Molecular BioSystems

PAPER

View Article Online
View Journal | View Issue

Cite this: *Mol. BioSyst.*, 2014, **10**, 1126

Drug repositioning by applying 'expression profiles' generated by integrating chemical structure similarity and gene semantic similarity†

Fujian Tan,‡[a] Ruizhi Yang,‡[a] Xiaoxue Xu,‡[a] Xiujie Chen,‡*[a] Yunfeng Wang,[a] Hongzhe Ma,[a] Xiangqiong Liu,[a] Xin Wu,[b] Yuelong Chen,[a] Lei Liu[a] and Xiaodong Jia[a]

Received 16th December 2013,
Accepted 13th February 2014

DOI: 10.1039/c3mb70554d

www.rsc.org/molecularbiosystems

Drug repositioning, also known as drug repurposing or reprofiling, is the process of finding new indications for established drugs. Because drug repositioning can reduce costs and enhance the efficiency of drug development, it is of paramount importance in medical research. Here, we present a systematic computational method to identify potential novel indications for a given drug. This method utilizes some prior knowledge such as 3D drug chemical structure information, drug–target interactions and gene semantic similarity information. Its prediction is based on another form of 'expression profile', which contains scores ranging from −1 to 1, reflecting the consensus response scores (CRSs) between each drug of 965 and 1560 proteins. The CRS integrates chemical structure similarity and gene semantic similarity information. We define the degree of similarity between two drugs as the absolute value of their correlation coefficients. Finally, we establish a drug similarity network (DSN) and obtain 33 modules of drugs with similar modes of action, determining their common indications. Using these modules, we predict new indications for 143 drugs and identify previously unknown indications for 42 drugs without ATC codes. This method overcomes the instability of gene expression profiling derived from experiments due to experimental conditions, and predicts indications for a new compound feasibly, requiring only the 3D structure of the compound. In addition, the high literature validation rate of 71.8% also suggests that our method has the potential to discover novel drug indications for existing drugs.

Introduction

Drug repositioning, also known as drug repurposing or reprofiling, is the process of finding previously unknown indications for existing drugs.[1] One significant advantage of drug repositioning over traditional drug development is that repositioned drugs have already passed a significant number of toxicity and other tests, their safety is known, and the risk of failure for reasons related to toxicology is reduced.[2] In addition, repurposing drugs is less expensive and less time consuming than developing novel drugs, which will help to decrease the high costs of pharmaceutical R&D. Therefore, drug repositioning is of paramount importance in medical research.

In recent years, chemical structures and drug phenotypes have been widely used in drug repositioning studies. Chemical structure-based approaches[3,4] search for similar drugs based on the assumption that structurally similar drugs will tend to share common indications. Although some progress in drug repositioning has been made using these methods, chemical structure or complexity does not always coincide with function. For example, Yildirim *et al.*[5] provided evidence that most of the drugs that target the same proteins have distinct chemical structures, and Keiser *et al.*[6,7] demonstrated that structurally similar drugs may bind proteins with dissimilar functions. Therefore, drug repositioning based on chemical structure alone is insufficient. Another drug repositioning strategy is based on comparing phenotypes, such as expression profiles and side effects. For example, several drug repositioning studies have computed the similarity of gene expression profiles of different diseases,[8] captured similarities and differences in pharmacological effects and modes of action (MoAs) based on transcriptional responses[9] and used the signatures of gene expression profiles to connect small molecules.[10] Signature reversion approaches compare drug and disease's signatures. If their signatures are sufficiently negatively correlated then the

[a] *College of Bioinformatics Science and Technology, Harbin Medical University, 194 Xuefu Road, Harbin, Heilongjiang 150081, PR China. E-mail: chenxiujie@ems.hrbmu.edu.cn; Fax: +86-451-86615922; Tel: +86-451-86615922*

[b] *The Third Affiliated Hospital, Harbin Medical University, Harbin, Heilongjiang 150081, PR China*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c3mb70554d

‡ These authors contributed equally to this work.

drug is predicted to be effective against the disease.[11–13] Guilt-by-association approaches predict that drugs which share similar signatures have similar therapeutic applications.[14–17] All these approaches use the expression profiles derived from microarray experiments to generate drug action signatures.[7] Although these studies have shown many encouraging results in finding novel applications for existing drugs, there remains room for improvement. For instance, in general, studies based on expression profiles examine those genes (called "signature genes") that show statistically significant differences in expression in cells treated with the drugs (*i.e.*, in different cell lines, with different drug concentrations, and so on). This is typically a small fraction of the genome, which cannot reflect the functional alteration of the whole genome for several reasons. First, the genes expressed in a specific cell line or tissue represent only a subset of the whole genome, as a direct consequence of cellular differentiation; many genes are turned off and cannot be evaluated in this expression profiles even though they may have important functions in biological pathways. Second, many genes encode proteins that are required for survival in very specific amounts and thus remain stably expressed most of the time. As a consequence, these genes are usually missed by analysis methods based on expression profiles. Third, the drug-induced modulation of protein function may affect not only gene expression levels, which are regulated by miRNA or transcription factors, but also structural modifications, such as the methylation or phosphorylation of proteins whose transcript levels are not altered.[18] That is, the relatively short length of gene lists (signatures) published from expression profile experiments may not fully reflect the unique pattern of gene expression for a given cell or tissue. These gene signature-based methods may produce different expression patterns due to the different signature lengths, cell lines, drug dosages, experimental batches and thresholds used. Although Iorio *et al.*[15] sought to eliminate the effects caused by various cell lines, drug dosages and experimental batches by aggregating all the ranked lists obtained by treating cells with that drug, their results may still be unstable for different signature lengths.

Gene semantic similarity provides a functional similarity measure for a pair of proteins based on the semantic similarity of their GO term annotations. Keiser and Schuffenhauer *et al.*[19,20] reported that drugs with similar chemical structures usually bind functionally related proteins. Here, to reflect that structure influences function and to follow the principle that structure is consistent with function, we choose gene semantic similarity as a measure of the functional similarity of each protein pair instead of examining sequence similarity or structural similarity. Although the sequences or structures of two gene products can be compared directly through alignment algorithms, their functions cannot; the same is not true of their functional aspects; sequences and structures have an objective representation and measurable properties, whereas functional aspects have neither.[21] Also Couto *et al.*[22] indicated, many applications in bioinformatics would benefit from comparing proteins based on their biological roles rather than their sequences.

With the development of bioinformatics, systems biology has become more important in drug development, particularly in drug repositioning.[7,23] Hopkins *et al.*,[24] suggested that researchers should develop drug discovery approaches that utilise network pharmacology generated by integrating a system-wide view of drug interactions and phenotype data. Moreover, network-based approaches are likely to make key contributions to drug repositioning.[25]

Considering these issues, we integrated multidimensional information and heterogeneous data to develop a network-based method for identifying new applications for existing drugs. This method followed the principle that structure is consistent with function and integrated gene semantic similarity with chemical structure information. Here, we calculated a consensus response score (CRS) for every protein to each drug to quantify the degree to which the chemical structure was consistent with the function of the protein. These CRSs composed the response matrix for all proteins to all the drugs. The CRS in the matrix ranged from 1 to −1 and described the importance of the protein to the activity of the drug; the larger the absolute value of the CRS, the more important the protein is to the drug. Hence, this matrix reflected the characteristics of traditional expression profiles that have been processed and could be defined as another form of 'expression profile'. Based on the assumption that the MoAs of two drugs will be similar if their CRSs to all of the proteins are similar, we defined the absolute value of correlation coefficients as the degree of similarity between drugs. We constructed a drug similarity network (DSN) in which two drugs are connected to each other if their corresponding similarity scores were above a defined threshold. Finally, we identified modules consisting of very similar drugs. The workflow of our method is illustrated in Fig. 1. By analysing these modules, we were able to not only capture similarities in pharmacological effects but also discover previously unreported indications for well-known drugs.

