JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

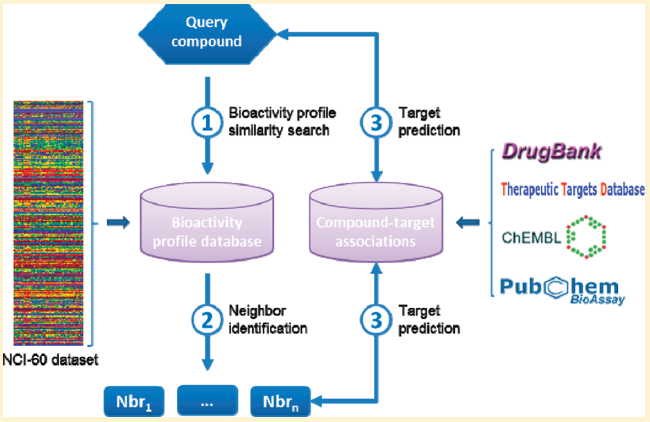
# Identifying Compound-Target Associations by Combining Bioactivity Profile Similarity Search and Public Databases Mining

Tiejun Cheng,[†] Qingliang Li,[†] Yanli Wang,[*,†] and Stephen H. Bryant[*,†]

[†]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, Maryland 20894, United States

Supporting Information

**ABSTRACT:** Molecular target identification is of central importance to drug discovery. Here, we developed a computational approach, named bioactivity profile similarity search (BASS), for associating targets to small molecules by using the known target annotations of related compounds from public databases. To evaluate BASS, a bioactivity profile database was constructed using 4296 compounds that were commonly tested in the US National Cancer Institute 60 human tumor cell line anticancer drug screen (NCI-60). Each compound was used as a query to search against the entire bioactivity profile database, and reference compounds with similar bioactivity profiles above a threshold of 0.75 were considered as neighbor compounds of the query. Potential targets were subsequently linked to the identified neighbor compounds by using the known targets of the query compound. About 45% of the predicted compound-target associations were successfully verified retrospectively, suggesting the possible application of BASS in identifying the targets of uncharacterized compounds and thus providing insight into the study of promiscuity and polypharmacology. Furthermore, BASS identified a significant fraction of structurally diverse compounds with similar bioactivities, indicating its feasibility of "scaffold hopping" in searching novel molecules against the target of interest.




## INTRODUCTION

A number of bioactive compounds of current interest are discovered by phenotypic screening,[1,2] most of which are functional in nature through analyzing the compound-induced effects in cells, tissues, and model organisms. These assays, however, can hardly provide immediate target information for tested compounds, imposing grand challenges on follow-up target identification for drug discovery.[3−5] The recent findings that many drugs act on multiple physiological targets to exert therapeutic effects and/or side effects have attracted intensive interest in exploring the promiscuity and polypharmacology of drugs,[6,7] in which identifying compound-target associations is a premise.

Experimentally, two major techniques are used for target identification.[3] Direct techniques, such as affinity chromatography[8,9] and protein microarray,[10] detect the binding of a compound to its target. Their applications are often hampered by the need to label a compound without affecting its functionality. Indirect techniques infer targets from the compound-induced cellular or physiological patterns through genomics,[11,12] proteomics,[13] metabolite profiling,[14] and other technologies. However, genome-wide or proteome-wide data could be very difficult and expensive to obtain.

Moreover, wet-lab experiments for target identification are often slow, whereas computational approaches can be efficient complements.[15] For example, molecular modeling studies have been reported for target prediction by virtually docking a compound of interest to a list of potential targets with known three-dimensional (3D) structures.[16,17] The primary limitation of this method is the need for high-resolution 3D structures of targets as well as accurate docking/scoring algorithms.[18,19] Statistical models also have been built for target prediction employing various machine learning methods including Bayesian analysis[20,21] and Support Vector Machines.[22] The common drawbacks of these models are that the real predictability beyond training space cannot always be guaranteed. In addition, the similarity principle,[23,24] despite its exceptions,[25] has been the basis for target identification using similarity metrics such as ligand chemical similarity[5,7,26] and drug side effects similarity.[4] On the other hand, with the rapid growth of public biological databases, such as the Protein Data Bank[27] (PDB), PubChem,[28] ChEMBL (http://www.ebi.ac.uk/chembl), DrugBank,[29,30] and Therapeutic Targets Database[31,32] (TTD), abundant bioactivity data of small molecules and their targets are now available to the entire research community. It is thus getting critical to develop *in silico* methods to identify compound-target associations and

**Received:** April 29, 2011
**Published:** August 11, 2011

infer targets for drugs and bioactive compounds by aggregating and integrating valuable target information from multiple resources.

