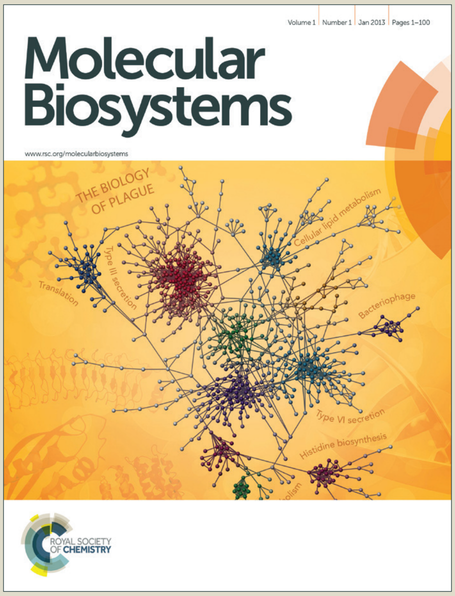
Molecular
BioSystems

Accepted Manuscript


Volume 1 | Number 1 | Jan 2013 | Pages 1–100
Molecular
Biosystems
www.rsc.org/molecularbiosystems
THE BIOLOGY OF PLAGUE
Cellular lipid metabolism
Type III secretion
Translation
Bacteriophage
Type VI secretion
Histidine biosynthesis
ROYAL SOCIETY OF CHEMISTRY

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Journal Name RSCPublishing

ARTICLE

Identifying the causative proteins of similar side effect pairs to explore the common molecular basis of these side effects

Cite this: DOI: 10.1039/x0xx00000x

Received 00th January 2012,
Accepted 00th January 2012

DOI: 10.1039/x0xx00000x

www.rsc.org/

Yunfeng Wang[a], Xiujie Chen[*], Ruizhi yang[b], Lei Liu[b], Yuelong Chen[b], Xiangqiong Liu[b]

Abstract

Drug side effects, or adverse drug reactions (ADR), have become a major public health concern and often caused drug development failure and withdrawal. Some ADRs always occur concomitantly. Therefore, identifying these ADRs and their common molecular basis can better promote the prevention and treatment. In this paper we predicted the potential proteins for similar mechanism ADR pairs based on three layers of information. i) The drug co-occurrence between a pair of ADRs. ii) The correlation between a protein and an ADR-pairs based on the co-occurrence of drugs. iii) The interaction between these proteins within the protein-protein interaction (PPI) network. Methods of randomization and functional annotation are used to investigate and analyze the relation between causative proteins and similar ADR pairs. The prediction accuracy of relation between similar ADR pairs and related proteins reached 80%, and it increases more with the number of drugs shared by the ADR pairs. From the ADR network made of single ADRs from predicted similar ADR pairs, we found some ADRs are involved in multiple ADR pairs. Functional analysis of these ADR related proteins suggests the similar molecular basis is shared by multiple ADR pairs containing the same ADR. And these ADR pairs are almost caused by the same drug sets. The results of this study are reliable and can provide theoretical basis for better prevention and treatment of the ADRs always occurring concomitantly.

Introduction

Drug side effects, or adverse drug reactions (ADR), have become a major public health concern and caused drug development failure and withdrawal[1]. How to prevent and avoid side effects induced by drugs is a challenging issue in drug development and clinical practices. The key problem is to understand the pathogenesis, namely which proteins and/or biological pathways correlate to the side effect. Currently, there is little knowledge about the association between ADRs and ADR-related proteins[2]. What's more, the related proteins of some ADRs are completely lack. Therefore, we urgently need to systematically identify ADR related proteins to promote experimental researches.

Some strategies in this field have been studied and achieved remarkable successes. The docking analysis of drug structure binding sites was one of the early predicting method for potential drug targets. Xie et al[3] generated off-target binding networks by comparing the structure of ligand-binding sites in all known protein structures. Using this analysis, the authors identified possible off-target for torcetrapib even though the binding site of ligand is not fully described. System biology analysis was another efficient method to predict relation between drug, target and side effects using qualitative graphical models or quantitative mathematical model[4]. In a recent study, Philip Bourne and colleagues[3] have used a chemical systems biology approach to explain the serious side-effects of a drug that was being trialed for prevention of cardiovascular disease. First, they can provide more detailed descriptions (even signatures) of drug effects, and second, they can provide a framework for the design of novel therapeutic strategies[5]. Griet Laenen[6] also used graph model of protein interaction network to predict drug targets, combined with drug gene expression profile. This approach relies on the analysis of gene expression following drug treatment in the context of a functional protein association network. Eugen Lounkine[7] developed an association metric to prioritize predicted off-targets, creating a drug–target–adverse drug reaction network. Both Griet Laenen and Eugen Lounkine's work provide insight on the importance of network relation among drugs, targets and side effects. In addition, many other studies regard the similarity as a measure to predict relation among drugs, targets and side effects. Liat Perlman[8] introduced a novel framework -- Similarity-based Inference of drug-TARgets (SITAR) -- for incorporating multiple drug-drug and gene-gene similarity measures for drug target prediction. Lucas Brouwers[9] took similarity between side effects into account and found that side-effect similarity of drugs could be caused by overlapping of drug targets and the close neighbor targets in network.