Results and discussion

'Expression profile' of drugs to proteins

We produced a matrix of more than 1.5 million elements with scores from −1 to 1 by computing the CRS between 965 drugs and 1560 proteins. In the matrix, the CRS describes the importance of the protein to the activity of the drug; the larger the absolute value of the CRS, the more important the protein is to the drug. The CRS reflects that structure influences function and can be interpreted as the extent of the consistency between the structure and the function. If the CRS is positive (negative), then the drug has a positive (negative) effect on the protein. Therefore, the CRS matrix can be considered as another form of 'expression profile', after collapsing the probe sets and normalising the expression values in the raw expression profiles. The 'expression profile' is provided in Additional file 4 (ESI†). We hypothesised that drugs associated with similar patterns in protein expression would have similar indications. Here, we chose to examine 1560 proteins' CRSs responding to a drug instead of a signature of traditional expression profile. This 'expression profile' thus overcomes the insufficiencies caused

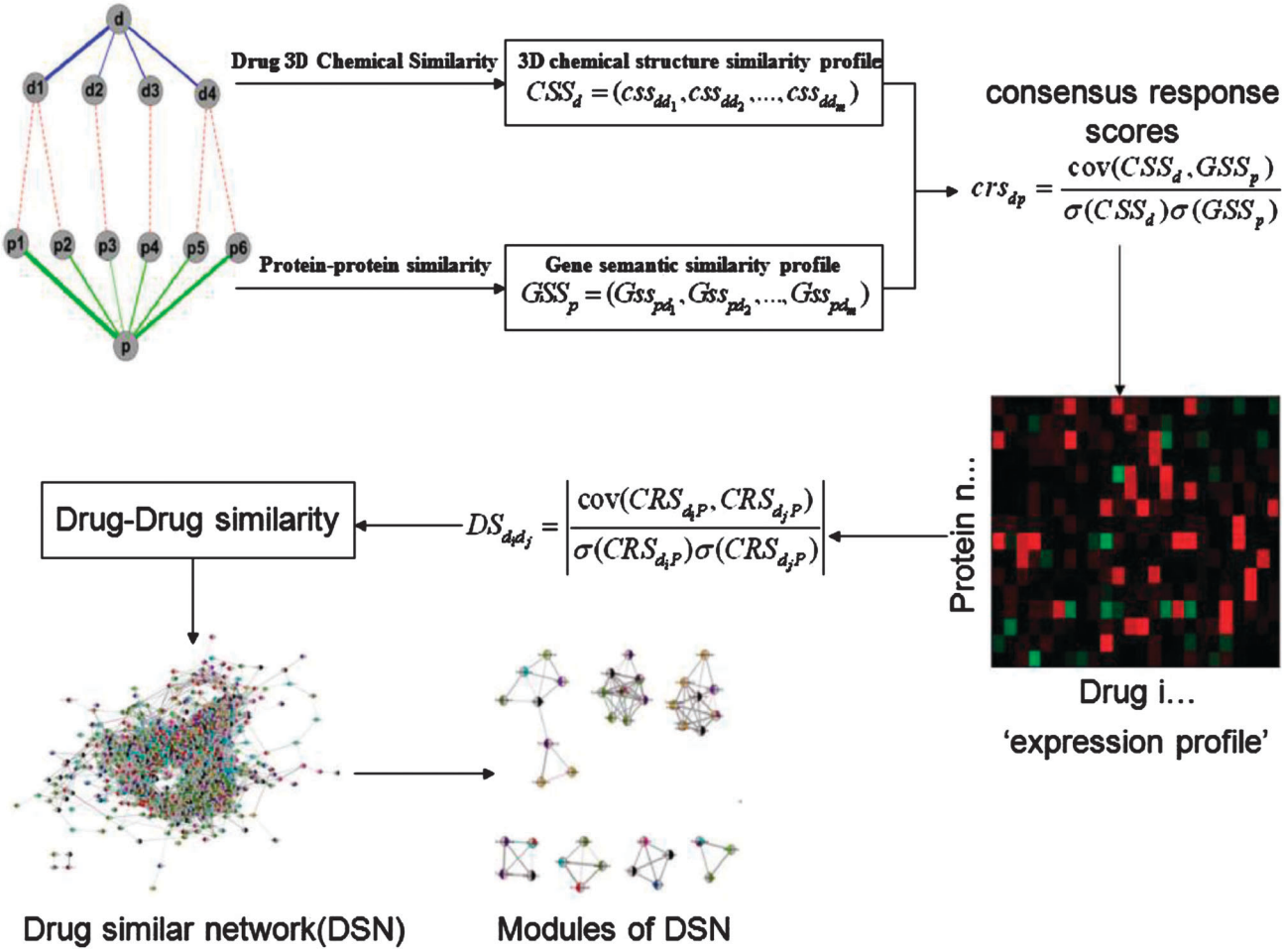

Fig. 1 The workflow of our method. First, we integrated drug 3D chemical similarity and gene semantic similarity information to generate an 'expression profile' by computing the CRS between drugs and proteins. Then, we defined the absolute value of correlation coefficients as the degree of similarity (DS) between every pair of drugs. We constructed the DSN by connecting two drugs if their DS was above a threshold that we defined. Finally, we capture modules consisting of similar drugs with common indications by mining the DSN.

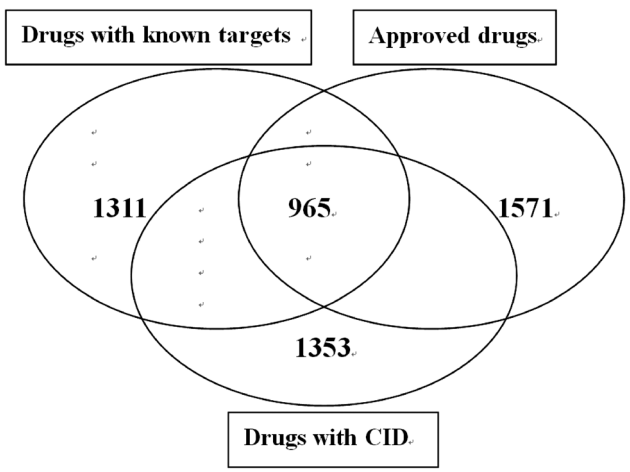

Fig. 2 Data sources. We extracted (a) 1571 FDA-approved drugs, (b) 1311 drugs with at least one known target and (c) 1353 drugs with chemical structure information recorded in the PubChem database. Overall, 965 drugs with 3D conformers were analysed.

by choosing a short list of genes that could not represent the whole genome, and avoided missing significant genes with importantly biological roles in drug actions simply because the expression of these genes did not change after treating cells with the drug. Therefore, after obtaining the 'expression profile', which in fact is the CRS matrix of the proteins and drugs, we determined the absolute value of correlation coefficients of expression values among all the proteins for each pair of drugs to indicate their degree of similarity. The larger the absolute value, the more similar the two drugs.

Constructing DSN and mining modules

In this study, we computed the similarity for each pair of the 965 drugs, for a total of 465 130 pairwise comparisons. Next, to construct the DSN, we used each drug as a node in a network and connected two nodes with a weighted edge according to their similarity beyond a threshold significance value. To determine this threshold, considering the large number of 465 130 pairwise similarity values, we chose to use a percentile within these data

to estimate a significance threshold. We selected the 95th percentile (0.8970; p = 0.05) as the significance threshold; *i.e.*, we connected drugs if their similarity values were greater than 0.8970.

The DSN has a major connected component with 911 nodes and 23 241 edges. In the following step, to visualise our similarity network, we loaded all significant drug pairs into Cytoscape,[26,27] a well-known open source system biology software platform for analysing and visualising complex interaction networks. Within Cytoscape, we first coloured drug nodes according to the Anatomical Therapeutic Chemical (ATC) classification[28] using Cytoscape plugin MultiColoredNodes,[29] which allowed us to visualise several node attributes by colour simultaneously. Second, we used the Cytoscape plug-in NetworkAnalyzer[30] to visualise the similarity values using the edge size; *i.e.*, larger similarity values were indicated by larger edge sizes. Third, we applied a force-directed layout algorithm to visualise the relationships between drugs. Although this algorithm is for drawing fine-looking heuristic graphs by mapping a particular graph layout to an energy value, the DSN layout was constructed independent of any knowledge about drug classes. The DSN is shown in Fig. 3. To further analyse the established network, we used the Cytoscape plugin MCODE[31] to mine highly connected regions in our network. The detailed module information is shown in Additional file 1 (ESI†).