End points of bioactivity data obtained from a panel of assays (i.e., bioactivity profile) may provide distinct insight to the biological function of compounds and their targets. For example, the COMPARE algorithm,[33] by the Developmental Therapeutics Program (DTP) of the US National Cancer Institute (NCI), could be used to suggest possible mechanism of action for a respective compound from related compounds or identify novel compounds that act by a similar mechanism of interest.[34−36] This tool compares the bioactivity patterns derived from the anticancer drug screening data across 60 human tumor cell lines (commonly known as the NCI-60 data set). By incorporating additional gene expression data, target information may be inferred.[34]

The NCI-60 data set was also used in our previous work,[37] where we observed in a few model systems that the target networks of small molecules were well-correlated with their bioactivity profiles. Here, given the rapid growth in available compound-target annotations in several public databases, we further investigated whether such correlations could be utilized to benefit the identification of new targets for drugs and bioactive compounds on a larger scale. To this end, we first constructed a database of bioactivity profiles for 4296 compounds tested in the NCI-60 data set. Second, we used each compound as a query to search against the entire bioactivity profile database to identify neighbor compounds with similar bioactivity profiles. Third, we collected target information from four public databases (DrugBank, TTD, ChEMBL and PubChem) for both query compounds and their neighbor compounds to evaluate our approach for predicting compound-target associations. The underlying assumption is that compounds with similar bioactivity profiles may share common targets. We were able to verify a remarkable portion of our predictions retrospectively.

## METHODS

**Construction of Bioactivity Profile Database.** The NCI-60 data set contains anticancer screening results for more than 40,000 compounds. It is publicly available in the PubChem BioAssay database[38] as 73 bioassays with the name of "*NCI human tumor cell line growth inhibition assay*" under the "*DTP/NCI*" data source. In this work, only the top 60 bioassays (referred hereafter as NCI-60) with the largest number of tested compounds were selected (Supporting Information, Table S1). Relevant bioactivity data were downloaded at the PubChem FTP site (ftp://ftp.ncbi.nlm.nih.gov/pubchem/Bioassay, accessed on December 9, 2010). A total of 5083 compounds were found commonly tested in all of the 60 bioassays. The bioactivity profile of each compound was derived by extracting the log($GI_{50}$) values obtained from the NCI-60 cell lines, where $GI_{50}$ is the concentration required for the 50% growth inhibition of tumor cells. 631 compounds with missing log($GI_{50}$) value in one or more of the NCI-60 cell lines were discarded. Additionally, 156 compounds were further discarded, because they exhibited identical bioactivity in all NCI-60 cell lines, which made them less informative and unsuitable for bioactivity profile similarity calculation (see below). As a result, 4296 compounds were collected and used for constructing the bioactivity profile database. The original bioactivity profile data for these compounds are available in Supporting Information, Table S2. Additional data set characteristics are summarized in Supporting Information, Figure S1 with respect to six physiochemical properties: molecular weight, octanol−water partition coefficient,[23] number of hydrogen bond donors, number of hydrogen bond acceptors, number of rotatable bonds, and topological polar surface area.

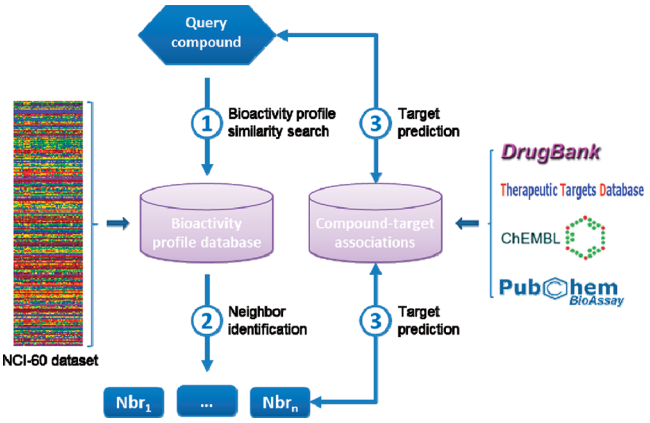



**Figure 1.** Schematic overview of the bioactivity profile similarity search (BASS) strategy. Target prediction can be bidirectional, that is, the known targets of a query compound can be predicted as the potential targets of its neighbor compound, or vice versa.

**BioActivity Profile Similarity Search (BASS).** The BASS approach consists of three major steps (Figure 1). For a given query compound in the NCI-60 data set, we first searched against the entire bioactivity profile database and calculated pairwise bioactivity profile similarity for each reference compound in the data set and the query compound. Second, a neighbor compound was identified if its bioactivity profile similarity is above a selected threshold. Finally, the known target of the query compound is predicted as the potential target of its neighbor compounds or vice versa. A critical step of BASS is to identify the neighbor compounds for a given query compound based on the similarity of bioactivity profiles ($Sim_{bio}$), which is defined as Pearson correlation coefficient ($R_p$)

$$Sim_{bio} = R_p = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \quad (1)$$

where $N$ equals 60, and $x_i$ and $y_i$ are the log($GI_{50}$) values of the $i^{th}$ NCI-60 cell line for compound $Q$ and compound $S$, respectively. In this work, $S$ is considered as a neighbor compound of $Q$ when $Sim_{bio}$ is above 0.75. This similarity threshold was chosen based on a statistical test, which was carried out by randomly selecting two compounds from the entire bioactivity profile database for 100,000 times and recording each time the bioactivity profile similarity. A probability ($p$-value) was subsequently calculated for obtaining a bioactivity profile similarity above a certain threshold. For the similarity threshold of 0.75 ($p$-value = 2.28e-3), we found a good balance between prediction accuracy and the number of predictions.

