Molecular BioSystems

Cite this: *Mol. BioSyst.*, 2012, **8**, 1528–1534

www.rsc.org/molecularbiosystems

PAPER

Predicting drug targets based on protein domains†

Yin-Ying Wang,[ab] Jose C. Nacher[c] and Xing-Ming Zhao*[b]

Received 1st November 2011, Accepted 14th February 2012
DOI: 10.1039/c2mb05450g

The identification of interactions between drugs and proteins plays key roles in understanding mechanisms underlying drug actions and can lead to new drug design strategies. Here, we present a novel statistical approach, namely PDTD (Predicting Drug Targets with Domains), to predict potential target proteins of new drugs based on derived interactions between drugs and protein domains. The known target proteins of those drugs that have similar therapeutic effects allow us to infer interactions between drugs and protein domains which in turn leads to identification of potential drug–protein interactions. Benchmarking with known drug–protein interactions shows that our proposed methodology outperforms previous methods that exploit either protein sequences or compound structures to predict drug targets, which demonstrates the predictive power of our proposed PDTD method.

Introduction

Most drugs act by binding to specific proteins, thereby changing their biochemical or biophysical activities with desired consequences.[1] Therefore, proteins targeted by drugs are important to understanding the mechanisms of action of drugs as well as designing new drugs that can bind to these therapeutic targets. However, the target proteins of many drugs are not complete or even not known, which hampers the discovery of new drugs. Although high-throughput screening is useful to identify drug–protein interactions, it is impractical to screen all possible drug–protein interactions due to the costly and time-consuming experiments. Therefore, it is necessary to develop fast and reliable computational methods to identify target proteins of new drugs.[2–7]

In the literature, there are many computational approaches that have been developed to predict the target proteins of new drugs. For example, utilizing the information about ligands binding to proteins, Keiser *et al.* predicted new targets for known drugs based on the chemical similarity between drugs and ligands.[8] However, the performance of this approach depends on the information about ligands binding to proteins, which may be not publicly available. Since compounds generally bind to specific pockets of a protein, docking simulation was widely used to identify those compounds that can bind to the pockets of known target proteins.[9] Unfortunately, the incompleteness of protein 3D structures limits the application of these methods. With the assumption that drugs having similar side effects generally target the same proteins, the side effect information associated with drugs was utilized to predict potential drug targets.[10,11] Nevertheless, the scarceness of drug side effect information limits this promising approach to those well studied drugs. Because drugs with similar chemical structures tend to target the same proteins and proteins with similar functions tend to be bound by the same drugs, machine learning approaches were recently proposed to predict new drug–protein interactions based on chemical structures and protein sequences.[12,13] However, this kind of approaches measure the similarity between target proteins based on global protein sequences, whereas protein functions are generally determined by local structures. Beyond chemical structures and protein sequences, pharmacological information was also explored to predict novel drug–protein interactions based on the assumption that drugs with similar therapeutic effects tend to target the same proteins.[14,15] More recently, Folger *et al.* utilized a metabolic network model to predict possible targets by simulating the metabolic flux distribution in the network, and successfully identified targets of anticancer drugs.[16]

Despite the success obtained by the methods mentioned above, there is still much room for improvement in the prediction accuracy. Recently, it has been found that protein–protein interactions are dominated by domain–domain interactions, which in turn could be used to predict new protein–protein interactions.[17,18] Therefore, we suspect that the specificity of drug–protein interactions is possibly determined by drug–domain interactions even though the drugs do not physically bind to protein domains. Most recently, it is reported that there are indeed some interactions between compound substructures and protein domains, and a set of chemical substructures shared by drugs are able to bind to a set of protein domains,[19] which confirms our hypotheses to some extent.

[a] *Department of Mathematics, Shanghai University, Shanghai 200444, China*
[b] *Institute of Systems Biology, Shanghai University, Shanghai 200444, China. E-mail: zhaoxingming@gmail.com*
[c] *Department of Complex and Intelligent Systems, Future University-Hakodate, Hokkaido 041-8655, Japan*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c2mb05450g

On the other hand, network analysis has recently provided a novel framework and a set of tools to investigate drug actions. For example, drug-therapy interaction networks have been explored and provide a new perspective of drug design.[20] As will be shown later, we also take advantage of network tools to support our findings in the present study.

In this work, we propose a novel approach to predict drug targets based on protein domains. Unlike the methods proposed by Yamanishi *et al.*,[19] in this work, we assumed that the drugs interacting with the same domain(s) tend to share therapeutic effects instead of chemical substructures. Firstly, we identified drug–domain interactions by investigating the domain components of proteins bound by those drugs that have similar therapeutic effects. Secondly, new drug–protein interactions were predicted based on our inferred drug–domain interactions. Benchmarking with known drug–protein interactions, our method significantly outperforms existing methods that utilize global protein sequences to predict drug targets. Further analysis of a drug–drug association network constructed based on our identified drug–domain interactions demonstrates that the drugs interacting with the same domain(s) tend to have similar therapeutic effects, which implies that domain building blocks of target proteins can provide insights into drug actions.

