PubChem as a Source of Polypharmacology

Bin Chen,[‡] David Wild,[‡] and Rajarshi Guha*,[†]

School of Informatics, Indiana University, Bloomington, Indiana 47408, and

and NIH Chemical Genomics Center, Rockville, Maryland 20850

Received June 1, 2009

Polypharmacology provides a new way to address the issue of high attrition rates arising from lack of efficacy and toxicity. However, the development of polypharmacology is hampered by the incomplete SAR data and limited resources for validating target combinations. The PubChem bioassay collection, reporting the activity of compounds in multiple assays, allows us to study polypharmacological behavior in the PubChem collection via cross-assay analysis. In this paper, we developed a network representation of the assay collection and then applied a bipartite mapping between this network and various biological networks (i.e., PPI, pathway) as well as artificial networks (i.e., drug-target network). Mapping to a drug-target network allows us to prioritize new selective compounds, while mapping to other biological networks enable us to observe interesting target pairs and their associated compounds in the context of biological systems. Our results indicate this approach could be a useful way to investigate polypharmacology in the PubChem bioassay collection.

1. INTRODUCTION

During the last 20 years, relatively few drugs have reached the market due to high attrition rates.[1] Lack of efficacy and toxicity are the two main causes, accounting for 60% of the failures. This may to some extent be attributed to the current drug-development paradigm, which aims to discover highly selective ligands for a single target. Its process starts with finding and validating a target of clinical relevance, followed by identifying hits from high-throughput screening. Interesting hits are analyzed to direct the design of potential leads, which will go to pharmacological and clinical testing. During this process, promiscuous compounds that interact with multiple targets are considered as undesirable and are removed as early as possible.

Many lead compounds are found to be unexpectedly ineffective or toxic when they are tested *in vivo*. When biological systems are viewed in a network paradigm, it is well-known that inhibition of a single target does not necessarily have an effect on the system overall. In these cases, alternate "paths" through the system are able to take over the loss incurred by the inhibition of the target.[2] This robustness to perturbations (via inhibition or even deletion of individual targets) has been suggested to be a byproduct of the scale-free property of biological networks.[3] On the other hand, the simultaneous targeting of multiple nodes can be lethal or deleterious to the system.[4] These observations suggest that the development of drugs that *simultaneously* target multiple receptors in a rational fashion could increase their efficacy, particularly in the treatment of a complex disease like cancer and central nervous system disorders.[5,6] This approach has been termed polypharmacology.[7] For example, Imatinib (Gleevec) inhibiting BCR-ABL, c-KIT, and PDGFRs has high efficacy in the treatment of chronic myeloid leukemia.[8]

Three strategies have been proposed to design multitarget therapies, but all of them have particular disadvantages.[7] Multidrug combinations, prescribing multiple individual medications, must consider patent compliance and drug−drug interactions. The development of multicomponent drugs that consider pharmacological intervention of several compounds that interact with multiple targets is hindered by the limited pharmacopoeia available currently.[6] Besides, both methods must consider balancing the dose of every ingredient involved. The third method, aiming to design a single compound with selective polypharmacology, could avoid these problems and have lower regulatory barriers for approval than other methods, but it is challenged by the validation of target combinations and incomplete SAR data.

Recent high-throughput technologies have allowed the elucidation of complex biological networks such as protein−protein interaction, metabolic, and gene regulatory networks. With such network data, one can apply network and pathway analysis techniques to drug discovery and development.[9] Early network analysis indicated a high correlation between lethality and the degree of nodes in a biological network. Later, edge betweenness[10] was found to exhibit a higher correlation to lethality in regulatory networks. The study of such properties directs the search of cellular drug targets when designing multitarget drugs.[11]

A biological network only characterizes the relation between protein and protein or protein and gene association. In order to identify candidate targets for polypharmacological drugs, proteins or genes must be connected to small (druglike) molecules. Paolini presented a global mapping of pharmacological space by the integration of SAR data from diverse sources.[12] Many attempts have been made to study target-compound association either from experimental results or from predictive models.[13] For example, the development

* Corresponding author e-mail: rguha@indiana.edu.
[‡] Indiana University.
[†] NIH Chemical Genomics Center.

10.1021/ci9001876 CCC: $40.75
Published on Web 08/27/2009

of binding assay panels like BioPrint provides a large number of experimental results. Bayesian models have been used to predict targets based on the fingerprint of structures.[14] Other nonstructural properties such as side effects have also been employed to identify drug targets.[15] According to the target-compound matrix described by Campillos et al., promiscuous compounds could be identified and their properties such as molecular weight and lipophilicity used to direct the rational design of polypharmacological drugs.[16] Yildirim[17] constructed a drug-target network and demonstrated most of the drugs have multiple targets via network analysis. Moreover, the network was mapped to a disease network to indicate that drug targets are often involved in multiple diseases. Using the network paradigm to study polypharmacological behavior, also named network pharmacology, was reviewed by Hopkins.[7] However, the development of polypharmacology is limited by the incomplete SAR data. Furthermore, the ability to validate target combinations is hampered by limited resources.

1.1. Goals. PubChem[18] is a public repository of chemical information covering both chemical structure as well as their activity in a variety of biological assays. Of particular interest is the bioassay collection, which collects the results of high throughout screens performed under the aegis of the Molecular Libraries Screening Centers Network (MLSCN).[19] As of July 2008, the repository contained the results of 1133 biological assays and 662,908 compounds tested, of which 139,326 compounds have a result for at least one bioassay.

The fact that the PubChem bioassay collection reports the activity of compounds in multiple assays suggests that it might allow us to identify polypharmacological behavior of compounds in the PubChem collection. In the current study we investigated this possibility by developing a network representation of the assay collection and then applying a bipartite mapping between this network and various biological networks. The overall goal is to identify compounds in the assay collection that are active against multiple targets and then characterize how they might interfere with a biological network, either by disrupting a pairwise interaction or by perturbing indirectly connected nodes. In Section 2, we discuss the design of the assay network, including the process of filtering for promiscuous compounds. We also describe the procedure by which the assay network is mapped to biological networks. In Section 3 we present applications of this methodology. Finally, Section 4 summarizes our findings and discusses the implications of the PubChem as a source of polypharmacology.

2. METHODS

2.1. Data Sets. We considered the PubChem assay collection as of November 9, 2008 and extracted data from 1223 assays. From each assay we extracted the Compound ID (CID) as well as the activity outcome and activity score fields. We noted that the activity score is quite noisy and, in general, not very reliable for modeling purposes. However, our goal was to be able to perform cross-assay analyzes, and the activity score is the only way to achieve this in an automated fashion. In addition to the assay data themselves we also retrieved the assay descriptions and protein targets, when available. In the current study we ignored assays that did not have an associated target. All the data were incorporated into a local PostgreSQL database, and the scripts required to regenerate the database are provided as Supporting Information.

The final database contained 602 assays containing 506,190 distinct compounds, of which 90,290 compounds were active in at least one assay. The assays represented 258 unique protein targets, each target being tested in multiple assays.

Promiscuity is a well-known problem associated with HTS assays. In this study, promiscuous compounds can be especially misleading, and so we filtered the compounds noted above to remove promiscuous entries. We did not consider promiscuity in HTS resulting from aggregation.[20] Given that promiscuity can arise from multiple factors,[21,22] a number of approaches have been described in identifying promiscuous compounds.[23,24] We employed the two-step procedure described by Paolini et al.[12] First, P_{active} is the number of targets that a compound was active against. Second, P_i is P_{active}/P_{test}, where P_{test} is the number of targets that the compound was tested against. Since a target might be tested in multiple assays, a compound that was active in any of those assays was considered active against that target. With this formulation, we define two constraints. A compound is deemed promiscuous if either $P_{active} \geq 5$ and $P_i = 1$ or if $P_{active} \geq 10$. This procedure identified 354 compounds, which were eliminated during network construction.

2.2. Assay Network Construction. Given the assay collection, we then constructed an assay network, in which each node represents an assay and two assays are connected by an edge. In the current study we considered two assays to be connected (i.e., an edge between two nodes) if they shared one or more active compounds.

There are two important features of the network. First, there may be several assays that are centered on the same target protein (e.g., primary and secondary assay). Given that primary assays can contain a large number of false positives, we decided to focus only on secondary assays when both primary and secondary assays were available for a target. Since there is no formal annotation indicating whether an assay is primary or secondary (i.e., confirmatory), we selected the smaller of two assays targeting the same protein. Second, while the assay data from PubChem provides an activity classification, we were interested primarily in the highly active compounds. Thus we considered only those compounds which had an activity score greater than 90 (on a range of 0 to 100).

