Insights into polypharmacology from drug-domain associations

Aurelio A. Moya-García* and Juan A. G. Ranea

Department of Molecular Biology and Biochemistry, University of Malaga, 29071 Málaga, Spain

Associate Editor: Olga Troyanskaya

ABSTRACT

Motivation: Polypharmacology (the ability of a single drug to affect multiple targets) is a key feature that may explain part of the decreasing success of conventional drug discovery strategies driven by the quest for drugs to act selectively on a single target. Most drug targets are proteins that are composed of domains (their structural and functional building blocks).

Results: In this work, we model drug–domain networks to explore the role of protein domains as drug targets and to explain drug polypharmacology in terms of the interactions between drugs and protein domains. We find that drugs are organized around a privileged set of druggable domains.

Conclusions: Protein domains are a good proxy for drug targets, and drug polypharmacology emerges as a consequence of the multi-domain composition of proteins.

Contact: amoyag@uma.es

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The decrease in the translation of drug candidates into effective therapies during the past two decades has occurred concurrently with the goal of rational drug design of developing selective ligands that act on a single target (Hopkins, 2008). This 'classical' perspective is an oversimplification challenged by a growing body of evidence showing that there are many drugs for each target and that a single drug can affect multiple targets (Goh et al., 2007; Mestres et al., 2009; Yildirim et al., 2007).

The emergent concept of polypharmacology bears this out, both from a target and a drug perspective. On one hand, we can obtain the correct combination of targets from analyses of the biological networks associated with a given disease, and on the other hand, we can select or design 'magic shotguns', i.e. drugs that are able to bind multiple targets with low specificity (Zanzoni et al., 2009). The network perspective from the study by Yildirim et al. (2007) illustrates this situation well. They organized all approved drugs reported by DrugBank into a drug–target network in which the drugs are depicted as nodes that are connected if the drugs share a protein target, and they also generated a target–protein network in which the proteins are nodes that are connected if the proteins are targeted by the same drug. More than half of the drugs in the drug–target network formed a giant interconnected cluster, and in both networks, the majority of the nodes were connected to at least one other drug/target, that is drugs share protein targets and proteins are targeted by more than one drug. Although Yildirim et al. bring up polypharmacology as a general attribute of drugs, their results also illustrate an important bias of drug design, which is that many approved drugs are based on the same therapeutic targets. Mestres et al. pursue this topic further and show that because of limited time and resources, small molecules are not screened systematically through a large and complete panel of proteins. Thus, drug–target networks derived on the basis of public bibliographic sources may be more representative of the target space explored by the pharmaceutical industry rather than being a true reflection of drug polypharmacology (Mestres et al., 2008). When they take into account drug affinity data by including targets with no particular therapeutic interest, they find that drugs target even more proteins, illustrating that drug polypharmacology is the rule rather than the exception.

Polypharmacology has recently been explained in terms of protein modularity within protein interaction networks. Within any protein interaction network, a module represents a densely connected group of nodes (and these may be either proteins or functional motifs/domains) that are weakly connected to the remaining network. A rich variety of global measures based on reverse engineering of network topologies have been suggested to uncover the organizing principles behind complex networks. Such reverse engineering of network topologies-based measures reveal that networked structures can emerge at different levels, from single-node characteristics and the tendency of pairs of nodes to connect to each other, to the patterns exhibited by associations of three or more nodes known as motifs (Nacher and Schwartz, 2012). Here, we aim to re-orientate the field to the fundamental principle of protein science, the premise that the functional units of proteins are domains, to explain polypharmacology.

