A New Method to Identify Repositioned Drugs for Prostate Cancer

Zikai Wu
Business School
University of shanghai for Science and Technology
Shanghai, 200093 China
Email: zkwu2011@gmail.com

Yong Wang
Academy of Mathematics and Systems Science
Chinese Academy of Sciences
Beijing 100190
Email: ywang@amss.ac.cn

Luonan Chen
Key Laboratory of Systems Biology
Shanghai Institutes for Biological Sciences
Chinese Academy of Sciences
Shanghai 200031
Email: lnchen@sibs.ac.cn

***Abstract*—With the merits of faster development time and reduced risk, identifying new indications for marketed drugs draws more and more attention. In particular, repositioning drugs with known indications has become an hot topic in the area of computational systems biology. However, one of the common shortcomings for most of the previous methods is the ignorance of side effect, i.e., drug through primary targets and off targets might induce both desired and unintended effects respectively, which could not appropriately evaluated in most of existing methods. In this paper with a new measure considering both efficacy and side effect, we developed a new method for identifying the repositioned drugs against prostate cancer by evaluating the mutual relations of the gene expression levels between prostate cancer samples and those induced by bioactive compounds. In this measure, the overlap between gene sets that were oppositely regulated in disease state and drug treatment state was quantified by jaccard index as drug's efficacy while the overlap between essential genes and positively correlated genes (or regulated just after drug treatment) was quantified by jaccard index as drug's side effect, which were balanced with a parameter λ. The preliminary results on repositioning drugs for prostate cancer verify the effectiveness and efficiency of the new method.**

I. Introduction

People worldwide are still threatened by various complex diseases such as cancer, diabetes and so on. To combat these diseases, much effort was made to develop effective drugs against them. Generally, it may take about eight to ten years with over one billion dollars spent to develop a new drug de novo. However, with the cost and time to develop a new drug continuing to increase, most of drug candidates were given up in different Research&Development stages for insufficient efficacy or serious side effects. From 1996 to 2011, the number of drugs approved by FDA (U.S. Food and Drug Administration) declined steadily [1].

One possible approach for circumventing this situation is so called drug repositioning, which attempts to mine the potential of marketed drugs with well known safety profile and identify new indications besides the designed indications for them. As the safety profiles of marketed drugs are generally known, drug repositioning has the advantage of mitigating the costs and risks associated with early development stage and shortening routes to approval for therapeutic indications. Due to this fact, drug repositioning draws more and more attention recently. Successful examples of drug repositioning include the indication of sildenafil for erecile dysfuction and pulmonary hypertension, thalidomide for severe erythema nodosum leprosum, and retinoic acid for acute promyelocytic leukemia[2], [3].

Most of the above mentioned successful examples were discovered clinically or experimentally, but it is still unclear for their molecular mechanisms against various diseases. Therefore, computational methods that can effectively reposition drugs against various diseases in a large scale are greatly needed. On the other hand, huge amount of high-throughput data related to drugs at various levels are rapidly accumulated. Now, it is possible to model cellular systems and uncover the mechanism underlying manifested phenotypes by exploring such system-wide data. In particular, repositioning drugs with known indications computationally becomes an hot topic in the field of computational systems biology.

Recently, many computational methods or approaches were proposed to reposition drugs against various diseases. Based on data sources utilized, these methods maily fall into two categories. In the first category, various static prior information is combined and utilized with different approaches for predicting new indications for drugs. The basic idea of them is, if two drugs with known indications have a sufficient high similarity in some attributes, their known indications would be exchanged each other. In some methods, drug's off-target set is predicted first, then the overlap or sequence similarity among drug's total target set (primary target plus off-target) is used for transferring indication among drugs [4], [5], [6], [7], [8], [9], [10], [11]. With more types of data related to drugs available, multiple data sources are integrated to reposition drugs in

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

different methods. In Gottlieb et al's work, structure, function information of target protein were employed to measure the similarity among drugs while disease gene's topological and functional properties and disease's phenotype description were used to evaluate the similarity among diseases and the known indication of drugs were transferred based on both drug's similarity and disease's similarity[12]. More recently, in Yang and Agarwal's work, indications were transferred among drugs guided by their similarity in the side effect profile[13]. In [14], heterogeneous modules composed of drugs, diseases and genes were extracted from background biomolecular interaction network and drug were repositioned based on common membership.