2.3. Bipartite Network Mapping. While the assay network is an interesting representation of the assay collection, its utility beyond visualization depends on connecting it to biological systems. While the assay network may highlight compounds exhibiting polypharmacological behavior, its real value arises from the ability to integrate this information with a real biological network. More specifically, given a compound active against two targets as indicated in the assay network, we are interested in understanding this behavior in the context of a biological system.

Here, biological system refers to some form of network, including metabolic networks and regulatory networks as well as artificial constructs such as disease networks and drug-target networks. To achieve this, we must somehow map nodes in the assay network to nodes in these biological networks. This procedure is highlighted in Figure 4.

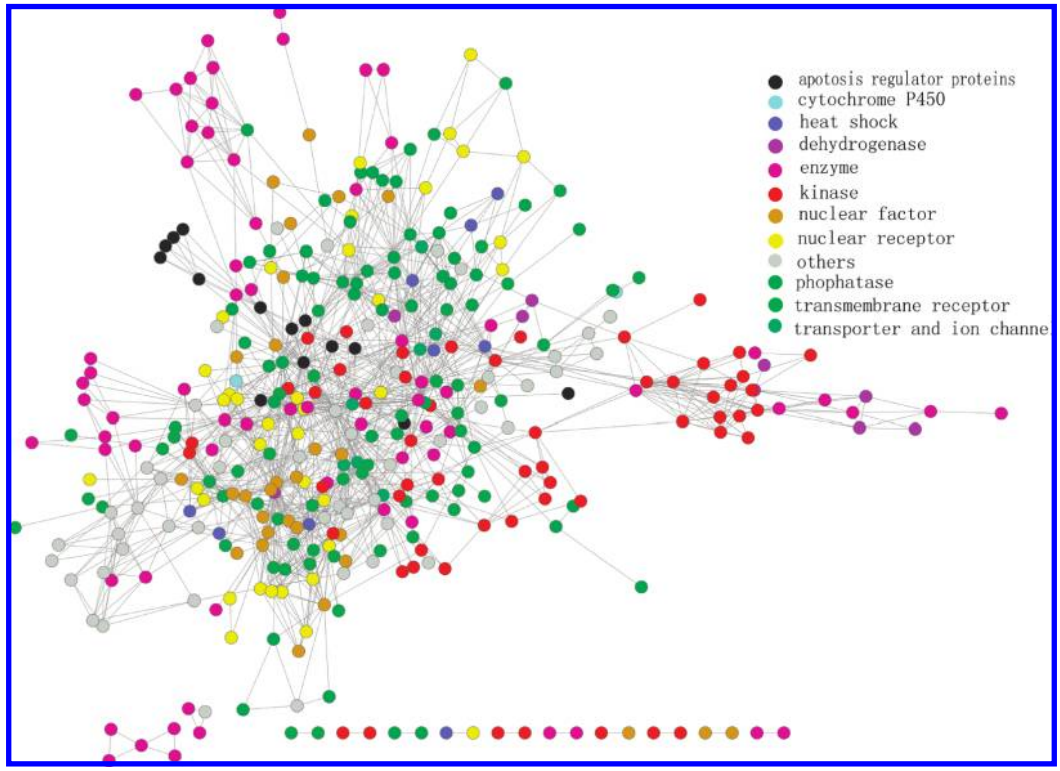

Figure 1. Bioassay network. Assays are represented as nodes; two assays are linked if the pair shares at least one active compound. Nodes are colored by the gene family of their corresponding targets.

The core of the mapping procedure is the mapping function. The simplest approach is to map a node from the assay network to one in the biological network if the two nodes correspond to the same protein. This can be termed the "identity mapping". Mathematically, we can write this as

$$F_{identity}({}^{A}N_i, {}^{B}N_j) = 1 \qquad \text{if } P_i = P_j \qquad (1)$$

$$= 0 \qquad \text{otherwise} \qquad (2)$$

where ${}^{A}N_i$ and ${}^{B}N_j$ represent a node i and j from the assay network and the biological network, respectively. P_i and P_j represent the protein associated with ${}^{A}N_i$ and ${}^{B}N_j$, respectively. If $F_{identity} = 1$, this indicates an edge between the node in the assay network and the node in the biological network. While simple, this approach is limited. If there is no protein in the biological network that is an exact match to one in the assay network, there will be no mapping.

Thus, we next consider a similarity mapping, in which two nodes are mapped if the similarity between their protein targets exceeds some cutoff. More formally

$$F_{similarity}({}^{A}N_i, {}^{B}N_j) = 1 \qquad \text{if } S(P_i, P_j) \geq S_c \qquad (3)$$

$$= 0 \qquad \text{otherwise} \qquad (4)$$

where $S(P_i,P_j)$ is a similarity function defined on the proteins corresponding to the two nodes, and S_c is some similarity cutoff. In this study we employed BLAST similarities to evaluate $S(P_i,P_j)$, and the similarity cutoff was defined in terms of the E-value. That is, if the BLAST similarity between the two proteins was less than 1×10^{-20}, we considered the assay node and the biological network node to be mapped.

While this discussion has focused on a node by node mapping, we can extend these definitions to map edges from one network to another. As a result, we obtained four possible mapping schemes:

• Exact Node Mapping (ENM) - One node in the assay network is mapped to one node in the biological network if the two nodes have an identical protein.

• Exact Edge Mapping (EEM) - One edge in the assay network is mapped to one edge in the biological network if the two edges connect to identical proteins.

• Similar Node Mapping (SNM) - One node in the assay network is mapped to one node in the biological network if the two nodes have similar sequences (BLAST). This contains ENM.

• Similar Edge Mapping (SEM) - One edge in the assay network is mapped to one edge in the biological network if the nodes in each edge are similar each other, in terms of BLAST results. This contains EEM.

2.4. Biological Networks. As discussed above, the assay network is a useful visual representation, but its utility becomes clearer when we map it to some biological networks. We considered three such networks.

Drug-Target Network. This type of network was described by Yildirim et al.[17] We constructed a similar network in which the nodes are protein targets, and two targets are connected by an edge if they are both targeted by one or more of the same drugs. The data for this network were taken from DrugBank[25] (accessed on February 2008). This database is a comprehensive resource that combines drugs and associated target information. We extracted 4674 drugs and 4552 targets, identified by their Drug IDs and Swiss-Prot IDs, respectively. We excluded drugs (such as ethylene glycol) with more than 50 targets, since that led to a particularly noisy target network. Furthermore, we only

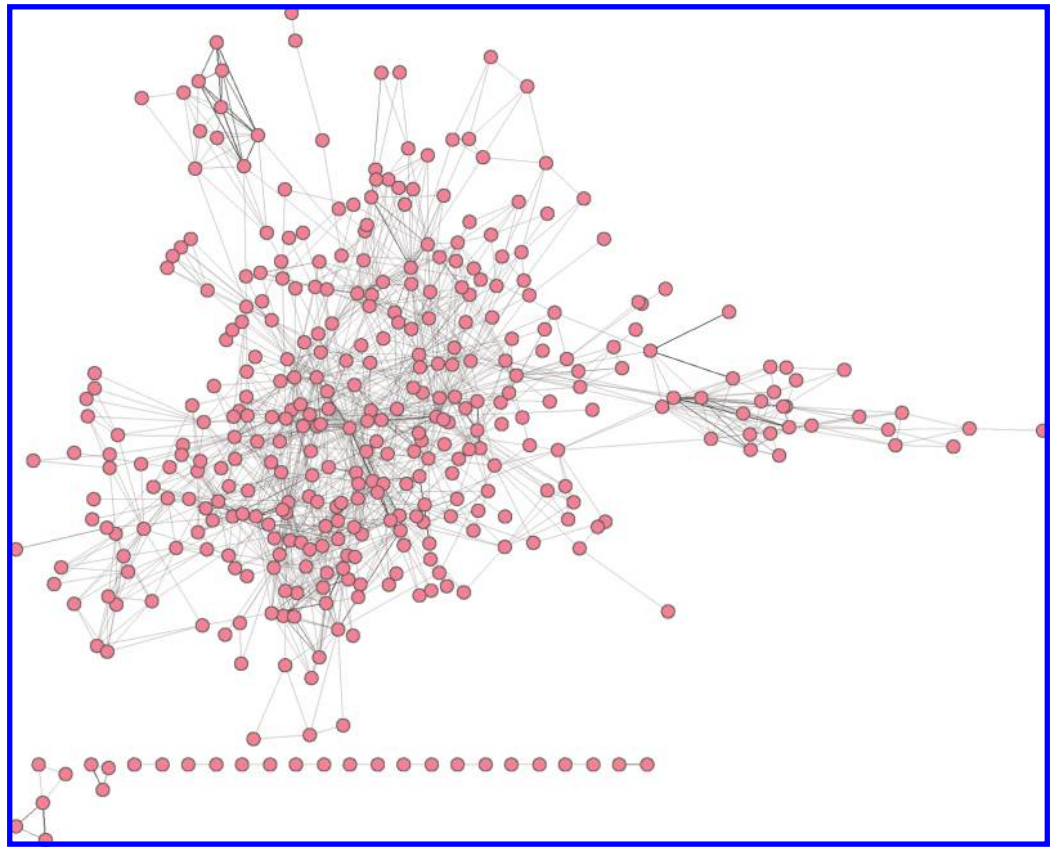

Figure 2. Bioassay network. Assays are represented as nodes; two assays are linked if the pair shares at least one active compound. The edge opacity is represented by the number of shared compounds.

considered human proteins as identified by Swiss-Prot IDs. As a result, 1973 drugs and 1414 targets were used to construct the drug-target network (Figure 11).