Results and discussion

Identification of drug–domain interactions

We investigated the domain composition of target proteins bound by drugs with similar therapeutic effects, which were found to bind proteins with similar functions.[15] The drug–protein interactions were obtained from the DrugBank database.[21] We assumed that the domains that occur frequently in drug targets may determine the binding between drugs and proteins. In this way, we identified a set of drug–domain interactions with different probabilities based on eqn (1) (see Methods). Fig. 1 shows a matrix of drug–domain interactions highlighting that there are indeed enriched drug–domain interactions. The detailed drug–domain interactions with corresponding probabilities can be found in File S1 (ESI†). Especially, those interactions with probability above 0.60 were treated as possible drug–domain interactions hereinafter, which results in 557 drug–domain interactions in total.

Based on the drug–domain interactions derived above, we constructed a drug–drug association network (Fig. 2), where one edge was laid between two drugs if these two drugs interact with at least one same domain. The drug association network was visualized with Cytoscape[22] software. In total, the drug association network consists of 253 nodes and 2303 edges, where the nodes were colored according to the first level of the Anatomical Therapeutic Chemical (ATC) classification system. It can be easily seen from Fig. 2 that drugs interacting with the same domain(s) tend to have the same therapeutic effects and are more likely to be clustered together. This finding implies that drugs with similar therapeutic effects may bind proteins with a similar domain composition.

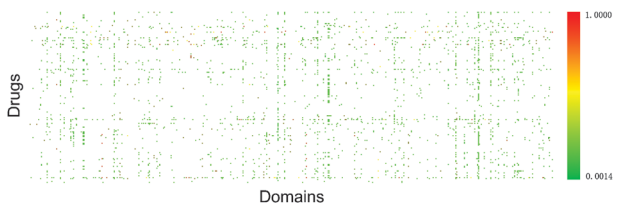

Fig. 1 The drug–domain interaction matrix, where each row stands for the drugs belonging to a therapeutic category (represented as one ATC code in the third level), each column stands for a domain, and each element in the matrix denotes the interaction probability (white means zero, and the gradient color from green to red corresponds to increasing probability).

Furthermore, we investigated the topology of the drug association network. In the literature, modular structures were often observed in many complex networks from natural to engineered systems,[23] where the topological modules consisting of highly dense interacting nodes are usually related to specific functions.[24–26] In the drug association network, we identified modules that consist of highly interlinked drugs (Fig. 2) using a simulated annealing algorithm.[26] Compared with 20 random networks generated by shuffling edges while preserving the node degree, the drug association network shows high modularity of 0.739 compared with 0.190 for randomized networks with a significant p-value of 1.13×10^{-17}. Table 1 lists the 15 modules and their corresponding dominant therapeutic effects. The detailed module information with corresponding memberships can be found in File S2 (ESI†). We observed that most of the drugs in the drug association network can be grouped into specific clusters while each cluster consists of drugs with similar therapeutic effects, which demonstrates that the drugs with similar therapeutic effects indeed tend to target the same protein domains.

Investigating the number of drugs that can interact with each domain, we found that about 70% of domains interact with no more than 3 drugs (Fig. 3), which possibly implies that domain specificity determines drug selectivity. Moreover, we investigated the functions of domains involved in drug–domain interactions according to InterPro database[27] (Fig. 4A). We found that these domains have diverse functions, where lyase activity is the most dominant function and oxidoreductase activity is the second. In addition, we investigated the molecular functions of those proteins that contain these domains. Fig. 4B shows the functional distribution of the target proteins that contain those domains interacting with drugs. Note that the functional distribution of domains is different from that of corresponding proteins containing these domains, which indicates that the functions of these target proteins are possibly determined by other domains or functional sites rather than the domains interacting with drugs.

Next, we explored the sequence length of those domains that interact with drugs, and also the protein sequences in which they are located. Fig. 5 shows the distribution of domains according to the ratios between their amino acid sequence lengths and those of the corresponding protein sequences in which they are located. The results show that only about 25% of the domains occupy most part (ratio larger than 0.6) of the corresponding protein sequences. That is, the functions of most drug targets may not be determined by their component domains that interact with drugs, which is consistent with the conclusion drawn from functional distributions.