A total of 42 clusters were identified, ranging in size from 3 to 122 and with corresponding edges ranging from 3 to 3795. Among these clusters, modules 1 and 2 contain more than 100 nodes and are so large that they capture extensive drug indications rather than specific indications. Therefore, our subsequent analyses focused mainly on the remaining 40 modules. These 40 modules can be divided into two categories: the first 20 clusters contain the drugs whose ATC codes are all known, and the second 20 clusters contain some drugs with unknown ATC codes. We calculated the proportion of all attributes (the attribute represents the first level of the ATC code) in each of the 40 modules. We assumed that the common effects of drugs within a module would be the top 1 or 2 indications based on attribute proportions; if these attributes' proportions were equal or close in a cluster, then the cluster's drugs have no common indications. In addition, in the case of modules with a dominant attribute, if the other attributes' proportions are

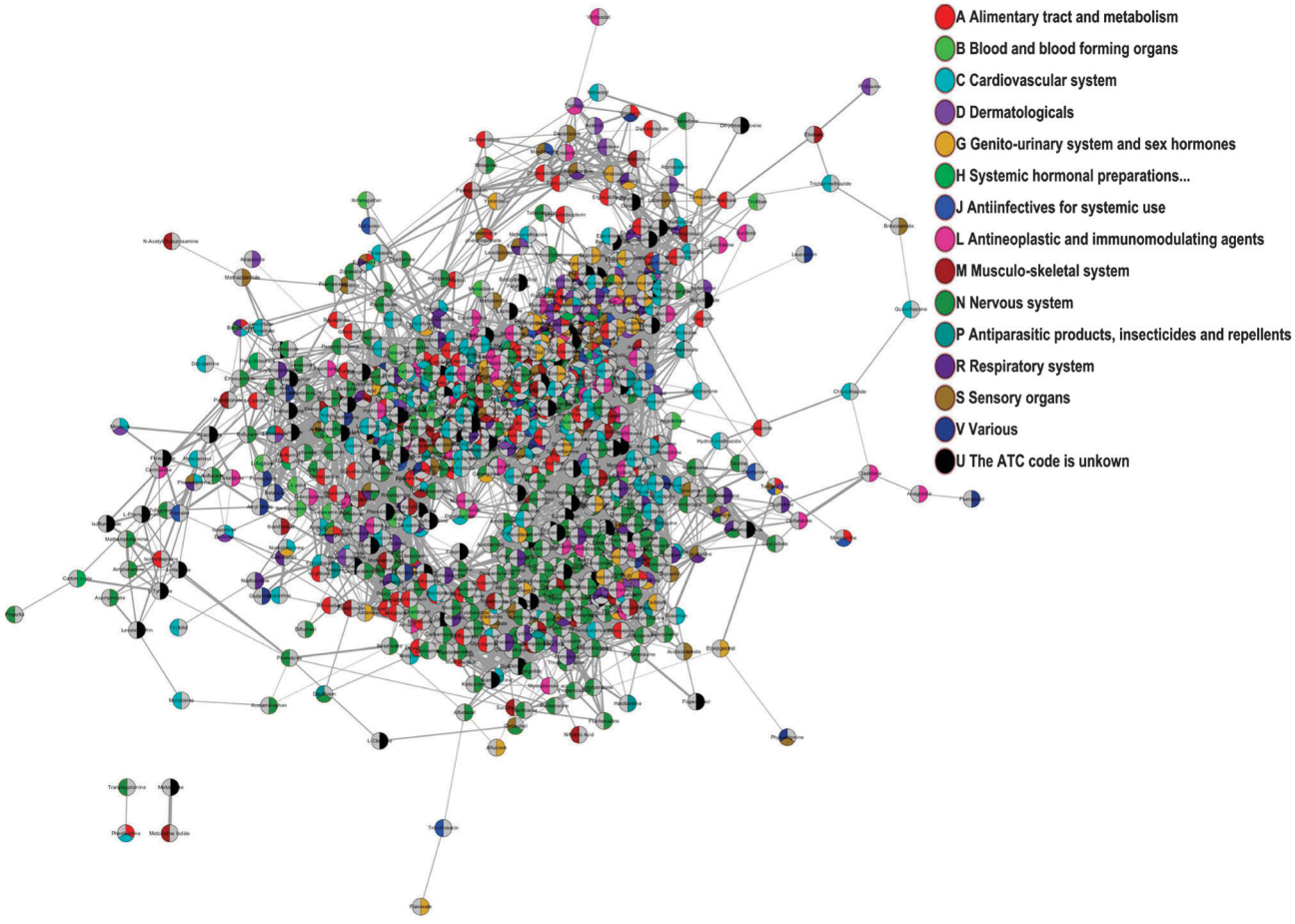

Fig. 3 The DSN. In the DSN, the nodes represent the drugs, and the size of each edge is proportional to the similarity of the connected drugs. We coloured the drug nodes according to the Anatomical Therapeutic Chemical (ATC) classification, and colour codes are given in the legend. We visualised drugs with several ACT codes by using multiple colours simultaneously; that is, each slice in the pie chart corresponds to one drug ATC code. The grey slice represents drugs with no attributes.

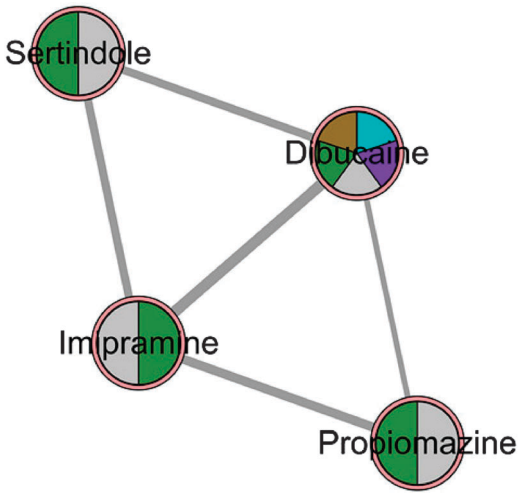

Fig. 4 Cluster 27.

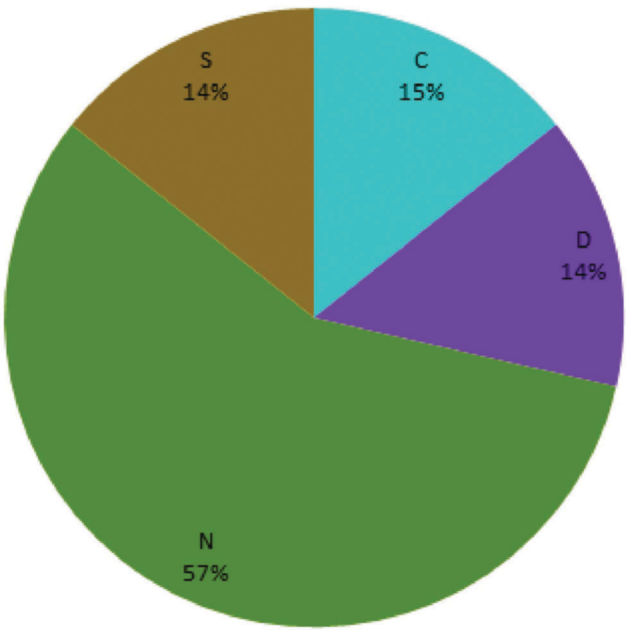

Fig. 5 The proportions of every attribute among all attributes in cluster 27.

equal or similar, then the common effects of the drugs within the module only tend to hold for the dominant indication, *e.g.*, in cluster 27 (Fig. 5). Based on this assumption, common indications were predicted for 33 modules, while the remaining 7 modules had no common indications. Through these common indications, we predicted new indications for 42 drugs with unknown ATC codes (Additional file 5, ESI†).

Examples of drugs with common indications in a cluster of the first category

Among the modules in the first category, we found three types of modules according to the proportion of the ATC code. In 2 modules, one attribute percentage accounted for 100% (module 33, 39); in 14 modules, some attribute was dominant; and in 4 modules, multiple attributes exhibited similar percentages. Additional file 5 (ESI†) provided the details of 20 clusters in the first category.

For example, in cluster 33, lumiracoxib, ketoprofen and mefenamic acid (Additional file 2, ESI†) all affect the musculoskeletal system (attribute percentage 100%) according to the first level of the ATC code (Table 1). More specifically, lumiracoxib is indicated for the acute and chronic treatment of the signs and symptoms of osteoarthritis of the knee in adults. Ketoprofen is indicated for the symptomatic treatment of acute and chronic rheumatoid arthritis, osteoarthritis, ankylosing spondylitis, primary dysmenorrhea, mild to moderate pain associated with musculotendinous trauma (sprains and strains), and postoperative (including dental surgery) or postpartum pain. Mefenamic acid is indicated for the treatment of rheumatoid arthritis, osteoarthritis, dysmenorrhea, mild to moderate pain, inflammation, and fever. These three drugs are all cyclooxygenase inhibitors; all are used for the treatment of osteoarthritis, and all of them exhibit analgesic effects.

Cluster 27 (Fig. 4) is an example of a module in which one attribute is dominant. In this case, attribute N was the top attribute (Fig. 5), and the other attributes' proportions were very similar; therefore, N is the common indication shared by all drugs in this module. The details of the results for this module are provided in Table 2. All four of the drugs in this module were used for the treatment of nervous system diseases such as schizophrenia, insomnia and depression, which are caused by a loss of the equilibrium between dopamine and 5-hydroxytryptamine. The main mechanism of these drugs is to target the 5-HTR enzyme and regulate the content of 5-HR in the central nervous system.