Protein−Protein Interaction Network (PPI). The Human Protein Reference Database (HPRD)[26] contains direct binary protein−protein interactions manually extracted from published experimental evidence. Interactions are indicated as *in vivo*, *in vitro*, or yeast two-hybrid. While a number of PPI databases are available, the HPRD has been reported to have a higher overall quality.[27] We downloaded HPRD data in March 2008. Data were cleansed by ignoring entries without a gene symbol and entries exhibiting an interaction within the same gene. This resulted in 34,119 interactions between 9021 proteins. In this network, the proteins are nodes and an edge exists between two nodes if the HPRD data indicate an interaction between those nodes.

Pathway. Given the importance of systems level of various molecular pathways,[9,28] we considered all the pathways in KEGG including metabolic pathway, signal transduction, cellular processes, and human diseases.[29] We retrieved pathways using proteins (Gi number) through the KEGG API in October 2008. We also downloaded all the protein sequences from KEGG while doing similarity mapping. Only the pathways involving at least one mapped result either from exact mapping or from similar mapping were considered. Since no explicit data are provided to annotate the location of proteins in a pathway, we assume that every protein is connected to each other within a pathway. So in EEMs and SEMs, once the pair of associated targets in the assay network occurs in the same pathway, we call this pair is mapped to a pathway.

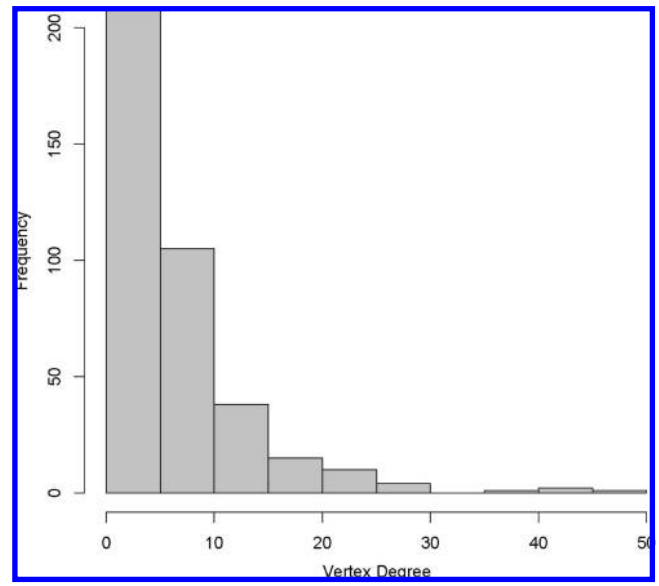

Figure 3. Vertex degree distribution. The power-law distribution exponent is 1.5.

3. RESULTS

3.1. Assay Network Properties. The assay network constructed according to Section 2 contained 384 nodes and 1304 edges, involving 212 targets. We visualized the network using Cytoscape.[30] Figure 1 displays the assay network color coded by Pfam classification of the targets. We considered ten groups: enzyme, kinase, phosphatase, transmembrane receptor, transporter and ion channel, nuclear receptor, nuclear factor, apoptosis regulator proteins, heat shock, and others. Figure 2 shows the same network where the edge

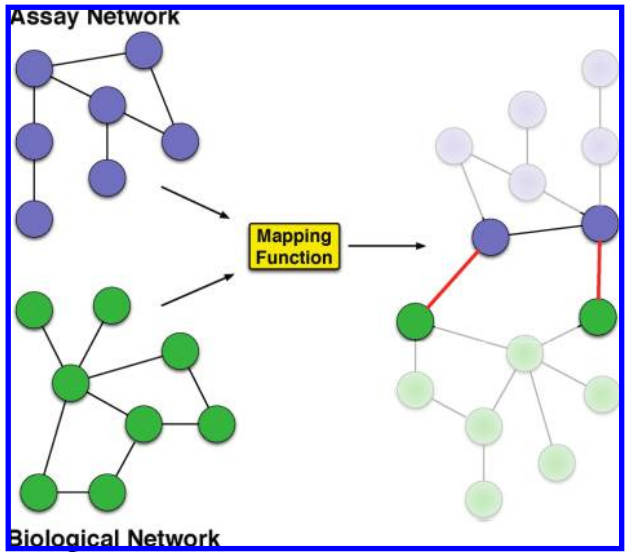

Figure 4. A schematic diagram of the bipartite mapping between the assay network and an arbitrary assay network. In general, the nodes in the biological network represent proteins. An assay node is mapped to a biological node based on a mapping function. Possibilities include the identity function or some similarity metric. In this diagram, the red edges indicate a mapping between the assay network and the biological network. If two nodes of an edge in the assay network are mapped to two nodes of an edge in biological network, the edges are mapped. The thicker edges are mapped in the figure.

opacity indicates the number of active compounds common between any two assays. This view highlights the clustering in this assay network.

While the assay network is artificial in nature, it is useful to consider some network metrics. The clustering coefficient[31] is 0.37 and the average degree is 6.8. Figure 3 suggests that the assay network is scale free[31] in nature, with a degree distribution that approximates a power law distribution ($\gamma = 1.5$).

Unsurprisingly, assays with either the same targets or with targets belonging to the same gene family are prone to cluster together. However, a number of assays share common active compounds even though the targets are different or belong to different gene families. In Figure 5, Assay **388** (NAD^+-dependent 15-hydroxyprostaglandin dehydrogenase) shares active compounds with 8 assays, of which targets are among several gene families. The common active compounds of the pairs between AID **388** and its neighbors vary largely, and most of the common compounds are not promiscuous. For example, CID 6246323 shared by AID **388** and AID **392** is tested in 136 assays, but it is active in only 3 assays. It is also of interest to study the compound targeting multiple proteins. For instance, genistein (CID 5280961) shared by AID **388** and AID **505**, as a well-known protein-tyrosine kinase inhibitor, has active result in AID **505**, where a serine/threonine kinase pim-2 oncogene is tested. Genistein is also active in AID **388**; this might be explained by the fact that genistein could directly inhibit 3- and 17-beta-hydroxysteroid dehydrogenase activity.[32]

But more importantly, the assay network is an artificial construct, and, as such, network metrics, likely, do not have physical significance. For example, the node degree is highly dependent on the number of tested compounds and the selection of compounds for testing. Thus we do not focus on the network properties, but instead consider the use of the assay network as a tool to study other biological networks.

3.2. Mapping. *Drug-Target Network.* Table 1 indicates that 126 assays map to 55 proteins in the drug-target network and 8 pairs composed of 14 unique targets are mapped as EEMs. Node mapping results increase a lot while similar mapping. The average degree of the drug-target network is 13; in contrast, the average degree of mapped targets is 17. It shows that the mapped targets are more promiscuous. In Figure 11, most of the mapped results are located in the center of giant components.