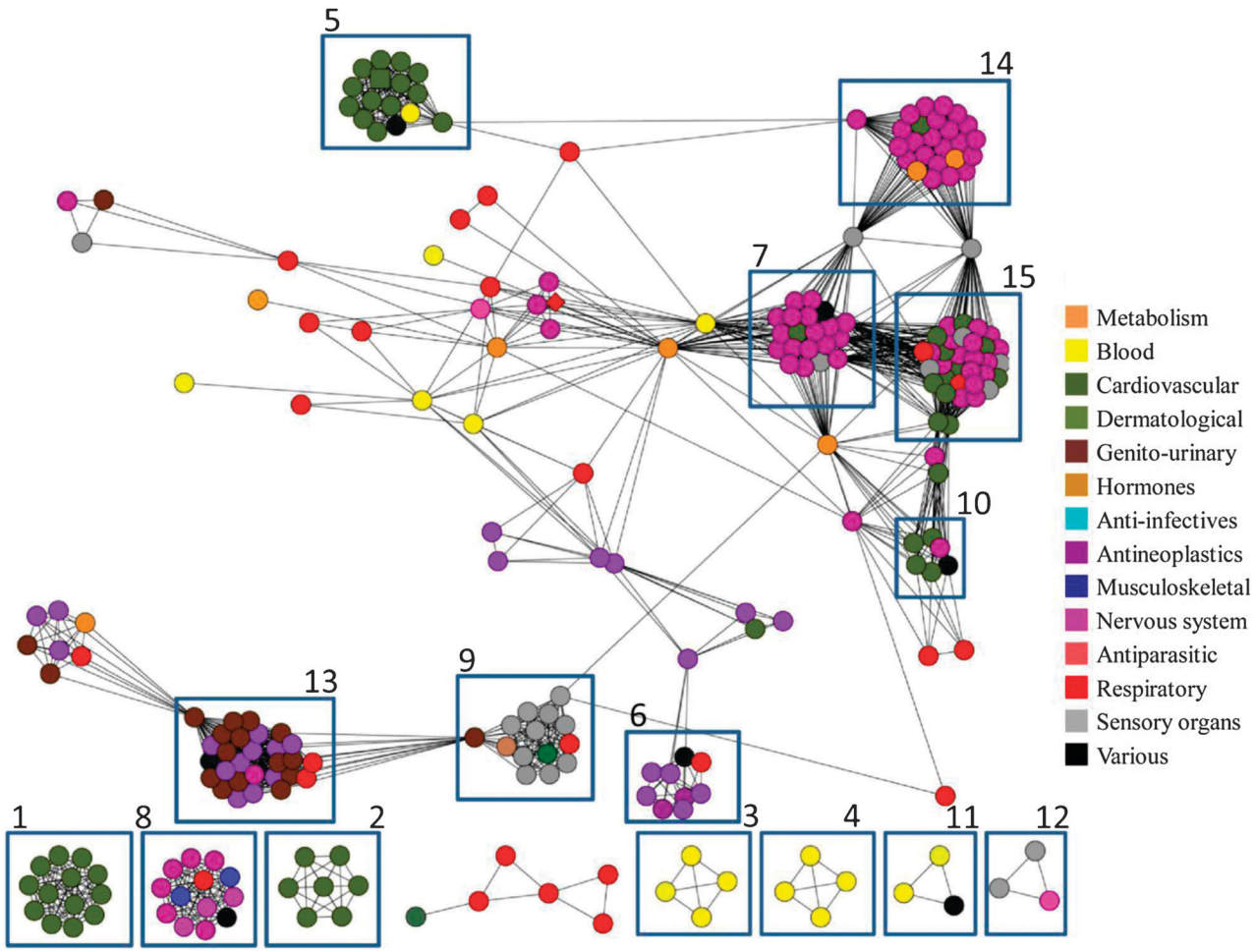

Fig. 2 The drug association network. Nodes represent drugs and are colored according to their first level ATC annotation, and two drugs are linked if they interact with at least one same domain.

Table 1 The modules detected by a simulated annealing algorithm from the drug association network, and their corresponding dominant therapeutic effects represented as the first level of ATC code

Module	Number of drugs	1st level ATC code	Proportion[a]
1	13	C	1.0000
2	7	C	1.0000
3	4	B	1.0000
4	4	B	1.0000
5	16	C	0.9375
6	9	L	0.8889
7	26	N	0.8462
8	13	N	0.6923
9	15	D	0.6667
10	6	C	0.6667
11	3	L	0.6667
12	3	B	0.6667
13	35	G	0.6571
14	37	N	0.6486
15	62	N	0.3710

[a] Proportion means the fraction of drugs that are annotated with the corresponding ATC code, *i.e.* the dominant therapeutic category, in the third column.

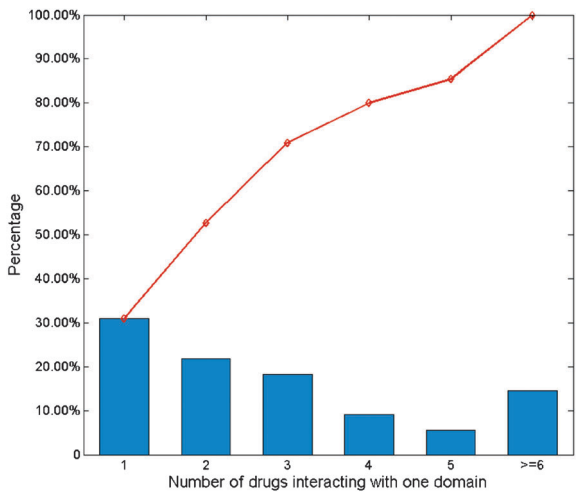

Fig. 3 The percentage of domains that interact with different numbers of drugs. The red curve denotes the cumulative percentage.

