Integrative Biology

Dynamic Article Links

Cite this: *Integr. Biol.*, 2012, **4**, 778–788

www.rsc.org/ibiology

PAPER

Drug repositioning through incomplete bi-cliques in an integrated drug–target–disease network†‡

Simone Daminelli,§ V. Joachim Haupt,§ Matthias Reimann and Michael Schroeder*

Received 4th November 2011, Accepted 14th March 2012
DOI: 10.1039/c2ib00154c

Recently, there has been much interest in gene–disease networks and polypharmacology as a basis for drug repositioning. Here, we integrate data from structural and chemical databases to create a drug–target–disease network for 147 promiscuous drugs, their 553 protein targets, and 44 disease indications. Visualizing and analyzing such complex networks is still an open problem. We approach it by mining the network for network motifs of bi-cliques. In our case, a bi-clique is a subnetwork in which every drug is linked to every target and disease. Since the data are incomplete, we identify incomplete bi-cliques, whose completion introduces novel, predicted links from drugs to targets and diseases. We demonstrate the power of this approach by repositioning cardiovascular drugs to parasitic diseases, by predicting the cancer-related kinase PIK3CG as a novel target of resveratrol, and by identifying for five drugs a shared binding site in four serine proteases and novel links to cancer, cardiovascular, and parasitic diseases.

Introduction

Drug repositioning

In the last few decades of drug discovery, investments in research and development have been constantly increasing. It was estimated in 2004 that the cost of bringing a single drug to the market was up to 800 million US$,[1] while more recently the estimation increased to 1.2 billion US$.[2] In contrast, the number of new approved drugs per year follows an opposite trend,[3] and the total amount of approved drugs still remains low. From 1950 to 2008, the FDA approved only a total of 1222 new molecular entities.[3]

Besides many strategies to lower costs and speed-up drug development, such as internal drug discovery or in-licensing, one of the most promising methods is drug repositioning.[4,5] The repositioning of existing drugs for new indications can count on many successful examples, such as sildenafil (Viagra) and Celecoxib (Celebrex) from Pfizer, or Duloxetine (Cymbalta) and Raloxifene (Evista) from Eli Lilly.[6]

Computational approaches to drug repositioning

Several computational methods have been applied to drug repositioning. A recent review[7] suggests a classification into "drug-based" and "disease-based" strategies. However, there is a lack of methods that combine both strategies, and integrated approaches[8,9] are rare. Databases, such as the PDB[10] for biological macromolecule structures, DrugBank,[11,12] PubChem[13,14] for compounds, SIDER[15] for drug side effects, ChEMBL[16] for biological activity data and the TTD,[17] or the CTD[18,19] that integrate information about gene/protein–drug–disease associations, are valuable resources to develop integrated approaches.

Many recent studies combined prior knowledge and different computational methods to clarify molecular mechanisms and genetic correlation of pathologies,[20–22] to associate protein

Insight, innovation, integration

During the last decade, drug repositioning became an integral part of drug discovery. Among the computational approaches to drug repositioning, integrated approaches are rare. Here, we integrate data from various resources to construct a high confidence drug–target/disease network.

By analyzing the network structure to identify so-called "incomplete bi-cliques", we describe a novel approach to predict new links between drugs, targets, and/or diseases. These predictions are solely based on the network topology. We demonstrate the validity of our approach by evaluating three case studies and discussing the influence of perturbation on the network structure.

BIOTEC, TU Dresden, Germany. E-mail: ms@biotec.tu-dresden.de; Tel: +49 351 463 400 60

† Published as part of an iBiology themed issue entitled "Computational Integrative Biology" Guest Editor: Prof. Jan Baumbach.

‡ Electronic supplementary information (ESI) available. See DOI: 10.1039/c2ib00154c

§ These authors contributed equally to this work.

complexes with diseases,[23,24] to improve the drug discovery process,[25–27] or to propose the re-positioning of known drugs.[28–30]

Structural bioinformatics or chemoinformatic approaches to drug repositioning comprise chemo(or ligand)-centric[31] and binding site-centric[31,32] strategies. Chemo-centric methods exploit chemical similarities among drug ligands to find new targets for known drugs, which is demonstrated to be successful by Keiser *et al.*[33] and Kinnings and Jackson.[34] However, their potential is limited since chemical similarity contributes weakly to the biological activity profile of a drug.[26,35,36] Binding site-centric approaches account for similarities in key physico-chemical features in ligand binding sites of proteins. Exploiting these similarities, new targets for existing drugs could be uncovered.[32] These include celecoxib,[37] ibuprofen,[38] entacapone,[39] and a trypanocidal agent.[40] The antiviral brivudine could be shown to bind to the 27 Da heat shock protein due to binding site similarity to the viral thymidine kinase. Subsequently, it was successfully tested as a modulator of chemoresistance in pancreatic cancer.[41]

Other computational approaches to drug repositioning are ontological profiles from text mining,[42] similarities in pharmacological effect,[26] in side effect,[28] in therapeutic application,[43] or the analysis of gene expression data.[30]

The vast amount of data available to pharmaceutical research still needs an integrated approach.[44] For that reason, approaches on the network level gained popularity in the last few years and might become the new paradigm in drug discovery.[45] Exploiting the wealth of available data (ligand/protein structures, drug–target relations, drug pharmacological profiles, clinical trials, ...) from various sources and skillfully combining them to networks hold a great potential for drug repositioning[46–48] and for discovering new drug relations.

Hence, we construct a complex drug–target–disease network and show how to mine it using network motifs. First, we introduce the network, then we develop an approach to predict novel drug–target and drug–disease associations by completing incomplete bi-cliques in the network.

Since the predictions are solely based on the network structure—reflecting biological relations in the data—the predictions are not ultimately regarded true, but rather further validated using external information.

Finally, we demonstrate the power of the approach by repositioning cardiovascular drugs to parasitic diseases and *vice versa* and by predicting novel drug targets.

Results and discussion

An integrated drug–target–disease network

We constructed a drug–target–disease network in two steps:

1. Drug–target links: first, we selected 147 promiscuous drugs (three or more known targets) from over 3000 drugs in over 65 000 protein structures in PDB. Based on the PDB, we added 553 targets and 759 drug–target links to the network. Each drug had at least three targets with an average of 5.16 and a maximum of 31 for staurosporine (see Table 1).

2. Drug–disease links: second, we mined the drug databases CTD, TTD, and PubChem for drug–disease links by mapping a drug's terminology to relevant terms from the medical subject headings MeSH. This way, we added 44 disease terms and pharmacological actions, and 592 drug–disease links to the network. On average, a drug had 4 terms.

Table 1 The most promiscuous drugs and diseases covered by the network

Drug	Targets	Diseases
Staurosporine	**31**	3
Methotrexate	**17**	9
Acarbose	**16**	3
Niacinamide	**16**	3
Quercetin	**15**	18
Indomethacin	8	**22**
Quercetin	15	**18**
Resveratrol	9	**18**
Tretinoin	14	**18**
Zoledronic Acid	5	**14**

Disease	Drugs
Neoplasms	42
Immune system diseases	38
Pathological conditions, signs and symptoms	31
Digestive system diseases	28
Male urogenital diseases	27
Nervous system diseases	27
Skin and connective tissue diseases	24
Cardiovascular diseases	24
Respiratory tract diseases	23
Bacterial infections and mycoses	23

The top promiscuous drug—in terms of disease indications—is indomethacin, an anti-inflammatory drug inhibiting cyclooxygenases.[50] Indomethacin was shown to be effective in different neoplasms,[51] parasitic and virus infections (Leishmaniasis[50] or Hepatitis C[52]), mental disorder (Alzheimer's disease[53]) and many other diseases. More than 80 distinct applications were mapped to 22 top level terms in the MeSH hierarchy, demonstrating the promiscuity of indomethacin in terms of its applications.

Another interesting example is quercetin, a flavonoid, which has 15 known targets including oxidoreductases, transferases, hydrolases and hydro-lyases, and 18 disease indications ranging from cardiovascular diseases (hypertension[54,55] or myocardial infarction[56]) to neoplasms (colorectal,[57] liver[58] or pancreatic[59]) or virus diseases (Influenza[60,61]).

For details on the network construction, see the Materials and methods section.

Visual network analysis: from graph to power graph

A first step in network analysis is visualization. Fig. 1B shows the network, which was described above, visualized in Cytoscape.[49] Overall, the network is a "fur ball", so that it is too complex for direct visualization. One approach to overcome this problem is the identification of network motifs and their exploitation for an improved visualization. To this end, power graphs have been introduced.[62] Power graphs identify cliques and bi-cliques in networks and represent them as power nodes with power edges.