Protein domains are compact and functional structures that can be considered the building blocks of proteins. The definition of protein domains overlaps the structural and functional views of proteins (Koonin et al., 2002). Because domains are units of structure (Orengo et al., 1997) and there is a limited repertoire of domain types (Wolf et al., 2000), protein domains are combined to form different proteins with different functions (Kummerfeld and Teichmann, 2009). Furthermore, protein–protein interactions are dominated by domain–domain interactions (Pang et al., 2012). Therefore, we think that drug–target interactions may be mediated by drug–domain interactions; hence, we argue that protein domains can be major contributors to polypharmacology. In other words, because proteins have a modular structure and the same domains can be repeatedly found in different proteins, the reason why a drug binds multiple protein targets

---

*To whom correspondence should be addressed.

may be related to the proteins sharing a common domain that the drug may be targeting. In this work, we attempt to unravel the association between drugs and protein domains to address the hypothesis that protein domains are the areas on which drugs act and are an important factor in the polypharmacological behaviour of most drugs.

2 METHODS

2.1 Data retrieval

We compiled a drug–target interaction dataset for human proteins using interaction data from the databases DrugBank (Knox *et al.*, 2010), STITCH (Kuhn *et al.*, 2008) and MATADOR (Günther *et al.*, 2008). We filtered the drug–target interactions from DrugBank to exclude nutraceutical compounds and to include only approved and experimental drugs labelled with a PubChem Compound identifier. High-confidence drug–target interactions from STITCH and MATADOR were added to our data to obtain a comprehensive non-redundant set of drug–target interactions for human proteins. The CATH and PFAM domain compositions for each target protein were obtained from Gene3D (release 10.2) and BioMart (Guberman *et al.*, 2011), respectively.

2.2 Scoring and evaluation of drug–domain associations

Each drug that interacts with a protein in our initial dataset is associated with all the domains that comprise the protein. We formulated a scoring scheme based on the information content of a drug associated with a domain, i.e. the self-information of each drug–domain pair from all the possible drug–domain pairs. The self-information for a drug–domain pair, taken from the mathematical theory of information, is defined as

$$I(dD) = -\log_2 P(d, D); \; P(d, D) = p(d) \cdot p(D) = \left(\frac{nd}{\sum d}\right) \cdot \left(\frac{nD}{\sum D}\right)$$

where $P(d,D)$ is the probability of finding a drug *(d)* associated to a domain *(D)*, and *nd* and *nD* are the number of times that drugs and domains, respectively, appear in the dataset. We calculate the score (s) for each drug–domain pair in our network as

$$s(dD) = N(dD) \cdot I(dD)$$

where *N(dD)* is the number of times the pair *dD* is in the network.

Random drug–domain networks were computed by shuffling the initial drug–target interactions and keeping the same CATH and PFAM domain composition for each target. For each drug–domain pair in our network, we obtained a distribution of scores computed from the random networks in which the particular drug–domain pair was found. Thus, for each drug–domain pair, we have a score and a distribution of 25 random scores. We computed the probability that each score was included in the distribution of random scores using a statistical *t*-test. Using a *P*-value threshold of 0.001, we could assess whether a drug–domain association was likely to occur by chance.

2.3 Domain co-occurrence network

Two domain co-occurrence networks were built using the CATH and PFAM domain composition for each protein in the drug–protein interaction dataset. Two domains in the network are linked if they appear in the same protein.

2.4 Evolutionary analysis of drug targets

Protein sequences of human GPCRs and kinases were obtained based on the presence of characteristic CATH domains in the human proteome and aligned with ClustalW (Larkin *et al.*, 2007). We obtained the phylogenetic distance from each tree and built one network for each target class

in the sequence similarity space. In these networks, nodes represent protein sequences, and two nodes are linked if the phylogenetic distance between the proteins is <0.75.

Data retrieval, network building and analysis were performed using the Ruby and Python scripting languages and the Networkx Python module (http://networkx.lanl.gov/). Bipartite graph analyses were performed with Cytoscape and with the BipartiteNet plugin developed by our group.

3 RESULTS

3.1 Modelling the drug–domain networks

Our initial dataset comprises 5531 drugs, 3580 proteins and 12 754 drug–protein interactions. We have considered two complementary perspectives on protein domains: the structural perspective provided by CATH (Orengo *et al.*, 1997) and the function-oriented perspective offered by PFAM (Finn *et al.*, 2010).