**Compilation of Target Information.** Target annotations for all the compounds in the bioactivity profile database were primarily collected from four public databases: DrugBank, TTD, ChEMBL, and PubChem. For DrugBank (http://www.drugbank.ca) and TTD (http://bidd.nus.edu.sg/group/cjttd/TTD_HOME.asp),

**Table 1. Summary of the 237 Compounds with Target Annotations in Relevant Public Databases**

| | DrugBank | TTD | ChEMBL | PubChem |
|---|---|---|---|---|
| no. of compounds | 28 | 33 | 23 | 215 |
| no. of target annotations | 44 | 50 | 67 | 1046 |

compound-target associations were downloaded from original Web sites (both accessed on December 9, 2010). For ChEMBL, the mirrored version of ChEMBL_08 in PubChem was used (http://pubchem.ncbi.nlm.nih.gov, accessed on December 9, 2010), and we considered a compound-target association when a respective compound exhibited an effective activity concentration $\leq 1\,\mu$M against its directly assigned target. For PubChem, the bioactivity outcome specifications from original bioassay depositors were adopted to establish compound-target associations. Additionally, we also manually collected the target annotations for a number of compounds from precedent literatures using the 'Literature Keyword Mining Tool' provided at PubChem. From a list of MeSH terms (http://www.ncbi.nlm.nih.gov/mesh) returned by this tool, we looked into the most relevant ones to the compound and/or target of interest and then followed the links to full-text literature and extracted evidence therein whenever possible. All protein targets were uniformly stored as UniProtKB identifiers (http://www.uniprot.org, accessed on February 4, 2011). Other molecular targets, such as DNA and RNA, were stored as target names. As a result, 237 compounds with known target annotations in one or more of the above four databases were identified (Table 1).

## RESULTS

**Evaluation of BASS for Target Identification.** Using the above 237 compounds with known target annotations as queries, BASS predicted a total of 4693 compound-target associations for neighbor compounds, i.e., the known targets of a respective query compound were considered as the potential targets of its neighbor compounds. In this work, if at least one potential target was also annotated in any of the above four databases, a successful prediction of the compound-target association was counted. It should be noted that only a part of such predictions could be evaluated when both query compound and neighbor compound had target annotations available. 634 out of the 4693 compound-target associations turned out to be verifiable. For a systematic evaluation of the predicted associations, a stringent criterion was first used by checking the identity of targets of the query compound and its neighbor compound. As a result, a success rate of 44.8% (284 successful predictions) was achieved, which accounted for 103 out of the 237 query compounds. When the identified targets were proteins and there was no exact match among that of a respective compound and its neighbor compound, a less stringent criterion of target identity was applied if protein target sequences were significantly related. In this work, two protein targets that showed an E-value <1e-12 in the BLAST[39] protein−protein sequence alignment were considered as biologically related. Under these conditions, the performance was further improved to 48.6% (308 predictions in total), which accounted for 108 out of the 237 query compounds. The above evaluation suggested that BASS, when combined with searching target information using public databases, may be used to identify targets for biological neighbor compounds with similar bioactivity profiles to a query compound. Detailed results are described for the following examples, with the complete results provided in Supporting Information, Table S3.

**Microtubule as a New Target.** Microtubules are composed of α- and β-tubulin heterodimers. They are cytoskeletal elements involved in many cellular processes, such as mitosis, cytokinesis, and vesicular transport.[40−42] Small molecules that bind to tubulin can interfere with microtubule dynamics, resulting in microtubule stabilization or destabilization, which induces cell cycle arrest and ultimately leads to apoptosis. Out of the 15 new molecular entities approved by FDA in 2010, two are targeting microtubule.[43] Considering its key roles in mitosis and cell division, microtubule continues to be a very important chemotherapeutic target of anticancer drugs.[44]

According to DrugBank (primary accession number, PAN: DB01229), Paclitaxel (PubChem Compound identifier, CID: 36314) is an FDA-approved drug to treat various cancers, including ovarian cancer and breast cancer. It promotes the assembly of microtubules from tubulin dimers and stabilizes them by preventing depolymerization. In this work, using Paclitaxel as a query for BASS retrieved seven neighbor compounds (Figure 2A). These included five closely related analogues of Paclitaxel, showing an average two-dimensional chemical similarity ($Sim_{chem}$) of 0.924 as characterized by PubChem fingerprint (ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt) and Tanimoto score.[45] This is consistent with previous observations that structurally similar compound may exhibit comparative bioactivities.[46,47] However, due to limited target annotations available to us at the time, we were not able to verify tubulin as a target for these structural analogs.