A bi-clique is a network motif in which two sets of nodes all mutually interact with each other. Consider the graph below with two drugs (triangles) and three targets (circles). The drugs

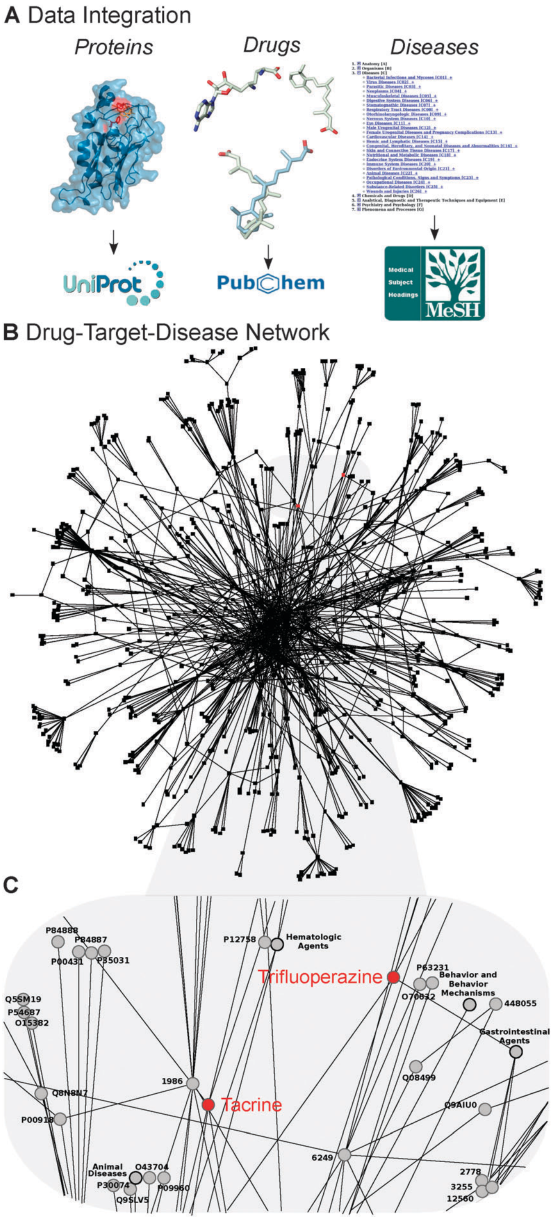
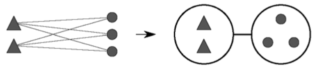


Fig. 1 A complex drug–target–disease network. (A) Data on proteins, drugs, and diseases are integrated into a network with 744 nodes and 1351 edges. (B) Network visualization in Cytoscape.[49] (C) Close-up: the two drugs tacrine and trifluoperazine are unrelated in such a classical visualization.

and targets form a bi-clique since all drugs interact with all targets.



Bi-cliques are named in graph theory $K_{n,m}$, where n and m are the sizes of the two sets of nodes. *I.e.*, the above bi-clique is called $K_{2,3}$. Power graphs subsume the above six interactions by introducing one power node for the drugs linked by one power edge to a power node for the targets. Thus, the identification of such a bi-clique reduces the number of edges from six to one.

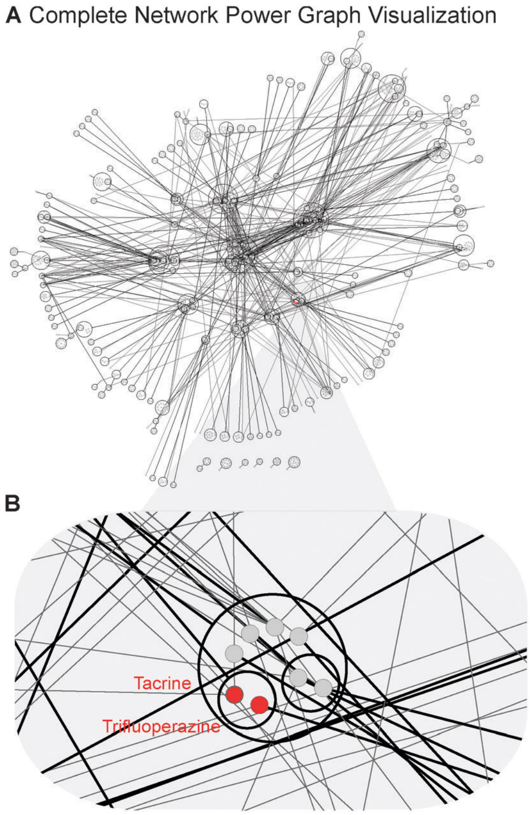


Fig. 2 (A) Power graph of Fig. 1B. (B) Tacrine and trifluoperazine are related due to common targets and diseases.

Table 2 Size and complexity of the original network and its power graph. Edges are reduced by 65%

Visualization	Nodes	Edges	Avg. degree
Original graph	744	1351	3.6
Power graph	944	471	1.0

Fig. 2A shows the power graph for the drug network. The complexity has been substantially reduced. As summarized in Table 2, the 1351 edges of the original network have been reduced by 65% to 471. Despite this reduction, no information is lost, since power edges summarize edges in the original graph.

Fig. 1C shows that the original network does not relate tacrine and trifluoperazine. However, in the power graph representation in Fig. 2B, they are grouped since they form a bi-clique with central nervous system terms. This grouping is meaningful since they are both applied to treat mental disorders and nervous system diseases.

The focus of power graphs on bi-cliques as motifs for abstraction is particularly advantageous for the drug–target–disease network which is a bi-partite network. It consists only of links from drugs to targets and to diseases, but not directly among drugs or among targets or diseases. This structure enforces that power graphs group indirectly related drugs.

Incomplete bi-cliques

Biological networks are inherently incomplete due to the limits of the originally integrated data sources. If edges are missing from a bi-clique, it falls apart into two bicliques. Consider Fig. 3. The top

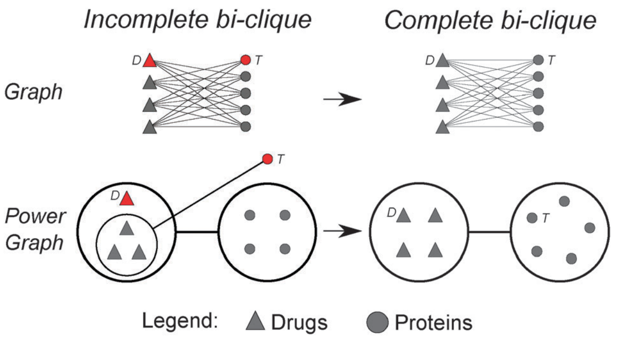


Fig. 3 Incomplete bi-cliques. Adding the edge D–T completes the $K_{4,5}$ bi-clique. Top: Graph representation. Bottom: Power graph representation. In power graphs, completing bi-cliques further reduces the number of edges.

left shows a network of four drugs and five targets. They all mutually interact apart from D and T. Thus, there is an incomplete $K_{4,5}$ bi-clique. Adding the edge D–T completes this incomplete bi-clique. In the power graph representation, completing incomplete bi-cliques further reduces the number of edges, as shown in Fig. 3.

The key contribution of this paper is the prediction of novel links from drugs to targets and diseases by completing incomplete bi-cliques.

Next, we will show that the power graph structure has a biological grounding, and that the identified bi-cliques hold a signal that is not random.

We will demonstrate the power of this approach by measuring the biological information in terms of shared binding site similarity (see Materials and methods for details) among proteins clustered together in power nodes.

The biological content will be compared between power graphs with randomly removed edges, and power graphs with edges randomly removed in such a way that specifically present bi-cliques are disrupted.

In a similar way, we will measure the influence of randomly added edges *versus* edges that increase the completeness of bi-cliques, as previously schematized in Fig. 3.

As a proof of concept, we will show in detail and manually validate three examples concerning how to reposition cardiovascular drugs to anti-parasitic diseases and how to predict novel targets or applications for known drugs.

Power graphs bi-cliques have a biological meaning

We systematically analyzed whether target proteins that are grouped in power nodes share a binding site and thus have a biological grounding. To further validate the incomplete bi-clique hypothesis, we investigated if completing or disrupting bi-cliques (Fig. 4, dashed lines) is influencing the grouping more than introducing random noise (Fig. 4, solid lines).

We considered the following different conditions:

- randomly removing edges (RR—Random Removing).
- removing edges increasing the incompleteness of present bi-cliques (DB—Disrupting Bi-cliques).
- randomly adding edges (RA—Random Adding).
- adding edges increasing the completeness of present bi-cliques (CB—Completing Bi-cliques).