There are another two representative methods. The first one is proposed by Mizutani S' group which identified correlated sets of side effects and proteins. They proposed an algorithm using sparse canonical correlation analysis (SCCA) based on the drug co-occurrence of drugs in protein-binding profiles and side effect profiles[2]. The second one is proposed by Kuhn M and his colleagues. Kuhn M utilized the fisher's exact test to identify the significant relations between ADRs and proteins based on drug-protein and drug-ADR profiles[10]. Kuhn's group also classified predicted proteins into categories and clustered them based on the co-occurrence of drugs in the protein binding profiles and side effect profiles. Nearly all these methods focused on single ADRs, however, many ADRs are not independent and they often occur concomitantly[11]. For instance, sweat increasing, lachrymation and pinpoint pupil always occur concomitantly after accidental administration of organophosphorus pesticide. Studies have shown that the sweat increasing, lachrymation and pinpoint pupil occur together because these ADRs are associated with the same ADR related protein cholinesterase (ChE)[12]. The pesticide inhibits the activation of ChE, and then acetylcholine (Ach) cumulates which leads to the ADRs above occurring at the same time. Therefore, identifying these ADRs and their common related proteins is important to the prevention and treatment. However, the surveys on this direction remain lacking.

Studies have shown that drugs inducing similar ADRs have common substructure and similar protein profiles[13, 14]. Therefore, the co-occurrence of drugs in ADR pairs reflects the correlation between their some related proteins and similar ADR pairs.

Some works have shown that proteins with interaction and close distance across the PPI network tend to share the same pathways[6, 9]. And drug targets with similar pharmacological action tend to have interaction with one another across the PPI network[15]. We can gain insight from these studies that side effects similarity and protein network contribute to the prediction of relation between side effect and proteins. Therefore, we supposed proteins inducing a pair of ADRs tend to interact with each other directly in the PPI network.

In this paper we assessed the similarity between two ADRs with the common drugs (also called co-occurrence drugs) and found the candidate protein set for ADR pair based on co-occurrence of drugs. Then we screened proteins with direct interactions across the PPI network as predicted related proteins set of ADR pairs.

Our purpose is to find proteins shared by an ADR pair with the similar pathogenic mechanism. It can firstly help explain the mechanism of two ADRs always occurring concomitantly, and may also help study single ADRs without related proteins information so far. Finally the shared proteins provide more effective treatment schemes on ADRs occurring concomitantly.

Methods

(1) DATA

(1.1) Discovery data The ADR-drug relations (frequency > 0.01) were extracted from the SIDER 2 database[16]. The higher frequency in patients indicates a closer relationship between drugs and ADRs. The influence of rare variance and epigenetic effect were eliminated via limiting the frequency (frequency > 0.01) of ADRs in SIDER. We also removed ADRs with ambiguous names such as pain, ache and so on, because these ADRs have unclear descriptions and always correlate with too many drugs. And then we standardized ADR terms and drug names according to MedDAR[17] dictionary and PubChem[18], respectively. Drug-related proteins were extracted from three databases (SUPER TARGET and Matador[19], STITCH[20] and PROMISCUOUS[21]), and then converted into gene symbol. To preserve the specific relation between predicted proteins and ADRs, we removed the top 10 common proteins (Figure1). As observed, all these common proteins belong to the kinase and metabolic enzymes. To guarantee the statistic performance, we removed ADRs that are related to less than three drugs. Finally, we obtained 909 side effects, 327 drugs, 6670 ADR-drug relations in total and 7297 drug-protein relations.

(1.2) Validation data The validation data consists of two parallel sections. The first section was text mining from published literatures and databases. If the relation between an ADR and a protein exists in literatures or databases, this predicted relation can be verified. The second section was mapping proteins to adverse outcome pathway (AOP)[22]. HTS assays are focused on different adverse outcome pathways and more than 10,000 small molecules were screened. To verify the pair of similar ADRs can possibly occur together and both caused by the same protein, we classified the ADR terms and proteins AOP terms based on ICD disease classification system[23].
The ICD main page can be got from this website (http://www.who.int/classifications/icd/en/).
The proteins related AOPs were downloads from AOPwiki (https://aopkb.org/aopwiki/index.php/Main_Page).

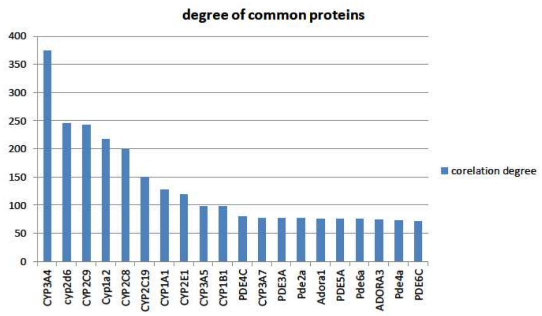


Figure 1 degree distribution of common proteins
The vertical axis represents the number of drugs related with the protein in horizontal axis. The horizontal axis represents the top 20 proteins with the largest degrees. All of these common proteins correlate with more than 50 drugs and the top 10 proteins are all metabolic enzymes. To preserve the specificity between proteins and ADRs we removed the top 10 common proteins from the drugs' related proteins.