Microarray data has been used as surrogate of transient cellular state under specific conditions. Naturally, in the second category, micoarray data are utilized to characterize cellular state and reposition drugs against various diseases. Methods of the second category follow the common assumption, that is, the aim of drug interfering is to restore the cellular state to normal state, and the changes of the transcriptional level induced by drug should reverse the changes of the transcriptional level under disease state. To this aspect, the basic idea of them is, if the differential expression profile under drug administration and disease states is anti-correlated significantly, the drug will has the potential to cure that disease. Therefore, in the second category, the key is how to measure the anti-correlation. In lamb's work, gene set enrichment analysis was employed to measure the correlation between expression profile under drug administration and that in disease state[15] while different modified versions of gene set enrichment analysis were used for measuring the relation between the two condition specific gene expression profiles[3], [16]. In Hu and Agarwal's work, simple Pearson correlation coefficient was employed to quantify the anti-correlation between two condition specific gene expression signatures[17] while the overlap between opposite regulated gene sets were quantified as the anti-correlation in [18].

It can be seen that the common idea of most of existing methods is to reposition drugs through measuring the potential efficacy of drug against disease by some similarity or inverse similarity in some attributes. However, a cellular system is a complex networked system. The perturbations on some cellular elements induced by drug will propagate through the related network. Therefore, drug can induce both desired effect and some unintended effect simultaneously. However, one of the common shortcomings for most of the previous methods is the ignorance of side effect.

With this in mind, in this paper, we developed a new method for repositioning drugs against prostate cancer by considering both drug efficacy and side-effect. The main characteristic is a new measure for evaluating the potential overall effect of drug against prostate cancer. In this measure, the overlap between gene sets that were oppositely regulated in disease state and drug treatment state is quantified by jaccard index as drug's efficacy while the overlap between essential genes and positively correlated genes (or regulated just after drug treatment) is quantified by jaccard index as drug's side effect, which are balanced with a parameter λ.

II. Materials and Methods

A. Materials

We combined data from publicly available microarray data set representing prostate cancer, gene expression profiles from human cell lines treated with drugs or small molecules and list of essential genes in human to reposition drug against prostate cancer.

The expression data set of prostate cancer was downloaded from National center for Biotechnology Information(NCBI) Gene Expression Omnibus(GEO) database [19] with accession number GDS1439 as it was already used in [18]. The data set comprise of six benign samples and thirteen disease samples.

As the annotation information, that is, the mapping between probe set and gene corresponding to each microarry platform updated regularly, we reannotated GDS1439 in soft format with latest corresponding GPL annotation file downloaded from AILUN's website[20]. Subsequently, in cases where multiple microarray probe sets mapped to the same Entrez GeneID, the mean expression value of them was assigned to the Entrez GeneID. Finally, the list of up and down-regulated genes was generated by comparing control samples to disease samples with R package named limma[21] and further ranked using t-statistic.

Gene expression profiles from human cell lines treated with drugs or small molecules were fetched from Connectivity Map 02[15]. More specifically, ratio matrix was downloaded from cmap02, in which each element denotes the ratio of probe set's expression in certain treatment instance.

Of all 6,100 treatment instances, 1741 instances were treated on PC3 prostate cancer cell line, which were retained for further analysis. At the same time, submatrix corresponding to this 1741 instances was extracted from ratio matrix. As this 1741 instances were conducted with two different affymetrix genechip platforms, that is, HG-U133A and HT_HG-U133A, the submatrix was separated into two matrices further based on platform. Subsequently, we annotated each matrix with latest corresponding GPL annotation file downloaded from AILUN's website[20] separately. In cases where multiple microarray probe sets mapped to the same Entrez GeneID, the geometric mean of ratio of them was assigned to the Entrez GeneID. Finally, this two annotated matrices were merged into a new matrix named PC ratio matrix by removing genes that absent in any matrix. In PC ratio matrix, the list of up and down-regulated genes for each instance were generated by the value of element in each column naturally.