Examples of drugs with common indications in a cluster of the second category

For those drugs whose ATC codes are unknown, we predicted their indications through those known drugs' indications whose attribute proportions are in top 1 and/or top 2. Additional file 5 (ESI†) provides the details of the 20 clusters in the second category. Among these modules, 3 have no common indications. To further demonstrate the efficacy of our approach, we used cluster 10 (Fig. 6) as an example; here, we identify indications of oxymorphone for which there are no ATC code information in DrugBank. Actually, oxymorphone has ATC code 'N02A' (http://en.wikipedia.org/wiki/Oxymorphone).

Within cluster 10, only oxymorphone lacked an ATC code, meaning that its indications are unclear. Fig. 7 shows the attribute percentages for the drugs in cluster 10. Based on our assumption, we inferred that oxymorphone should share the common indications of this module, namely, pain relief and cough suppression (Additional file 6, ESI†). Our results, predicted based on ATC codes, are consistent with the reported indications in the literature. As an opiate,[32,33] oxymorphone may have analgesic properties and be useful in treating complaints such as chronic pain[34,35] and respiratory depression[36,37] by acting on brain stem respiratory centres. Thus, we conclude that oxymorphone could be used to relieve moderate to severe pain

Table 1 Details of cluster 33

Drug	ATC code	Drug indications	Common indication
Lumiracoxib	M	For the acute and chronic treatment of the signs and symptoms of osteoarthritis of the knee in adults.	These three drugs are all cyclooxygenase inhibitors, used for the treatment of osteoarthritis disease and all of them exhibit analgesic effect.
Ketoprofen	M	For symptomatic treatment of acute and chronic rheumatoid arthritis, osteoarthritis, ankylosing spondylitis, primary dysmenorrhea and mild to moderate pain associated with musculotendinous trauma (sprains and strains), postoperative (including dental surgery) or postpartum pain.	
Mefenamic acid	M	For the treatment of rheumatoid arthritis, osteoarthritis, dysmenorrhea, and mild to moderate pain, inflammation, and fever.	

Table 2 Details of cluster 27

Drug	ATC code	Drug indications	Common indication
Lumiracoxib	M	For the acute and chronic treatment of the signs and symptoms of osteoarthritis of the knee in adults.	These three drugs are all cyclooxygenase inhibitors, used for the treatment of osteoarthritis disease and all of them exhibit analgesic effect.
Ketoprofen	M	For symptomatic treatment of acute and chronic rheumatoid arthritis, osteoarthritis, ankylosing spondylitis, primary dysmenorrhea and mild to moderate pain associated with musculotendinous trauma (sprains and strains), postoperative (including dental surgery) or postpartum pain.	
Mefenamic acid	M	For the treatment of rheumatoid arthritis, osteoarthritis, dysmenorrhea, and mild to moderate pain, inflammation, and fever.	

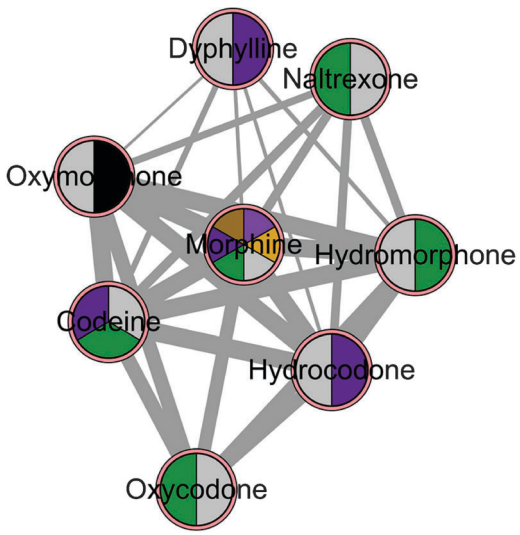

Fig. 6 Cluster 10. The cluster is composed of 8 drugs, among which naltrexone, hydromorphone and oxycodone belong to the nervous system, dyphylline and hydrocodone belong to the respiratory system, and codeine and morphine belonged to both the nervous system and respiratory system. Oxymorphone had no ATC code.

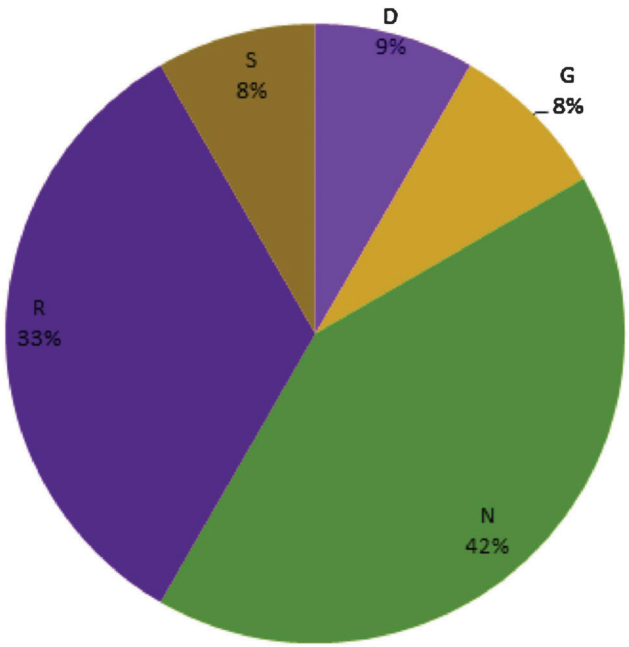

Fig. 7 The attribute percentages of the drugs in cluster 10.

caused by most diseases, including chronic osteoarthritis,[38] cancer and so on. Furthermore, oxymorphone could be used as a cough suppressant, similar to codeine and hydrocodone, or for its anaesthetic effects.[39] Hence, we predicated oxymorphone's ATC code is 'N'. Actually oxymorphone has ATC code 'N02A' (http://en.wikipedia.org/wiki/Oxymorphone) supporting our predication.

In addition, we found that some nodes that are more connected to each other inside a large module formed a small cluster with similar indications, such as in cluster 5 (Fig. 8). Inside module 5, nifedipine, nitrendipine, amlodipine, alprenolol and carvedilol were all categorised as having effects on the cardiovascular system. These drugs are all used for the treatment of hypertension and angina.

Comparison with the cMap online tool

To assess whether our approach was successful in finding the common indications of drugs within a module and predicting their unknown indications, we compared our results with those provided by the cMap online tool[11] which computes a

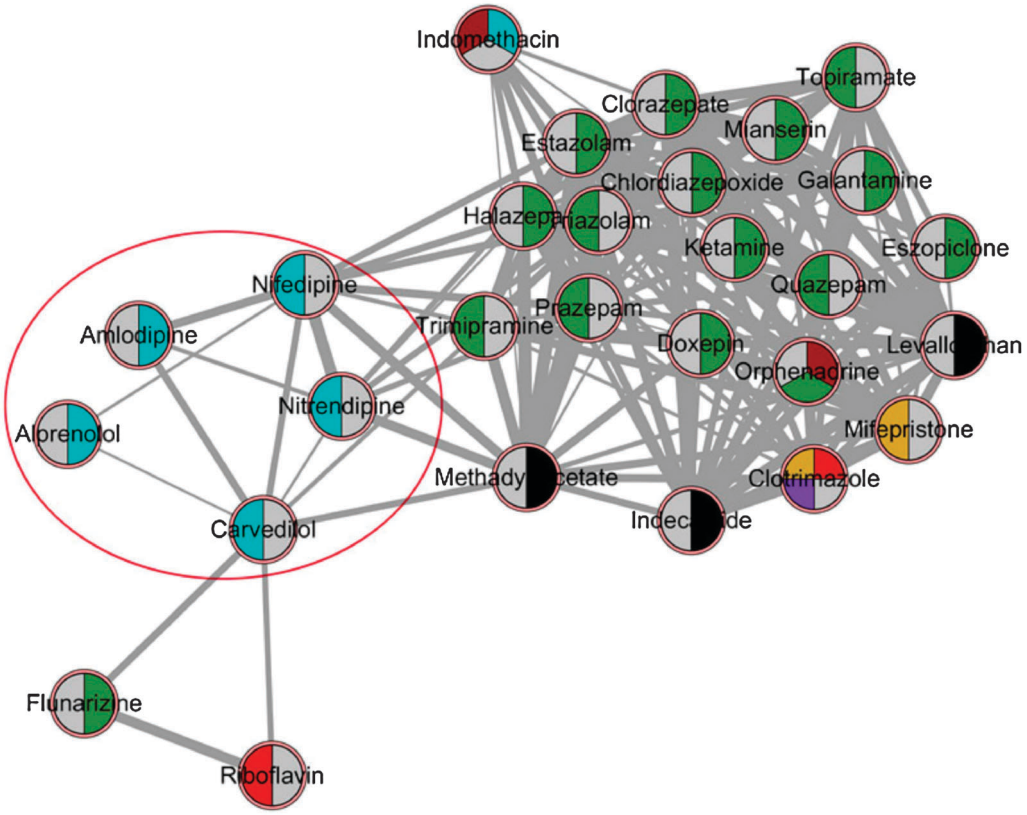

Fig. 8 Cluster 5. Nifedipine, nitrendipine, amlodipine, alprenolol and carvedilol were more connected to each other inside a large module and formed a small cluster of drugs with similar indications.