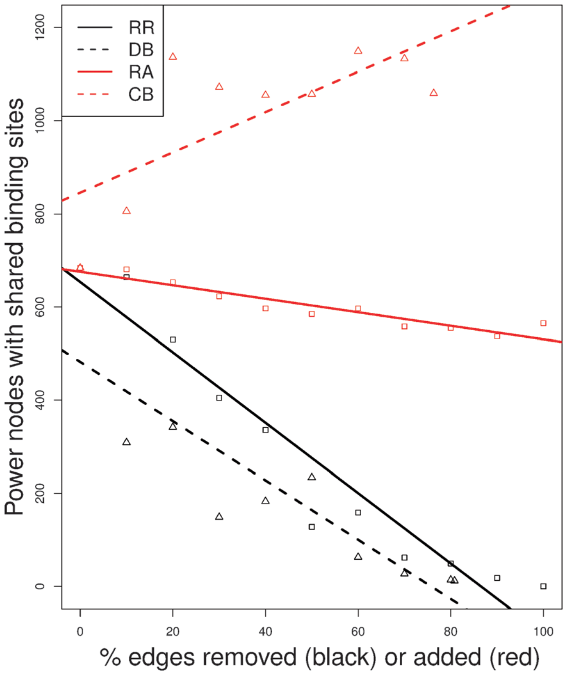


Fig. 4 Biological signal in power graphs. The number of protein couples sharing a similar binding site, grouped into the same power node, under different conditions and an increasing percentage of edges removed (black) or added (red). Disrupting bi-cliques (DB) is more detrimental for shared binding sites than random removal (RR). Completing bi-cliques (CB) is more beneficial for shared binding sites than random addition (RA).

We repeated the experiment considering an increasing percentage of removed or added edges (from 0% to 100%). As visualized in Fig. 4, when randomly removing edges we observed a decreasing number of protein couples that share a similar binding site, while grouped into the same power node. This indicates that the biological information reflected by the power graph structure is meaningful and not random.

When disrupting the identified incomplete bi-cliques (or decreasing the completeness of the bi-cliques), the biological similarity dropped more than when removing random edges. This confirms that the identified incomplete bi-cliques are biologically meaningful and that their breaking causes a higher loss in information than by chance.

Finally, systematically adding edges that increase the bi-clique completeness led to an increased biological knowledge, while adding random edges showed an opposite trend with a loss of information similar to the random removal.

This demonstrates that our method, with respect to the entire network, improves the biological content. However, it must be noted that part of the predictions are false positives, and therefore external validation is necessary.

Three examples were extracted from the network by visual inspection for further in-depth validation. The first example was selected by randomly picking two apparently unrelated diseases (cardiovascular and parasitic), the second by choosing two drugs among the top promiscuous (see Table 1) and the last by searching for a drug which has known targets but no indications present in our network.

Validation

Since the prediction of new links in the network—the completion of incomplete bi-cliques—is solely based on a graph theoretic approach, additional evidence is needed to finally set the predicted link. This is crucial to account for the biological meaning of the links.

To validate predicted links from the incomplete bi-clique approach, the literature is mined for drug–target and drug–disease links. Additionally, binding site similarity is computed for drug–target links. If it is known that a drug binds to target T_1 and our approach predicts a new target T_2, then the drug binding sites in T_1 and T_2 may be similar. Binding site similarity can be checked with tools such as ProBiS[63] or SMAP.[64] Besides literature evidence and binding site similarity, many examples can be verified by the similarity of disease terms that are grouped in the prediction.

As an example for the latter, consider Fig. 5A and B. The figure shows drugs involved in cardiovascular and parasitic diseases. Drugs are among others annotated with the terms cardiovascular diseases and/or cardiovascular agents. These two terms stem from different data sources and hence the drugs are not consistently annotated with both. Instead, Fig. 5 shows that they share 9 drugs, but have 15 and 6 drugs on their own, respectively. The incomplete bi-clique approach suggests that the 15 terms should also be linked to cardiovascular agents, and the 6 to cardiovascular diseases. Thus, when aggregating the two terms into one general term "cardiovascular disease", this creates a single bi-clique linking 30 drugs (Fig. 5B).

From cardiovascular to anti-parasitic drugs

While the above example has a simple linguistic validation, we can apply the same principle to the relation of parasitic and cardiovascular diseases. As shown in Fig. 5, cardiovascular and antiparasitic disease and agents share two drugs—indomethacin and quercetin. Applying the incomplete bi-clique principle in Fig. 5B, we hypothesize that the 30 cardiovascular drugs can also be applied in parasitic diseases. Similarly, we hypothesize that the 9 antiparasitic drugs have an indication in cardiovascular diseases. As validation we conducted a literature search (details in Materials and methods) and for 23 drugs, we found evidence supporting the hypotheses (for details see the Table S1, ESI‡):

- Antiparasitic: Aspirin,[65] Benzylamine,[66] Cocaine,[67] Wortmannin,[68] Zoledronic acid.[69]
- African trypanosomiasis: Bortezomib,[70] Caffeine,[71] Dopamine.[72]
- Malaria: Acetazolamide,[73] Calcitriol,[74] Estradiol,[75] Ethacrynic acid,[76] Resveratrol,[77] Roscovitine,[78] Sildenafil (Viagra),[79] Trichostatin A.[80]
- Leishmaniasis: Estradiol,[81] Resveratrol.[82]
- Babesiosis: Roscovitine.[83]
- Fasciolosis: Genistein.[84]
- Chagas' disease (American trypanosomiasis): Risedronic acid.[85]
- Cardiovascular diseases: Monorden,[86] Pentamidine,[87,88] Pepstatin,[89] Sinefungin,[90] Suramin.[91]

18 of the cardiovascular drugs were found to be applicable in parasitic diseases and 5 antiparasitic drugs in cardiovascular diseases. Thus in total, 67% or 20 out of 30 cardiovascular drugs are known to be applicable in parasitic diseases. While 64% or 7 out of 11 antiparasitic drugs are related to cardiovascular diseases.

Fig. 5C shows that, overall, 25 compounds share cardiovascular and parasitic disease indication, while, respectively, 10 and 4 drugs are potential candidates for repositioning. These 14 drugs represent valid candidates for drug repositioning, since no relations to parasitic diseases or cardiovascular diseases, respectively, were found in the literature.

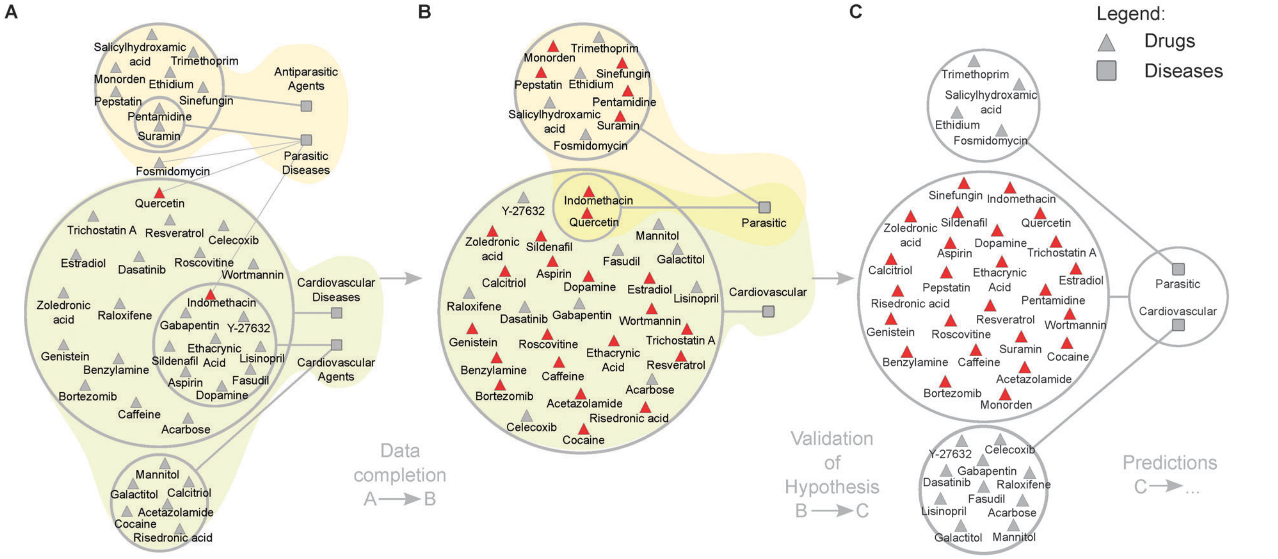


Fig. 5 From cardiovascular disease to antiparasitic effect and *vice versa*: (A) the incomplete bi-clique approach suggests a relation of cardiovascular disease, agents, parasitic disease, and antiparasitic agents. (B) In a first step, the respective disease and agent terms are merged. (C) In a second step, cardiovascular and parasitic diseases are grouped. Overall, a weak link of 2 cardiovascular and antiparasitic drugs in (B) could be extended by 25 validated links (red triangles in (C)). The drugs in the upper and lower power node represent predictions and thus candidates for repositioning.