In total, 38 identical drugs and 130 bioassay results are involved in the SEMs. We investigated the minimum, maximum, and average similarity between bioassay results and corresponding drugs using the Tanimoto coefficient with ECFP-6 (Table 2). Sterochemistry was not considered during the similarity calculation. Most of the drugs are significantly different with the corresponding assay results. For example, dexamethasone (CID 5743), an anti-inflammatory 9-fluoroglucocorticoid, targets NF-κB1 and NR3C1, which are tested in AIDs **895** and **450**, respectively. The two assays share two common active compounds hydrocortisone (CID 5754) and tocris-1126 (CID 6603742) in the assay network (Figure 7). It is not surprising that dexamethasone and tocris-1126 have similar activities in NF-κB1 and NR3C1, as they only have a slight difference in stereochemistry. The activity of

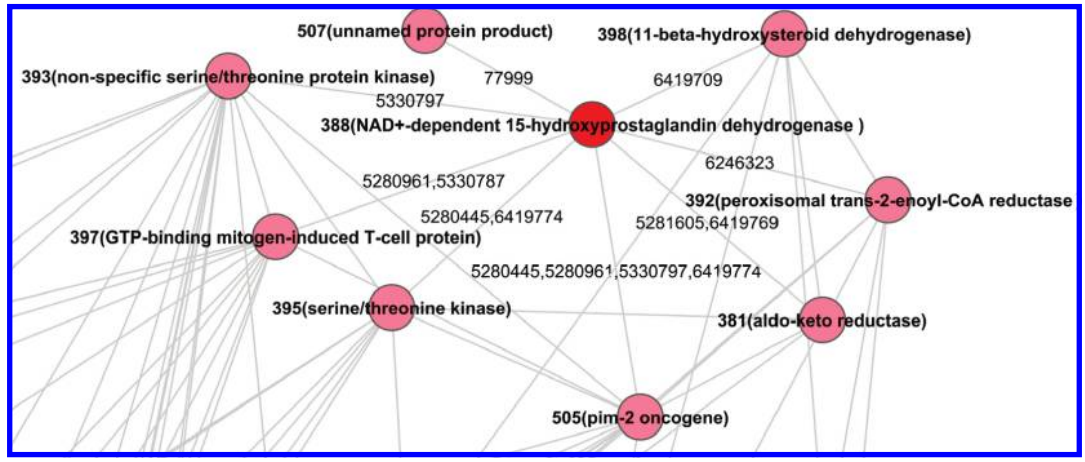

Figure 5. Nodes linking to AID **388** in the assay network. AID and its target name are used to label each node. Edge is labeled as the list of CIDs of common active compounds between two targets.

Table 1. Mapping Results[e]

	ENM[a]		EEM[b]		SNM[c]		SEM[d]	
	assays	targets	assays	targets	assays	targets	assays	targets
drug-target network	126	55	20	8	374	140	54	18
PPI	307	131	33	14	343	161	33	14
pathway	179	85	119	54	264	117	277	121

[a] Exact node mappings. [b] Exact edge mappings. [c] Similar edge mappings. [d] Similar edge mappings. [e] ENMs in assays means the number of mapped nodes in the assay network. ENMs in targets means the number of mapped nodes in the biological network. Theoretically, the number in assays is larger than that in targets, as multiple assays might share the same target.

Table 2. Similarity between Active Compounds and Corresponding Drugs[a]

target pairs	average similarity	min similarity	max similarity
ABCG2 ABCB1	8.40×10^{-2}	3.91×10^{-2}	1.56×10^{-1}
BCL2 ALB	5.97×10^{-2}	4.40×10^{-2}	9.04×10^{-2}
CHRM1 ABCB1	9.23×10^{-2}	9.09×10^{-2}	9.30×10^{-2}
CHRM4 CHRM1	5.18×10^{-2}	3.88×10^{-2}	9.68×10^{-2}
DRD5 DRD1	6.54×10^{-2}	2.14×10^{-2}	1.36×10^{-1}
ESR2 ESR1	7.00×10^{-2}	3.33×10^{-2}	1.14×10^{-1}
HTR1E HTR1A	8.92×10^{-2}	4.85×10^{-2}	2.08×10^{-1}
NR3C1 NFKB1	4.91×10^{-1}	2.37×10^{-1}	1.00
OPRM1 ALB	1.06×10^{-1}	1.06×10^{-1}	1.06×10^{-1}
PTK2B ESR2	5.04×10^{-2}	5.04×10^{-2}	5.04×10^{-2}
PTPN1 CDC25B	2.44×10^{-2}	1.41×10^{-2}	2.82×10^{-2}
PTPN1 CTSL1	4.05×10^{-2}	4.05×10^{-2}	4.05×10^{-2}
SELE REG1A	2.88×10^{-2}	2.88×10^{-2}	2.88×10^{-2}
STAT1 BCL2	4.77×10^{-2}	4.55×10^{-2}	4.88×10^{-2}

[a] ECFP-6 was used to calculate similarity.

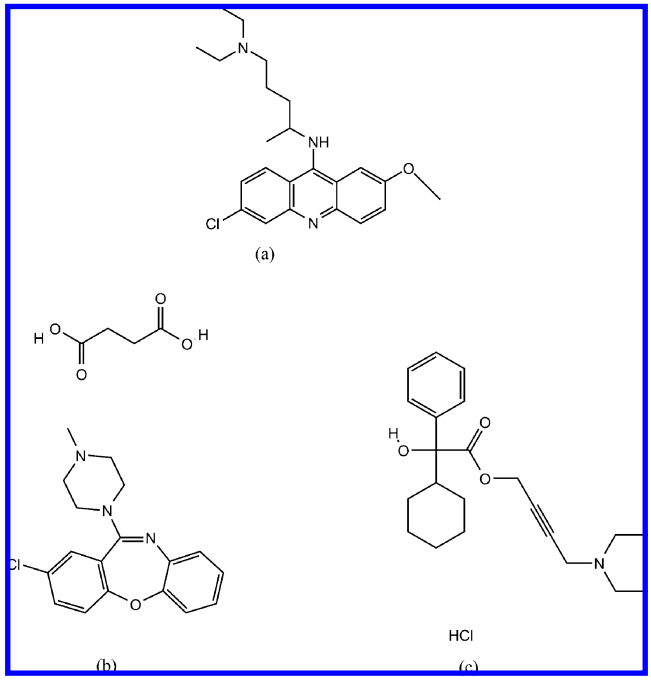

Figure 6. (a) Quinacrine, (b) loxapine, and (c) oxybutynin. Quinacrine targets ABCB1 and CHRM1. Loxapine and oxybutynin are both active in assay **377** and assay **859**, which correspond to ABCB and CHRM1, respectively. Loxapine and oxybutynin are drugs reported in DrugBank, but no record indicates either one would target ABCB1 and CHRM1 simultaneously.

dexamethasone is also similar to that of hydrocortisone. It shows that the addition of the methyl and flourine to hydrocortisone has no effect on the activity but improves its druglikeness. However, quinacrine has 7 targets in DrugBank, in which the ABCB1 and CHRM1 pair has been

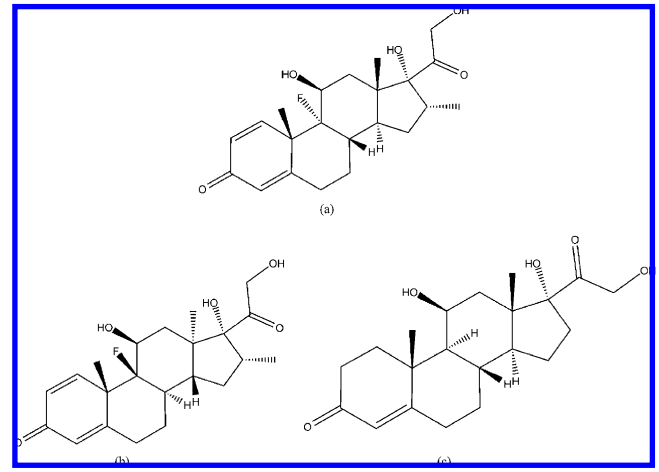

Figure 7. (a) Dexamethasone, (b) tocris-1126, and (c) hydrocortisone. Dexamethasone NF-κB1 and NR3C1, which are tested in AIDs **895** and **450**, respectively, where two common active compounds hydrocortisone and tocris-1126 are involved.

mapped (Figure 6). ABCB1 was tested in AID **377** and CHRM1 was tested in AID **859**. These two assays share two common active compounds, viz., loxapine (CID 71399) and oxybutynin (CID 91505). The fact that both loxapine and oxybutynin are active in AID **377** (ATP binding cassete) which identifies substrates (or inhibitors) for multidrug resistance transporter is not too surprising. Since such substrates tend to be hydrophobic molecules, both loxapine and oxybutynin are good candidates. Oxybutynin is an antagonist for M1 muscarinic receptor in AID **859**. Since AID **859** is a cellular assay, loxapine might gets metabolized to amoxapine which is a member of the tricyclic antidepressant family. Compounds in this family are thought to bind to muscarinic and histamine receptors. It is reported that amoxapine is a considerably weak antagonist of muscarinic cholinergic receptors *in vitro* assessment.[33]

PPI. While 80% of the nodes in the assay network could be mapped to the HPRD-derived PPI network, only 14 pairs exhibited common active compounds. It is of interest to analyze the statistical properties of the mapped results. For example, if a pair of targets (i.e., assays) in the assay network maps to targets in the PPI network exhibiting high node degree, this can suggest that these two nodes play a significant role in the small world structure (and hence, robustness) of the PPI network. Thus a compound identified from the assay network targeting these two nodes in the PPI network could act as a probe to investigate this property. We analyzed results in terms of vertex degree, betweenness centrality, and the length of shortest path. The degree distribution reveals the small number of proteins with the higher degree as hubs, which play an important role in the