traditional signature of differentially expressed genes for each gene expression profile. Because not all of the drugs in our modules were included in the original cMap dataset, we chose the drugs that were available in cMap to be tested. We randomly selected 28 of the presented drugs (whose attributes' proportions were dominant in their clusters) within each of 28 modules and compared the results obtained with our approach with those provided by the cMap tool by determining whether the obtained drugs have common indications. To avoid the unstable expression patterns (different signatures) caused by different signature lengths, different cell lines, different batches and so on, we used the Prototype Ranked Lists (PRLs) for each of the 28 drugs published by Iorio[15] which have eliminated the effects caused by differences in cell lines, dosages and batches by aggregating all the ranked lists that have been obtained from cells treated with that drug. To evaluate the potential differences due to different signature lengths, we chose signature lengths of 300 (up probes: 150 and down probes: 150) and 500 (up probes: 250 and down probes: 250), respectively (Additional file 7, ESI†) to query using the cMap online tool. In the output list of drugs connected to each of the signatures, we retained the drugs that were predicted to be positively connected to the signature of the query instance and whose p-value is lower than 0.05 (Additional file 8, ESI†). We defined the accuracy of each drug as the proportion of the drugs whose attributes were the same as that of the query drug out of the whole output list of that drug. For the output list of drugs, we only take into account the drugs whose attributes are known. As a result, 25 (of 28) drugs' predications are more accurate than cMap's (Fig. 9 and Additional file 3, ESI†). Moreover, the results are significantly different from those of cMap with the signature length 300 or 500 ($p = 2.74828 \times 10^{-05}$ and 3.0862×10^{-05}, respectively). In addition, the results obtained using the signature-based method (cMap) are unstable (Fig. 10). The coincidence rate obtained by querying

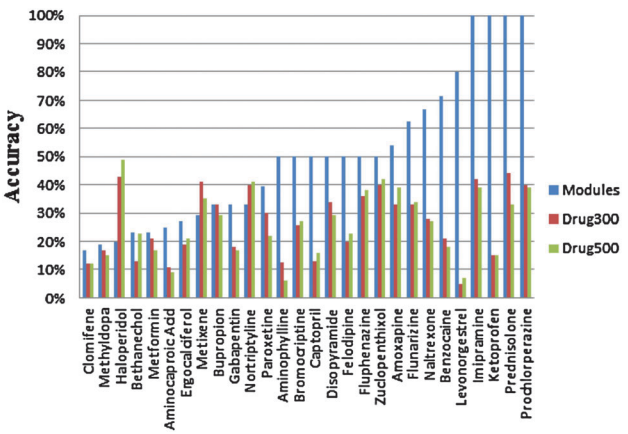

Fig. 9 The predictive accuracy of our method (Modules) and the cMap online tool for 28 drugs using signature lengths of 300 (Drug300) and 500 (Drug500), respectively.

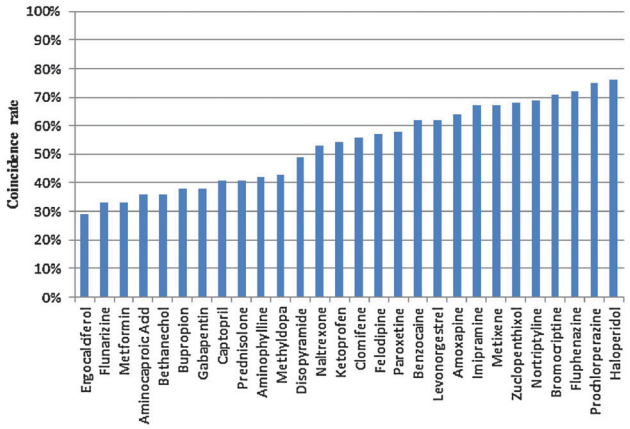

Fig. 10 The coincidence rate of the similar drugs obtained by querying cMap using the 300- and 500-gene signatures.

the 300- and 500-gene signatures for the same drug with cMap ranged from 29% to 76%, with a mean of only 53%. That is to say, compared with our method, the results provided by the cMap online tool were highly dependent on the size of the different gene expression profiles. Also, based on these 28 drugs our method provided more accuracy than the cMap online tool.

We took module 25 as an example to further assess our method from the biological function viewpoint. This module contains 6 drugs, including 3 drugs with ATC code N (fluphenazine, thiothixene and mesoridazine), 2 drugs with ATC code R (buclizine and astemizole), and 1 drug without any ATC code (carphenazine). Astemizole was once used to treat respiratory system diseases but has been withdrawn from the market. According to the principle of our method, the two respiratory system drugs and the drug without an ATC code should have pharmacological activities similar to those of the three nervous system drugs. Recent studies have indicated that all three nervous system drugs, fluphenazine, thiothixene and mesoridazine, target the dopamine receptor, inhibiting the signal transduction between synapses and neuron excitability, which sequentially achieve a therapeutic effect against mental diseases. These three drugs are widely used in the clinical treatment of mental diseases, and carphenazine has also been indicated by recent studies because of its similar pharmacological actions.[40] The other two drugs, buclizine and astemizole, are used for the clinical treatment of respiratory diseases. Both of these drugs act by binding to the histamine receptor, inhibiting the tracheal smooth musculature spasms caused by histamine, which achieves the therapeutic effect of dilating the bronchus and improving ventilation. Recent studies also demonstrated that buclizine can also target the muscarinic acetylcholine receptor M1, competitively inhibiting Ach binding to receptors and thereby inhibiting the function of the parasympathetic system.[41,42] Buclizine is also found to have adverse effects involving the central nervous system, such as drowsiness, dizziness, incoordination, stomach pain and diarrhea.[43] Studies also found that astemizole can interact with the potassium voltage-gated channel, which regulates voltage-gated signal transduction. Consequently, arrhythmia is induced when astemizole accumulates in the myocardial tissue and disturbs the signalling among myocardial cells, which was the main reason for its withdrawal.[44] Therefore, buclizine and astemizole, which are classified by the ATC as respiratory system drugs, and carphenazine, which has no ATC codes, all have functions in the central or peripheral nervous system, consistent with the predictions of our modules. Nevertheless, the potential function of a drug may induce adverse reactions, as observed for astemizole. Thus, the new drug functions predicted by our method require additional validation, especially for experimental drugs. A cMap query for drugs similar to fluphenazine only included 1 of the 6 drugs in module 25 (astemizole), indicating that our method has the potential ability to predict new drug indications.

Performance evaluation

In this section, we shall evaluate the performance of our methods. Among the 42 modules identified, we ignored 2 modules (modules 1 and 2) containing more than 100 nodes, 2 modules (modules 33 and 39) in which all drugs had the same attribute (ATC), 7 modules (modules 23, 31, 34, 35, 36, 37, and 42) in which all attributes exhibited equal proportions, and one module (module 27) in which each drug had the dominant attribute. The remaining 30 modules, including 269 drugs with ATC codes and 42 drugs without ATC codes, were used to further assess our method's accuracy by literature validation. At first, we employed our program (in Python) to identify the literature about drugs in our predication list. Then 3 experts with backgrounds in pharmacy and biology read the literature. Literature works considered to support our prediction by at least two experts were extracted. We supplied the prediction list and links of the literature in Additional file 9 (ESI†). For example, our predication that ritodrine has the new indication 'cardiac' was positive because in ref. 45 authors stated that ritodrine has adverse effects upon cardiovascular systems, including pulmonary oedema and myocardial ischemia. We had predicted indications for 143 drugs with ATC codes and 42 drugs without ATC codes. Overall, 71.8% (254/354) predications were validated (Additional file 9, ESI†).