Quercetin and resveratrol share a target

As a second example for the incomplete bi-clique approach, consider Fig. 6, which shows quercetin and resveratrol together with a selection of their common targets and diseases (also see the full network in Fig. S1, ESI‡). Both drugs are produced in plants and are among the top five promiscuous drugs in terms of known application. Resveratrol is a derivative of stilbene while quercetin is a flavonol with antioxidant properties.

Quercetin is also highly promiscuous in terms of its targets. In particular, it interacts with the phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit gamma isoform (PIK3GC) which regulates signaling cascades involved in cell growth, survival, proliferation, motility, and morphology. It is a candidate target in therapy for inflammatory diseases and pathologies such as asthma, rheumatoid arthritis, allergy, systemic lupus erythematosus, airway inflammation, lung injury, and pancreatitis.[92] Resveratrol interacts with the Leukotriene A-4 hydrolase (LTA4H) that catalyzes the final step in the biosynthesis of the proinflammatory mediator leukotriene B4.[93]

While differing in the targets, the two drugs both share links to neoplasms and cardiovascular diseases. Thus, the incomplete bi-clique approach suggests that both drugs interact with both targets as visualized in Fig. 6B.

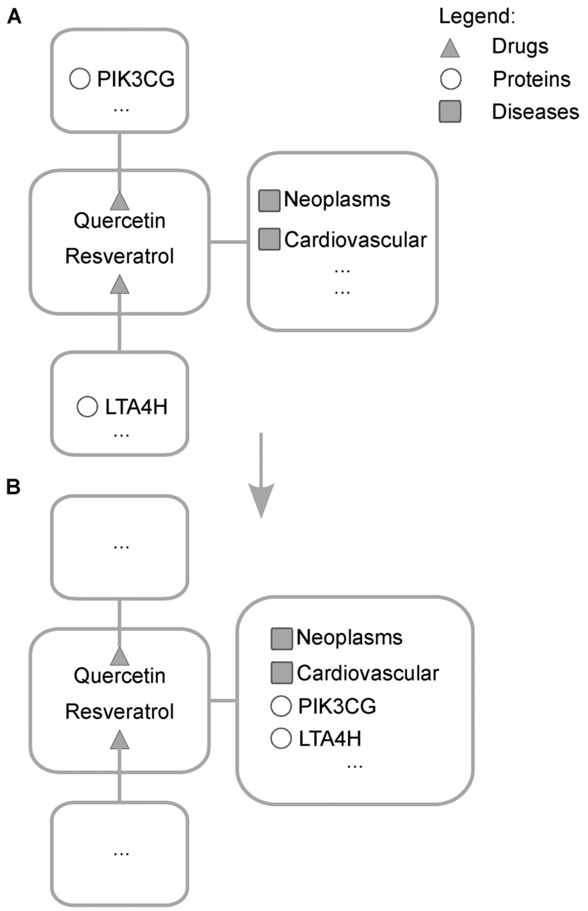


Fig. 6 Abstract representation of the network in Fig. S1 (ESI‡). (A) Quercetin and resveratrol with relevant links to two targets and two diseases. (B) The incomplete bi-clique approach suggests that quercetin and resveratrol bind both targets and implicitly that there is a link between the two targets and the two diseases.

We tested the two targets for binding site similarity, but none could be detected. However, recent studies showed that resveratrol, although not directly binding PIK3GC can inhibit its phosphorylation,[94] acts on the AKT/PIK3 pathway[95] and can induce apoptosis in different cancers.[96,97]

The binding of quercetin to LTA4H was confirmed *in vitro* by pull-down assay.[98] Both targets are involved in neoplasms and cardiovascular diseases, according to the TTD database. Moreover, although the Tanimoto coefficient between quercetin and resveratrol is not particularly high with 0.67, both drugs have two phenol subunits in a similar position.

Binding site similarity of four CID 1746 targets

As a third example, we consider the drug 4-iodine-benzo(*b*)-thiophene-2-carboxamidine (PubChem CID 1746), which has no therapeutic indications in our data. It binds three serine proteases, namely cationic trypsin, the urokinase-type plasminogen activator, and prothrombin. Some of its targets are promiscuously binding other drugs, therefore we tested if our approach of bi-cliques completion could suggest therapeutic indications or novel targets for this drug.

We extracted other drugs from our network that are sharing at least one target with CID 1746 (see Fig. 7 and for the full subnetwork Fig. S2, ESI‡). As shown, cationic trypsin is also the target of niacinamide, benzylamine, and pentamidine.

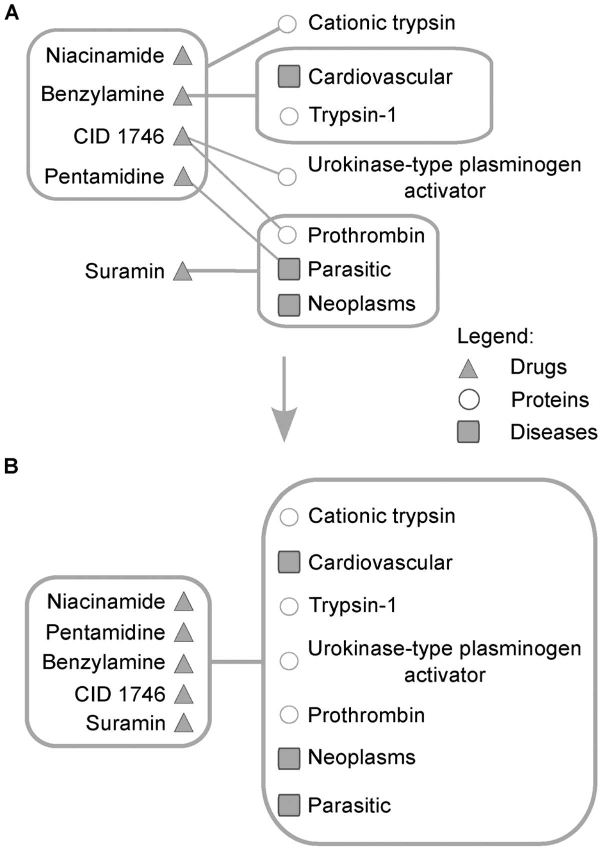


Fig. 7 Abstract representation of the network in Fig. S2 (ESI‡). (A) Five promiscuous drugs and their incomplete subnetwork. Several incomplete bi-cliques are present. (B) Our approach suggests the creation of a complete bi-clique, inferring novel targets and repositioning all the drugs to neoplasms, cardiovascular and parasitic diseases.

Thus, they form a power node together with CID 1746. The CID 1746 target prothrombin is shared with suramin. Benzylamine has trypsin-1 as target. CID 1746 has no disease indication, but benzylamine is linked to cardiovascular diseases, pentamidine is anti-parasitic and suramin is anti-parasitic, and linked to cancer.

As shown in Fig. 7B, the incomplete bi-clique approach suggests that the five drugs are all linked to the four targets and three diseases reducing the subnetwork to a single bi-clique. To be more precise, from 8 known drug–target links, 12 new ones are predicted, and from 4 known disease links, 11 new ones are predicted.

The grouping of the drugs is surprising as their chemical similarity is low (PubChem Tanimoto similarity score varies from 0.35 to 0.58). However, they share a benzylamine moiety. This moiety is always (except for suramin) placed in the same way in the binding sites (see Fig. 8). The four targets, cationic trypsin, urokinase-type plasminogen activator, prothrombin, and trypsin-1, on the other hand are all serine proteases with sequence similarity as low as 24% to 35%. Only cationic trypsin and trypsin-1 have a sequence identity of 63%. Nonetheless, the binding sites (especially the catalytic triad) are all highly conserved in sequence, as well as in structure, with significant *p*-values of 10^{-5} to 10^{-12} in the SMAP binding site comparison tool (see the Table S3, ESI‡). Fig. 8 shows the binding site regions in red and the conserved catalytic triad (Ser-His-Asp) in stick representation. Moreover, these proteins are highly similar in their overall structure (RMSD upon structure superposition between 0.45 and 0.9).

