drug development process is complicated, time-consuming and adds to low therapeutic efficacy and/or unacceptable toxicity. While the mechanism is triggered from mutated genes in disease or drugs therapies, the regulation of signaling molecules in protein network plays an important role in the success of understanding the dynamic behaviors under specific disease or the pharmacodynamics under drug exposure [2]. Suthram and colleagues also demonstrated that common functional modules underlie similar diseases, and also highlighted the therapeutic importance of those modules [3]. The structures and chemical properties of drugs are associated with the effective use as therapies and drugs with similar structures may share functions. Therefore, it is important to integrate the prior knowledge and to develop an alternative and efficient way to investigate the boost of drug discovery which finds the connections between existing drugs and disease phenotypes including negative and positive impacts.

Until recently, several studies attempt to discover the associations between drugs and diseases using manually collected or computational approaches. The Comparative Toxicogenomics Database (CTD; http://ctd.mdibl.org) as a unique tool to infer chemical–disease associations via chemical–gene interactions, and gene–disease relationships that are manually curated from the published literature using controlled vocabularies [4]. Li et al. developed a computational framework to build disease-specific drug-protein connectivity maps based on protein interaction networks (PIN) and literature mining [5]. In the past decade, genome-wide monitoring of gene expression, provided by high-throughput microarray technology, has been commonly used as an important tool for the investigation of drug-disease associations. Gene module can reveal specific functional linkages between complex genomic signatures in human disease to targeted therapeutic strategies. One of the most comprehensive and systematic approaches for drug-disease association is the Connectivity Map (Cmap) project [6]. The Connectivity Map currently provided a reference collection of gene expression based molecular activity profiles of each drug, which were obtained by systematically exposing to a few key cancer cell lines and measuring the genome-wide transcriptional response. Given the disease-specific genetic signature, these profiles can be used as the basis of comparison the gene expression profile to realize the association between drugs and diseases based on shared molecular activity [7]. In addition, given the treatment profiles, previous research used a "guilt by association" (GBA) approach to predict novel associations between drugs and diseases by assuming that the if two diseases share some similar therapies, then the drug used for only one of the two may also be therapeutic for the other [8]. With a gold standard set of the drug-disease associations, Gottlieb et al. developed a computational method, PREDICT, to identify drug-disease associations and also predict new drug indications based on their features including chemical structure, side effects, gene expression profile, and chemical-protein interactome [9]. Based on the different features to predict the infer drug-disease associations must have the positive and negative data to build the accurate prediction

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

model. It is technically difficult to determine negative data such like non-drug targets due to the lack of researchers who are interested in validating them. However, the knowledge learned from the limited training data could be biased and hard to predict the novel new associations. With the network-based analysis, Zhao and Li developed a Bayesian partition method called comCIPHER to identify drug-gene-disease co-modules underlying the closeness measurement and known drug-disease associations from CTD database [10]. The shortest path analysis considers only a single path which ignores the potential contribution of the other paths with longer length. On the other hand, they simply take all the interactions in PIN into consideration but do not further select active interactions among proteins while only a part of the interactions among a set of proteins may be active at a specific condition.

Although the links of disease-associated proteins and drug targets are often established, the associations between drugs and diseases are not clear and still need an integrated approach to infer the associations. Therefore, we propose a novel presentation of drug-protein/gene-disease relationships with a network propagation model, where genes with similar functional modules are related to not only the drugs but also the diseases phenotype.

II. METHODS

We integrate three homo-networks (phenotype, drug, protein networks) in an integrated network in Fig. 1. The bipartite network with edges capture interactions between two different types of the homo-networks is called hetero-network. Each homo-network is defined as undirected graph $G_i = (V_i, E_i)$ where V_i is the node set and E_i is the edge set in the homo-network *i*. Each hetero-network is defined as bipartite graph $G_{ij} = (V_i \cup V_j, E_{ij})$. Here, E_{ij} represents the set of edges connecting nodes between homo-networks *i* and *j*.

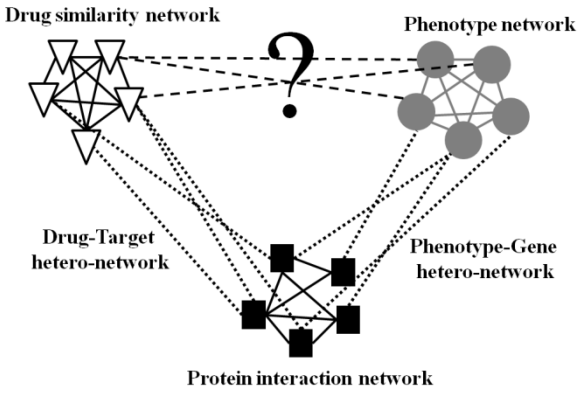

Figure 1. The idea of our proposed method.