As disease expression data set and PC ratio matrix were also conducted with different platforms, to remove potential bias, we retained only genes that are expressed in both.

In this work, list of essential genes in human was obtained from DEG database[22]. Subsequently, symbols of essential genes were transformed into Entrez GeneID by R package named biomaRt. Besides, we obtained the list of drugs that were approved by FDA or under clinical trial against prostate cancer from the supplementary materials of Jin et al's work[23] and took them as the gold standard set for prediction.

B. Methods

The idea behind the proposed method is that cellular systems is a complex networked system. Disease can be viewed as such a perturbed system. Significantly differentially expressed genes under disease state can be used as surrogate of cellular change to some extent. The aim of drug treatment is just removing this cellular changes and restoring cellular system to normal state. Therefore, number of abnormally regulated genes after drug treatment that were regulated oppositely under disease state can be used to measure the extent to which diseased cellular systems restored. On the other hand, some genes were newly regulated or further regulated in same direction after drug treatment, which may be the source of side effect. Essential genes are genes that are indispensable to support cellular life[22]. Their changes in transcription level may cause significant unfavorable phenotype variation, such as side effect. Therefore, the number of essential genes that were newly regulated or further regulated in same direction after drug treatment can be used to measure the extent to which side effect emerged to some extent. Based on this two measure, a new scheme could be developed to score and rank drug-disease association and reposition drugs against diseases. The details are addressed in the following.

1) Evaluating drug-disease associations based on microarray data: As mentioned in above section, two category of ranked lists of genes were prepared. The first category include the ranked list of up regulated genes and ranked list of down regulated genes in prostate cancer tissues from GEO data while the other category include the ranked list of up regulated genes and ranked list of down regulated genes of perturbed cell line genes obtained from CMAP02.

Motivated by Shigemiu's work[18], subsequently, the top and bottom k% genes were defined as up-regulated genes in prostate cancer (shortened as PCU) and down-regulated genes in prostate cancer (PCD) respectively. In the latter category, we defined the top and bottom k% genes as up-regulated genes by bioactive compounds (DU) and down-regulated genes by bioactive compounds (DD) respectively, where k ranges from 10 to 30 in increments of 5.

The overlap or similarity between DU and PCD was measured by Jaccard index [24] as efficacy score named $score^1_{up}$, that is,

$$score^1_{up} = \frac{|DU \cap PCD|}{|DU \cup PCD|} \tag{1}$$

On the other hand, we also employed Jaccard index to evaluate the overlap between set of genes that fall in DU while outside PCD and set of essential genes(EG) as side effect score named $score^2_{up}$,that is,

$$score^2_{up} = \frac{|(DU \setminus PCD) \cap EG|}{|(DU \setminus PCD) \cup EG|} \tag{2}$$

Further, effect of drug reflected in up regulated genes was measured as score_up, which combines score1 and score 2 with a balancing factor λ varying from 0 to 1.

$$score_{up} = \lambda * score^1_{up} - (1 - \lambda) * score^2_{up} \tag{3}$$

Similarly, the overlap or similarity between DD and PCU, the overlap between set of genes that fall in DD and outside PCU and set of essential genes were measured by Jaccard index as score3 and score 4 respectively,

$$score^1_{down} = \frac{|DD \cap PCU|}{|DD \cup PCU|} \tag{4}$$

$$score^2_{down} = \frac{|(DD \setminus PCU) \cap EG|}{|(DD \setminus PCU) \cup EG|} \tag{5}$$

Further, effect of drug reflected in up regulated genes was measured as score_down, which combines score3 and score 4 with a balancing factor lambda.

$$score_{down} = \lambda * score^1_{down} - (1 - \lambda) * score^2_{down} \tag{6}$$