In order to assess the predications which were not validated by the literature, we tested whether these predications are in accordance with the current experimental knowledge. We checked whether these predications appear in current clinical trials (http://clinicaltrials.gov/) similar to ref. 46. Overall, we acquired 33 (of 100) predications that are being investigated in clinical trials (Additional file 9, ESI†). To further assess the predications which were not validated by the literature, we checked tissue-specific expression information of these drugs' targets (http://www.genecards.org/) motivated by ref. 46. Overall, 54 (of 100) drugs' targets were expressed in the same tissues with predicted systems, supporting our predications (Additional file 9, ESI†). In particular, 63 of 100 indications (not validated by the literature) were supported by current clinical trials or by

target tissue-specific expression information, indicating our method's high accuracy.

Conclusion

Here, we take full advantage of the prior knowledge, such as the known drug–target associations, 3D chemical structure information and GO functional annotation data. We developed a novel method for capturing similarities in drug indications by applying another form of 'expression profile' and identifying dominant attributes within a module. This strategy has significant advantages in predicting the effects of a new drug because it merely uses chemical structure information initially. And it addresses inadequacies in predicting novel roles for a drug based simply on similarities in chemical structures or functions.

Through this 'expression profile', which consists of the CRS between drugs and proteins rather than raw expression data, some unexpected drug–protein relationships emerged, thus we were able to obtain valuable information about the relative importance of every protein to the activity of every drug. Furthermore, our method avoids some problems caused by analyses based on the raw experimental expression profile. For example, the use of different statistical methods (or signature lengths, cell lines, drug dosages, experimental batches, or thresholds) may yield different expression patterns (signature). These demonstrated that the integration of existing multi-dimensional information may generate additional knowledge and present an advantage over raw expression data.

Because the prediction of drug–drug similarity through this 'expression profile' is based on the MoAs of the drugs, the drugs in each module of the DSN will have similar pharmacological mechanisms. We defined the dominant indications within a module as the common indications likely shared by all drugs in this module. By applying our method, we successfully mined 33 modules of drugs with similar indications, found their common indications, and predicted indications for 143 drugs and 42 drugs without ATC codes. Overall 71.8% indications were confirmed in the literature and 63% of 100 indications (not validated by the literature) were supported by current clinical trials or by target tissue-specific expression information. The indications were confirmed and can be regarded as valuable direction for further research.

Despite the efficacy of our method, there are still several caveats. First, our approach requires some prior knowledge about the drugs, *e.g.*, its chemical structure and known drug–target associations. However, chemical structure information has been addressed extensively, and the methods for identifying the protein targets of a given drug have developed rapidly. Second, among the modules we mined, several modules (*e.g.*, clusters 1 and 2) were so large that they captured only the drugs' extensive indications rather than specific indications, owing to the integration of information about too many individual drugs. However, this problem can be eliminated neatly by examining the more tightly connected units (*e.g.*, cluster 5) inside these modules. Finally, to improve the overall performance, we may be able to consider this subject from other perspectives, such as drug therapeutic similarity and side effect similarity. Studies addressing the above-mentioned problems may represent a further step toward drug repositioning. In summary, we have developed a novel drug repositioning method through mining a DSN generated by integrating the drug chemical structure similarity and gene semantic similarity.

Materials and methods

Data sources

Drug information: ATC codes, CIDs (Compound IDs in the PubChem Compound Database[3]) and known drug–target interactions were obtained from DrugBank 3.0.[47–49] We extracted drugs that (a) were FDA approved, (b) had at least one known target and (c) had chemical structure information recorded in the PubChem database.[45] A total of 1045 drugs were obtained, together with 3869 drug–target interactions identified by mapping the gene symbol to the human Entrez gene ID. In this paper, 3D conformers were extracted to compute the 3D drug chemical similarity based on the Tanimoto coefficient.[50] However, 80 of 1045 compounds had no 3D conformers in the PubChem database. Therefore, 965 drugs (Fig. 2) were included in the final experimental dataset.

The protein dataset in our research was from the cellular signaling network published by Cui *et al.*,[51,52] as the protein dataset in our research. There are 1560 proteins after removing duplicate gene IDs and proteins with no gene IDs. We chose the proteins of the human cellular signaling network instead of all of the human proteins, because all human proteins are so large containing many proteins unrelated to drugs. Using all human proteins could interfere with the CRSs of proteins to drugs making it difficult to find out the slight shades of difference between the transcriptional responses to the drugs. Furthermore, signal transduction in the cellular signaling network describes the process of converting external signals to a specific internal cellular response (such as gene expression). Since most of the known diseases exhibit dysfunctional aspects in this network, identifying novel drug targets based on the signaling network has been a focal point.[53]

Drug chemical similarity

Drug 3D structure similarity was computed by PubChem database,[3] supplied three 3D similarity measures: shape-Tanimoto (ST), color-Tanimoto (CT), and Combo-Tanimoto (ComboT). The ST score is a measure of shape similarity and the CT score is a measure of feature similarity (color Tanimoto). ComboT integrates ST and CT measures by adding these two scores and ComboT scores range between 0 (for no similarity) and 2 (for identical molecules). Specifically we only chose a single theoretical conformer per compound (the ''default'' conformer provided by PubChem) and utilized ComboT score to compute the drug 3D structure similarity. When computing the CRS between drugs and proteins, we only selected drug pairs with

3D structure similarity scores greater than or equal to 0.4 (the reason that we chose 0.4 as threshold was supplied in Additional file 10, ESI†).

Gene semantic similarity

Gene semantic similarity can be computed by controlled biological or biochemical vocabularies, such as Gene Ontology (GO Terms). In this study, we captured the semantic similarity between every protein investigated in our protein dataset and each target corresponding to all 965 drugs by applying the GoSemSim package in R.[54] The GoSemSim package provided five methods.[55–59] In particular, Resnik, Jiang, Lin and Schlicker's methods mainly based on the information content of the most informative common ancestor node of two terms ignored the structure information of the GO directed acyclic graph.[60] Wang's similarity measure considered node and edge information. Particularly, it considered different types of GO relationships ('is-a' and 'part-of') and assigned weights to these different relationships. It has been shown to be consistent with human perspectives. Details about Wang's method can be seen in ref. 59. We used the function 'geneSim' of molecular function (ont = 'MF') by applying the measure based on Wang *et al.*[59] (measure = 'Wang') and adjusted to humans (organism = 'human').

Consensus response score and drug similarity

To obtain the CRS between a drug (d) and a protein (p), our model adopted the correlation coefficient as the CRS. In this model, $\mathrm{css}_{\mathrm{dd}'}$ represents the 3D chemical structure similarity between a query drug d and another drug d′, and gss(p,p′) is the gene semantic similarity between a protein p and another protein p′.

To quantify the correlation coefficient or the CRS between a drug d and a protein p, we first extracted the drugs whose 3D structure similarity scores with d were greater than 0.4 (the effect of different thresholds were supplied in Additional file 10, ESI†) and denoted them as $\mathrm{D} = (\mathrm{d}_1,\mathrm{d}_2,\ldots\mathrm{d}_m)$. Thus, we obtained the 3D chemical structure similarity profile of d as follows:

$$\mathrm{CSS_d} = (\mathrm{css}_{\mathrm{dd}_1},\mathrm{css}_{\mathrm{dd}_2},\ldots,\mathrm{css}_{\mathrm{dd}_m})$$

Then, we extracted the known target set T_i corresponding to drug d_i in drug set D, where $\mathrm{T}_i = (t_1^i,t_2^i,\ldots,t_k^i)$ and k is the total number of known targets of the drug d_i, to obtain a m-dimensional vector $\mathrm{TS} = (\mathrm{T}_1,\mathrm{T}_2,\ldots,\mathrm{T}_m)$. We next defined the functional relatedness between p and T_i as the accumulation of semantic similarities between protein p and each of the proteins in T_i, that is,

$$\mathrm{Gss}_{\mathrm{pd}_i} = \sum_{j=1}^{k} \mathrm{gss}\left(t_j^i, \mathrm{p}\right)$$

Thus, we can obtain the gene semantic similarity profile between p and TS, which is defined as

$$\mathrm{GSS_p} = (\mathrm{Gss}_{\mathrm{pd}_1},\mathrm{Gss}_{\mathrm{pd}_2},\ldots,\mathrm{Gss}_{\mathrm{pd}m}).$$

Finally, we defined the $\mathrm{crs}_{\mathrm{dp}}$ between a drug d and a protein p as the Pearson correlation coefficient

$$\mathrm{crs}_{\mathrm{dp}} = \frac{\mathrm{cov}(\mathrm{CSS_d}, \mathrm{GSS_p})}{\sigma(\mathrm{CSS_d})\sigma(\mathrm{GSS_p})}$$

where cov and σ are covariance and standard deviation, respectively. In a sense, we assume that the chemical structure similarity of every two drugs is related to the functional relatedness of their targets. Keiser *et al.*, have found that drugs with similar chemical structures usually bind functionally related proteins.