(2) Proteins prediction for similar ADR pairs

(2.1) Screening similar ADR pairs A Tanimoto coefficient (TC)[24] means the number of drugs shared by two ADRs by the total number of drugs inducing the two ADRs. We used the TC value to evaluate the similarity between any two ADRs. To confirm the threshold of the TC for each similar ADR pair, we utilized randomizations to screen the statistical significant ADR-ADR relations. First, drugs corresponding to an ADR were randomly replaced by the same number drugs from the whole drug group. Second, the new TC value was calculated for each pair of ADRs after once exchange. This above pipeline was repeated for 1,000 times and the P value was calculated from the distribution. ADR pairs with the P value less than 0.05 were supposed to be the similar ADR pairs.

(2.2) Screening specific proteins shared by similar ADR pairs The drugs' related proteins were supposed to be the potential causative proteins for the corresponding ADRs. To identify proteins significantly related to the ADR pairs, we randomly exchanged drugs with the same number of drugs from the whole drug sets, which were repeated 1000 times. The raw score is counted using the formula [Si=Ni/N, Si means the raw score of protein i, Ni means the number of drugs related to the protein i, N means the number of drugs inducing the ADR]. After generating 1000 times, we got the background distribution of new scores with the similar formula [Si=Ni'/N, Si means the random score of protein I, Ni' means the number of drugs related to protein i after exchange]. Last, we determined whether a protein correlates with an ADR pair by transforming the rank of raw score in distribution into P value. We chose the proteins with a P value less than 0.05 as the candidate proteins[25].

For each ADR pair, proteins with direct interactions within the candidate proteins tend to have similar biological functions and induce similar phenotypic effects. STRING 9.05[26] is used to identify the interaction relation between candidate proteins. STRING is a database of known and predicted protein interactions. The interactions include direct (physical) and indirect (functional) associations derived from four sources: genomic context, experiment, co-expression and previous knowledge. The interaction type was restricted to neighbor, database, and experiment with medium confidence 0.400. We filtered the independent proteins which were assumed to be low likelihood. Thus, the proteins with interactions were supposed to be the specific causative proteins of ADR pairs.

(2.3) Function analysis of predicted protein set for each pair of ADRs We analysed the functional specificity of each ADR pair via Gene Ontology. We performed two kinds of enrichment: biological process (BP) and molecular function (MF). Each protein set was enriched on the GO terms using DAVID[27]. BP and MF were restricted to the GOTERM_BP_FAT and GOTERM_MF_FAT, respectively.

(3) Results verification

The verification process included two parts, text mining and disease classification. Firstly, it's a true positive result if the relation between an ADR and protein exists in the published papers in the PubMed abstracts or in the disease related database such as DART and CTD. Secondly, we downloaded all the adverse outcome pathways (AOPs) of the predicted ADR related proteins from AOPwiki. Then all the adverse outcome pathways and adverse drug reactions were classified into different terms according to the ICD disease classification system. As the ADRs and AOPs both describe drugs or proteins adverse actions to body, both ADRs and AOPs were treated as disease terms and were manually classified according to the keywords in each ADR and AOP term. It's a true positive result if the ADR and protein's AOP were classified into a same disease term.

Result

1. Similar ADR pairs identification

909 ADRs with a frequency of more than 1% were extracted from the SIDER database and formed 412686 pairs. The TC value of each ADR pair was calculated based on the drug co-occurrence. A higher TC value suggests a higher similarity between two ADRs and this pair is more likely to share the same causal proteins. Randomization was generated to reorder drugs to confirm the TC threshold of each pair of ADRs. Finally, 319 similar ADR pairs were significantly similar with P values less than 0.05 (additional file 1).

We found that one ADR present in multiple similar ADR pairs. Therefore, we constructed an ADR-ADR network to investigate the number of each ADR composing similar ADR pairs. In this network, nodes represent ADRs and edges represent similar relations（figure 2）. The degree of each ADR in the network is different, ranging from 1 to 23 (Figure 3). The larger degree an ADR has in the network, the more similar ADR pairs it belongs to. There are totally 10 side effects (16.67%) with a degree larger than 10 such as paresthesias and oedema. And 15 side effects (25%) have the degree larger than 5 but less than 10, such as somnolence and indigestion. The ADR with the highest degree is dyspnea, which belongs to 23 similar ADR pairs.