Finally, the association strength between drug and prostate cancer was evaluated by combing score_up and score_down, that is,

$$score = score_{up} + score_{down} \tag{7}$$

2) Collapsing instances: In cmap02, multiple instances may correspond to same drug and even with same dose. We offered three different options for handling such cases. The first option was to simply retain the multiple instances. Accordingly, the number of true positive, false positive, true negative and false negative will be counted by instance in repositioning experiment. The second option is to calculate the maximum of individual instances's score that correspond to same drug with specified dose as the repositioned score of specified drug-dose pair. Accordingly, the number of true positive, false positive, true negative and false negative will be counted by drug-dose pair. The last option is to calculate the maximum of individual instances's score that correspond to same drug as the repositioned score of specified drug. Accordingly, the number of true positive, false positive, true negative and false negative will be counted by drug.

3) Choice of parameter: With fixed k and λ, the strength of association between prostate cancer and all compounds in PC ratio matrix were measured and ranked by score. The $F1$ score defined below was adopted as repositioning performance index

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{8}$$

where precision is the ratio of true positives in predicted positives and recall is the ratio of true positives that can be predicted correctly. The threshold above which the highest F1 score was achieved was used to make future prediction. We reposition a drug against prostate cancer if its score is above the threshold. Subsequently, the values of k and λ that give maximum F1 score will be identified.

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

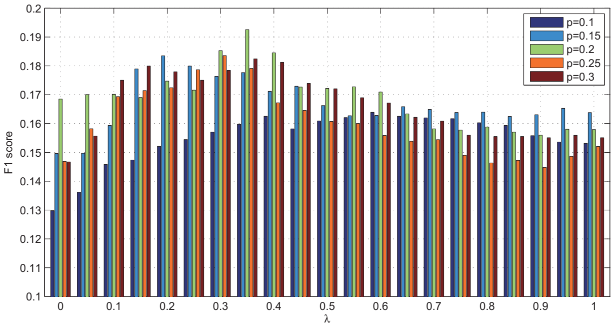

Fig. 1. F1 score of repositioning experiment under option 1 of collapsing instance, in which $p = \frac{k}{100}$

III. Results and discussion

As described in last section, different reposition efficacy measured by $F1$ score will be obtained with different combinations of k and λ. Besides, the option of collapsing instances will also affect the repositioning result. To explore the impact of parameters on drug repositioning, we performed 105 drug repositioning experiments under each option of collapsing instances with 105 combinations of parameters respectively, in which k vary from 10 to 30 by 5 and λ range from 0 to 1 by 0.05. All the experiments's F1 scores under each option of collapsing instance were summarized in figure1, figure2 and figure3 respectively.

It can be seen from this three figures that optimal $F1$ score always were obtained with $0 < \lambda < 1$ with fixed k. $\lambda = 1$ or $\lambda = 0$ means only efficacy score or side effect were utilized to measure drug-disease association respectively. Therefore, this results demonstrated the necessity of measuring drug-disease association with integrated efficacy and side effect information to some extent.

To further demonstrate the superiority of repositioning drug by integrating both efficacy and side effect measure, we took a close look at figure 3. It summarized repositioning effort with collapsing instance option 3. The global optimal $F1$ score was obtained with $k = 10$ and $\lambda = 0.2$. Under this parameter setting, we detected 132 unique bioactive compounds for prostate cancer. Of the 132 compounds, 16 of these are FDA approved or under clinical trial, which were summarized in table1. That also means we recovered 16 of the 45 compounds approved by FDA approved or under clinical trial for prostate cancer in PC ratio matrix. On the contrary, nine of this 45 compounds were among the top 132 compounds sorted by drug-disease score with $k = 10$ and $\lambda = 0$. With $k = 10$ and $\lambda = 1$, ten of this 45 compounds were among the top 132 compounds. Besides, five drugs, that is, azacitidine, dexamethasone,estradiol, metformin, tamoxifen were repositioned against prostate cancer successfully with $k = 10$ and $\lambda = 0.2$ and couldn't repositioned under the other two parameter settings, which imply the superiority of repositioning drug by integrating both efficacy and side effect measure further.