All drugs except suramin (Fig. 8D) can bind to these binding sites. This is because the binding sites are similar and the drugs share the benzylamine moiety. It is difficult to infer the other three serine proteases as a target for suramin, since its binding mode in prothrombin is at least partly due to crystallization effects.[99] But it was shown that suramin is capable of inhibiting the interaction between the urokinase-type plasminogen activator and its receptor.[100] Thus, the binding of suramin to trypsin is a valid candidate for testing.

To validate the 11 predicted drug–disease links, we checked TTD and searched the literature. TTD associates both urokinase-type plasminogen activator and prothrombin with neoplasms. Similarly, prothrombin is also related to cardiovascular diseases. Literature search confirmed 9 out of 11 drug–disease predictions (see the Table S2, ESI‡).

Two novel indications were predicted for CID 1746: cardiovascular and parasitic diseases. One is partially known, if the inferred indications in CTD are considered: the treatment of strokes (cardiovascular diseases) is suggested for CID 1746, with low score *via* the PLAU gene (urokinase-type plasminogen activator). No indication for CID 1746 in the treatment of parasitic diseases was found. The application of this drug in cardiovascular or parasitic diseases should now be validated *in vitro*. For details see the Table S2, ESI.‡

Conclusions

In this paper, we reported how to integrate prior knowledge, regarding drugs and their molecular targets as well as disease indications for known drugs, and how to use this in a systems approach to fill gaps in the data and gain new biological insights. We showed that this approach holds the potential to propose drugs for repositioning to novel applications.

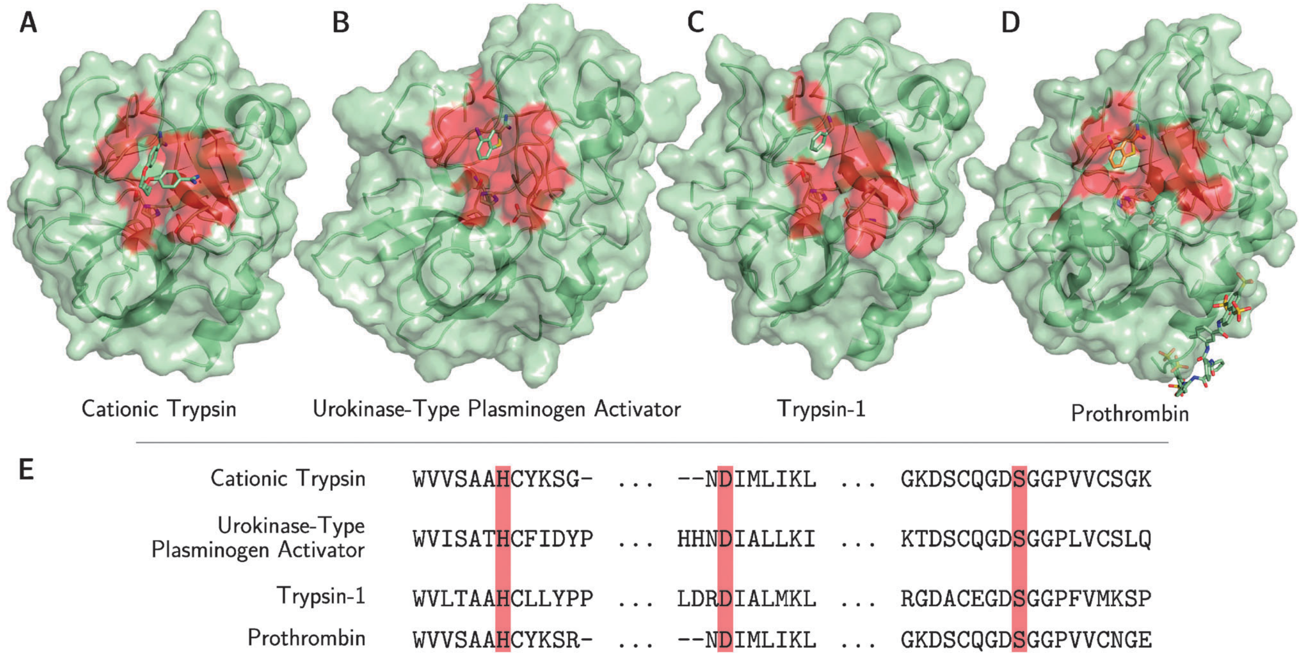

A
B
C
D
Cationic Trypsin
Urokinase-Type Plasminogen Activator
Trypsin-1
Prothrombin
E
Cationic Trypsin WVVSAAHCYKSG- ... --NDIMLIKL ... GKDSCQGDSGGPVVCSGK
Urokinase-Type Plasminogen Activator WVISATHCFIDYP ... HHNDIALLKI ... KTDSCQGDSGGPLVCSLQ
Trypsin-1 WVLTAAHCLLYPP ... LDRDIALMKL ... RGDACEGDSGGPFVMKSP
Prothrombin WVVSAAHCYKSR- ... --NDIMLIKL ... GKDSCQGDSGGPVVCNGE

Fig. 8 Binding sites (red) of the protein targets with their original ligand (green) and the catalytic triad shown in stick representation. (A) Cationic trypsin is shown with pentamidine (PDB ID 3GY3), (B) urokinase-type plasminogen activator with CID 1746 (1C5X), (C) trypsin-1 with benzylamine (1UTJ), and (D) prothrombin with suramin in green bound to the helix on bottom right (3BF6). Additionally, CID 1746 (from the prothrombin structure 1C5N) is shown in orange in the active site. (E) The sequence alignment shows the conserved binding site residues. The catalytic triad is highlighted in red (above in stick representation).

Information on drugs, targets and diseases is wide but sparse and therefore difficult to exploit. However, network based approaches allow a visual analysis of such knowledge. Power graphs can overcome the limitation of visualization due to network complexity, by node clustering and edge reduction.

Finally, we have shown that power graph analysis can facilitate the process of inferring new relationships between drugs, proteins, and diseases.

We proposed a method to complete missing information in drug–target–disease networks: through identification and completion of incomplete bi-cliques.

A global evaluation of the method was presented, showing that the biological significance—as binding site similarity among proteins—could be improved by increasing the completeness of identified bi-cliques by power graphs.

As we presented in three case studies, this method suggests novel targets for known drugs or finds new possible therapeutic applications for them.

The major limitation of this approach is its direct demand for external validation. If a new connection between two entities is inferred, to complete an incomplete bi-clique, then this is not necessarily a true positive. However, the confidence can be increased by using additional data, *e.g.*, literature or bio-assay data, to validate inferred connections.

Materials and methods

Database of drugs, proteins, diseases, pharmacological actions, and their relations

Difficulties rise from the inhomogeneity and incompleteness of the data as well as from the differences in levels of detail. With this as a starting point, the inference of new relations between drugs, targets and diseases is a challenging endeavor.

Two fundamental problems have to be solved: standardization and normalization of the data and creation of a structure within the data. The structure should only be determined by the relations between the entities.

Data from various resources were combined. The first problem of standardization was solved by mapping each type of information to its most representative resource: Proteins to UniProt,[101] drugs to PubChem,[102] and textual terms (*i.e.*, diseases) to the controlled vocabulary MeSH (Fig. 1A).

As a source for protein structural data, the PDB[103] as of 2010/10/29 was used. This dataset comprises 66 820 protein structures, containing 26 588 unique protein chains (clustered by 95% sequence identity as provided by the PDB). Unless explicitly stated otherwise, we will always refer to the representative of such a cluster when speaking about a protein. All cluster representatives were mapped to UniProt,[101] using its mapping service. Co-crystallized ligands were mapped to PubChem, using the SMILES of the compounds (as provided by the PDB) and the PubChem[102] PUG (PUG, http://pubchem.ncbi.nlm.nih.gov/pug/pughelp.html).

To get information about drugs, TTD[17] (version 4.3.01 as of 2011/07/01) was included. Subsequently, drugs were mapped to PubChem CIDs. This was done using the mapping of TTD and the PubChem PUG with InChIs[104] or SMILES. This only retains TTD drugs with a chemical structure present. The conversion of small molecule structures to InChI and SMILES was done using OpenBabel (version 2.2.3, http://openbabel.sourceforge.net)[105] and Pybel.[106] If an identical structure could not be found in PubChem during the mapping, a similarity search *via* the PUG was performed (chemical similarity (Tanimoto score) $T \geq 0.9$). This was implemented as a binary search starting from $T = 1.0$. The Tanimoto score was decreased ($T \in \{0.90, 0.91, \ldots, 1.0\}$) if no similar compound was found in PubChem. If such a search results in more than one compound, the smallest CID is used as a representative.