A. Phenotype network

A phenotype network consists of disease phenotypes as nodes and the phenotypic similarity as edges based on the Online Mendelian Inheritance in Man (OMIM) database [11]. However, there are some noise and informative problem in phenotype similarity score matrix. The similarity score of two diseases that falls in the range [0.6, 1] shows potentially relevant phenotypic similarities and significant functional similarity. On the other hand, the scores fall within [0, 0.3] are not informative. Therefore, we use the logistic function provided from [12] to convert the informative similarity scores of diseases to be close to 1 and the non-informative scores to be close to 0 over the phenotype similarity score matrix constructed by [13] via text mining techniques. The symmetric similarity matrix $W_p(p_i,p_j)$ in phenotype homo-network represents the phenotypic similarity score between phenotypes p_i and p_j.

B. Drug network using chemcial similarity

DrugBank database provides drug-related information and also stores the drug targets of a drug [14]. We extract the FDA-approved drugs and their canonical simplified molecular input line entry specification (SMILES) [15]. We calculate hashed fingerprints using the Chemical Development Kit (CDK) [16] and find common substructures between two hashed fingerprints using Tanimoto coefficient [17], that is, the size of the intersection over the union when viewing each fingerprint as a set of specified elements. If the similarity between two chemical structures of drug x and x' defined as $sim(x, x')=|x \cap x'|/|x \cup x'|$. The value of the $sim(x, x')$ falls in the range of zero (no bits in common) to unity (all bits the same). The symmetric chemical structures similarity matrix W_d in drug homo-network denotes the similarity of the chemical structures between any pair of the drugs.

C. Weighted protein interaction network using gene expression data

Since protein networks are the assembly of the protein signal cascades that transfer the biological function and information through the pathways and provide a comprehensive map of the functional interactions. Due to the increasing availability of human protein interaction networks, network-based analysis provided an opportunity on the discovery of relationships between the mechanism of the drugs and human disease phenotype. Prieto and De Las Rivas have shown a limited intersection and overlap between the five databases (HPRD, BIND, IntAct, MINT, and OPHID) [18]. The information contained in these databases is partly complementary and the knowledge of the protein interactions can be increased and improved by combining multiple databases. Current PIN only provides the functional relations among the products of the genes but they do not provide information about the conditions under which the interactions occur. We extract the log base 2 of each gene in experimental dataset and average the gene expression values for the same gene symbol. Since co-expressed genes are more likely to function together, we apply Pearson correlation coefficient for every pair-wise relations and the range of the value would be [-1,+1]. In general statistical usage, the positive value in Pearson correlation indicates an increasing linear relationship and negative value indicate a decreasing linear relationship. On the other hand, the

correlation approaches to zero shows there would be little or no association among the pair of genes. By visualizing gene expression as combinations of activated and deactivated functional modules, we take the absolute value of correlation value to capture inhibitory activity (negative correlation) as well as activation activity (positive correlation). Therefore, we define the weight function in (1) as the product of the absolute value of Pearson correlation and the sum of the absolute value of differential expression changes of the two corresponding genes between the normal and disease samples.

$$W_g(g_i, g_j) = \left|R^d_{g_i g_j}\right| \times \left(\left|E^d_{g_i} - E^n_{g_i}\right| + \left|E^d_{g_j} - E^n_{g_j}\right|\right) \quad (1)$$

where $W_g(g_i,g_j)$ denotes the weight function from gene g_i to node g_j. $\left|R^d_{g_i g_j}\right|$ is the absolute value of Pearson correlation coefficient for the interaction among gene g_i and g_j from disease samples. $E^d_{g_i}$ is the average gene expression value of gene g_i in disease microarray data and $E^n_{g_i}$ represents the average expression value in normal microarray data. The higher weight denotes the stronger correlation or differential expression exchanges between any pair of genes. We use the symmetric matrix W_g to store the weights of interactions.