In addition, we surprisingly found that those domains that interact with drugs belonging to the same therapeutic category significantly tend to interact with each other (p-value $= 4.38 \times 10^{-12}$, Fisher's exact test) according to the experimentally determined domain–domain interactions obtained from the DOMINE[28] database. For example, the drugs belonging to the Cardiovascular category (ATC code C)

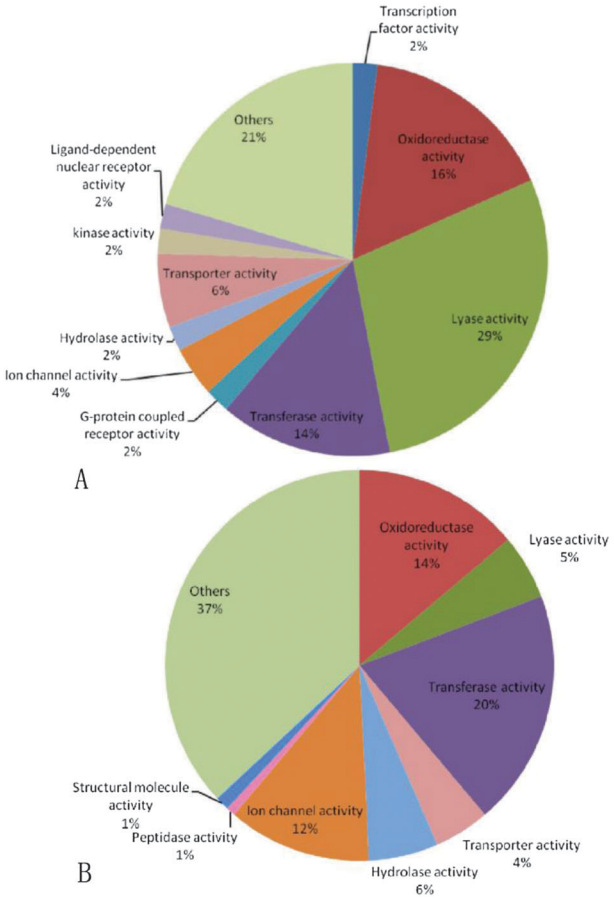

Fig. 4 (A) The distribution of domains that can interact with drugs according to their functions; (B) the distribution of proteins which contain the domains interacting with drugs according to their functions.

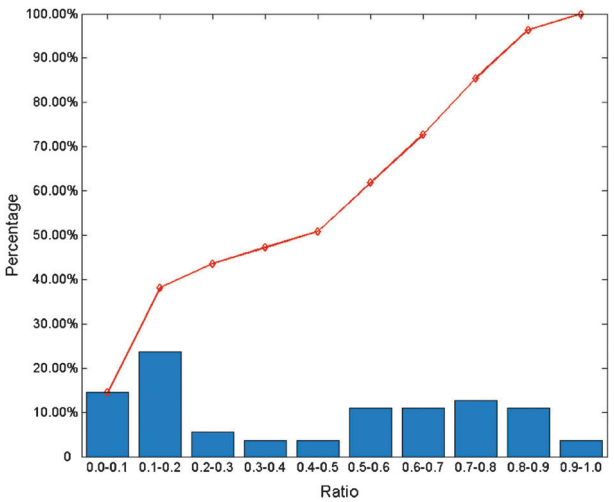

Fig. 5 The percentage of domains according to the ratio between their seuqence lengths and those of corresponding proteins in which they are located. The red curve denotes the cumulative percentage.

interact with 7 domains, among which 5 domains have interactions (Fig. 6). Moreover, we found that the interacting domain pairs tend to share functions. For example, the three domains

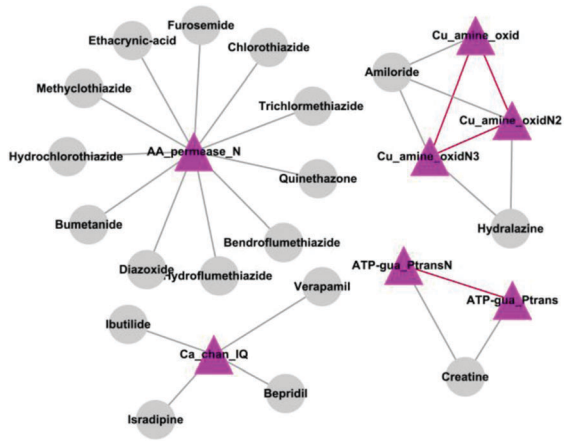

Fig. 6 The drug–domain and domain–domain interactions, where the gray circles denote drugs belonging to the cardiovascular category and the purple triangles denote domains, and the red edges denote the interactions between domains while the other edges denote drug–protein interactions.