TABLE I
Repositioned drugs that lie in gold standard set

Drug Name	Status
sirolimus	clinical trial
paclitaxel	clinical trial
phentolamine	clinical trial
tanespimycin	clinical trial
doxorubicin	clinical trial
methylprednisolone	clinical trial
estradiol	clinical trial
vinblastine	clinical trial
valproic acid	clinical trial
metformin	clinical trial
theophylline	clinical trial
diethylstilbestrol	clinical trial
tamoxifen	clinical trial
azacitidine	clinical trial
dexamethasone	clinical trial
fulvestrant	clinical trial

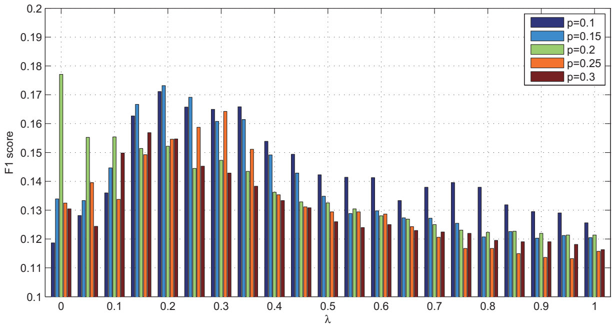

Fig. 2. F1 score of repositioning experiment under option 2 of collapsing instance, in which $p = \frac{k}{100}$

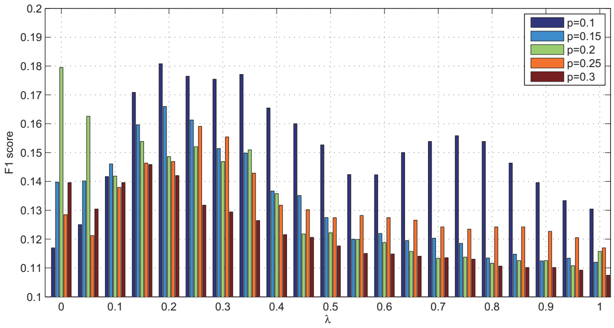

Fig. 3. F1 score of repositioning experiment under option 3 of collapsing instance, in which $p = \frac{k}{100}$

IV. Conclusion and future work

A cellular system is a complex networked system. Perturbation caused by drug will propagate through this networked system. Therefore, a drug can induce both desired effect and some unintended effect, which were ignored by most of the previous repositioning methods for measuring drug-disease potential associations. In this paper, we developed a new

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

method for identifying such repositioned drugs against prostate cancer based on a new measure. In this measure, number of abnormally regulated genes after drug treatment that were regulated oppositely under disease state was quantified as drug's efficacy while the number of essential genes that were newly regulated or further regulated in the same direction after drug treatment was quantified as drug's side effect, which were balanced with a parameter λ. The preliminary results on repositioning drugs for prostate cancer verify the effectiveness and efficiency of the new method.

However, there are several issues that limit the repositioning performance. First, the size of differentially expression genes that constitute disease signature or drug signature were chose empirically, which cannot guarantee the resulted signature's biological relevance. For this problem, we will integrate other data, such as biomolecular interaction data and develop new model and algorithm for extracting disease signature and drug signature with more biological relevance. Second, only overlap among genes that abnormally regulated in disease state and drug treatment and essential gene was quantified as drug's overall effect. The amplitude of differential expression and other information were not taken into account. With this in mind, we will further develop more relevant measure to define drug's efficacy and side effect by integrating multiple available information with disease signature and drug signature at hand.

ACKNOWLEDGMENT

Dr Zikai Wu is supported by PhD start funds of University of Shanghai for Science and Technology(Grant No. 1D-10-303-003) and Science Foundation for the Youth Scholars by Educational Commission of Shanghai(Grant No. 51-12-303-105). The authors would like to thank Xingming Zhao and Zhiping Liu for helpful discussion.

REFERENCES

[1] Mullard, A (2012). 2011 FDA drug approvals. Nature Reviews Drug Discovery 11: 91-94.

[2] Aronson, J(2007). Old drugs-New uses. British Journal of Clinical Phamacology 64: 563-565.