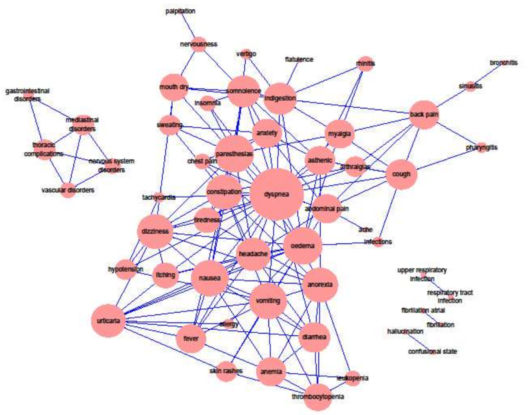


Figure 2 Network of single ADRs composing similar ADR pairs. The nodes represent ADRs and edges represent similar relations between ADRs. The node size represents the degree of node in the network. Some ADRs with high degree compose similar ADR pairs with many ADRs while some ADRs with low degree compose similar ADR pairs with only specific ADRs.

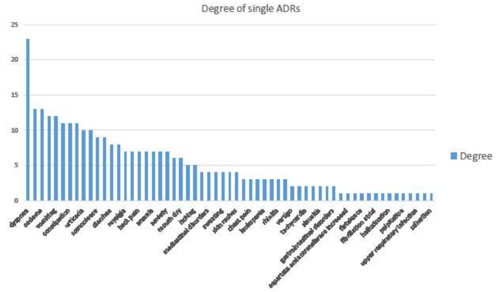


Figure 3 Degree of each ADR in the network

The degree of each ADR represents the number of ADRs with similar relations to it. All the single ADRs in the network are analysed and the degree distribution can be seen in Figure 3. The horizontal axis is the list of single ADRs and the vertical axis represents the degree of each ADR.

2. Significant protein sets for each ADR pair

Co-occurrence drugs were defined as drugs inducing either ADR in a similar pair. We obtained 683 drugs and their 814 related proteins (additional file 2). 8116 ADR pair-protein relations are significant after randomization. For a pair of ADR, we extracted proteins with direct interactions in PPI from its related proteins. Consequently, 1908 relations consisted of 153 ADR pairs and 301 related proteins were significant in both statistic and biology (additional file 3). We used GO enrichment analysis[28] to find the specific function of these specific proteins, and constructed the network consisting of ADR pairs and GO terms (figure 4). We found the functions of ADR pairs were completely overlapped or partially overlapped, which contained the same ADR. For example, vascular disorders, thoracic complications; gastrointestinal disorders, mediastinal disorders; gastrointestinal disorders, thoracic complications are three similar ADR pair. All these three similar ADR pairs are related with cysteine metabolic process, response to nitrosative stress and sulfur amino acid catabolic process. We used node size to indicate the degree of each node in the network. The big nodes in purple are ADR pairs involving broad biology activities, such as dizziness and hypotension, which suggests the complexity of this ADR pair's pathogenesis. The big nodes in red are biology processes involved in many different ADR pairs, such as estrogen receptor activity and phosphodiesterase activity, and we found such biology processes are always involved in ADR pairs consisting of common ADRs rather than serious ones. For instance, phosphodiesterase activity is involved in 38 similar ADR pairs, and these ADR pairs mainly concentrate on the ADR of anorexia, nausea, vomiting and other common symptoms. Meanwhile, some functions with low degree only correlate with specific ADR pairs, such as the function of down regulation of smooth muscle contraction that only correlates with the ADR pair of nervousness and palpitation. The palpitation is a severe side effect which may cause arrhythmia and even heart failure. The smooth muscle contraction abnormal induced by drugs has been proved to be a main reason of palpitation[29]. The abnormal release of neurotransmitters induced by dysfunction of smooth muscle contraction can also cause nervousness[30].

All these findings suggest that some functions are involved in many ADRs, and they usually regulate multiple biology activities and therefore they always correlate with common ADRs. In contrary, some functions only correlate with specific ADRs, which are usually severe ADRs. These functions often regulate specific biology processes and severe ADRs occur when these specific processes are abnormal.

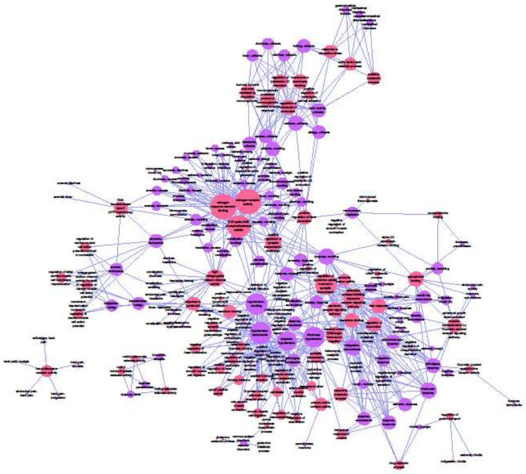


Figure 4 The network of similar ADR pairs and functional annotations

The purple and red nodes in the network represent ADR pairs and functional annotations, respectively. The node size represents the degree of each node in the network. So the big nodes in purple are ADR pairs involving various biology functions and big nodes in red are functions related with numerous ADR pairs.

3. Validation of the results

Two methods were used in our validation. Firstly, we searched the relation between an ADR and a protein in PubMed and related databases. If the relation was reported in published literatures or databases, it could be treated as a positive result. Secondly, we mapped proteins onto adverse outcome pathway (AOPs) to investigate whether the protein and ADR have functional relationship. If an ADR and the protein related AOP were classified in the same disease category by international classification of diseases (ICD), this relation could be treated as a positive result. Detailed biological interpretation were further produced on some notable results.