Cu_amine_oxid, Cu_amine_oxidN2 and Cu_amine_oxidN3 that interact with each other have completely the same functions, including copper ion binding, primary amine oxidase activity, *etc*. Another interacting domain pair, ATP-gua_PtransN and ATP-gua_Ptrans, has the same functions of kinase activity and transferase activity. Looking at the proteins that contain these interacting domain pairs, we observed that a pair of proteins tend to belong to the same pathway if they contain an interacting domain pair. For example, the four proteins, CKM, CKB, CKMT1 and CKMT2, that contain the interacting domain pair ATP-gua_Ptrans and ATP-gua_PtransN, belong to the arginine and proline metabolic pathway. We can conclude from these observations that even if the drugs belonging to the same therapeutic category bind to different domains, they may affect the same pathways and therefore have similar therapeutic effects on the biological systems.

Prediction of drug–protein interactions

After identifying drug–domain interactions as described above, we aimed to predict potential drug–protein interactions (eqn (2)) with the assumption that drug–protein interactions are driven by drug–domain interactions. Especially, we used the drug–protein interactions from DrugBank as the gold standard to evaluate our proposed method, where the drug–protein interactions from DrugBank were used as a positive set while other possible drug–protein pairs were used as a negative set.

Furthermore, we compared our method with two other approaches, *i.e*. NP_{SG} that integrates chemical similarity and protein sequence similarity and NP_{TG} that integrates drug therapeutic similarity and protein sequence similarity, where the two approaches integrate different similarities with the nearest profile technique proposed by Yamanishi *et al.*[12]

The above mentioned three methods were applied to predict drug–protein interactions from all possible drug–protein pairs between 883 drugs and 1046 proteins, where the drug–protein interactions from DrugBank were used as the gold standard.

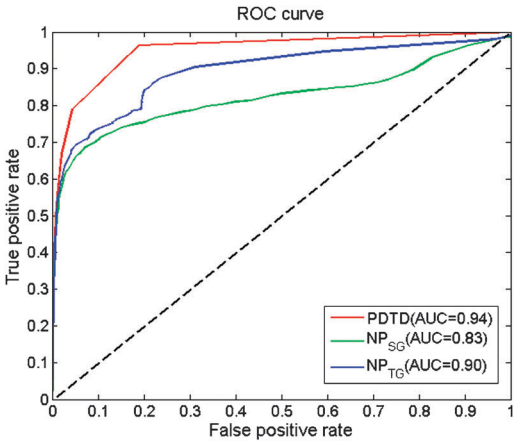

Fig. 7 The performance of different prediction approaches, where PDTD is our proposed method, NP_{SG} denotes the nearest profile integration of chemical similarity and protein sequence similarity, and NP_{TG} denotes the nearest profile integration of therapeutic similarity and and protein sequence similarity.

Fig. 7 shows the results obtained by different methods, from which we can see that our PDTD method significantly outperforms the other two methods with an AUC (area under the ROC curve) score of 0.94, in comparison with 0.90 for NP_{TG} and 0.83 for NP_{SG}. Our predicted drug–protein interactions can be found in File S3 (ESI†). These results suggest that our proposed method is indeed effective for predicting drug–protein interactions.

In addition, we found some of our predicted drug–protein interactions in other databases although they cannot be found in DrugBank. In particular, 6 out of 190 predicted drug–protein interactions can be found in the STITCH[29] database, a comprehensive drug–protein interaction database. For example, pentostatin is used to treat hairy cell leukemia. Utilizing the drug–domain interactions derived previously, we predict that pentostatin interacts with three proteins, including AMPD1, AMPD2, and AMPD3. These three drug–protein interactions have been reported in the STITCH database although they are not deposited in DrugBank. The overlap between our predictions and those reported in the literature demonstrates the predictive power of our proposed method.

Investigating the relationship between our predicted targets and those known ones bound by the same drug, we found that they either interact with each other or belong to the same pathway. Since interacting proteins or proteins in the same pathway generally share functions, the drugs that target these proteins may therefore have similar therapeutic effects, thereby validating our predictions to some extent. For example, the protein CPS1 was predicted to interact with L-glutamine (DB00130) based on the interactions between its two domains (CPSase_L_chain and CPSase_L_D2) and L-glutamine. CPS1 was found to interact with GLUL and GLS according to STRING[30] database, while GLUL and GLS were already known to be targeted by L-glutamine. In addition, these three proteins also share some pathways, such as alanine, aspartate, and glutamate metabolic pathways.