D. Integrated network

The gene-phenotype hetero-network is an undirected bipartite graph with disease phenotype vertices and disease-associated genes extracted from the OMIM database [11]. The disease-associated genes corresponding to their proteins can interact with other proteins to form functional modules in PIN. On the other hand, the drug target information was obtained from DrugBank database [14] and the target protein also can be map onto the PIN. We denote that asymmetric matrices W_{pg}, W_{dg} represent two hetero-networks standing for the adjacency matrices of link structures from phenotype-gene and drug-target relationships, respectively. If drug d_i has a target g_i, then $W_{dg}(d_i, g_j)$=1 otherwise $W_{dg}(d_i, g_j) = 0$. When a drug target or disease-associated gene has no linking relationship with other proteins in PIN, we set the probability of connection to any other protein as 1/(n-1), where n is the total number of proteins in PIN. The reason that we use random relationship instead of zero is to prevent a drug target or disease-associated gene becoming a "sink node" in PIN. This setting also allows its probability to be propagated in our method.

E. Network propagation in the integrated network

Given a query phenotype, we identify our problem of inferring drug-disease association as a network propagation which simulates a random walker move on query phenotype to its immediate neighbors randomly in the integrated network [19]. We adopt the idea from [20] which developed a label propagation algorithm for an integrated network.

For every pair of nodes *i* and *j* with interactions in the homo-network or hetero-network, we define a diagonal matrix $D_r(i,i)$ which means the sum of row *i* while $D_c(j,j)$ is the sum of column *j* in the similarity matrix *W* respectively. Since the similarity matrix of a homo-network is symmetric, the diagonal matrix $D_r(i,i)$ is equal to $D_c(j,j)$. Therefore, we normalize all the similarity matrices by computing (2) to get a normalized stochastic matrix. It is performed by representing each edge in terms of a transition probability based on the similarity of the immediate neighborhood. On the other hand, the asymmetric matrices of the hetero-networks are also normalized using (2) where *i* denotes the node in a homo-network and *j* denotes the node in the other homo-network.

$$S(i, j) = W(i, j) / \sqrt{D_r(i,i) D_c(j, j)} \quad (2)$$

Given the normalized matrix *S*, diffusion parameter α and the vector p^0 with the initial value of each node, the probability transition process in network propagation method on single network is defined as (3).

$$p^t = (1 - a)p^0 + aSp^{t-1} \quad (3)$$

Where *t* denotes the time step. The first term denotes the random walker can jump back to initial nodes with the probability of (1-α) at every time step and then "restarts" to propagation. The second term denotes an iterative walker's transition in the network. The transition will reach farther nodes in the network with a larger diffusion parameter α. Otherwise, the walker will be trapped at initial nodes if α is zero. Let P^t be a vector in which a node in the network holds the probability of finding itself in the iterative random walker process up to the step *t*. After certain steps, the probabilities will reach a steady state which is obtained by performing the iteration until the difference between P^t and P^{t-1} measured by *L2* norm falls below a very small number such as 10^{-9}.

We extend the network propagation to an integrated network. We assume that the nodes receive the probabilities from other nodes in the same homo-network, and also get the probabilities from nodes in other homo-networks through hetero-networks [21]. According to the recursive definition of importance, the probability *p* would be replaced by adding the additional probabilities from its immediate neighbors through hetero-networks such as the constraint introduced in [22]. It is too naive to assume that the information from different homo-networks and that from current homo-network be equally important. Therefore, the new initial probability vector p_i^0 in each homogeneous network i is proposed by changing into a weighted sum which is formulated as follows (4).

$$p_i^0 = a_i p_i^0 + b_{ij} \sum_{i \neq j} S_{ij} p_j^0 \quad (4)$$

Where a_i and b_{ij} denote the weights in homo-network *i* and between two homo-networks *i* and *j* respectively. S_{ij} denotes the normalized stochastic matrix of the hetero-network.

Those weights are greater than or equal to zero. The vector p_i denotes the probabilities of nodes in homo-network *i* and the vector p_j denotes the probabilities of nodes in other homo-network j. With the consideration of (3) and (4), to keep the sum of the probabilities of nodes in each probability vector be equal to one and we must further ensure the parameters a_i and b_{ij} in (4) to satisfy (5):

$$a_i + \sum_{i \neq j} b_{ij} = 1 \tag{5}$$

If homo-network *i* is connected to other *k* homo-networks, we adopt the weight of b_{ij} the same as diffusion parameter α_i to the immediate neighbors in the other homo-networks. According to (5), we calculate $a_i+k\alpha_i=1$ and obtain the weight $a_i=1-k\alpha_i$. Therefore, we further elaborate the network propagation method on a homo-network *i* in (3) into (6).

$$p_i^t = (1-a_i)\left[(1-ka_i)p_i^0 + a_i \sum_{i \neq j} S_{ij} p_j^0\right] + a_i S_i p_i^{t-1} \tag{6}$$

Thus, this network propagation is calculated with an enriched initialization from the other homo-networks through hetero-subnetworks and the proof of convergence is in [23].