With our scoring scheme based on the statistics of domain occurrence in the protein targets of each drug, we provide reliable drug–domain associations for 26.05% (CATH) and 27.78% (PFAM) of the drugs in the initial drug–protein interaction dataset. Our drug–domain network models include only the drug–domain associations that are more likely to appear than those expected at random (*P*-value = 0.001), based on the statistical meaning of the occurrence of every possible drug–domain pair.

To provide a meaningful comparison of the drug–domain associations with the drug–protein interactions, we built drug–protein subnetworks for the drugs that we can associate with the PFAM and CATH domains. Table 1 shows the global properties of the resulting drug–domain and drug–protein bipartite networks.

3.2 Druggable domains

The drug average degree indicates that drug–domain networks tend to be organized around particular druggable domains that draw together most of the drugs. We checked this finding by examining the network heterogeneity and the clustering coefficient of the drug–drug projections. Monopartite projections of the drug–domain and drug–protein networks (where drugs are nodes and two drugs are connected by an edge if they share at least one target in the corresponding bipartite network) contain information about how drugs are targeting the same proteins (or domains). Projections from drug–domain networks tend to have fewer hubs and more clusters than those from drug–protein networks, as shown in Table 2.

When we compared the drug–domain bipartite networks with the drug–protein bipartite subnetworks, it was clear that drugs revolve around a privileged set of targets to a greater extent in the former. We aimed to identify these druggable domains as proxies for the druggable genome. Hopkins and Groom (2002) used drug-binding domain annotations to identify the druggable protein families in the human genome. Our drug–domain associations revealed roughly the same description of the druggable genome.

Table 1. Properties of the drug–domain (dD) bipartite networks built with the PFAM and CATH protein domain definitions compared with the drug–protein (dP) networks

Network	Drugs	Targets	Pairs	Drug average degree	Target average degree
dDPFAM	1535	223	2846	1.85	12.76
dDCATH	1441	175	2829	1.96	16.17
dP subnet PFAM	1535	2236	6562	4.27	2.93
dP subnet CATH	1441	2309	6673	4.63	2.89
dP	5531	3580	12 754	2.30	3.56

Table 2. Heterogeneity and clustering coefficient for the drug–drug projections of the drug–target bipartite networks studied

Parameter	dD CATH	dP CATH	dD PFAM	dP PFAM
Heterogeneity	0.61	0.89	0.46	0.84
Cluster coefficient	0.96	0.83	0.97	0.85

3.3 Drug selectivity in drug–domain networks

Drug–domain associations are more selective than drug–protein interactions. The degree distributions of the bipartite networks clearly indicate the higher selectivity of drug–domain associations (Fig. 1).

The fraction of drugs targeting one domain is considerably higher than that targeting one protein in both the CATH and the PFAM drug–domain networks, whereas the fraction of drugs with more than one target is lower in both the CATH and the PFAM drug–domain networks except for drugs with two targets. In this particular case, there are more drugs targeting two domains than drugs targeting two proteins. When we inspected the co-occurrence of these domain pairs in the same proteins using domain co-occurrence networks, we found that 94.5% (CATH) and 87.5% (PFAM) of the pairs appear in the same protein.

Nacher and Scwartz (2012) showed that proteins linked by drug associations tend to interact preferably with one another than with other proteins. To check whether the domains associated with promiscuous drugs are clustered in modules, we mapped drugs associated with two or more PFAM domains to a protein–domain interaction network obtained from the DOMINE database (Yellaboina *et al.*, 2010). We identified 575 modules within this domain interaction network using ModuLand (Kovács *et al.*, 2010). We found that 85% of these modules contained up to five domains, and that 46.94% of the promiscuous drugs are associated with the PFAM domains contained in these modules.