We obtained the CRS between each drug (of 965 drugs) and each protein (of 1560 proteins) in our datasets. Because the CRS can be considered another type of expression response (*i.e.*, of proteins exposed to a drug), we can define the 'expression profile' of drug d_i, which is

$$\mathrm{CRS}_{\mathrm{d}_i\mathrm{P}} = (\mathrm{crs}_{\mathrm{d}_i\mathrm{P}_1},\mathrm{crs}_{\mathrm{d}_i\mathrm{P}_2},\ldots,\mathrm{crs}_{\mathrm{d}_i\mathrm{P}_{1560}}).$$

Based on the assumption that two drugs i and j with similar CRSs to all proteins will have similar MoAs, we define the absolute value of the Pearson correlation coefficients as the degree of similarity between drugs i and j, that is,

$$\mathrm{DS}_{\mathrm{d}_i\mathrm{d}_j} = \left|\frac{\mathrm{cov}(\mathrm{CRS}_{\mathrm{d}_i\mathrm{P}}, \mathrm{CRS}_{\mathrm{d}_j\mathrm{P}})}{\sigma(\mathrm{CRS}_{\mathrm{d}_i\mathrm{P}})\sigma(\mathrm{CRS}_{\mathrm{d}_j\mathrm{P}})}\right|$$

In our manuscript, we just used the Pearson correlation coefficient. We have assessed effects of different correlation coefficients (Pearson correlation coefficient, Spearman's rank correlation coefficient, Kendall rank correlation coefficient *etc.*) in Additional file 10 (ESI†). The results showed that the Pearson correlation coefficient is more suitable for our method.

Module mining

In this section, we used the Cytoscape-Plugin MCODE with its default parameters[30] to mine the highly connected regions in our DSN. MCODE is a very popular clustering algorithm which utilizes vertex weighting to grow clusters from a starting vertex of high local weight by iteratively adding neighboring vertices with similar weights. The effect of other clustering algorithms integrating with our method was supplied in Additional file 10 (ESI†).

Generating a prototype ranked list

In this section, we obtained the Prototype Ranked Lists (PRLs) from Iorio's work. These authors built a PRL for every drug by aggregating all ranked lists that had been obtained by treating cells with that drug (*i.e.*, in different cell lines, with different concentrations and so on). We used the same notation as Iorio, as follows:

D: the set of all the possible permutations of microarray probe-set identifiers (MPI);

X: a set of ranked lists of MPIs computed by sorting, in decreasing order, the genome-wide differential expression profiles obtained by treating cell lines with the same drug;

δ: $\mathrm{D}^2 \rightarrow \mathrm{N}$: Spearman's Footrule distance associating with each pair of ranked lists in X, a natural number quantifying the similarity between them;

B: $D^2 \rightarrow D$: the Borda Merging Function associating each pair of ranked lists in X with a new ranked list obtained by merging the lists with the Borda Merging Method.

For the input set X, if $|X| > 1$, then the two ranked lists x and y in X with the smallest Spearman's Footrule distance are identified and merged using the Borda Merging Method, producing the new ranked list of MPI denoted as z. Next, the two lists x and y are removed from X, and the new list (z) is added to generate a new set X. This process is repeated until $|X| = 1$, and the only list remaining in X is defined as the PRL of the drug.

Spearman's Footrule distance: Let r: $P \times D \rightarrow [1,\ldots,m]$ be a function defined for the set of MPI(P) and D, with values in the interval $[1,\ldots m]$, $m = |P|$. $r(i,x) = d$ represents the position (d) of MPI $i \in P$ in the ranked list D. Ignoring normalisation terms, we calculate Spearman's Footrule,

$$\delta(x,y) = \sum_{i=1}^{m} |r(i,x) - r(i,y)|, \quad \text{where } x,y \in X \subseteq D.$$

Borda Merging Method: the Borda Merging Function is defined as $B(x,y) = z$, $(x,y,z \in D)$. We calculate the list of values $P = [p_1,p_2,\ldots,p_m]$ as follows: $p_i = r(i,x) + r(i,y)$, where r is the previously defined function. A new ranked list z is obtained by sorting the values P in increasing order.

Acknowledgements

The authors would like to acknowledge the support of the Funds by the National Natural Science Foundation of China [Grant No. 61372188], the Graduate Innovation Foundation of Heilongjiang province, China [Grant No. YJSCX2012-223HLJ, YJSCX2012-341HLJ] and the Innovation Manpower Fund of Harbin Science and Technology Bureau, China [Grant No. 2010RFXXS053].

References

1 V. J. Haupt and M. Schroeder, Old friends in new guise: repositioning of known drugs with structural bioinformatics, *Briefings Bioinf.*, 2011, **12**(4), 312–326.
2 J. A. DiMasi, R. W. Hansen, H. G. Grabowski and L. Lasagna, Cost of innovation in the pharmaceutical industry, *J. Health Econ.*, 1991, **10**(2), 107–142.
3 E. E. Bolton, Y. Wang, P. A. Thiessen and S. H. Bryant, PubChem: Integrated Platform of Small Molecules and Biological Activities, Chapter 12, in Annual Reports in Computational Chemistry, 2008, 4(217–241).
4 M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijer, R. C. Matos and T. B. Tran, *et al.*, Predicting new molecular targets for known drugs, *Nat. Rev. Cancer*, 2009, **462**(7270), 175–181.
5 M. A. Yildirim, K. I. Goh, M. E. Cusick, A. L. Barabasi and M. Vidal, Drug-target network, *Nat. Biotechnol.*, 2007, **25**(10), 1119–1126.
6 M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin and B. K. Shoichet, Relating protein pharmacology by ligand chemistry, *Nat. Biotechnol.*, 2007, **25**(2), 197–206.
7 J. T. Dudley, E. Schadt, M. Sirota, A. J. Butte and E. Ashley, Drug discovery in a multidimensional world: systems, patterns, and networks, *J. Cardiovasc. Transl. Res.*, 2010, **3**(5), 438–447.
8 J. T. Dudley, E. Schadt, M. Sirota, A. J. Butte and E. Ashley, Drug discovery in a multidimensional world: systems, patterns, and networks, *J. Cardiovasc. Transl. Res.*, 2010, **3**(5), 438–447.
9 F. Iorio, R. Bosotti, E. Scacheri, V. Belcastro, P. Mithbaokar, R. Ferriero, L. Murino, R. Tagliaferri, N. Brunetti-Pierri and A. Isacchi, *et al.*, Discovery of drug mode of action and drug repositioning from transcriptional responses, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**(33), 14621–14626.
10 J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J. P. Brunet, A. Subramanian and K. N. Ross, *et al.*, The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease, *Science*, 2006, **313**(5795), 1929–1935.
11 J. Lamb, The Connectivity Map: a new tool for biomedical research, *Nat. Rev. Cancer*, 2007, **7**(1), 54–60.
12 M. Sirota, J. T. Dudley, J. Kim, A. P. Chiang, A. A. Morgan, A. Sweet-Cordero, J. Sage and A. J. Butte, Discovery and preclinical validation of drug indications using compendia of public gene expression data, *Sci. Transl. Med.*, 2011, **3**(96), 96ra77.
13 D. G. McArt and S. D. Zhang, Identification of candidate small-molecule therapeutics to cancer by gene-signature perturbation in connectivity mapping, *PLoS One*, 2011, **6**(1), e16382.
14 F. Iorio, A. Isacchi, D. di Bernardo and N. Brunetti-Pierri, Identification of small molecules enhancing autophagic function from drug network analysis, *Autophagy*, 2010, **6**(8), 1204–1205.
15 F. Iorio, R. Bosotti, E. Scacheri, V. Belcastro, P. Mithbaokar, R. Ferriero, L. Murino, R. Tagliaferri, N. Brunetti-Pierri and A. Isacchi, *et al.*, Discovery of drug mode of action and drug repositioning from transcriptional responses, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**(33), 14621–14626.
16 A. J. Wolpaw, K. Shimada, R. Skouta, M. E. Welsch, U. D. Akavia, D. Pe'er, F. Shaik, J. C. Bulinski and B. R. Stockwell, Modulatory profiling identifies mechanisms of small molecule-induced cell death, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**(39), E771–E780.
17 G. Hu and P. Agarwal, Human disease-drug network based on genomic expression profiles, *PLoS One*, 2009, **4**(8), e6536.
18 Gene expression profile: http://en.wikipedia.org/wiki/Gene_expression_profiling#Limitations.
19 M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin and B. K. Shoichet, Relating protein pharmacology by ligand chemistry, *Nat. Biotechnol.*, 2007, **25**(2), 197–206.
20 A. Schuffenhauer, P. Floersheim, P. Acklin and E. Jacoby, *J. Chem. Inf. Comput. Sci.*, 2003, **43**(2), 391–405.