The good performance of our proposed PDTD method implies that drug–protein interactions are indeed determined by drug–domain interactions which sometimes cannot be captured by the global similarity between proteins, although the drug–domain interactions are not necessarily physical binding here. For example, domain A_deaminase (PF00962) interacts with pentostatin (DB00552), an antineoplastic agent. With this drug–domain interaction, we can successfully predict ADA and AMPD1 as the targets of the drugs belonging to the antineoplastic category. However, the above target information of pentostatin cannot be detected with global protein sequences, where the sequence similarity between ADA and AMPD1 is only 0.021996 (normalized Smith–Waterman score[12]).

Concluding remarks

In this article, a novel approach was presented to predict the interactions between drugs and proteins with the assumption that drug–protein interactions are determined by drug–domain interactions. A simple method was proposed to infer the drug–domain interactions based on known drug–protein interactions, which in turn were used to predict drug–protein interactions. A drug association network constructed based on the derived drug–domain interactions exhibits a highly modular structure, with most of the drugs grouped into clusters with similar therapeutic effects. This finding confirms that drugs interacting with the same domains indeed tend to have similar therapeutic effects, *i.e.* the same ATC codes, which proves the effectiveness of our proposed method. In addition, benchmarking with known drug–protein interactions indicates that our proposed approach outperforms existing approaches that utilize protein sequences to predict drug–protein interactions, which demonstrates the predictive power of our approach. The success of the proposed method also confirms the hypotheses that the drug–protein interactions are determined by local protein structures and drugs with similar therapeutic effects tend to target the same domains.

In addition, we analyzed the domain compositions of all drug target proteins (Fig. S1A, ESI†) and our predicted ones (Fig. S1B, ESI†). In general, drugs tend to target proteins that contain single domains, which implies that the interactions between single-domain proteins and drugs are more likely accomplished through drug–domain interactions. We also found that our predicted targets tend to contain multiple domains compared with all known targets, which possibly indicates that the functions of our predicted targets are not determined by the domains involved in drug–domain interactions.

We also noticed that there is room to improve our proposed method. In this work, we limited drugs belonging to the same therapeutic category (the same ATC code) as therapeutically similar drugs. If more advanced therapeutic similarity measures were adopted, *e.g.* therapeutic similarity proposed by Zhao and Li,[15] the performance of our proposed method could be improved. In addition, it is possible to improve our approach if more complicated techniques were utilized to integrate drug–domain interactions and therapeutic similarity.

Despite possible limitations, we believe that our methodology provides an alternative way to finding new targets of known drugs, which in turn can help to provide insights into mechanisms of action of drugs and possibly propose novel indications for old drugs.[31]

Materials and methods

Data sources

All the drugs and known drug–protein interactions were retrieved from the DrugBank database (Version 2.5), which is one of the most complete chemo-informatics resources.[21] In this data set, there are about 4800 drug entries, some of which have associated therapy information represented as the Anatomic Therapeutic Chemical (ATC) classification system. Here, only those approved human drugs with ATC annotations were considered. In addition, we further require the drugs to have corresponding target protein annotations in DrugBank. Especially, only the drugs whose target proteins that can be found in the Uniprot database[32] were considered so that some unknown proteins were removed. Consequently, 883 drugs with both target and therapy information and 1046 proteins were kept for further analysis, where each drug has at least one target protein.

The amino acid sequences of all human proteins were obtained from the Uniprot database. The domain annotations for these proteins were obtained from InterPro[27] where available.

Identification of drug–domain interactions

To identify drug–domain interactions, we investigated the domain compositions of proteins bound by drugs belonging to the same therapeutic category. Here, the same therapeutic category means the same ATC code at the third level. In the ATC classification system, the therapeutic effects are divided into five levels from general to specific, where the first level is the most general level. The hierarchical structure of the ATC code provides an ideal framework for analyzing the relationships between drugs and therapeutic effect at different resolutions.[20]

We assumed that the drug–protein interactions are actually accomplished through drug–domain interactions. Therefore, we investigated whether there are enriched domains in the targets of drugs that have the same therapeutic effect, and these enriched domains are regarded as the ones that interact with drugs in the therapeutic category. The probability of a domain m_i interacting with drugs having a similar therapeutic effect $d_{\mathrm{ATC}(j)}$ is defined as follows:

$$P(m_i\_d_{\mathrm{ATC}_{(j)}}) = \frac{N(p|m_i)}{N(p'|m_i)} \quad (1)$$

where $\mathrm{ATC}_{(j)}$ means ATC code j, $P(m_i\_d_{\mathrm{ATC}(j)})$ is the probability that domain m_i interacts with drugs annotated with $\mathrm{ATC}_{(j)}$, $N(p|m_{(i)})$ denotes the number of proteins that are both bound by drugs belonging to $\mathrm{ATC}_{(j)}$ and contain domain m_i, and $N(p'|m_{(i)})$ is the number of all human proteins that contain domain m_i.