On the other hand, tubulin was verified as a target for one neighbor compound Vinblastine (CID: 241902; $Sim_{bio}$ = 0.785; *p*-value = 1.31e-3) which was structurally unrelated to Paclitaxel ($Sim_{chem}$ = 0.560, Figure 2A). Vinblastine is an approved anticancer drug (PAN: DB00570) which is thought to play a key role in mitosis inhibition at metaphase via its interaction with tubulin. The crystal structure of Vinblastine-tubulin complex reveals that Vinblastine binds at the interface between two tubulin heterodimers,[40] in contrast to Paclitaxel which binds at the taxol site of β-tubulin.[42] Furthermore, using Vinblastine as a query, BASS identified a number of neighbor compounds that were common to those of Paclitaxel. Interestingly, this second search identified two additional neighbor compounds which were previously reported as tubulin inhibitors (CID: 249332[48,49] and 347381;[50] $Sim_{bio}$ = 0.753 and 0.756; *p*-value = 2.25e-3 and 2.12e-3; $Sim_{chem}$ = 0.984 and 0.526, respectively). In addition, BASS identified another non-Paclitaxel neighbor compound NSC355256 (CID: 434718; $Sim_{bio}$ = 0.789; *p*-value =1.14e-3; $Sim_{chem}$ = 0.671) using Paclitaxel as a query (Figure 2A). Due to limited target annotation available to us, we were unable to verify tubulin as a target for this compound. However, we noticed that it shared the chemical scaffold of an approved drug Colchicine (PAN: DB01394; CID: 6167) with a significant structural similarity ($Sim_{chem}$ = 0.878). As indicated by the crystal structure of Colchicine-tubulin complex, Colchicine binds to the β-tubulin subunit of microtubule at the interface with α-tubulin.[41] This example indicated that BASS had the potential to discover novel inhibitors and explore new starting points for lead optimization, demonstrating the advantage of BASS for identifying compounds with various chemical scaffolds, which may provide insight to 'scaffold hopping' against the target of interest.[51,52]

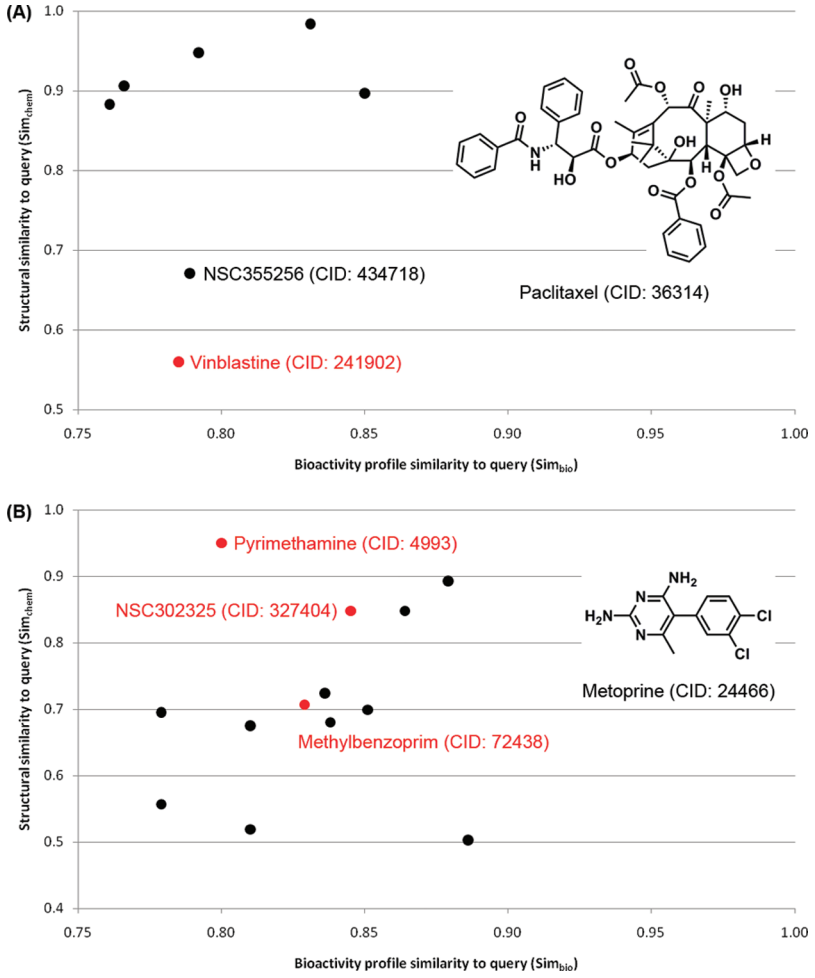



**Figure 2.** Chemical similarity as a function of biological similarity for the neighbor compounds of (A) Paclitaxel and (B) Metoprine retrieved by bioactivity profile similarity search, respectively. The red-labeled neighbor compounds were those verified to share the common target with Paclitaxel (microtubule) and Metoprine (dihydrofolate reductase), respectively.