First, we set the initial probability distribution over nodes constructed in such a way that the probabilities to our query disease are set as one and other nodes in the other homo-networks are set as zero. Second, we apply network propagation method in (6) on each homo-network iteratively until the probaility converges. Finally, we use the coverged probability of each node in each homo-network as initail value and repeated the network propagation method again until all homo-networks converge to a fnal probability. Take an example, if a drug communicates with the query disease with a high similarity score, it will receive the highest probability in network propagation. Similarly, a drug with a lower similarity score will receive lower node visitation to the query disease.

F. Evaluation of association specificity between drug and disease

In our method, we bias the network propagation to chemical similarity, gene expression, and phenotype similarity data. The transition from the network propagation may skew the visitation frequencies towards those supplied data values. The frequencies of node visitations may be highly biased based on the network topology and the probability of the node may directly reflect the relative centralities in the network which was proved by [24]. In order to control the topological biases in PIN, we calculated the reference visitation frequencies without considering the gene expression data values. By setting all the similarity scores among genes equal to one in PIN, we run our method to get a set of reference probabilities P_i^{ref} for each homo-network *i* using (6) and then evaluate the specificity of the probability using Z-score as defined in (7).

$$Z_i(v) = \frac{P_i(v) - avg(P_i^{ref})}{std\left(P_i^{ref}\right)} \tag{7}$$

Here, $P_i(v)$ denote the probability of node *v* in homo-network *i* using genomic data calculated by our method. The function of *avg* and *std* denote the average and standard deviation for the set of reference probabilities P_i^{ref} in homo-network *i*.

III. EXPERIMENTS AND RESULTS

A. Material and benchmark

Prostate cancer is a leading cancer and aggressive metastasis disease worldwide and it is the second common cancer-death among men. We adopt microarray data taken from [25] that consists of 62 primary prostate tumors and 41 normal tissues from Stanford Microarray Database (SMD) [26]. We map the UniProt protein ID to the human Entrez gene ID, erase the duplicated interaction pairs, and successfully obtained 137,037 interactions for 13,388 unique genes from current PIN databases. Then, we extract 33,930 corresponding protein interactions in PIN among 6,206 genes in microarray data. The phenotype network contains 5,080 diseases and gene-phenotype hetero-network contains 275 disease phenotypes and 649 genes from 877 relations while mapping the genes in microarray data and PIN. We collect 1,571 FDA-approved drugs and 1,410 of them with available SMILES data in DrugBank database. There are 3,702 relations between 723 drugs and 1,041 targets to be the drug-target hetero-network. We extract 53 drug-prostate cancer associations in CTD database as our benchmark.

B. Potential drugs for prostate cancer

After running our method with diffusion parameters α_d, α_g, and α_p as 0.1, 0.7 and 0.3 in drug, gene/protein, and phenotype homo-networks, we display 18 drug-prostate cancer associations on the basis of the one-tailed test at a 0.01 level of significance with Z-score > 2.33 in Table I. Since the predictions are not regarded as true and need further validated using external information supported. There are 7 FDA-proved drugs treated with prostate cancer and 7 drugs have relations with literature supported. In the Fluoxymesterone precautions lists, geriatric men are at higher risk for developing enlarged prostates or prostate cancer. Previous studies reported that arsenic trioxide (As_2O_3) provides a novel, safe approach or in combination with other conventional chemotherapeutic agents as a new agent to treat not only for androgen-dependent prostate cancer but also for androgen-independent prostate cancer [27, 28]. As_2O_3 also has been shown to be effective in leukemia and it has a potential to be a drug for prostate cancer treatment in our results. The finding of novel drug As_2O_3 for prostate cancer treatment shows that the evidence related to the concept of the drug reposition, which aims to discover new indications for existing drugs [29]. The tetracycline family including Minocycline has been used as antibiotics for decades and