After obtaining the probability of drug–domain interactions, we can set a threshold to determine whether a pair of drug and domain interacts, where those drug–domain pairs with probabilities above the threshold were treated as drug–domain interactions. In addition, we can predict drug–protein interactions based on identified drug–domain interactions as follows:

$$P(p_i\_d_{\mathrm{ATC}(j)}) = 1 - \prod_{m_k \in p_i} (1 - P(m_k\_d_{\mathrm{ATC}(j)})) \quad (2)$$

where $P(p_i\_d_{\mathrm{ATC}(j)})$ is the probability of protein p_i interacting with drugs belonging to $\mathrm{ATC}_{(j)}$, $P(m_k\_d_{\mathrm{ATC}(j)})$ is the probability that domain m_k interacts with drugs from $\mathrm{ATC}_{(j)}$, and protein p_i contains domain m_k.

Comparison with other methods

In order to validate our results, we compared our method with two versions of the nearest profile method proposed by Yamanishi *et al.*[12] The first one integrates drug therapeutic similarity and protein sequence similarity, while the second one integrates compound structure similarity and protein sequence similarity.

The protein sequence similarity $S(P, P')$ between two proteins (P, P') was calculated as the normalized Smith–Waterman score[33] defined as follows:

$$S(P, P') = \frac{SS(P, P')}{\sqrt{SS(P, P)SS(P', P')}} \quad (3)$$

where $\mathrm{SS}(\cdot,\cdot)$ denotes the original Smith–Waterman score.[12] The drug therapeutic similarity was calculated as its longest matched prefix based on the hierarchical structure of ATC codes as described by Zhao and Li.[15] The chemical similarity was calculated as the Tanimoto similarity by utilizing the Chemistry Development Kit.[34]

To integrate different similarities, the nearest neighbor profile technique proposed by Yamanishi *et al.*[12] was adopted here. For drugs with either chemical structures or therapy information, the targets of a new drug D_{new} can be predicted as follows:

$$D_{P,\mathrm{new}} = S(D_{\mathrm{new}}, D_{\mathrm{nearest}})D_{P,\mathrm{nearest}} \quad (4)$$

where D_{nearest} is the compound that is most similar to D_{new}, $D_{P,\mathrm{nearest}}$ is a binary interaction profile vector for D_{nearest} with 1 denoting the interaction between D_{nearest} and the corresponding protein, while 0 denotes no interaction. $D_{P,\mathrm{new}}$ is the interaction profile for D_{new}, and $S(D_{\mathrm{new}}, D_{\mathrm{nearest}})$ denotes the similarity between D_{new} and D_{nearest}. Here, the similarity can be chemical structural similarity or therapeutic similarity.

Similarly, given a new protein P_{new}, we can predict the drugs that interact with this protein as follows:

$$P_{D,\mathrm{new}} = S(P_{\mathrm{new}}, P_{\mathrm{nearest}})P_{D,\mathrm{nearest}} \quad (5)$$

where P_{nearest} is the protein that is most similar to P_{new}, $P_{D,\mathrm{nearest}}$ is a binary interaction profile vector for P_{nearest} with 1 denoting the interaction between P_{nearest} and the corresponding drug, while 0 denotes no interaction, $P_{D,\mathrm{new}}$ is the interaction profile for P_{new}, and $S(P_{\mathrm{new}}, P_{\mathrm{nearest}})$ denotes the sequence similarity between P_{new} and P_{nearest} based on eqn (3).

Acknowledgements

This work was partly supported by the Innovation Program of Shanghai Municipal Education Commission (10YZ01), Shanghai Rising-Star Program (10QA1402700), and National Natural Science Foundation of China (61103075, 91130032).