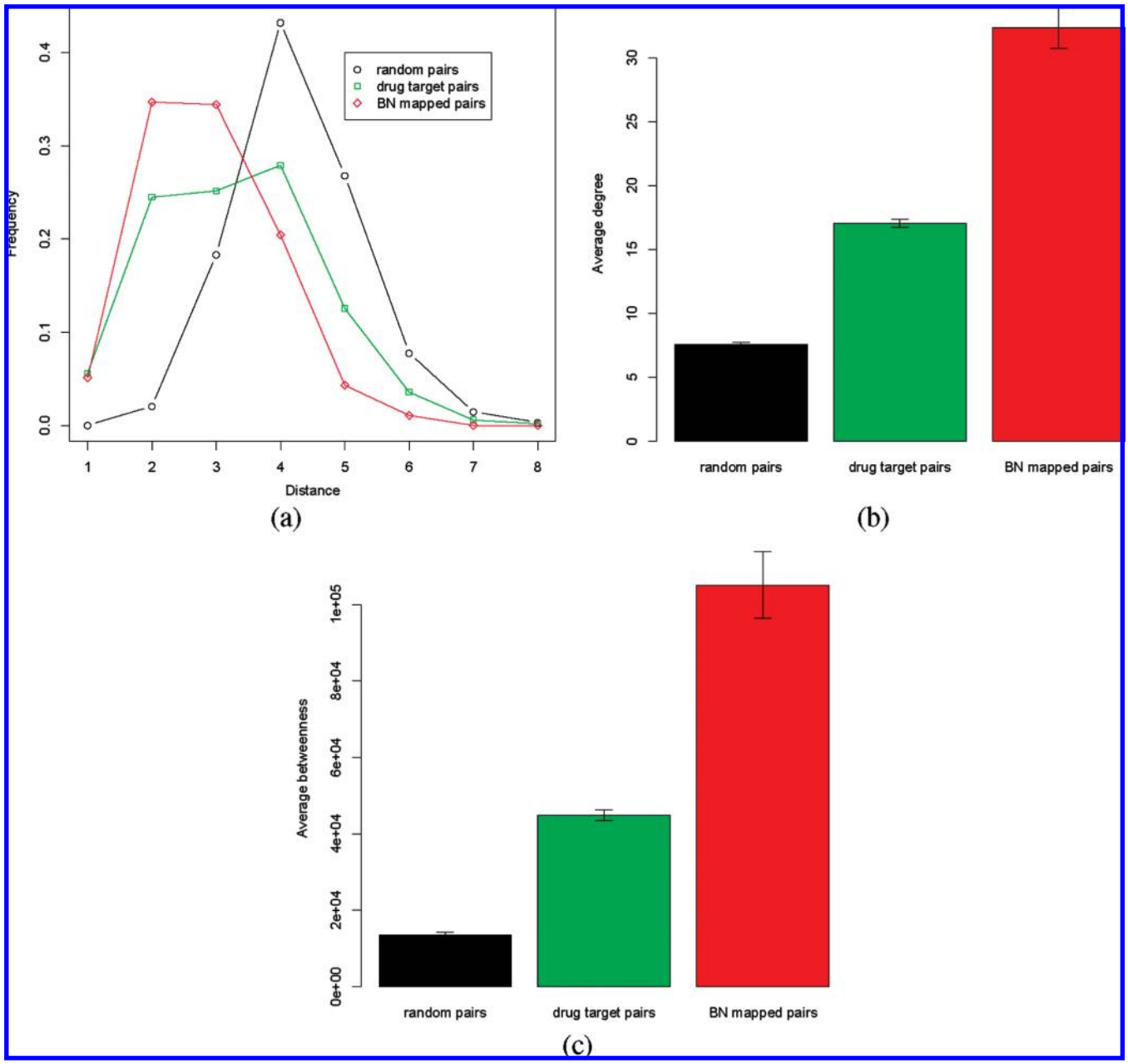

Figure 8. Mapping drug target pairs and assay network pairs to the PPI. BN means bioassay network. (a) Fractions of the distance of the mapped pairs in PPI. (b) Average degree of the pairs in PPI. (c) Average betweenness of the pairs in PPI.

robustness of the network. Another important factor is the bottleneck measured by betweenness which could identify nonhub nodes that also play a role in robustness. The degree and betweenness evaluate only an individual node; the relation of two nodes could be evaluated by the shortest path.

The average degree, node betweenness, and shortest path of all the nodes in the PPI were 7.8, 1.4×10^4, and 4.3, respectively, while the average degree, node betweenness, and shortest path of ENMs were 15.7, 3.5×10^4, and 3.7, respectively. The higher degree and node betweenness show that the proteins involved in the assay network are more "important" than average. Given that assays are selected for screening based on biological relevance, this observation likely applies more to the social process of assay development.

The mapped results have overall higher centrality than average in PPI in terms of basic network properties, but the only comparison of individuals is unable to identify the pairs suitable for further observation. The drug-target network[17] not only shows that many drugs have multiple targets but also shows that lots of the targets are really not essential proteins in the PPI network, although their average degree is larger than that of random ones. Furthermore, it is observed that some drugs do not interact with disease causing proteins directly; instead, there are certain distances between drug targets and disease causing proteins within PPI. It is possible to reason that the action of a drug might result from the combination of targets instead of a single target. Hence, we studied the pairs of drug targets in PPI and tested whether they have significant difference with randoms ones in terms of network statistic properties. Using the drug target pairs, we questioned whether the mapped pairs from assay network are potentially "druggable". Furthermore, we mapped associated proteins of some selected diseases onto the PPI

network and calculated the distance between drug target pairs and diseases associated proteins.

The drug-target network and the assay network were first mapped to the PPI network via ENM separately. The length of the shortest path between the mapped pairs in the PPI network, their average degree, and their average betweenness were calculated, and the unreachable pairs, of which no path is found, were ignored. The shortest path means the least number of steps from one node to another node in a network. We compared the results of drug target pairs, assay network mapped pairs, and random pairs. In Figure 8, the random pairs are normally distributed at the center 4. Both drug target pairs and assay network mapped pairs shift to the left side. Meanwhile, both have almost the same frequency at distance 1, but mapped pairs from the assay network are more likely 2 or 3 steps away. Figure 8(b) shows assay network mapped pairs are more likely to be hubs comparing to drug target pairs, although either one has a higher degree than random pairs. Figure 8(c) also shows that the assay network mapped pairs are more significant than drug target pairs. The fact that drug targets are not actually essential proteins implies that the target pairs need not be essential. The overlap of drug target pairs and assay network mapped pairs suggests that a portion of assay network mapped pairs are of interest to further investigate.

In addition, we selected top 7 diseases with the number of drugs involved in DrugBank greater than 22 (we took them directly from the paper[17]). Their drugs, associated drug targets, and associated causing proteins were identified in DrugBank and OMIM. Only the drugs with more than one target that can reach to any disease associated protein of a corresponding disease were considered. For each drug, its targets were combined to generate all the possible pairs. The distance between one pair to the disease associated proteins is defined as the average distance between the targets of the pair to the disease associated proteins. The distance between a target to the disease associated proteins is measured by the minimum shortest paths from the target to all the disease associated proteins.

Some drugs might have common target pairs, resulting in duplicated target pairs, so we removed them to get unique drug target pairs before analysis. Meanwhile, we also calculated the distance of assay network mapped pairs as well as random pairs and refined the results using the same approach as drug target pairs. Except for HIV with only 10 pairs, other diseases have a number of available drug target pairs. For instance, leukemia has 149 drug target pairs. Like the distance distribution in Figure 8(a), assay network mapped pairs and drug target pairs are more likely to shift to the left in most of the sample diseases. Most of the drug target pairs have distance 2. However, in hypertension, the distribution of drug target pairs is quite similar to that of the random pairs; it suggests that distance is unable to differentiate druggable targets in hypertension. From the degree distribution (Figure 10(g)), it is also suggested that the selection of druggable pairs should be careful. For example, the average degree of drug target pairs in either pulmonary disease or rheumatoid arthritis is greatly different from that of assay network mapped pairs.