previous reports showed the potential use of tetracycline family in early combination therapy for prostate cancer patients to reduce bone metastasis [30]. According to the results of a study published in [31], Clodronate reduces the risk of death by 23% in men with metastatic prostate cancer. In addition, Melatonin may have the ability to induce the death of cancerous cells of the prostate in vivo and in vitro, by suppressing a substance called Sirt1 which is involved in the development of prostate cancer [32, 33]. Lapatinib is a tyrosine kinase inhibitor of Epidermal growth factor receptor (EGFR) and human epidermal growth factor receptor 2 (HER2) and it has implicit antitumor effects in prostate cancer in a phase II trial [34]. Sorafenib is relatively well tolerated in androgen-independent prostate cancer showing evidence of improved bony metastatic lesions [35].

TABLE I. DRUG-PROSTATE CANCER ASSOCIATIONS

Drug ID	Drug Name	Z-score
DB01128[a]	Bicalutamide	47.43
DB00665[a]	Nilutamide	47.43
DB00499[a]	Flutamide	34.33
DB00421[a]	Spironolactone	34.02
DB01395[a]	Drospirenone	27.91
DB01185[b]	Fluoxymesterone	25.86
DB00367[a]	Levonorgestrel	24.87
DB01406[a]	Danazol	20.55
DB01169[b]	Arsenic trioxide	6.10
DB00626	Bacitracin	5.48
DB01017[b]	Minocycline	5.27
DB00851	Dacarbazine	5.08
DB00720[b]	Clodronate	3.67
DB00157	NADH	3.52
DB00697	Tizanidine	3.11
DB01065[b]	Melatonin	2.98
DB01259[b]	Lapatinib	2.75
DB00398[b]	Sorafenib	2.50

a. Primary therapeutic function approved by our benchmark; b. supported by literature

C. The performance of our method

In the experiments, we compare our method with Cmap [7]. We obtain 23 over-expressed and 63 under-expressed gene signatures from microarray data using 1-fold changes. We select the results including negative enrichment (the signature has opposite effect to the drug expression profile) as drugs for disease therapy and the other positive enrichment as toxic drugs. The results from Cmap only cover 29 the drug-disease associations in our benchmark. With the ranked lists extracted by Cmap and our methods, we compute sensitivity and specificity and observed the area under the receiver operator characteristic (ROC) curve (AUC) for analyzing the quality of performance in Fig. 2. The Cmap method obtained an AUC of 0.443 and our method obtains 0.985. Our results show that only depending on the drug response expression is incapable to capture the drug–disease associations due to the profiles generated under different conditions. Netowork-based analysis has a good chance to discover the drug-disease associations and its performance is consistent with previous study in [36]. With the whole associations in our benchmark, we obtain an AUC of 0.978.

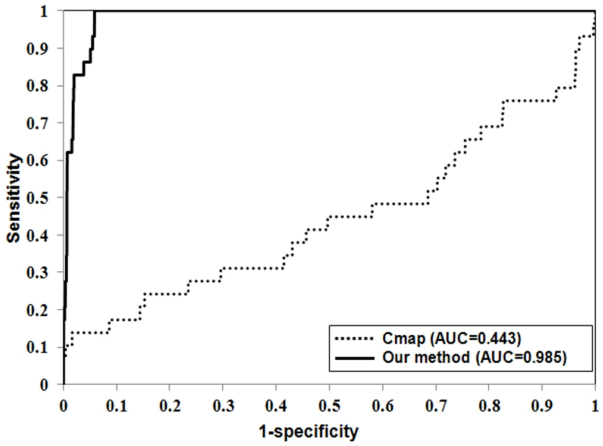

Figure 2. ROC curve among Cmap and our method.

IV. CONCLUSION

In this paper, we represent a network propagation approach for inferring and evaluating relationships between drugs and query disease in integrated network including drug–drug similarities, protein–protein interactions, and phenotype–phenotype similarities. In our experiment, we adopt the prostate cancer as our case study and the results clearly outperform previous Cmap method. We successfully rank the well known drug–disease associations higher and also infer more potential associations with literature supported. The success of our methods can be attributed as follows: First, we integrate heterogeneous data and knowledge into our model. Second, our network propagation method combines the information derived from the other connected networks to infer the drug-disease association. We believe that the combination of network and data source could help us to infer the drug-disease association. Our method can easily extend to other disease phenotype queries. The limitation of our approach is the difficulty in distinguishing the positive and negative associations between drug and disease. The other limitation is the chemical similarity approach which cannot represent the physiological effects. We can add the additional information such as pharmacological data and expression profiles to reduce the limitations of the drug-drug similarities by using Tanimoto coefficient and also to increase the yield of the analyses of drug–disease associations.