References

1 M. A. Yildirim, K. I. Goh, M. E. Cusick, A. L. Barabasi and M. Vidal, *Nat. Biotechnol.*, 2007, **25**, 1119–1126.
2 J. Bajorath, *Curr. Opin. Chem. Biol.*, 2008, **12**, 352–358.
3 T. I. Oprea, A. Tropsha, J. L. Faulon and M. D. Rintoul, *Nat. Chem. Biol.*, 2007, **3**, 447–450.
4 M. G. Siegel and M. Vieth, *Drug Discovery Today*, 2007, **12**, 71–79.
5 J. R. Miller, S. Dunham, I. Mochalkin, C. Banotai, M. Bowman, S. Buist, B. Dunkle, D. Hanna, H. J. Harwood, M. D. Huband, A. Karnovsky, M. Kuhn, C. Limberakis, J. Y. Liu, S. Mehrens, W. T. Mueller, L. Narasimhan, A. Ogden, J. Ohren, J. V. N. V. Prasad, J. A. Shelly, L. Skerlos, M. Sulavik, V. H. Thomas, S. VanderRoest, L. A. Wang, Z. G. Wang, A. Whitton, T. Zhu and C. K. Stover, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 1737–1742.
6 C. T. Walsh and M. A. Fischbach, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 1689–1690.
7 P. Imming, C. Sinning and A. Meyer, *Nat. Rev. Drug Discovery*, 2006, **5**, 821–834.
8 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, T. B. Tran, R. Whaley, R. A. Glennon, J. Hert, K. L. Thomas, D. D. Edwards, B. K. Shoichet and B. L. Roth, *Nature*, 2009, **462**, 175–181.
9 A. C. Cheng, R. G. Coleman, K. T. Smyth, Q. Cao, P. Soulard, D. R. Caffrey, A. C. Salzberg and E. S. Huang, *Nat. Biotechnol.*, 2007, **25**, 71–75.
10 M. Campillos, M. Kuhn, A. C. Gavin, L. J. Jensen and P. Bork, *Science*, 2008, **321**, 263–266.
11 M. Iskar, G. Zeller, X. M. Zhao, V. van Noort and P. Bork, *Curr. Opin. Biotechnol.*, 2011, **23**, 1–8.
12 Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda and M. Kanehisa, *Bioinformatics*, 2008, **24**, i232–i240.
13 H. Yabuuchi, S. Niijima, H. Takematsu, T. Ida, T. Hirokawa, T. Hara, T. Ogawa, Y. Minowa, G. Tsujimoto and Y. Okuno, *Mol. Syst. Biol.*, 2011, **7**, 472.
14 Y. Yamanishi, M. Kotera, M. Kanehisa and S. Goto, *Bioinformatics*, 2010, **26**, i246–i254.
15 S. Zhao and S. Li, *PLoS One*, 2010, **5**, e11764.
16 O. Folger, L. Jerby, C. Frezza, E. Gottlieb, E. Ruppin and T. Shlomi, *Mol. Syst. Biol.*, 2011, **7**, 501.
17 X. M. Zhao, Y. Wang, L. Chen and K. Aihara, *Proteins*, 2008, **72**, 461–473.
18 X. M. Zhao, L. Chen and K. Aihara, *Proteins*, 2010, **78**, 1243–1253.
19 Y. Yamanishi, E. Pauwels, H. Saigo and V. Stoven, *J. Chem. Inf. Model.*, 2011, **51**, 1183–1194.
20 J. C. Nacher and J. M. Schwartz, *BMC Pharmacol.*, 2008, **8**, 5.
21 D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang and J. Woolsey, *Nucleic Acids Res.*, 2006, **34**, D668–D672.
22 P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker, *Genome Res.*, 2003, **13**, 2498–2504.
23 R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii and U. Alon, *Science*, 2002, **298**, 824–827.
24 E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai and A. L. Barabasi, *Science*, 2002, **297**, 1551–1555.
25 M. E. Newman, *Phys. Rev. E: Stat. Nonlinear Soft Matter Phys.*, 2004, **69**, 066133.
26 R. Guimera and L. A. Nunes Amaral, *Nature*, 2005, **433**, 895–900.
27 S. Hunter, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, A. Laugraud, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, N. Mulder, D. Natale, C. Orengo, A. F. Quinn, J. D. Selengut, C. J. Sigrist, M. Thimma, P. D. Thomas, F. Valentin, D. Wilson, C. H. Wu and C. Yeats, *Nucleic Acids Res.*, 2009, **37**, D211–D215.
28 S. Yellaboina, A. Tasneem, D. V. Zaykin, B. Raghavachari and R. Jothi, *Nucleic Acids Res.*, 2011, **39**, D730–D735.
29 M. Kuhn, D. Szklarczyk, A. Franceschini, M. Campillos, C. von Mering, L. J. Jensen, A. Beyer and P. Bork, *Nucleic Acids Res.*, 2010, **38**, D552–D556.
30 D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen and C. von Mering, *Nucleic Acids Res.*, 2011, **39**, D561–D568.
31 X. M. Zhao, M. Iskar, G. Zeller, M. Kuhn, V. van Noort and P. Bork, *PLoS Comput. Biol.*, 2011, **7**, e1002323.
32 C. H. Wu, L. S. L. Yeh, H. Z. Huang, L. Arminski, J. Castro-Alvear, Y. X. Chen, Z. Z. Hu, P. Kourtesis, R. S. Ledley, B. E. Suzek, C. R. Vinayaka, J. Zhang and W. C. Barker, *Nucleic Acids Res.*, 2003, **31**, 345–347.
33 T. F. Smith and M. S. Waterman, *J. Mol. Biol.*, 1981, **147**, 195–197.
34 C. Steinbeck, C. Hoppe, S. Kuhn, M. Floris, R. Guha and E. L. Willighagen, *Curr. Pharm. Des.*, 2006, **12**, 2111–2120.