From the ligands, all small compounds (five non-hydrogen atoms or less), common cofactors, detergents and solutes as well as the compounds listed in the article by Gold and Jackson[107] were blacklisted and removed to reduce noise in the data set. The filtering of small compounds was performed with OpenBabel by counting the non-hydrogen atoms. The reduction of the compound dataset led from a total of 3087 TTD drugs to 1322 approved, 516 clinical trial, and 1069 experimental drugs (2907 in total). Those drugs being co-crystallized with three or more different proteins in the PDB (promiscuous drugs) are retained. In total, 147 of such drugs are present (see Table 1 for examples).

Drug–protein target associations were extracted from PDB structures using mmLib.[108]

Chemical–disease relationships were retrieved from CTD (as of 2011/07/05)[18,19] and TTD.[17] The CTD chemicals were mapped to PubChem CIDs using the provided MeSH IDs and the mapping from MeSH terms to CIDs as provided by PubChem. The CTD diseases could be directly mapped to MeSH. Whereas the diseases stored in TTD had to be mapped to MeSH using synonyms and word combinations. Subsequently, this mapping was manually checked for correctness. Pharmacological actions (as MeSH terms) of drugs were retrieved from PubChem (as of 2011/07/04) and MeSH (as of 2011/01/25).

The second issue of different levels of detail was addressed—to some extent—by exploiting the hierarchical structure of MeSH. MeSH (http://www.nlm.nih.gov/mesh) is organized hierarchically as a tree, such that general terms (*e.g.*, "anti-infective agents") are found near the root and the most precise terms are found in the leaves (*e.g.*, "HIV protease inhibitors"). This organization made possible a mapping from the low-level (specific) terms to more general high-level terms (*e.g.*, "anti-HIV agents" is mapped to "antiviral agents"). Thus, chemicals can now be grouped according to their high-level pharmacological action or disease term (see Table 3).

The above raw data are available upon request.

Network analysis using power graphs

A drug–target–disease network of the 147 promiscuous drugs, 553 protein targets, 27 diseases, and 17 pharmacological actions was created. By clustering nodes sharing edges, the power graph[62] of this network is created. See the transition from Fig. 1 and 2 and Table 2. Nodes are clustered to so-called power nodes, which are connected by power edges. Thus, the classical edges can be omitted without losing information. This brings structure to the data and reveals relations, which were not apparent before.

Table 3 The high level MeSH pharmacological action and disease terms used in this study, according to the MeSH tree. All of their child terms were mapped to their corresponding parent (high-level) term

MeSH term (high level)	MeSH tree number
Anti-allergic agents	D27.505.954.016
Anti-bacterial agents	D27.505.954.122.085
Antifungal agents	D27.505.954.122.136
Anti-infective agents, local	D27.505.954.122.187
Anti-infective agents, urinary	D27.505.954.122.237
Anti-inflammatory agents	D27.505.954.158
Antineoplastic agents	D27.505.954.248
Antiparasitic agents	D27.505.954.122.250
Antirheumatic agents	D27.505.954.329
Antiviral agents	D27.505.954.122.388
Cardiovascular agents	D27.505.954.411
Central nervous system sgents	D27.505.954.427
Dermatologic agents	D27.505.954.444
Disinfectants	D27.505.954.122.425
Gastrointestinal agents	D27.505.954.483
Hematologic agents	D27.505.954.502
Lipid regulating agents	D27.505.954.557
Renal agents	D27.505.954.613
Reproductive control agents	D27.505.954.705
Respiratory system agents	D27.505.954.796
Stimulants, historical	D27.505.954.888
Animal diseases	C22
Bacterial infections and mycoses	C01
Cardiovascular diseases	C14
Congenital, hereditary, and neonatal diseases and abnormalities	C16
Digestive system diseases	C06
Disorders of environmental origin	C21
Endocrine system diseases	C19
Eye diseases	C11
Female urogenital diseases and pregnancy complications	C13
Hemic and lymphatic diseases	C15
Immune system diseases	C20
Male urogenital diseases	C12
Musculoskeletal diseases	C05
Neoplasms	C04
Nervous system diseases	C10
Nutritional and metabolic diseases	C18
Occupational diseases	C24
Otorhinolaryngologic diseases	C09
Parasitic diseases	C03
Pathological conditions, signs and symptoms	C23
Respiratory tract diseases	C08
Skin and connective tissue diseases	C17
Stomatognathic diseases	C07
Substance-related disorders	C25
Virus diseases	C02
Wounds and injuries	C26

Our algorithm for computing power graph representations from graphs supports weighted graphs and a minimum similarity threshold. The algorithm consists of a first phase that collects candidate power nodes (clusters of nodes) and a second phase that uses these to search for power edges (edges between power nodes).

In the first phase, candidate power nodes are identified with hierarchical clustering based on neighborhood similarity. A candidate power node is a set of nodes that have neighbors in common. The three basic motifs recognized by power graphs are the star, the clique and the biclique, and constitute the basic abstractions when transforming the original graph into a power graph.

In the second phase, power edges are searched. The minimal power graph problem has to be seen as an optimization problem to find the power graph achieving the highest edge reduction. Among the candidate power nodes found in phase one, each pair that forms a complete bipartite subgraph (bi-clique) is a candidate power edge. Power graphs offer up to 90% compression of the original network structure, allowing for efficient visualization.[62]

Visual inspection of the network (see Table 2 for the network characteristics) is performed with Cytoscape (http://www.cytoscape.org)[49] and the power graph plug in. The power graph software can be found at http://www.biotec.tu-dresden.de/research/schroeder/powergraphs.

Validation of predictions and used software

Literature search was performed with GoPubMed.org.[109] *E.g.*, to validate one of the hypotheses from Fig. 5 "Roscovitine is an antiparasitic agent.", the following query on GoPubMed.org was used: roscovitine AND "Antiparasitic Agents" [mesh]. This returns an article by Harmse *et al.*,[78] showing the antimalarial activity of roscovitine.

To validate new targets, the binding site alignment tool SMAP (version 2.0, http://funsite.sdsc.edu)[40,110–112] in the version of the downloadable binary package was employed. An alignment was considered to be significant for a p-value $\leq 10^{-3}$. The alignment was restricted to the ligand binding sites of the current ligand and the default parameters of SMAP were used.

Protein structures were visualized and superposed with PyMol (http://www.pymol.org).

All computations performed for this work—if not stated otherwise—were scripted in Python using among others BioPython (http://biopython.org) on Linux 2.6.

Acronyms

CTD	Comparative Toxicogenomics Database
InChI	International Chemical Identifier
MeSH	Medical Subject Headings
PDB	Protein Data Bank
PUG	Power User Gateway
SMILES	Simplified Molecular Input Line Entry Specification
TTD	Therapeutic Targets Database

Acknowledgements

We thank Julia Thomas for checking the mapping of TTD to MeSH and PubChem. Maria Kissa is kindly acknowledged for the fruitful discussions and for suggesting the paper of Oi *et al.* Funding by GeneCloud and Ponte is kindly acknowledged.

References

1 M. D. Rawlins, *Nat. Rev. Drug Discovery*, 2004, **3**, 360–364.
2 S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg and A. L. Schacht, *Nat. Rev. Drug Discovery*, 2010, **9**, 203–214.
3 B. Munos, *Nat. Rev. Drug Discovery*, 2009, **8**, 959–968.
4 K. A. O'Connor and B. L. Roth, *Nat. Rev. Drug Discovery*, 2005, **4**, 1005–1014.
5 C. R. Chong and D. J. Sullivan, *Nature*, 2007, **448**, 645–646.

6 T. T. Ashburn and K. B. Thor, *Nat. Rev. Drug Discovery*, 2004, **3**, 673–683.

7 J. T. Dudley, T. Deshpande and A. J. Butte, *Briefings Bioinf.*, 2011, **12**, 303–311.

8 S. Günther, M. Kuhn, M. Dunkel, M. Campillos, C. Senger, E. Petsalaki, J. Ahmed, E. G. Urdiales, A. Gewiess, L. J. Jensen, R. Schneider, R. Skoblo, R. B. Russell, P. E. Bourne, P. Bork and R. Preissner, *Nucleic Acids Res.*, 2008, **36**, D919–D922.

9 J. von Eichborn, M. S. Murgueitio, M. Dunkel, S. Koerner, P. E. Bourne and R. Preissner, *Nucleic Acids Res.*, 2011, **39**, D1060–D1066.

10 H. M. Berman, T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J. D. Westbrook and C. Zardecki, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2002, **58**, 899–907.

11 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.