ACKNOWLEDGMENT

This research is partially supported by the Bioresources Collection and Research Center of Linko Chang Gung Memorial Hospital and National Tsing Hua University of Taiwan R. O. C. under the grant number 101N2761E1.

REFERENCES

[1] K.I. Goh, M.E. Cusick, D. Valle, B. Childs, M. Vidal and A.L. Barabási, "The human disease network," Proc Natl Acad Sci USA, vol. 104, no. 21, 2007, pp. 8685-90.

[2] M.A. Yildirim, K.I. Goh, M.E. Cusick, A.L. Barabasi, M. Vidal,"Drug-target network," Nat Biotechnol. vol. 25, no. 10, 2007, pp. 1119-26.

[3] S. Suthram, J.T. Dudley, A.P. Chiang, R. Chen, T.J. Hastie and A.J. Butte, "Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets," PLoS Comput Biol, vol. 6, 2010, pp. e1000662.

[4] C.J. Mattingly, M.C. Rosenstein, G.T. Colby, J.N. Forrest Jr and J.L. Boyer, "The Comparative toxicogenomics database (CTD): a resource for comparative toxicological studies," J. Exp. Zoolog. A Comp. Exp. Biol, vol. 305, 2006, pp. 689-692.

[5] J. Li, X. Zhu and J.Y. Chen, "Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts," PLoS Comput Biol, vol. 5, 2009, pp. e1000450

[6] J. Lamb, "The Connectivity Map: a new tool for biomedical research," Nat Rev Cancer, vol. 7, 2007, pp. 54–60.

[7] J. Lamb, E.D. Crawford, D. Peck, J.W. Modell, I.C. Blat, M.J. Wrobel, and et al., "The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease," Science, vol. 313, 2006, pp. 1929–1935.

[8] A.P. Chiang and A.J. Butte, "Systematic evaluation of drug-disease relationships to identify leads for novel drug uses," Clin Pharmacol Ther, vol. 86, 2009, pp. 507–510.

[9] A. Gottlieb, G. Y Stein, E. Ruppin and R. Sharana, "PREDICT: a method for inferring novel drug indications with application to personalized medicine," Mol Syst Biol, vol. 7, 2011, pp. 496.

[10] S. Zhao and and S. Li, "A co-module approach for elucidating drug-disease associations and revealing their molecular basis," Bioinformatics, vol. 28, no. 7, 2012, pp. 955-61.

[11] A. Hamosh, A.F. Scott, J. Amberger, C. Bocchini, D. Valle and V.A. McKusick, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," Nucleic Acids Res, vol. 30, 2002, pp. 52–55.

[12] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi and R. Sharan, "Associating Genes and Protein Complexes with Disease via Network Propagation," PLoS Comput Biol. vol. 6, no. 1, 2010, pp. e1000641.

[13] M.A. van Driel, J. Bruggeman, G. Vriend, H.G. Brunner and J.A. Leunissen, "A text-mining analysis of the human phenome," Eur J Hum Genet, vol. 14, 2006, pp. 535–542.

[14] D.S. Wishart, C. Knox, A.C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam and M. Hassanali, "DrugBank: a knowledgebase for drugs, drug actions and drug targets," Nucleic Acids Res, vol. 36, 2008, pp. D901–D906.

[15] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," J Chem Inf Comput Sci, vol. 28, 1988, pp. 31–36.

[16] C. Steinbeck, C. Hoppe, S. Kuhn, M. Floris, R. Guha and E.L. Willighagen, "Recent developments of the chemistry development kit (CDK)—an open-source java library for chemo- and bioinformatics," Curr Pharm Des, vol. 12, 2006, pp. 2111–2120.

[17] T. Tanimoto, Internal report: IBM Technical Report Series, November, 1957.

[18] C. Prieto and J. De Las Rivas, "APID: agile protein interaction data analyzer," Nucleic Acids Research, vol. 34(Web Server issue), 2006, pp. W298-302.

[19] S. Kohler, S. Bauer, D. Horn, P.N. Robinson, "Walking the interactome for prioritization of candidate disease genes," The American Journal of Human Genetics, vol. 82, no. 4, 2008, pp. 949-958.