**Dihydrofolate Reductase As a New Target.** In the above example, we demonstrated that the targets of biological neighbor compounds could be inferred from the known targets of a drug molecule. It would be more practical and interesting to investigate, from a reverse perspective, whether BASS could be used to suggest new targets for a drug molecule by gathering known target information from its neighbor compounds (Figure 1). Dihydrofolate reductase (DHFR) converts dihydrofolate into tetrahydrofolate. The latter is a methyl group shuttle required for the *de novo* biosynthesis of purines, thymidylates, and certain amino acids, which are essential for DNA synthesis and cell multiplication.[53]

In this example, we used the experimental drug Metoprine (CID: 24466) as a query. According to DrugBank (PAN: DB04655), its annotated target is Histamine N-methyltransferase (HNMT).[54] For its 13 neighbor compounds identified by BASS (Figure 2B), none was found targeting HNMT according to the available target annotations. On the other hand, further investigation indicated that three neighbor compounds, Pyrimethamine (CID: 4993), NSC302325 (CID: 327404), and Methylbenzoprim (CID: 72438[55,56]) had been previously reported targeting DHFR. According to DrugBank (PAN: DB00205), Pyrimethamine was an FDA-approved antimalarial drug through a mode of action by inhibiting DHFR.[53] Based on ChEMBL annotation (PubChem BioAssay identifier, AID: 55830), NSC302325 was a DHFR inhibitor with an $IC_{50}$ of 0.85 $\mu$M.[57] The direct annotation of DHFR as a target of Methylbenzoprim was not available in any of the above four databases. However, its annotated target in ChEMBL (AID: 56179 and 56314), bifunctional dihydrofolate reductase-thymidylate synthase (DHFR-TS), was found to be closely related with DHFR (BLAST E-value = 6e-136). The binding of Methylbenzoprim to DHFR was further supported by previous NMR experiments[55] as well as molecular modeling studies.[56] Using either one of the three compounds Pyrimethamine, NSC302325 and Methylbenzoprim as a query, BASS could identify Metoprine as a neighbor compound ($Sim_{bio}$ = 0.800, 0.845, and 0.829; *p*-value =9.4e-4, 3.7e-4, and 4.8e-4, respectively). Moreover, all three compounds were structurally related to the query Metoprine ($Sim_{chem}$ = 0.950, 0.707, and 0.848, respectively). Therefore, it is natural to consider DHFR as a potential target of Metoprine, which was confirmed by further investigation into the target annotation in TTD (DrugID: DCL000304) and precedent literatures.[58−60]

**Predicting Polypharmacology.** Polypharmacology is receiving increasing attention in drug discovery for exploring both side effects and new therapeutic opportunities.[61] As a step forward, BASS can be readily applied for predicting the polypharmacology of a given compound by collecting known targets from its

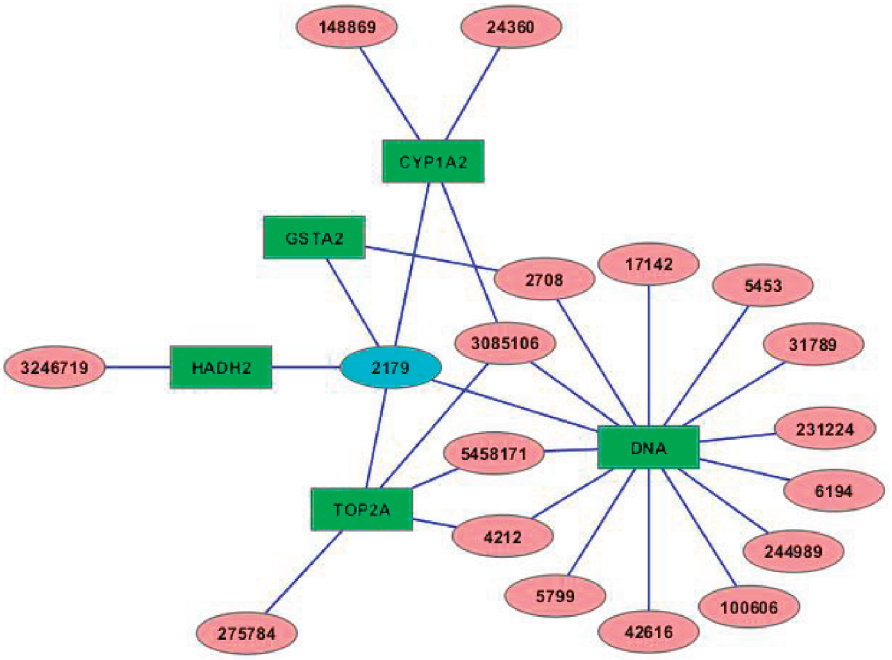



**Figure 3.** Polypharmacology of Amsacrine. Compounds (labeled with PubChem compound identifier, CID) and targets are denoted as ellipses and rectangles, respectively. The edge linking between indicates that there is a compound-target association. The query compound Amsacrine (CID: 2179) is colored with cyan. DNA: DNA; TOP2A: topoisomerase, type II alpha; HADH2: hydroxyacyl-coenzyme A dehydrogenase, type II; GSTA2: glutathione S-transferase A2; CYP1A2: cytochrome P450, family 1, subfamily A, polypeptide 2.

neighbor compounds. Here, we presented such an example using the approved drug Amsacrine (CID: 2179) as a query (Figure 3). A total of 67 neighbor compounds were identified by BASS. More than a dozen of them were known DNA intercalators or cross-linkers according to DrugBank annotations and/or precedent literatures. There were also several neighbor compounds that were previously reported as inhibitors of topoisomerase, type II alpha (TOP2A). The two targets of DNA and TOP2A were also annotated for Amsacrine in DrugBank (PAN: DB00276). Additionally, Amsacrine together with one neighbor compound (CID: 2708; PAN: DB00291) were confirmed to interact with the enzyme glutathione S-transferase A2 (GSTA2). In a quantitative high-throughput screening assay (AID: 886) launched by the US National Institutes of Health Chemical Genomics Center (NCGC), both Amsacrine and its neighbor compound (CID: 3246719) demonstrated inhibitory activity against hydroxyacl-coenzyme A dehydrogenase, type II (HADH2). In another bioassay (AID: 410) conducted by NCGC, Amsacrine and two neighbor compounds (CID: 24360 and 148869) were both found active against cytochrome P450, family 1, subfamily A, polypeptide 2 (CYP1A2). Therefore, it is straightforward to depict a polypharmacological graph of Amsacrine by gathering available target information predicted from its neighbor compounds (Figure 3).