Pathways. If the targets of one pair of nodes in the assay network occur in a pathway, the pair is considered to be mapped to the pathway regardless of the distance of two

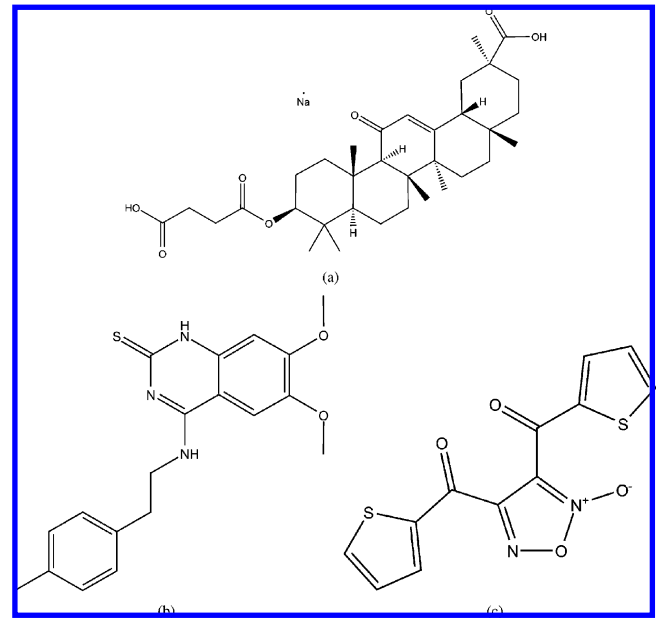

Figure 9. Structures in pathway mapping example. (a) CID 6419769, (b) CID 15996835, and (c) CID 573747.

mapped targets in the pathway. Hence, it is expected that there are more mapped results in the pathway than that in other networks. Table 1 shows that more than 20% of the pairs are mapped in the similar mapping methods. Thirty-six pathways are involved in the exact mapping.

Divergent and redundant pathways enable a system to keep functioning if one pathway is blocked as there is an alternative pathway to compensate. For example, transforming growth factor-beta-induced SMAD2- and SMAD3-mediated transcriptional activation can compensate for each other, if either one is inhibited.[6] It is also inappropriate to inhibit the upstream signal that has other downstreams like MAPK signaling other than SMAD2 and SMAD3. Therefore, it is expected that a drug that blocks *both* SMAD2 and SMAD3 would be effective while blocking transforming growth factor-beta signaling.

In the VEGF signaling pathway (Table 3), the binding of VEGF to VEGFR-2 activates PI3K, leading to activation of small GTP-binding protein Rac, which promotes actin reorganization and further regulates cellular migration. Actin reorganization is also regulated through triggering the sequential activation of Cdc42, SAPK2/p38, and HSP27.[34,35] Inhibiting either side of the pathway is not able to totally regulate actin reorganization. CID 15996835, which is not reported in the literature, is able to block both Rac and Cdc42, which belong to the Rho family of GPCRs. Inhibition of both would affect the function of actin reorganization according to the analysis of the KEGG pathway.

MAPK[36,37] mainly consists of three subfamilies, the p38 MAPK, ERK, and JNK/SAPK. The activation of ERK tends to induce cell proliferation, whereas activation of JNK and p38 favors the induction of cell death. Its pathway coordinates cell proliferation, differentiation, and death to maintain homeostasis through a sequential protein kinase cascade which regulates a variety of substrates such as transcriptional factors, cytoskeletal elements, and other protein kinases. The inhibitors of proteins involved in any level could be studied for many diseases, particularly cancer. But as a number of divergent and redundant paths involved make this small-

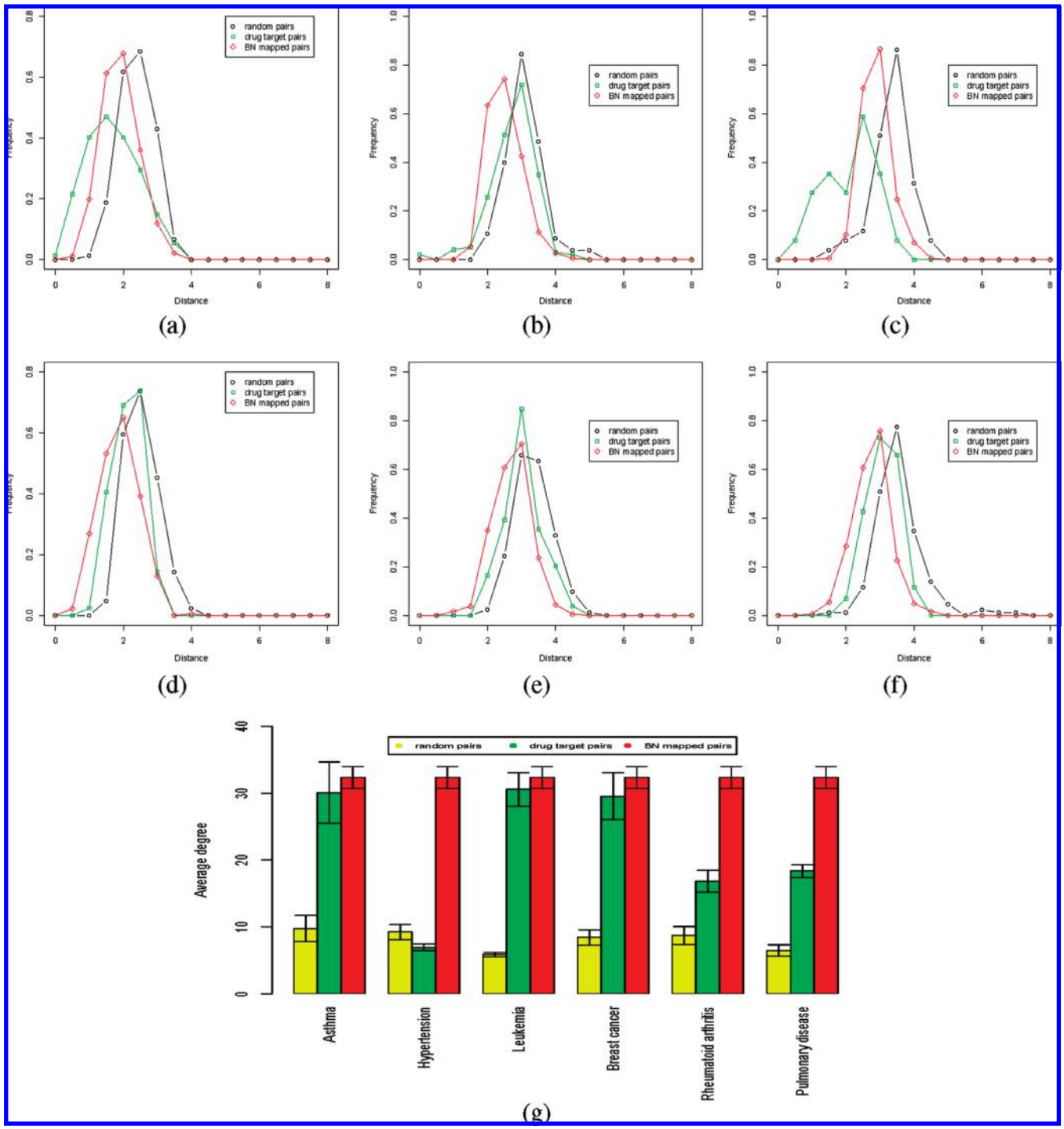

Figure 10. Compare assay network mapped pairs and drug target pairs for particular diseases in PPI. BN means bioassay network. (a) Fractions of the distance between the pairs to leukemia associated proteins. (b) Fractions of the distance between the pairs to hypertension associated proteins. (c) Fractions of the distance between the pairs to asthma associated proteins. (d) Fractions of the distance between the pairs to breast cancer associated proteins. (e) Fractions of the distance between the pairs to rheumatoid arthritis associated proteins. (f) Fractions of the distance between the pairs to pulmonary disease associated proteins. (g) Degree distribution of the pairs for different diseases in PPI.

world network very robust, it restricts to seek for highly selective inhibitors, and its robustness might also lead to drug resistance. Cisplatin is used to treat ovarian cancer, but the development of a resistant cell population limits its efficiency in long-term trials.[37] It is believed that cisplatin-induced cancer cell apoptosis is involved in the activation of the MAPK pathways, but the induction of MKP-1 might also result in cisplatin resistance. It is reported that MKP-1 might be involved in the ERK-MKP-1 signaling pathway that protects cells from cisplatin-induced apoptosis, which leads to cisplatin resistance.[38] It is suggested that targeting ERK-MKP-1 could destroy this pathway and further overcome cisplatin resistance in human ovarian cancer treatment. In our experiment (Table 3), we found CID 573747 can both inhibit ERK2 and MKP-1; it is of interest to use this compound to do further experiments, although there is no literature report for this compound. This compound