3.1 Text mining We utilized text mining to validate the results of ADR pairs from published literatures and databases. 80% of 1908 significant ADR pair-protein relations were validated by text mining.

The validation results were classified into three categories. In 22% of the results, both ADRs in a similar pair can be proved related with the predicted protein. In 58% of the results, either ADR can be proved to associate with the predicted protein. ADR pair-protein relations in the rest 20% are novel predictions which haven't been researched or verified (additional file 4).

In order to discover the influence of drug numbers to predicting accuracy, we analysed the correlation between sharing drug numbers and predicting accuracy (Figure 5). We found the more drugs a pair of similar ADRs shares, the higher prediction accuracy it reaches.

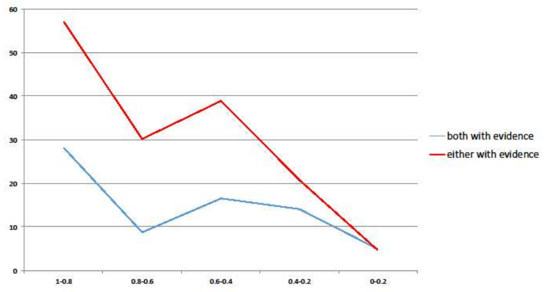


Figure 5 The correlation between the number of drugs shared by two similar ADRs and the accuracy
The horizontal axis indicates the intervals of accuracy and vertical axis indicates the number of drugs shared by two ADRs. The red curve indicates either ADR-protein relation in an ADR pair can be proved. The blue curve indicates that both ADR-protein relations in an ADR pair can be proved.

3.2 Disease classification All the ADR terms and proteins related adverse outcome pathways were all classified according to the ICD system. Finally, in 812 (43%) of 1908 results both of the two ADRs were classified into the same disease classification with the predicted proteins' AOPs. In another 620 (32%) results, either of the ADR were classified into the same classification with predicted proteins' AOPs. In addition, 366 (19%) results were not verified because the proteins were failed to map onto AOPs. The rest 110 (6%) results were unverified or two ADRs belong to diverse disease classifications. (additional file 4)

Integrating the text mining and disease classification result, 54% of predicted relations which present between both two ADRs and the protein were proved by either text mining or disease classification. And 5% of the relations between two ADRs and the protein were proved by neither of the two methods. The detailed verification ratio of each categories were shown in Table 1.

Table 1 verification ratio of each predicted relation categories.

categories	text mining	disease classification	integration
both ADRs proved	22%	43%	54%
single ADR proved	58%	32%	41%
neither ADR proved	20%	25%	5%

3.3 Biology interpretation We further illustrated the three kinds of validated ADR pairs via biological interpretation, respectively.

Taking anemia, anorexia（C0002871, C0003123）as an example, we predicted 14 related proteins, in which 6 proteins (ADA，CACNA1I，ESR2，CACNA1G，PDE4C，PDE4D) have been proved related to the two ADRs. We further analysed the biological mechanism of ADA inducing these two ADRs. Many literatures have showed that anemia and anorexia always occur concomitantly in clinic[31]. In our findings, anemia and anorexia is a pair of similar ADRs (TC=0.86) and they share 17 approved drugs (Figure 6(A)), 12 of which are classified by ATC as Immunomodulatory drugs. The rest drugs are two anti-infection drugs, two nervous system drugs and one antiparasitic drug. It should be pointed that, as for the 17 related drugs, even though they have diverse indications, most of them target ADA. This gene encodes adenosine deaminase that catalyzes the hydrolysis of adenosine to inosine. Raised levels of this enzyme have been proved to associate with congenital hemolytic anemia[32]. Abnormity of this enzyme causes immunodeficiency disease, in which both B and T lymphocytes are impaired. Studies also found that anorexia is associated with weight loss in patients with acquired immune deficiency syndrome[33]. Therefore, when drugs perturb ADA's function, abnormal adenosines metabolite makes damage to T and B immune cells, and eventually cause immune system disease with the symptom of anemia and anorexia.

Though reported in clinical adverse reaction monitoring center and literatures, many ADRs' molecular mechanisms are still unclear. In 58% of our predictions, only one ADR in a similar pair can be validated. These related proteins can be inferred as related proteins for the other ADR due to the similar mechanism within the same ADR pairs.