## ■ DISCUSSION

The promising results from the overall evaluation of the predicted compound-target associations and those shown in the above examples demonstrated that bioactivity profile similarity search (BASS) may be applied to predict new targets for drugs and bioactive compounds from the target annotations of their neighbor compounds that are available in public databases. Nevertheless, for a larger number of target predictions, we were not able to verify them due to insufficient target annotations in public databases or due to difficulty in literature searching. It thus remains interesting for further (experimental) studies to verify the targets predicted here, especially for those resulting from significant bioactivity profile similarity. For those completely uncharacterized bioactive compounds, BASS may also be helpful to target identification by suggesting potential targets aggregated from their biological neighbor compounds. To facilitate the readers of interest, we included a list of query compounds which yet have no target annotation in any of the above four public databases or precedent literatures and their neighbor compounds with known target annotations (Supporting Information, Table S4).

It should be mentioned that the compound-target associations identified in this work were verified retrospectively by taking advantage of the target annotations derived from public databases or by literature searching, and we emphasize that this work could not have been done without the open access to public databases which now contain vast amount of chemical biology data. For a number of cases (e.g., microtubule example), the predictions were strongly convincing as supported by the crystal structures of ligand-target complexes. Nevertheless, for other cases, the reported compound-target annotations in relevant databases or literatures may require further investigation to better understand the underlying mechanism of binding. For example, though Paclitaxel and Vinblastine both bind to microtubule, they actually bind at very different sites, which may be responsible for their different modes of action. To address these issues, structural biology studies, such as NMR or X-ray diffraction experiments, would be particularly persuasive. With the growing availability of public databases containing ligand-target annotations, such as DrugBank, TTD, ChEMBL, and PubChem, the accuracy of BASS may be further improved.