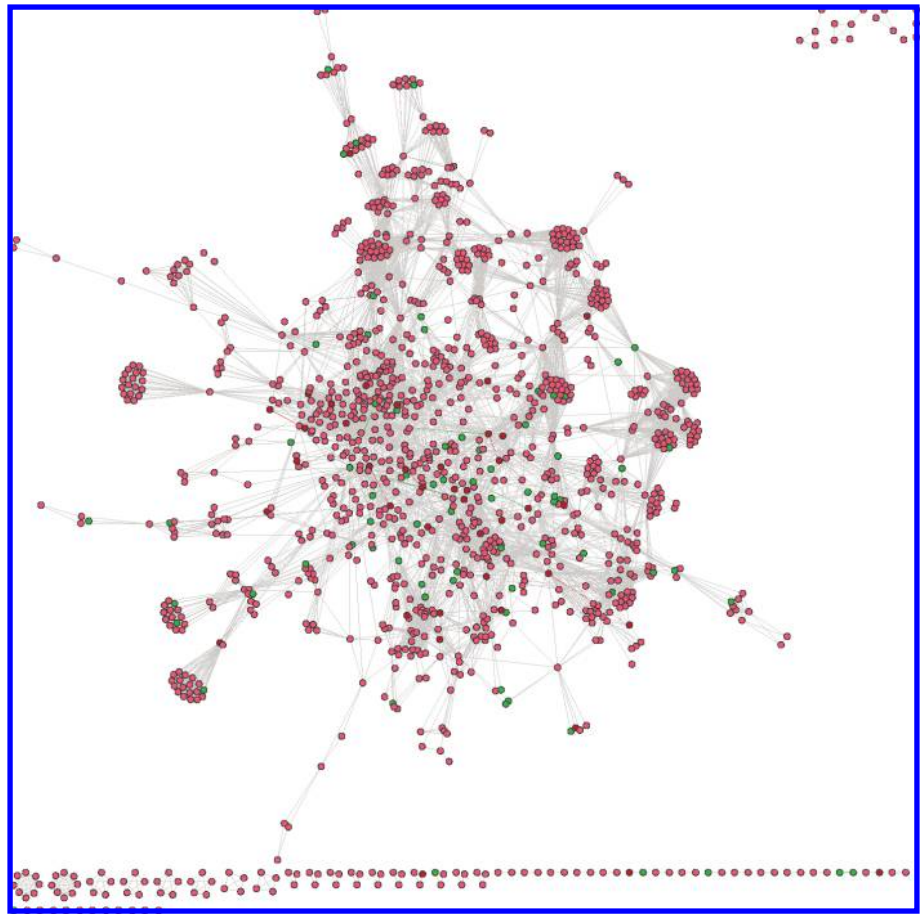

Figure 11. Mapped drug target network. Node represents target; two targets are linked if at least one drug targets both of them. ENMs and EEMs are colored by red, and SNMs and SEMs are colored by green.

Table 3. Examples of Results in Pathway Mapping

target 1	target 2	AID 1	AID 2	one pathway	one common compound (CID)[a]
Cdc42	Rac1	761	1340	VEGF signaling pathway	15996835
HSD11B1	AKR1C4	398	381	C21-steroid hormone metabolism	6419769
MPK1 (ERK2)	DUSP1	995	563	MAPK signaling pathway	573747

[a] Structures are shown in Figure 9.

provides a direction to design inhibitors of both ERK2 and MKP-1 to reduce cisplatin resistance.

In the C21-steroid hormone metabolism pathway (Table 3), HSD11B1 mainly catalyzes the interconversion of cortisone and cortisol that plays a central role in the metabolic syndrome like tetrad of obesity, insulin resistance, and so on. AKR1C4, locating at the downstreams of the pathway, catalyzes many important intermediates. Once AKR1C4 is inhibited, the activitation of steroid hormones will be affected. Carbenoxolone (CID 6419769) could inhibit both enzymes; it is not surprising since one of its well-known targets is HSD11B1 and another proved target 3-alpha-(or 20-beta)-hydroxysteroid dehydrogenase is very similar with AKR1C4. No report shows the consequence of inhibiting *both* HSD11B1 and AKR1C4; however, mapping in this pathway might be able to partially explain why the side effect of carbenoxolone is frequent.[39]

4. DISCUSSION

The current approach to construct the assay network employs the PubChem activity scores, which are known to be noisy. An alternative would be to consider the use of Z-scores or percentiles. Furthermore, our current approach involves an arbitrary activity score cutoff. While this allows us to focus on compounds that are likely true actives, the resultant network can be relatively sparse. To investigate the effect of cutoff we reconstructed the assay network with an activity cutoff of 80 and mapped the assay network to the drug-target network. The resultant mapping was not significantly more detailed than that described previously. One significant drawback of our current approach is that we are limited to assays with a reported protein target. Thus, phenotypic assays and others not involving a specific target cannot be employed. Future work will consider other approaches such as the use of semantic analysis of assay descriptions to derive a "semantic similarity score". A related problem is that while many network databases are available (such as Reactome, Pathway Interaction Database,[40] and KEGG), it is not always trivial to completely map pathways from different databases. A large reason for this is the use of different identifiers for genes and proteins. However, this can be circumvented by prior normalization of network data.

In the current study this is implicitly performed by use of the NCBI protein identifiers. More importantly is the occurrence of noncomparable entities in different networks. For example, KEGG pathways will include chemical compounds as nodes. Clearly, mapping such nodes to nodes in a bioassay network does not make sense. On the other hand, one could consider the mapping of orthologs from a KEGG pathway to nodes in a bioassay network (which would be similar to the use of the semantic similarity score described above). As a result of these issues, care must be taken to preprocess network data for biological networks to normalize identifiers and also ensure that nodes used in the bipartite mapping are actually comparable.

One problem that was not considered in this analysis was the fact that certain compound classes exhibit a preference for certain targets - an example being the preference of dibenzocycloheptadiene derivatives for the dopamine and histamine receptors. Thus, such compounds showing up as active against a variety of assays focusing on such preferred targets is not surprising. Finally, promiscuity is problematic - currently our filters are coarse and likely miss a number of promiscuous compounds. A more sophisticated approach would be to consider a predictive model to filter out promiscuous compounds.[24]

While the bioassay network covers a number of targets, one must keep in mind that the PubChem bioassay collection focuses on specific biological areas. This suggests that the specific assay network we constructed here is not a completely general tool. Given the large number of drug targets, the coverage of targets in the assay network is relatively small. As a result, identity mappings are not very comprehensive. Thus similarity mappings are more useful. Note that even if target coverage in the assay network is expanded, similarity mapping is still valuable as it is unlikely that all possible protein targets will be represented in the assay collection.

When mapping the assay network to the drug target network, we identified a number of cases where assayed compounds were determined to be active against a target, but structurally dissimilar known inhibitors of that target. The approach thus prioritizes new compounds, though since the current approach only considers pairs of targets, it is somewhat biased; many drugs are known to be nonselective and hit multiple targets.

Given that a manual literature search for every pair of targets and their common active compounds is quite tedious, we employed the mapping to the drug targets to reduce the search space - focusing only on some interesting mapped pairs. In general, we observed that the pairs of assay nodes mapped to the PPI network were more significant in terms of centrality than pairs of nodes in the drug target network. In both cases, they were significantly different compared to random pairs.

While talking about polypharmacology, one tends to think of the pairs with very high degree. However, drug targets are not really essential. Our results indicate that it would be more reasonable to consider only pairs with degrees lying in a certain range and with a certain shortest path between two nodes. Comparing their distance to disease causing proteins against the distance of drug target pairs to disease causing proteins also provides a direction to limit the selection of candidates while studying some complex diseases.

5. CONCLUSIONS

Most analyses of PubChem data have focused on individual assays. We have presented a technique that allows cross-assay analyses to be linked to external biological data, using a network construct. Our results indicate this approach could be a useful way to investigate polypharmacology in the PubChem bioassay collection. While the current work does not address a significantly large number of targets, we believe that the mapping procedure coupled to (automated) literature searches could lead to an efficient way to link the wealth of information in the bioassay collection to biological networks.