[20] T.H. Hwang and R. Kuang, "A heterogeneous label propagation algorithm for disease gene discovery," Proc. of SIAM International Conference on Data Mining (SDM), 2010, pp. 583-594.

[21] W. Xi, B. Zhang, Z. Chen, Y. Lu, S. Yan, W.Y. Ma and E.A. Fox, "Link fusion: a unified link analysis framework for multi-type interrelated data objects," World Wide Web Conference Series - WWW, 2004, pp. 319-327.

[22] G. Pinski and N. Narin, "Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics," Information Process and Management, vol. 12, 1976, pp. 297-312.

[23] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Scholkopf, "Learning with local and global consistency," In NIPS, vol. 16, 2004, pp. 321-328.

[24] K. Komurov, M.A. White and P.T. Ram1, "Use of data-biased random walks on graphs for the retrieval of context-specific networks from genomic data," PLoS Comput Biol., vol. 6, no. 8, 2010, pp. e1000889.

[25] J. Lapointe, C. Li, J.P. Higgins, M. van-de-Rijn, E. Bair, K. Montgomery, et al., "Gene expression profiling identifies clinically relevant subtypes of prostate cancer," In Proceedings of the National Academy of Sciences USA, vol. 101, no. 3, 2004, pp. 811-816.

[26] G. Sherlock, T. Hernandez-Boussard, A. Kasarskis, G. Binkley, J.C. Matese, S.S. Dwight, and et al., "The stanford microarray database," Nucleic Acids Research, vol. 29, 2001, pp. 152–155.

[27] R. Uslu, UA. Sanli, C. Sezgin, B. Karabulut, E. Terzioglu, S.B. Omay and E. Goker, "Arsenic Trioxide-mediated Cytotoxicity and Apoptosis in Prostate and Ovarian Carcinoma Cell Lines," Clin Cancer Res., vol. 6, no. 12, 2000, pp. 4957-64.

[28] H. Maeda, S. Hori, H. Nishitoh, H. Ichijo, O. Ogawa, Y. Kakehi and A. Kakizuka, "Tumor growth inhibition by arsenic trioxide (As_2O_3) in the orthotopic metastasis model of androgen-independent prostate cancer," Cancer Res., vol. 61, no. 14, 2001, pp. 5432-40.

[29] H. Ye, L.L. Yang, Z.W. Cao, K.L. Tang and Y.X. Li, "A pathway profile-based method for drug repositioning," Chinese Science Bulletin , vol. 57, 2012.

[30] Z. Saikali and G. Singh, "Doxycycline and other tetracyclines in the treatment of bone metastasis," Anticancer Drugs, vol. 14, no. 10, 2003, pp. 773-8.

[31] D.P. Dearnaley, M.D. Mason, M.K. Parmar, K. Sanders and M.R. Sydes, "Adjuvant therapy with oral sodium clodronate in locally advanced and metastatic prostate cancer: long-term overall survival results from the MRC PR04 and PR05 randomized controlled trials," Lancet Oncol. vol. 10, no. 9, 2009, pp. 872-6.

[32] S.S. Joo and Y.M. Yoo, "Melatonin induces apoptotic death in LNCaP cells via p38 and JNK pathways: therapeutic implications for prostate cancer," J Pineal Res., vol. 47, no. 1, 2009, pp. 8-14.

[33] B. Jung-Hynes, T.L. Schmit, S.R. Reagan-Shaw, I.A. Siddiqui, H. Mukhtar and N. Ahmad, "Melatonin, a novel Sirt1 inhibitor, imparts Antiproliferative effects against prostate cancer in vitro in culture and in vivo in TRAMP model," J Pineal Res., vol. 50, no. 2, 2011, pp.140-149.

[34] Y.E. Whang, C.N. Moore, A.J. Armstrong, W.K. Rathmell, P.A. Godley, J.M. Crane, and et al., "A phase II trial of lapatinib in hormone refractory prostate cancer," J Clin Oncol, vol. 26, 2008

[35] W.L. Dahut, C. Scripture, E. Posadas, L. Jain, J.L. Gulley, P.M. Arlen, and et al., "A phase II clinical trial of Sorafenib in androgen-independent prostate cancer," Clin Cancer Res., vol. 14, no. 1, 2008, pp. 209-14.

[36] G. Hu and P. Agarwal, "Human disease-drug network based on genomic expression profiles," PLoS ONE, vol. 4, no. 8, 2009, pp. e6536.