[3] Sirota,M et al(2011). Discovery and Preclinical Validation of Drug Indications Using Compendia of Public Gene Expression Data. Science Translational Medicine 3(96): 96ra77.

[4] Lum, P, Derry,J and Schadt,E(2009). Integrative genomics and drug development. Pharmacogenomics 10:203-212.

[5] Luo,H et al.(2011). DRAR-CPI: a server for identifying drug repositioning potential and adverse drugreactions via the chemical-protein interactome. Nucleic Acids Research 1(7): doi: 10.1093/nar/gkr299.

[6] Yang,L et al.(2011).Exploring Off-Targets and Off-Systems for Adverse Drug Reactions via Chemical-Protein Interactome-Clozapine-Induced A-granulocytosis as a Case Study. PLoS Computational Biology 7(3): e10002016.

[7] Schadt,E, Friend,S and Shaywitz,D(2009). A network view of disease and compound screening. Nature Reviews Drug Discovery 8: 286-295.

[8] Qu,X, Gudivala,R et al(2009). Inferring novel disease indications for known drugs by semantically linking drug action and disease mechanism relationships. BMC Bioinformatics 10(S5): S4.

[9] Keiser,M et al(2009). Predicting new molecular targets for known drugs. Nature 462: 171-181.

[10] Xie,L, Li,J, Xie,L and Bourne,P(2009). Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of CETP inhibitors. Plos Computational biology 5: e1000387.

[11] Kinnings,S et al(2009). Repositioning the safe medicine Comtan to treat multi-drug and extensively drug resistant tuberculosis. Plos Computational biology 5: e1000423.

[12] Gottlieb,A, Stein,G, Ruppin,E and Sharan,R(2011). PREDICT: a method for inferring novel drug indications with application to personalized medicine. Molecular Systems Biology 7: 496.

[13] Yang,L and Agarwal(2011). Systematic Drug Repositioning Based on Clinical Side Effects.Plos One 6(12): e28025.

[14] Zhao,SW and Li,S(2012). A co-module approach for elucidating drug-disease associations and revealing their molecular basis. Bioinformatics doi:10.1093/bioinformatics/bts057.

[15] Lamb,J et al(2006). The Connectivity Map: Using gene expression signatures to connect small molecular, genes and disease. Science 313: 1929-1935.

[16] Iorio,F et al(2010). Discovery of drug mode of action and drug repositioning from transcriptional responses. Proceedings of National Academic Science(USA) 107(33): 14621-14626.

[17] Hu,G and Agarwal,P(2009). Human Disease-Drug based on Genomic Expression Profiles. Plos one 4(8): e6536.

[18] Shigemiu,D, Hu,ZJ, Hung,JH, Huang,CL, Wang,YJ and DeLisi,C(2012). Using Functional Signatures to Identify Repositioned Drugs for Breast, Myelogenous Leukemia and Prostate Cancer. Plos Computational Biology 8(2): e1002347.

[19] Barrett,T, Suzek,TO, Troup,DB, Wilhite,SE, Ngau,WC et al(2005). NCBI GEO: mining millions of expression profiles-databases and tools. Nucleic Acid Research 33: D562-566.

[20] Chen,R, Li,L and Butte,J(2007). AILUN: Reannotating gene expression data automatically. Nature Methods 4: 879.

[21] Smyth,GK(2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology 3(1): 3.

[22] Zhang,R, Ou,HY and Zhang,CT(2004). DEG: a database of essential genes. Nucleic Acid Research 32(s1): D271-272.

[23] Jin,G, Fu,C, Zhao,H, Cui,K, Chang,J and Wong,S(2012). A Novel Method of Transcriptional Response Analysis to Facilitate Drug Repositioning for Cancer Therapy. Cancer Research 72(1): 33-44.

[24] Lipkus,AH(1999). A proof of the triangle inequality for the Tanimoto distance. Journal of Mathematical Chemistry 26(1-3): 263-265.

978-1-4673-4398-5/12/$31.00 ©2012 IEEE