Supporting Information Available: Scripts required to regenerate the PostgreSQL database. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Kola, I.; Landis, J. Can the Pharmaceutical Industry Reduce Attrition Rates. *Nat. Rev. Drug. Discovery* **2004**, *3*, 711–715.
(2) Hartman, J. L.; Garvik, B.; Hartwell, L. Cell Biology - Principles for the Buffering of Genetic Variation. *Science* **2001**, *291*, 1001–1004.
(3) Albert, R.; Jeong, H.; Barabasi, A. L. Error and Attack Tolerance of Complex Networks. *Nature* **2000**, *406*, 378–382.
(4) Ooi, S. L.; Pan, X. W.; Peyser, B. D.; Ye, P.; Meluh, P. B.; Yuan, D. S.; Irizarry, R. A.; Bader, J. S.; Spencer, F. A.; Boeke, J. D. Global Synthetic-Lethality Analysis and Yeast Functional Profiling. *Trends Biochem. Sci.* **2006**, *22*, 56–63.
(5) Roth, B. L.; Sheffler, D. J.; Kroeze, W. K. Magic Shotguns Versus Magic Bullets: Selectively Non-Selective Drugs for Mood Disorders and Schizophrenia. *Nat. Rev. Drug. Discovery* **2004**, *3*, 353–359.
(6) Keith, C. T.; Borisy, A. A.; Stockwell, B. R. Multicomponent Therapeutics for Networked Systems. *Nat. Rev. Drug. Discovery* **2005**, *4*, 71–U10.
(7) Hopkins, A. L. Network Pharmacology: The Next Paradigm in Drug Discovery. *Nat. Chem. Biol.* **2008**, *4*, 682–690.
(8) Druker, B. J.; Talpaz, M.; Resta, D. J.; Peng, B.; Buchdunger, E.; Ford, J. M.; Lydon, N. B.; Kantarjian, H.; Capdeville, R.; Ohno-Jones, S.; Sawyers, C. L. Efficacy and Safety of a Specific Inhibitor of the BCR-ABL Tyrosine Kinase in Chronic Myeloid Leukemia. *N. Engl. J. Med.* **2001**, *344*, 1031–1037.
(9) Ganter, B.; Giroux, C. N. Emerging Applications of Network and Pathway Analysis in Drug Discovery and Development. *Curr. Opin. Drug Discovery Dev.* **2008**, *11*, 86–94.
(10) Yu, H.; Kim, P. M.; Sprecher, E.; Trifonov, V.; Gerstein, M. The Importance of Bottlenecks in Protein Networks: Correlation with Gene Essentiality and Expression Dynamics. *PLoS Comput. Biol.* **2007**, *3*, 713–720.
(11) Csermely, P.; Agoston, V.; Pongor, S. The Efficiency of Multi-Target Drugs: The Network Approach Might Help Drug Design. *Trends Pharmacol. Sci.* **2005**, *26*, 178–182.
(12) Paolini, G. V.; Shapland, R. H. B.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. Global Mapping of Pharmacological Space. *Nat. Biotechnol.* **2006**, *24*, 805–815.
(13) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating Protein Pharmacology by Ligand Chemistry. *Nat. Biotechnol.* **2007**, *25*, 197–206.
(14) Nidhi; Glick, M.; Davies, J. W.; Jenkins, J. L. Prediction of Biological Targets for Compounds Using Multiple-Category Bayesian Models Trained on Chemogenomics Databases. *J. Chem. Inf. Model.* **2006**, *46*, 1124–1133.
(15) Campillos, M.; Kuhn, M.; Gavin, A. C.; Jensen, L. J.; Bork, P. Drug Target Identification Using Side-Effect Similarity. *Science* **2008**, *321*, 263–266.
(16) Morphy, R.; Rankovic, Z. Fragments, Network Biology and Designing Multiple Ligands. *Drug Discovery Today* **2007**, *12*, 156–160.
(17) Yildirim, M. A.; Goh, K. I.; Cusick, M. E.; Barabasi, A. L.; Vidal, M. Drug-Target Network. *Nat. Biotechnol.* **2007**, *25*, 1119–1126.
(18) PubChem. The PubChem Project; 2009. http://pubchem.ncbi.nlm.nih.gov(accessed July 1, 2009).
(19) Austin, C. P.; Brady, L. S.; Insel, T. R.; Collins, F. S. NIH Molecular Libraries Initiative. *Science* **2004**, *306*, 1138–1139.
(20) Feng, B. Y.; Shelat, A.; Doman, T. N.; Guy, R. K.; Shoichet, B. K. High-Throughput Assays for Promiscuous Inhibitors. *Nat. Chem. Biol.* **2005**, *1*, 146–148.

(21) Babaoglu, K.; Simeonov, A.; Irwin, J. J.; Nelson, M. E.; Feng, B.; Thomas, C. J.; Cancian, L.; Costi, M. P.; Maltby, D. A.; Jadhav, A.; Inglese, J.; Austin, C. P.; Shoichet, B. K. Comprehensive Mechanistic Analysis of Hits from High-Throughput and Docking Screens against β-Lactamase. *J. Med. Chem.* **2008**, *51*, 2502–2511.
(22) Shoichet, B. K. Interpreting Steep Dose-Response Curves in Early Inhibitor Discovery. *J. Med. Chem.* **2006**, *49*, 7274–7277.
(23) Seidler, J.; McGovern, S. L.; Doman, T. N.; Shoichet, B. K. Identification and Prediction of Promiscuous Aggregating Inhibitors Among Known Drugs. *J. Med. Chem.* **2003**, *46*, 4477–4486.
(24) Roche, O.; et al. Development of a Virtual Screening Method for Identification of "Frequent Hitters" in Compound Libraries. *J. Med. Chem.* **2002**, *45*, 137–142.
(25) Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: A Comprehensive Resource for in Silico Drug Discovery and Exploration. *Nucleic Acids Res.* **2006**, *34*, D668–D672.
(26) Mishra, G. R.; et al. Human Protein Reference Database - 2006 Update. *Nucleic Acids Res.* **2006**, *34*, D411–D414.
(27) Mathivanan, S.; Periaswamy, B.; Gandhi, T. K. B.; Kandasamy, K.; Suresh, S.; Mohmood, R.; Ramachandra, Y. L.; Pandey, A. An Evaluation of Human Protein-Protein Interaction Data in the Public Domain. *BMC Bioinf.* **2006**, *7*, xx.
(28) Ma, H.; Goryanin, I. Human Metabolic Network Reconstruction and its Impact on Drug Discovery and Development. *Drug Discovery Today* **2008**, *13*, 402–408.
(29) Kanehisa, M.; Araki, M.; Goto, S.; Hattori, M.; Hirakawa, M.; Itoh, M.; Katayama, T.; Kawashima, S.; Okuda, S.; Tokimatsu, T.; Yamanishi, Y. KEGG for Linking Genomes to Life and the Environment. *Nucleic Acids Res.* **2008**, *36*, D480–D484.
(30) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13*, 2498–2504.
(31) Albert, R.; Barabasi, A. L. Statistical Mechanics of Complex Networks. *Rev. Mod. Phys.* **2002**, *74*, 47–97.
(32) Whitehead, S. A.; Cross, J. W.; Burden, C.; Lacey, M. Acute and Chronic Effects of Genistein, Tyrphostin and Lavendustin A on Steroid Synthesis in Luteinized Human Granulosa Cells. *Hum. Reprod.* **2002**, *17*, 589–594.
(33) Coupet, J.; Fisher, S. K.; Rauh, C. E.; Lai, F.; Beer, B. Interaction of Amoxapine with Muscarinic Cholinergic Receptors - an in Vitro Assessment. *Eur. J. Pharmacol.* **1985**, *112*, 231–235.
(34) Hoeben, A.; Landuyt, B.; Highley, M. S.; Wildiers, H.; Van Oosterom, A. T.; De Bruijn, E. A. Vascular Endothelial Growth Factor and Angiogenesis. *Pharmacol. Rev.* **2004**, *56*, 549–580.
(35) Lamalice, L.; Houle, F.; Jourdan, G.; Huot, J. Phosphorylation of Tyrosine 1214 on VEGFR2 is Required for VEGF-induced Activation of Cdc42 Upstream of SAPK2/p38. *Oncogene* **2004**, *23*, 434–445.
(36) Chen, Z.; Gibson, T. B.; Robinson, F.; Silvestro, L.; Pearson, G.; Xu, B. E.; Wright, A.; Vanderbilt, C.; Cobb, M. H. MAP Kinases. *Chem. Rev.* **2001**, *101*, 2449–2476.
(37) Wang, D.; Boerner, S. A.; Winkler, J. D.; LoRusso, P. M. Clinical Experience of MEK Inhibitors in Cancer Therapy. *Biochim. Biophys. Acta* **2007**, *1773*, 1248–1255.
(38) Wang, J.; Zhou, J. Y.; Wu, G. S. ERK-Dependent MKP-1-Mediated Cisplatin Resistance in Human Ovarian Cancer Cells. *Cancer Res.* **2007**, *67*, 11933–11941.
(39) Andrews, R. C.; Rooyackers, O.; Walker, B. R. Effects of the 11 Beta-hydroxysteroid Dehydrogrenase Inhibitor Carbenoxolone on Insulin Sensitivity in Men with Type 2 Diabetes. *J. Clin. Endocrinol. Metab.* **2003**, *88*, 285–291.
(40) Schaefer, C.; Anthony, K.; Krupa, S.; Buchoff, J.; Day, M.; Hannay, T.; Buetow, K. PID: The Pathway Interaction Database. *Nucleic Acids Res.* **2009**, *37*, D674–9.

CI9001876