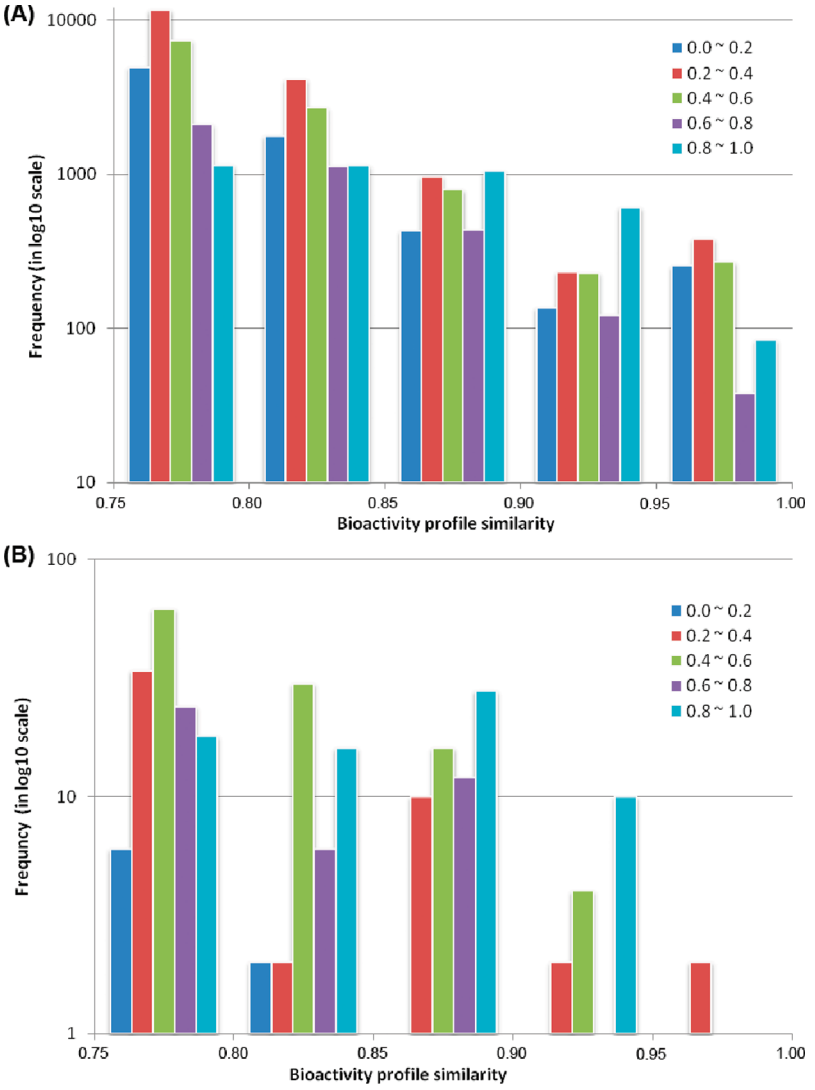



**Figure 4.** The number of neighbor compounds within certain range of chemical similarity as a function of bioactivity profile similarity for the query-neighbor pairs of (A) the 2335 query compounds and their neighbor compounds (44,368 in total) and (B) the 284 verified target predictions, respectively. The five columns (from left to right) in each bin of bioactivity profile similarity correspond to the chemical similarity range of [0.0−0.2), [0.2−0.4), [0.4−0.6), [0.6−0.8), and [0.8−1.0], respectively.

Lead optimization based on chemical scaffold has been broadly embraced by medicinal chemists[62] as a central guiding principle to design ligands with higher potency and/or more desirable physicochemical properties.[63] It will be interesting to look into the chemical space of the neighbor compounds identified by BASS (Figure 4 and Supporting Information, Figure S2). Figure 4A shows the number of neighbor compounds within certain range of chemical similarity as a function of bioactivity profile similarity using all query-neighbor pairs in the bioactivity profile database. As one can see, BASS was able to identify not only structurally similar neighbor compounds but also a considerable number of structurally dissimilar ones with related bioactivities. These may provide novel molecules or new starting points for future ligand design, which would not have been discovered by conventional medicinal chemistry efforts. Therefore, BASS has the appealing capability of 'scaffold hopping', as demonstrated in the above microtubule example and the data shown in Figure 4B as a whole. It thus represents a new strategy for identifying candidate compounds with diverse chemical scaffolds that are biologically relevant to the aimed target. It is worth stressing that the threshold of bioactivity profile similarity for defining a neighbor compound is user adjustable, though a conservative threshold of 0.75 was used in this work. In fact, when a less stringent threshold of 0.70 ($p$-value = 5.22e-3) or lower was applied in BASS, we could still verify a number of predictions.

The idea of using bioactivity profile (pattern or fingerprint) is, of course, not entirely new. Other similar ideas have been proposed. Nevertheless, computational approaches making use of different profiling data may vary, in particular, toward achieving different research goals. For example, the "Connectivity Map" approach developed by Lamb et al.[12] employs mRNA expression profiles to establish connections between small molecules and with diseases. The "biospectra analysis" approach by Fliri et al.[64,65] aims to group compounds with related inhibitory bioactivities against a panel of protein targets and correlate to biological functions. Our specific goal in this work is to associate compounds with targets based on the similar NCI-60 cell lines

bioactivity profiles of small-molecule compounds and their target annotations in public databases. We anticipate that BASS may be of benefit to the target identification for anticancer drug discovery. Analysis using BASS could generate hypothesis to understand both the mode of action and mechanism of binding for bioactive compounds by suggesting new targets from well-characterized neighbor compounds. Our work could contribute to the target prediction and the state-of-art drug repositioning. The free-of-charge screening service provided by the DTP/NCI would make BASS more appealing. By submitting their own compounds of interest, researchers could obtain high-quality and confidential bioactivity profiles, which in turn can be used as inputs for BASS to identify potential targets by consulting the known targets of compounds in the bioactivity profile database. Nevertheless, before additional experiments are done, it should be mentioned that BASS may only be applicable for identifying the targets of the compounds which can cause cellular responses in the NCI-60 cell lines.

## CONCLUSIONS

We have presented a computational approach, BASS, for mutually identifying compound-target associations by comparing the bioactivity profiles that are derived from the NCI-60 cell lines. When two compounds share similar bioactivity profiles, the targets of either compound may be considered as the potential targets for the other compound. To evaluate BASS, each compound in the bioactivity profile database was used as a query to search against the entire database for neighbor compounds that may share common targets. An overall success rate of 44.8% was achieved for the predicted compound-target associations by using the prior knowledge of target annotations from public databases, and it was further improved to nearly 50% when considering related protein targets. Analysis shows that BASS not only could identify structurally similar bioactive compounds that are biological relevant to the target of interest but also had the power of suggesting novel chemical scaffolds for the aimed target. Moreover, BASS may represent an efficient strategy for integrating experimental data and target information newly emerged for any of the neighbor compounds. Therefore, BASS may be applied to suggest new targets for old drugs and provide insight into anticancer drug discovery, facilitating the study of the toxicity, promiscuity, and polypharmacology of drugs and bioactive compounds.

## ASSOCIATED CONTENT

**Supporting Information.** The 60 bioassays (NCI-60) used in this study. The original data of bioactivity profile for 4296 compounds. The complete results of the 284 predicted compound-target associations using the 237 compounds with known targets as queries. The predicted compound-target associations for the query compounds which have no target annotations. The characteristics of the data set used in this study. The global view of chemical similarity as a function of bioactivity profile similarity. This material is available free of charge via the Internet at http://pubs.acs.org.

## AUTHOR INFORMATION

### Corresponding Author

*E-mail: ywang@ncbi.nlm.nih.gov (Y.W.), bryant@ncbi.nlm.nih.gov (S.H.B.).