3.4 Phylogenetic relationships between druggable domains

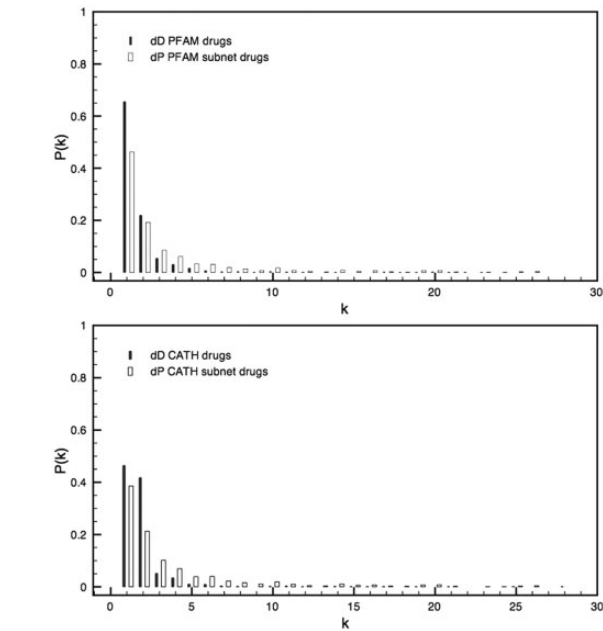To get insight into the evolutionary relationships between domains associated with polypharmacological drugs, we selected two examples of drugs targeting the main classes of pharmacological targets (GPCRs and kinases): doxepin (PubChem ID 667 477), a modulator of a number of serotonin, adrenergic and histamine receptor subtypes, and staurosporine (PubChem ID 44 120 114), a protein kinase inhibitor. In Supplementary Figure S1, we represent the phylogenetic trees for GPCRs and kinases as networks in the sequence similarity space, where proteins with high sequence similarity appear linked together in modules. Doxepin targets are closely related to each other, whereas staurosporine targets show a much lower sequence similarity.

4 DISCUSSION

Our work is not the first to attempt to associate drugs with protein domains. Wang *et al.* (2012) developed a statistical approach to predict potential targets for new drugs based on interactions between drugs and protein domains. Interestingly, they assumed that drugs act by binding to specific proteins, contrary to the idea proposed by drug polypharmacology. Moreover, they relied on the idea that drugs interacting with the same domain tend to share therapeutic effects; hence, they associate domains with therapeutic categories using the anatomical therapeutic chemical (ATC) codes, that is, a domain is associated with a group of drugs sharing the same ATC code. As far as we know, the work described in this report is the first attempt to identify direct drug–domain associations in a systematic and comprehensive manner.

Other previous research has focused on drug–domain relationships as well. Hopkins and Groom (2002) and Russ and Lampel (2005) relied on manual annotations of drug-binding domain to explore the druggable genome. We obtained roughly the same set

Fig. 1. Drugs degree distributions. Drugs degree distributions in the drug–domain (dD; filled bars) and drug–protein (dP; empty bars) bipartite networks. PFAM, upper panel. CATH, lower panel

of druggable domains with our automatic method. We found that drugs tend to revolve around a privileged set of domains, as can be observed in the drug–domain bipartite networks and can be deduced from the monopartite projections of the drug–domain and drug–protein networks. Network heterogeneity reflects the tendency of a network to contain hub nodes; a drug–drug projection of a bipartite network with a low heterogeneity reflects a trend in which drugs are organized around a privileged set of targets. This is the case for the drug–drug projections of both drug–domain networks when compared with their corresponding drug–protein subnetworks, as shown in Table 2.

Drug degree distributions (particularly in drug–domain networks), after we corrected for the co-occurrence of domains in the same proteins, clearly show that drugs target domains in a more specific way than they target proteins. Different domain targets for a polypharmacological drug tend to cluster in the same modules, within domain interaction networks such as DOMINE. These domains, and in particular co-occurring domains, can be interpreted as a single module or as a single pharmacological target, in the same sense that protein complexes are considered as drug targets in the work of Nacher and Schwartz (2012).