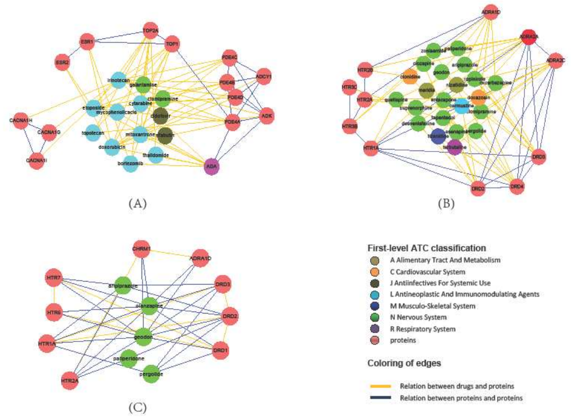

(A)
(B)
(C)
CHRM1
ADRA1D
HTR7
DRD3
HTR2C
DRD2
HTR1A
DRD1
HTR2A
First-level ATC classification
A Alimentary Tract And Metabolism
C Cardiovascular System
J Antiinfectives For Systemic Use
L Antineoplastic And Immunomodulating Agents
M Musculo-Skeletal System
N Nervous System
R Respiratory System
proteins
Coloring of edges
Relation between drugs and proteins
Relation between proteins and proteins

Figure 6 category annotations of drugs and proteins
There are two kinds of nodes in Figure 6, drugs and proteins. Edge in blue represents the interaction between proteins and edge in yellow represents the relation between drugs and proteins. All drugs can be classified to different categories according to the ATC code. And proteins can also be divided to several categories. (A) Drugs and proteins of anemia, anorexia. (B) Drugs and proteins of anxiety, somnolence. (C) Drugs and proteins of indigestion, dyspnea.

ADRA2A was predicted to be the causative gene of anxiety and somnolence (C0003467, C2830004) (Figure 6(B)). Adrenine and its receptors (ADRAs) play crucial roles in brain development and regulation of mood[34]. The relation between ADRA2A and somnolence hasn't been researched yet. However, dysfunction in neurotransmitter and hormone signal contribute to this disease process[35]. Adrenergic receptor ADRA2A impacts on the neurotransmitter content in central nervous system. Therefore, drugs targeting this protein will induce various neural symptoms such as somnolence.

There are also 20% of the ADR pair-protein relations with no evidence. These relations are significant in both statistics and biology. We found biological correlation between the predicted protein and ADR pair using NCBI GENE database (or molecular function from GO term).

Taking the indigestion, dyspnea (C0013395, C0013404) for instance, records from clinic have showed patients with gastrointestinal symptoms such as indigestion may also have pulmonary symptoms like dyspnea[36]. And we predicted 10 proteins, containing five 5-HT receptors, three DA receptors, one cholinergic receptor and one adrenergic receptor. Only DAD2 and HTR2C can be proved inducing dyspnea. Even though the other genes haven't been validated to cause this pair of ADRs, we found all these genes participate in the biological process of indigestion and dyspnea.

These proteins are important neurotransmitter mainly distributing on the gastrointestinal and tracheal smooth muscle such as HTRs and cholinergic receptor [37]. DA receptors mainly distribute in the coronary vascular, gastrointestinal vascular[37]. These proteins regulate the bronchial caliber and blood flow in cardiovascular and gastrointestinal tracts. When drugs target these proteins, indigestion and dyspnea may occur as a result of the caliber and blood flow change in gastrointestinal tracts and bronchus.

Discussion

In clinical practices, we find some ADRs are not independent but correlating with each other. Hence, the potential relation between ADRs is an important aspect to consider when studying the mechanism of ADR occurrences. Nevertheless, the studies on this direction remain rare and their related proteins remain undefined, which inhibits the researches of these ADRs occurrence mechanism, failing to prevent and treat these ADRs.

In this paper, we integrated sources of drug-protein and drug-ADR relations. Causative proteins were predicted according to the hypothesis that ADRs occur not relying on the isolated proteins but on the proteins with interactions[9]. We predicted causative proteins of similar ADR pairs based on three layers information: i) The drug co-occurrence between a pair of ADRs. ii) The correlation between a related protein and an ADR pairs based on the co-occurrence of drugs. iii) The interaction between these proteins within the protein-protein interaction (PPI) network. Randomization was used to identify significant ADR pair-protein relations. All the relations with P value less than 0.05 were significant which suggests that the relation between ADR pair and proteins was not randomly generated. Then we identified the protein set in which proteins interact with each other as predicted proteins set.

This study predicted similar ADR pairs based on drugs co-occurrence, which lowered the promiscuity of drug chemical structure, highlighted the common substructural features of the co-occurrence drugs. We also predicted ADR pairs' causative proteins according to the interactions among them, which further highlights the linkage characteristics among proteins in certain functions. Compared with the results of single ADR related proteins prediction and the previous methods[2, 10], the outcome turned out better that the accuracy (65%) of single ADRs equals with previous studies and the accuracy of ADR pairs reached 80%, which suggests the forementioned two layer information (the common substructural features of the co-occurrence drugs and the linkage characteristics among proteins in certain functions) enhance the performance in similar ADR pair related proteins prediction .

Functional enrichments found that the functions of multiple similar ADR pairs containing the same ADR are involved identical or similar biology functions. The similar molecular basis of multiple ADR pairs explains the reason that some ADRs occur simultaneously in clinical practices.