12 C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. C. Guo and D. S. Wishart, *Nucleic Acids Res.*, 2011, **39**, D1035–D1041.

13 Q. Li, T. Cheng, Y. Wang and S. H. Bryant, *Drug Discovery Today*, 2010, **15**, 1052–1057.

14 E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, S. Federhen, M. Feolo, I. M. Fingerman, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, A. Marchler-Bauer, V. Miller, I. Mizrachi, J. Ostell, A. Panchenko, L. Phan, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. J. Wilbur, E. Yaschenko and J. Ye, *Nucleic Acids Res.*, 2011, **39**, D38–D51.

15 M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen and P. Bork, *Mol. Syst. Biol.*, 2010, **6**, 343.

16 A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani and J. P. Overington, *Nucleic Acids Res.*, 2012, **40**, D1100–D1107.

17 F. Zhu, B. Han, P. Kumar, X. Liu, X. Ma, X. Wei, L. Huang, Y. Guo, L. Han, C. Zheng and Y. Chen, *Nucleic Acids Res.*, 2010, **38**, D787–D791.

18 A. P. Davis, B. L. King, S. Mockus, C. G. Murphy, C. Saraceni-Richards, M. Rosenstein, T. Wiegers and C. J. Mattingly, *Nucleic Acids Res.*, 2011, **39**, D1067–D1072.

19 A. P. Davis, C. G. Murphy, C. A. Saraceni-Richards, M. C. Rosenstein, T. C. Wiegers and C. J. Mattingly, *Nucleic Acids Res.*, 2009, **37**, D786–D792.

20 K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal and A.-L. Barabsi, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 8685–8690.

21 I. Feldman, A. Rzhetsky and D. Vitkup, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 4323–4328.

22 S. Navlakha and C. Kingsford, *Bioinformatics*, 2010, **26**, 1057–1063.

23 K. Lage, E. O. Karlberg, Z. M. Strling, P. I. Olason, A. G. Pedersen, O. Rigina, A. M. Hinsby, Z. Tmer, F. Pociot, N. Tommerup, Y. Moreau and S. Brunak, *Nat. Biotechnol.*, 2007, **25**, 309–316.

24 T.-P. Nguyen and F. Jordan, *BMC Syst. Biol.*, 2010, **4**, 103.

25 S. Zhao and S. Li, *PLoS One*, 2010, **5**, e11764.

26 Y. Yamanishi, M. Kotera, M. Kanehisa and S. Goto, *Bioinformatics*, 2010, **26**, i246–i254.

27 F. Hormozdiari, R. Salari, V. Bafna and S. C. Sahinalp, *J. Comput. Biol.*, 2010, **17**, 669–684.

28 M. Campillos, M. Kuhn, A.-C. Gavin, L. J. Jensen and P. Bork, *Science*, 2008, **321**, 263–266.

29 E. Kotelnikova, A. Yuryev, I. Mazo and N. Daraselia, *J. Bioinf. Comput. Biol.*, 2010, **8**, 593–606.

30 F. Iorio, R. Bosotti, E. Scacheri, V. Belcastro, P. Mithbaokar, R. Ferriero, L. Murino, R. Tagliaferri, N. Brunetti-Pierri, A. Isacchi and D. di Bernardo, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 14621–14626.

31 S. Prot, O. Sperandio, M. A. Miteva, A.-C. Camproux and B. O. Villoutreix, *Drug Discovery Today*, 2010, **15**, 656–667.

32 V. J. Haupt and M. Schroeder, *Briefings Bioinf.*, 2011, **12**, 312–326.

33 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. H. Thomas, D. D. Edwards, B. K. Shoichet and B. L. Roth, *Nature*, 2009, **462**, 175–181.

34 S. L. Kinnings and R. M. Jackson, *J. Chem. Inf. Model.*, 2011, **51**, 624–634.

35 Y. C. Martin, J. L. Kofron and L. M. Traphagen, *J. Med. Chem.*, 2002, **45**, 4350–4358.

36 J. D. Ferreira and F. M. Couto, *PLoS Comput. Biol.*, 2010, **6**, e1000937.

37 A. Weber, A. Casini, A. Heine, D. Kuhn, C. T. Supuran, A. Scozzafava and G. Klebe, *J. Med. Chem.*, 2004, **47**, 550–557.

38 R. Minai, Y. Matsuo, H. Onuki and H. Hirota, *Proteins: Struct., Funct., Bioinf.*, 2008, **72**, 367–381.

39 S. L. Kinnings, N. Liu, N. Buchmeier, P. J. Tonge, L. Xie and P. E. Bourne, *PLoS Comput. Biol.*, 2009, **5**, e1000423.

40 J. D. Durrant, R. E. Amaro, L. Xie, M. D. Urbaniak, M. A. J. Ferguson, A. Haapalainen, Z. Chen, A. M. Di Guilmi, F. Wunder, P. E. Bourne and J. A. McCammon, *PLoS Comput. Biol.*, 2010, **6**, e1000648.

41 J.-C. Heinrich, A. Tuukkanen, M. Schroeder, T. Fahrig and R. Fahrig, *J. Cancer Res. Clin. Oncol.*, 2011, **137**, 1349–1361.

42 C. Plake and M. Schroeder, *Curr. Pharm. Biotechnol.*, 2011, **12**, 449–457.

43 A. P. Chiang and A. J. Butte, *Clin. Pharmacol. Ther.*, 2009, **86**, 507–510.

44 S. J. Campbell, A. Gaulton, J. Marshall, D. Bichko, S. Martin, C. Brouwer and L. Harland, *Drug Discovery Today*, 2010, **15**, 3–15.

45 A. L. Hopkins, *Nat. Chem. Biol.*, 2008, **4**, 682–690.

46 A. Pujol, R. Mosca, J. Farrs and P. Aloy, *Trends Pharmacol. Sci.*, 2010, **31**, 115–123.

47 S. J. Cockell, J. Weile, P. Lord, C. Wipat, D. Andriychenko, M. Pocock, D. Wilkinson, M. Young and A. Wipat, *J. Integr. Bioinform.*, 2010, **7**, 116.

48 J. von Eichborn, M. S. Murgueitio, M. Dunkel, S. Koerner, P. E. Bourne and R. Preissner, *Nucleic Acids Res.*, 2011, **39**, D1060–D1066.

49 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.

50 H. Nahrevanian, R. Hajihosseini, M. Arjmand, M. Farahmand and F. Ghasemi, *Southeast Asian J. Trop. Med. Public Health*, 2009, **40**, 1188–1198.

51 H. Axelsson, C. Lnnroth, M. Andersson and K. Lundholm, *Int. J. Oncol.*, 2010, **37**, 1143–1152.

52 P. Andreone, A. Gramenzi, E. Loggi, L. Favarelli, C. Cursaro, M. Margotti, M. Biselli, S. Lorenzini and M. Bernardi, *Cytokine*, 2004, **26**, 95–101.

53 E. G. McGeer and P. L. McGeer, *Expert Rev. Neurother.*, 2001, **1**, 53–60.

54 S. Saponara, E. Carosati, P. Mugnai, G. Sgaragli and F. Fusi, *Br. J. Pharmacol.*, 2011, **164**, 1684–1697.

55 M. F. Montenegro, E. M. Neto-Neves, C. A. Dias-Junior, C. S. Ceron, M. M. Castro, V. A. Gomes, A. Kanashiro and J. E. Tanus-Santos, *Naunyn-Schmiedebergs Arch. Pharmacol.*, 2010, **382**, 293–301.

56 V. R. Punithavathi and P. S. M. Prince, *J. Biochem. Mol. Toxicol.*, 2011, **25**, 28–40.

57 A. Singhal, H. Jain, V. Singhal, E. J. Elias and A. Showkat, *Pharmacognosy Research*, 2011, **3**, 35–39.

58 C. Chi, Y. Chang, Y. Ou, C. Hsieh, W. Lui, F. Peng and T. Liu, *Oncol. Rep.*, 1997, **4**, 1021–1024.

59 S. Borska, M. Drag-Zalesinska, T. Wysocka, M. Sopel, M. Dumanska, M. Zabel and P. Dziegiel, *Folia Histochem. Cytobiol.*, 2010, **48**, 222–229.