21 C. Pesquita, D. Faria, A. O. Falcao, P. Lord and F. M. Couto, Semantic similarity in biomedical ontologies, *PLoS Comput. Biol.*, 2009, **5**(7), e1000443.
22 F. M. Couto, M. J. Silva and P. M. Coutinho, Measuring semantic similarity between Gene Ontology terms, *Data & Knowledge Engineering*, 2007, **61**(1), 137–152.
23 A. Schrattenholz, K. Groebe and V. Soskic, Systems biology approaches and tools for analysis of interactomes and multi-target drugs, *Methods Mol. Biol.*, 2010, **662**, 29–58.
24 A. L. Hopkins, Network pharmacology: the next paradigm in drug discovery, *Nat. Chem. Biol.*, 2008, **4**(11), 682–690.
25 A. Pujol, R. Mosca, J. Farres and P. Aloy, Unveiling the role of network and systems biology in drug discovery, *Trends Pharmacol. Sci.*, 2010, **31**(3), 115–123.
26 P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks, *Genome Res.*, 2003, **13**(11), 2498–2504.
27 M. E. Smoot, K. Ono, J. Ruscheinski, P. L. Wang and T. Ideker, Cytoscape 2.8: new features for data integration and network visualization, *Bioinformatics*, 2011, **27**(3), 431–432.
28 A. Skrbo, B. Begovic and S. Skrbo, Classfication of drugs using the ATC system (Anatomic, Therapeutic, Chemical Classification) and the latest changes, *Med. Arh.*, 2004, **58**(1), 138–141.
29 G. Warsow, B. Greber, S. S. Falk, C. Harder, M. Siatkowski, S. Schordan, A. Som, N. Endlich, H. Scholer and D. Repsilber, *et al.*, ExprEssence–revealing the essence of differential experimental data in the context of an interaction/regulation net-work, *BMC Syst. Biol.*, 2010, **4**, 164.
30 Y. Assenov, F. Ramirez, S. E. Schelhorn, T. Lengauer and M. Albrecht, Computing topological parameters of biological networks, *Bioinformatics*, 2008, **24**(2), 282–284.
31 G. D. Bader and C. W. Hogue, An automated method for finding molecular complexes in large protein interaction networks, *BMC Bioinf.*, 2003, **4**, 2.
32 A. D. Kaye, A. Baluch and J. T. Scott, Pain management in the elderly population: a review, *Ochsner J.*, 2010, **10**(3), 179–187.
33 B. G. Chen, S. M. Wang and R. H. Liu, GC-MS analysis of multiply derivatized opioids in urine, *J. Mass Spectrom.*, 2007, **42**(8), 1012–1023.
34 F. Mayyas, P. Fayers, S. Kaasa and O. Dale, A systematic review of oxymorphone in the management of chronic pain, *J. Pain Symptom Manage.*, 2010, **39**(2), 296–308.
35 A. K. Matsumoto, Oral extended-release oxymorphone: a new choice for chronic pain relief, *Expert Opin. Pharmacother.*, 2007, **8**(10), 1515–1527.
36 R. B. Patt, Delayed postoperative respiratory depression associated with oxymorphone, *Anesth. Analg.*, 1988, **67**(4), 403–404.
37 W. D. Fiske, J. Jobes, Q. Xiang, S. C. Chang and I. H. Benedek, The effects of ethanol on the bioavailability of oxymorphone extended-release tablets and oxymorphone crush-resistant extended-release tablets, *J. Pain*, 2012, **13**(1), 90–99.
38 H. McIlwain and H. Ahdieh, Safety, tolerability, and effectiveness of oxymorphone extended release for moderate to severe osteoarthritis pain: a one-year study, *Am. J. Ther.*, 2005, **12**(2), 106–112.
39 K. W. Chamberlin, M. Cottle, R. Neville and J. Tan, Oral oxymorphone for pain management, *Ann. Pharmacother.*, 2007, **41**(7), 1144–1152.
40 O. Vinar, M. Formankova, D. Taussigova and S. Ruzicka, Carphenazine in the treatment of schizophrenic psychoses. Controlled clinical trial, *Act. Nerv. Super.*, 1967, **9**(4), 353–355.
41 J. P. Overington, B. Al-Lazikani and A. L. Hopkins, How many drug targets are there?, *Nat. Rev. Drug Discovery*, 2006, **5**(12), 6.
42 P. Imming, C. Sinning and A. Meyer, Drugs, their targets and the nature and number of drug targets, *Nat. Rev. Drug Discovery*, 2006, **5**(10), 821–834.
43 G. K. McEvoy, Dose adjustment in renal impairment: response from AHFS Drug Information, *BMJ*, 2005, **331**(7511), 293.
44 J. S. Silvestre and J. R. Prous, Comparative evaluation of hERG potassium channel blockade by antipsychotics, *Methods Find. Exp. Clin. Pharmacol.*, 2007, **29**(7), 457–465.
45 B. Dordevic, M. P. Stojiljkovic, T. Phtpara, D. Loncar-Stojiljkovic and L. Vojvodic, Successful resuscitation of a patient with prolonged tocolytic therapy and an emergency vesarean section, *Vojnosanitetski pregled military-medical and pharmaceutical review*, 2002, **59**(3), 325–328.
46 A. Gottlieb, G. Y. Stein, E. Ruppin and R. Sharan, PREDICT: a method for inferring novel drug indications with application to personalized medicine, *Mol. Syst. Biol.*, 2011, **7**, 496.
47 C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak and V. Neveu, *et al.*, DrugBank 3.0: a comprehensive resource for 'omics' research on drugs, *Nucleic Acids Res.*, 2011, **39**(Database issue), D1035–D1041.
48 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.*, 2008, **36**(Database issue), D901–D906.
49 D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang and J. Woolsey, DrugBank: a comprehensive resource for *in silico* drug discovery and exploration, *Nucleic Acids Res.*, 2006, **34**(Database issue), D668–D672.
50 P. Willett, V. Winterman and D. Bawden, Implementation of nearest-neighbor searching in an online chemical structure search system, *J. Chem. Inf. Comput. Sci.*, 1986, **26**, 36–41.
51 Q. Cui, A network of cancer genes with co-occurring and anti-*co*-occurring mutations, *PLoS One*, 2010, **5**(10), e13180.
52 Q. Cui, Y. Ma, M. Jaramillo, H. Bari, A. Awan, S. Yang, S. Zhang, L. Liu, M. Lu and M. O'Connor-McCourt, *et al.*, A map of human cancer signaling, *Mol. Syst. Biol.*, 2007, **3**, 152.

53 A. Persidis, Signal transduction as a drug-discovery platform, *Nat. Biotechnol.*, 1998, **16**(11), 1082–1083.
54 G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu and S. Wang, GOSemSim: an R package for measuring semantic similarity among GO terms and gene products, *Bioinformatics*, 2010, **26**(7), 976–978.
55 R. Philip, Semantic similarity in a taxonomy: An Information-Based measure and its application to problems of ambiguity in natural language, *J. Artif. Intell. Res.*, 1999, **11**, 95–130.
56 D. W. Conrath and J. J. Jiang, Semantic similarity based on corpus statistics and lexical taxonomy, *Proceedings of 1st International Conference on Research In Computational Linguistics*, 1997.
57 D. Lin, An information-theoretic definition of similarity, in *In Proceedings of the 15th International Conference on Machine Learning*, 1998, pp. 296–304.
58 A. Schlicker, F. S. Domingues, J. Rahnenfuhrer and T. Lengauer, A new measure for functional similarity of gene products based on Gene Ontology, *BMC Bioinf.*, 2006, **7**, 302.
59 J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu and C. F. Chen, A new method to measure the semantic similarity of GO terms, *Bioinformatics*, 2007, **23**(10), 1274–1281.
60 X. Chen, R. Yang, J. Xu, H. Ma, S. Chen, X. Bian and L. Liu, A sensitive method for computing GO-based functional similarities among genes with 'shallow annotation', *Gene*, 2012, **509**(1), 131–135.