We also found the functional differences between a single ADR and the ADR pairs including the single ADR such as back pain. Back pain induced by 219 drugs is related with 10 biology functions (additional files 4). When back pain composes similar ADR pairs with abdominal pain, indigestion, myalgia, sinusitis, and arthralgias, they are all associated with benzodiazepine receptor activity, and the drug sets inducing back pain and the five ADRs stated above are similar. This phenomenon further suggests that drug co-occurrence highlights the drugs' common substructures inducing similar ADR pairs via similar mechanism. (Figure 7)

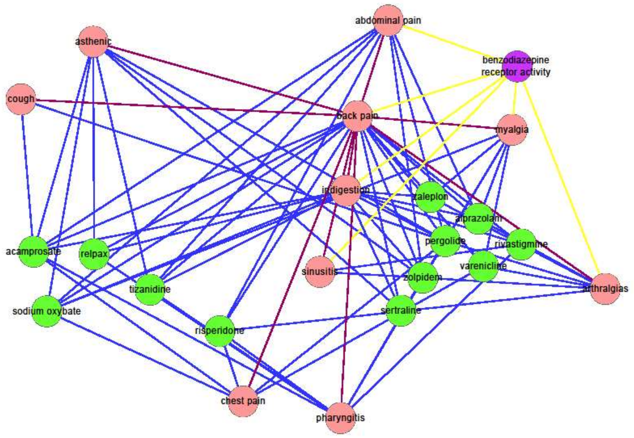


Figure 7 network of the function of back pain and its similar ADRs The network consist of 9 ADRs (in red) similar to back pain and drugs (in green) shared by these ADR pairs. The function of benzodiazepine receptor activity is enriched using the predicted proteins from these ADR pairs. Edges in yellow indicate the relation between an ADR and the function. Edges in blue represent the relation between drugs and ADRs. Same drug set is shared by these ADRs with relation to the benzodiazepine receptor activity.

It's noteworthy that some ADRs only occur in specific populations which are mostly severe adverse effects without definite mechanism. Recent studies tend to explain these ADRs at the genetic level such as mutation and epigenetic[38, 39]. As we filtered out the influences of rare variance and epigenetic effect via limiting the frequency (frequency > 0.01) of ADRs in SIDER, our method is not applicable to predict the related proteins of these ADRs resulting from the mutation and epigenetic. In the following study, we may focus on the relation between ADR and gene alteration.

Conclusions

Similar side effects always occur concomitantly in clinic and compromise therapeutic efficacy. These ADRs share the similar biological processes and pathways, therefore, identifying the causal protein shared by both ADRs will uncover the molecular basis of ADR pairs occurring concomitantly and be helpful to better prevent and treat ADRs by targeting their sharing biological process and pathway.

In this paper we predict potential related proteins of ADR pairs based on three layers of information. This proposed method is expected to be helpful in other ways. Firstly, we can identify the related proteins of both single ADR and similar ADR pairs. Secondly, the causal proteins shared by two similar ADRs may be regarded as the biomarker or a new therapeutic target. Thirdly, drugs targeting the causal proteins may induce the similar ADRs occurrence, thus these drugs should be used more carefully or choose alternative medicine. Identifying the causal proteins of similar ADRs provides a new insight to elucidate the mechanism of occurring concomitantly and optimize therapeutic medicine choice.

Acknowledgements

The authors would like to acknowledge the support of the Funds by the National Natural Science Foundation of China [Grant No. 61372188], and Provincial Education Department Project of Heilongjiang, China [Grant No. 12541331].

Notes and references

[a] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, Heilongjiang 150081, PR China

† additional file 1.xlsx

† additional file 2.xlsx

† additional file 3.xlsx

† additional file 4.xlsx

1. K. M. Giacomini, R. M. Krauss, D. M. Roden, M. Eichelbaum, M. R. Hayden and Y. Nakamura, *Nature*, 2007, 446, 975-977.
2. S. Mizutani, E. Pauwels, V. Stoven, S. Goto and Y. Yamanishi, *Bioinformatics*, 2012, 28, i522-i528.
3. L. Xie, J. Li, L. Xie and P. E. Bourne, *PLoS computational biology*, 2009, 5, e1000387.
4. R. T. Peterson, *Nature chemical biology*, 2008, 4, 635-638.
5. N. Plant, *Toxicology*, 2008, 254, 164-169.
6. G. Laenen, L. Thorrez, D. Bornigen and Y. Moreau, *Molecular bioSystems*, 2013, 9, 1676-1685.
7. E. Lounkine, M. J. Keiser, S. Whitebread, D. Mikhailov, J. Hamon, J. L. Jenkins, P. Lavan, E. Weber, A. K. Doak, S. Cote, B. K. Shoichet and L. Urban, *Nature*, 2012, 486, 361-367.
8. L. Perlman, A. Gottlieb, N. Atias, E. Ruppin and R. Sharan, *Journal of computational biology : a journal of computational molecular cell biology*, 2011, 18, 133-145.
9. L. Brouwers, M. Iskar, G. Zeller, V. van Noort and P. Bork, *PloS one*, 2011, 6, e22187.
10. M. Kuhn, M. Al Banchaabouchi, M. Campillos, L. J. Jensen, C. Gross, A. C. Gavin and P. Bork, *Molecular systems biology*, 2013, 9, 663.
11. X. Chen, X. Liu, X. Jia, F. Tan, R. Yang, S. Chen, L. Liu, Y. Wang and Y. Chen, *Sci Rep*, 2013, 3, 1744.
12. V. Dhull, A. Gahlaut, N. Dilbaghi and V. Hooda, *Biochemistry research international*, 2013, 2013, 731501.
13. M. Campillos, M. Kuhn, A. C. Gavin, L. J. Jensen and P. Bork, *Science*, 2008, 321, 263-266.
14. A. F. Fliri, W. T. Loging, P. F. Thadeio and R. A. Volkmann, *Nat Chem Biol*, 2005, 1, 389-397.