To get a hint of the evolutionary perspective of druggable domains, we selected polypharmacological drugs that target the two main families of druggable proteins, as described by Hopkins *et al.* (2002) in the druggable genome, i.e. GPCRs and kinases. Although both doxepin and staurosporine can be associated with protein domains that are archetypical for each protein class, the targets of doxepin are closely related in the GPCRs phylogenetic tree, whereas the targets of staurosporine are widespread in the human kinome. These two examples illustrate a difference in the subfamily targeting specificity of the selected drugs and suggest further studies, such as the implications in drug polypharmacology of the evolutionary conservation of structural domains (with particular focus on binding site conservation) and the structural similarity of domains for distinct targets.

Our results support the idea that protein domains are a good proxy for drug targets when considering the structural and functional perspectives of their definition. Hence, drug polypharmacology, understood as the ability of a single drug to bind several proteins, emerges as a consequence of the multidomain composition of proteins.

ACKNOWLEDGEMENTS

REFERENCES

Finn,R.D. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.

Goh,K.I. *et al.* (2007) The human disease network. *Proc. Natl. Acad. Sci. USA*, **104**, 8685–8690.

Guberman,J.M. *et al.* (2011) BioMart Central Portal: an open database network for the biological community. *Database (Oxford)*, **2011**, bar041.

Günther,S. *et al.* (2008) SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res.*, **36**, D919–D922.

Hopkins,A.L. and Groom,C.R. (2002) The druggable genome. *Nat. Rev. Drug. Discov.*, **1**, 727–730.

Hopkins,A.L. (2008) Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.*, **4**, 682–690.

Knox,C. *et al.* (2010) DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. *Nucleic Acids Res.*, **39**, D1035–D1041.

Koonin,E.V. *et al.* (2002) The structure of the protein universe and genome evolution. *Nature*, **420**, 218–223.

Kovács,I.A. *et al.* (2010) Community landscapes: an integrative approach to determine overlapping network module hierarchy, identify key nodes and predict network dynamics. *PLoS One*, **5**, e12528.

Kuhn,M. *et al.* (2008) STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res.*, **36**, D684–D688.

Kummerfeld,S.K. and Teichmann,S.A. (2009) Protein domain organisation: adding order. *BMC Bioinformatics*, **10**, 39.

Larkin,M.A. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

Mestres,J. *et al.* (2008) Data completeness—the Achilles heel of drug-target networks. *Nat. Biotechnol.*, **26**, 983–984.

Mestres,J. *et al.* (2009) The topology of drug-target interaction networks: implicit dependence on drug properties and target families. *Mol. Biosyst.*, **5**, 1051–1057.

Nacher,J.C. and Schwartz,J.M. (2012) Modularity in protein complex and drug interactions reveals new polypharmacological properties. *PLoS One*, **7**, e30028.

Orengo,C.A. *et al.* (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.

Pang,E. *et al.* (2012) Promiscuous domains: facilitating stability of the yeast protein–protein interaction network. *Mol. Biosyst.*, **8**, 766–771.

Russ,A.P. and Lampel,S. (2005) The druggable genome: an update. *Drug Discov. Today*, **10**, 1607–1610.

Wang,Y.Y. *et al.* (2012) Predicting drug targets based on protein domains. *Mol. Biosyst.*, **8**, 1528–1534.

Wolf,Y.I. *et al.* (2000) Estimating the number of protein folds and families from complete genome data. *J. Mol. Biol.*, **299**, 897–905.

Yellaboina,S. *et al.* (2010) DOMINE: a comprehensive collection of known and predicted domain-domain interactions. *Nucleic Acids Res.*, **39**, D730–D735.

Yildirim,M.A. *et al.* (2007) Drug-target network. *Nat. Biotechnol.*, **25**, 1119–1126.

Zanzoni,A. *et al.* (2009) A network medicine approach to human disease. *FEBS Lett.*, **583**, 1759–1765.