60 L. Zhang, Y.-X. Cheng, A.-L. Liu, H.-D. Wang, Y.-L. Wang and G.-H. Du, *Molecules*, 2010, **15**, 8507–8517.

61 Y. Kim, S. Narayanan and K.-O. Chang, *Antiviral Res.*, 2010, **88**, 227–235.

62 L. Royer, M. Reimann, B. Andreopoulos and M. Schroeder, *PLoS Comput. Biol.*, 2008, **4**, e1000108.

63 J. Konc and D. Janežič, *Nucleic Acids Res.*, 2010, **38**, W436–W440.

64 J. Ren, L. Xie, W. W. Li and P. E. Bourne, *Nucleic Acids Res.*, 2010, **38**, W441–W444.

65 R. Lpez-Muoz, M. Fandez, S. Klein, S. Escanilla, G. Torres, D. Lee-Liu, J. Ferreira, U. Kemmerling, M. Orellana, A. Morello, A. Ferreira and J. D. Maya, *Exp. Parasitol.*, 2010, **124**, 167–171.

66 V. V. Kouznetsov, J. R. Castro, C. O. Puentes, E. E. Stashenko, J. R. Martnez, C. Ochoa, D. M. Pereira, J. J. N. Ruiz, C. F. Portillo, S. M. Serrano, A. G. Barrio, A. Bahsas and J. Amaro-Luis, *Arch. Pharm.*, 2005, **338**, 32–37.

67 R. F. Squires and E. Saederup, *Vopr. Med. Khim.*, 1997, **43**, 576–583.

68 M. V. Braga and W. de Souza, *FEMS Microbiol. Lett.*, 2006, **256**, 209–216.

69 C. L. Byington, R. Dunbrack, Jr., F. G. Whitby, F. E. Cohen and N. Agabian, *Exp. Parasitol.*, 1997, **87**, 194–202.

70 D. Steverding and X. Wang, *Parasites & Vectors*, 2009, **2**, 29.

71 C. H. Steeves and S. L. Bearne, *Bioorg. Med. Chem. Lett.*, 2011, **21**, 5188–5190.

72 O. A. Owolabi, C. Wilson, D. H. Molyneux and V. W. Pentreath, *Ann. Trop. Med. Parasitol.*, 1990, **84**, 127–131.

73 J. Krungkrai, A. Scozzafava, S. Reungprapavut, S. R. Krungkrai, R. Rattanajak, S. Kamchonwongpaisan and C. T. Supuran, *Bioorg. Med. Chem.*, 2005, **13**, 483–489.

74 H. J. Vial, M. J. Thuet and J. R. Philippot, *Mol. Biochem. Parasitol.*, 1982, **5**, 189–198.

75 P. W. Klein, J. D. Easterbrook, E. N. Lalime and S. L. Klein, *Gender Medicine*, 2008, **5**, 423–433.

76 M.-A. Dude, U. Kaeppler, M. Herb, M. Schiller, F. Schulz, B. Vedder, S. Heppner, G. Pradel, J. Gut, P. J. Rosenthal, T. Schirmeister, M. Leippe and C. Gelhaus, *Molecules*, 2009, **14**, 19–35.

77 H. I. Moon and J. Sim, *Ann. Trop. Med. Parasitol.*, 2008, **102**, 447–450.

78 L. Harmse, R. van Zyl, N. Gray, P. Schultz, S. Leclerc, L. Meijer, C. Doerig and I. Havlik, *Biochem. Pharmacol.*, 2001, **62**, 341–348.

79 B. L. Howard, P. E. Thompson and D. T. Manallack, *J. Comput.-Aided Mol. Des.*, 2011, **25**, 753–762.

80 K. T. Andrews, A. Walduck, M. J. Kelso, D. P. Fairlie, A. Saul and P. G. Parsons, *Int. J. Parasitol.*, 2000, **30**, 761–768.

81 H. Snider, C. Lezama-Davila, J. Alexander and A. R. Satoskar, *NeuroImmunoModulation*, 2009, **16**, 106–113.

82 L. Kedzierski, J. M. Curtis, M. Kaminska, J. Jodynis-Liebert and M. Murias, *Parasitol. Res.*, 2007, **102**, 91–97.

83 K. Nakamura, N. Yokoyama and I. Igarashi, *Parasitology*, 2007, **134**, 1347–1353.

84 E. Toner, G. P. Brennan, K. Wells, J. G. McGeown and I. Fairweather, *Parasitology*, 2008, **135**, 1189–1203.

85 L. R. Garzoni, M. C. Waghabi, M. M. Baptista, S. L. de Castro, M. d. N. L. Meirelles, C. C. Britto, R. Docampo, E. Oldfield and J. A. Urbina, *Int. J. Antimicrob. Agents*, 2004, **23**, 286–290.

86 T. M. Griffin, T. V. Valdez and R. Mestril, *Am. J. Physiol.: Heart Circ. Physiol.*, 2004, **287**, H1081–H1088.

87 J. P. Fauchier, L. Fauchier, D. Babuty, J. C. Breuillac, P. Cosnay and P. Rouesnel, *Arch. Mal. Coeur Vaiss.*, 1993, **86**, 757–767.

88 J. Simk, A. Csilek, J. Karszi and I. Lorincz, *Infection*, 2008, **36**, 194–206.

89 R. Sodi, S. M. Darn, A. S. Davison, A. Stott and A. Shenkin, *Ann. Clin. Biochem.*, 2006, **43**, 49–56.

90 N. Sirinupong, J. Brunzelle, J. Ye, A. Pirzada, L. Nico and Z. Yang, *J. Biol. Chem.*, 2010, **285**, 40635–40644.

91 J. J. Galligan, M. C. Hess, S. B. Miller and G. D. Fink, *J. Pharmacol. Exp. Ther.*, 2001, **296**, 478–485.

92 L. Barberis and E. Hirsch, *Thromb. Haemostasis*, 2008, **99**, 279–285.

93 B. Odlander, H. E. Claesson, T. Bergman, O. Rdmark, H. Jrnvall and J. Z. Haeggstrm, *Arch. Biochem. Biophys.*, 1991, **287**, 167–174.

94 Q. Chen, S. Ganapathy, K. P. Singh, S. Shankar and R. K. Srivastava, *PLoS One*, 2010, **5**, e15288.

95 J. Tian, J. Gao, J. Chen, F. Li, X. Xie, J. Du, J. Wang and N. Mao, *Zhongguo Zhongyao Zazhi*, 2010, **35**, 1878–1882.

96 Y. Wang, T. Romigh, X. He, M. S. Orloff, R. H. Silverman, W. D. Heston and C. Eng, *Hum. Mol. Genet.*, 2010, **19**, 4319–4329.

97 X. He, Y. Wang, J. Zhu, M. Orloff and C. Eng, *Cancer Lett.*, 2011, **301**, 168–176.

98 N. Oi, C.-H. Jeong, J. Nadas, Y.-Y. Cho, A. Pugliese, A. M. Bode and Z. Dong, *Cancer Res.*, 2010, **70**, 9755–9764.

99 L. M. T. R. Lima, C. F. Becker, G. M. Giesel, A. F. Marques, M. T. Cargnelutti, M. de Oliveira Neto, R. Q. Monteiro, H. Verli and I. Polikarpov, *Biochim. Biophys. Acta, Proteins Proteomics*, 2009, **1794**, 873–881.

100 N. Behrendt, E. Rønne and K. Danø, *J. Biol. Chem.*, 1993, **268**, 5985–5989.

101 M. Magrane, Uniprot Consortium, Database (Oxford), 2011, 2011, bar009.

102 E. E. Bolton, Y. Wang, P. A. Thiessen and S. H. Bryant, in *Annual Reports in Computational Chemistry*, ed. R. A. Wheeler and D. C. Spellmeyer, Elsevier, 2008, vol. 4, ch. 12, pp. 217–241.

103 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Res.*, 2000, **28**, 235–242.

104 S. Heller and A. McNaught, *Chem. Int.*, 2009, **31**, 7–9.

105 R. Guha, M. T. Howard, G. R. Hutchison, P. Murray-Rust, H. Rzepa, C. Steinbeck, J. Wegner and E. L. Willighagen, *J. Chem. Inf. Model.*, 2006, **46**, 991–998.

106 N. M. O'Boyle, C. Morley and G. R. Hutchison, *Chem. Cent. J.*, 2008, **2**, 5.

107 N. D. Gold and R. M. Jackson, *J. Mol. Biol.*, 2006, **355**, 1112–1124.

108 J. Painter and E. A. Merritt, *J. Appl. Crystallogr.*, 2004, **37**, 174–178.

109 A. Doms and M. Schroeder, *Nucleic Acids Res.*, 2005, **33**, W783–W786.

110 L. Xie and P. E. Bourne, *BMC Bioinformatics*, 2007, **8**(Suppl 4), S9.

111 L. Xie and P. E. Bourne, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 5441–5446.

112 L. Xie, L. Xie and P. E. Bourne, *Bioinformatics*, 2009, **25**, i305–i312.