15. F. Napolitano, Y. Zhao, V. M. Moreira, R. Tagliaferri, J. Kere, M. D'Amato and D. Greco, *J Cheminform*, 2013, 5, 30.

16. M. J. Jahid and J. Ruan, *Proceedings. IEEE International Conference on Bioinformatics and Biomedicine*, 2013, DOI: 10.1109/BIBM.2013.6732532, 440-445.

17. E. G. Brown, *Drug safety*, 2003, 26, 145-158.

18. Q. Li, T. Cheng, Y. Wang and S. H. Bryant, *Drug discovery today*, 2010, 15, 1052-1057.

19. S. Gunther, M. Kuhn, M. Dunkel, M. Campillos, C. Senger, E. Petsalaki, J. Ahmed, E. G. Urdiales, A. Gewiess, L. J. Jensen, R. Schneider, R. Skoblo, R. B. Russell, P. E. Bourne, P. Bork and R. Preissner, *Nucleic acids research*, 2008, 36, D919-922.

20. M. Kuhn, D. Szklarczyk, A. Franceschini, C. von Mering, L. J. Jensen and P. Bork, *Nucleic acids research*, 2012, 40, D876-880.

21. J. von Eichborn, M. S. Murgueitio, M. Dunkel, S. Koerner, P. E. Bourne and R. Preissner, *Nucleic acids research*, 2011, 39, D1060-1066.

22. C. Piroird, J. M. Ovigne, F. Rousset, S. M. Teissier, C. Gomes, J. Cotovio and N. Alepee, *Toxicology in vitro : an international journal published in association with BIBRA*, 2015, DOI: 10.1016/j.tiv.2015.03.009.

23. V. A. Wood, D. T. Wade, R. L. Hewer and M. J. Campbell, *Journal of neurology, neurosurgery, and psychiatry*, 1989, 52, 449-458.

24. J. W. Godden, L. Xue and J. Bajorath, *Journal of chemical information and computer sciences*, 2000, 40, 163-166.

25. S. Lee, K. H. Lee, M. Song and D. Lee, *BMC bioinformatics*, 2011, 12 Suppl 2, S2.

26. A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. von Mering and L. J. Jensen, *Nucleic acids research*, 2013, 41, D808-815.

27. X. Jiao, B. T. Sherman, W. Huang da, R. Stephens, M. W. Baseler, H. C. Lane and R. A. Lempicki, *Bioinformatics*, 2012, 28, 1805-1806.

28. E. Camon, D. Barrell, V. Lee, E. Dimmer and R. Apweiler, *In Silico Biol*, 2004, 4, 5-6.

29. F. Pelliccia, P. Gallo, C. Cianfrocca, G. d'Amati, P. Bernucci and A. Reale, *Int J Cardiol*, 1990, 29, 47-54.

30. X. G. Gan, R. H. An and D. B. Zhong, *Zhonghua Nan Ke Xue*, 2006, 12, 175-177.

31. A. Maccio, C. Madeddu and G. Mantovani, *Expert Opin Pharmacother*, 2012, 13, 2453-2472.

32. V. Casanova, I. Naval-Macabuhay, M. Massanella, M. Rodriguez-Garcia, J. Blanco, J. M. Gatell, F. Garcia, T. Gallart, C. Lluis, J. Mallol, R. Franco, N. Climent and P. J. McCormick, *PLoS One*, 2012, 7, e51287.

33. L. M. Borgelt, K. L. Franson, A. M. Nussbaum and G. S. Wang, *Pharmacotherapy*, 2013, 33, 195-209.

34. E. K. Lambe, S. G. Fillman, M. J. Webster and C. Shannon Weickert, *PLoS One*, 2011, 6, e22799.

35. S. M. Rothman and M. P. Mattson, *Neuromolecular medicine*, 2012, 14, 194-204.

36. T. Al-Abdoulsalam and M. A. Anselmo, *Canadian respiratory journal : journal of the Canadian Thoracic Society*, 2011, 18, 81-83.

37. J. M. van Rooyen and J. Offermeier, *South African medical journal = Suid-Afrikaanse tydskrif vir geneeskunde*, 1981, 59, 329-332.

38. S. Leone, M. L. Zuccoli, C. Fucile, S. Storace, A. Martelli and F. Mattioli, *Journal of clinical pharmacy and therapeutics*, 2012, 37, 733-735.

39. A. B. Csoka and M. Szyf, *Medical hypotheses*, 2009, 73, 770-780.