## ACKNOWLEDGMENT

We thank the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM) for funding support. We also thank the NIH Fellows Editorial Board (FEB) for manuscript revision.

## REFERENCES

(1) Swinney, D. C.; Anthony, J. How were new medicines discovered? *Nat. Rev. Drug Discovery* **2011**, *10*, 507–519.

(2) Kodadek, T. Rethinking screening. *Nat. Chem. Biol.* **2010**, *6*, 162–165.

(3) Lomenick, B.; Hao, R.; Jonai, N.; Chin, R. M.; Aghajan, M.; Warburton, S.; Wang, J.; Wu, R. P.; Gomez, F.; Loo, J. A.; Wohlschlegel, J. A.; Vondriska, T. M.; Pelletier, J.; Herschman, H. R.; Clardy, J.; Clarke, C. F.; Huang, J. Target identification using drug affinity responsive target stability (DARTS). *Proc. Natl. Acad. Sci.* **2009**, *106*, 21984–21989.

(4) Campillos, M.; Kuhn, M.; Gavin, A.-C.; Jensen, L. J.; Bork, P. Drug Target Identification Using Side-Effect Similarity. *Science* **2008**, *321*, 263–266.

(5) Keiser, M. J.; Setola, V.; Irwin, J. J.; Laggner, C.; Abbas, A. I.; Hufeisen, S. J.; Jensen, N. H.; Kuijer, M. B.; Matos, R. C.; Tran, T. B.; Whaley, R.; Glennon, R. A.; Hert, J.; Thomas, K. L. H.; Edwards, D. D.; Shoichet, B. K.; Roth, B. L. Predicting new molecular targets for known drugs. *Nature* **2009**, *462*, 175–181.

(6) Hopkins, A. L. Predicting promiscuity. *Nature* **2009**, *462*, 167–168.

(7) 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.

(8) Sato, S.-i.; Murata, A.; Shirakawa, T.; Uesugi, M. Biochemical Target Isolation for Novices: Affinity-Based Strategies. *Chem. Biol.* **2010**, *17*, 616–623.

(9) Sleno, L.; Emili, A. Proteomic methods for drug target discovery. *Curr. Opin. Chem. Biol.* **2008**, *12*, 46–54.

(10) Zhu, H.; Bilgin, M.; Bangham, R.; Hall, D.; Casamayor, A.; Bertone, P.; Lan, N.; Jansen, R.; Bidlingmaier, S.; Houfek, T.; Mitchell, T.; Miller, P.; Dean, R. A.; Gerstein, M.; Snyder, M. Global Analysis of Protein Activities Using Proteome Chips. *Science* **2001**, *293*, 2101–2105.

(11) Hughes, T. R.; Marton, M. J.; Jones, A. R.; Roberts, C. J.; Stoughton, R.; Armour, C. D.; Bennett, H. A.; Coffey, E.; Dai, H.; He, Y. D.; Kidd, M. J.; King, A. M.; Meyer, M. R.; Slade, D.; Lum, P. Y.; Stepaniants, S. B.; Shoemaker, D. D.; Gachotte, D.; Chakraburtty, K.; Simon, J.; Bard, M.; Friend, S. H. Functional Discovery via a Compendium of Expression Profiles. *Cell* **2000**, *102*, 109–126.

(12) Lamb, J.; Crawford, E. D.; Peck, D.; Modell, J. W.; Blat, I. C.; Wrobel, M. J.; Lerner, J.; Brunet, J.-P.; Subramanian, A.; Ross, K. N.; Reich, M.; Hieronymus, H.; Wei, G.; Armstrong, S. A.; Haggarty, S. J.; Clemons, P. A.; Wei, R.; Carr, S. A.; Lander, E. S.; Golub, T. R. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science* **2006**, *313*, 1929–1935.

(13) Towbin, H.; Bair, K. W.; DeCaprio, J. A.; Eck, M. J.; Kim, S.; Kinder, F. R.; Morollo, A.; Mueller, D. R.; Schindler, P.; Song, H. K.; van Oostrum, J.; Versace, R. W.; Voshol, H.; Wood, J.; Zabludoff, S.; Phillips, P. E. Proteomics-based Target Identification. *J. Biol. Chem.* **2003**, *278*, 52964–52971.

(14) Watkins, S.; German, J. Metabolomics and biochemical profiling in drug discovery and development. *Curr. Opin. Mol. Ther.* **2002**, *4*, 224–228.

(15) Jenkins, J. L.; Bender, A.; Davies, J. W. In silico target fishing: Predicting biological targets from chemical structure. *Drug Discovery Today: Technol.* **2006**, *3*, 413–421.

(16) Li, H.; Gao, Z.; Kang, L.; Zhang, H.; Yang, K.; Yu, K.; Luo, X.; Zhu, W.; Chen, K.; Shen, J.; Wang, X.; Jiang, H. TarFisDock: a web server for identifying drug targets with docking approach. *Nucleic Acids Res.* **2006**, *34*, W219–W224.

(17) Chen, Y. Z.; Zhi, D. G. Ligand−protein inverse docking and its potential use in the computer search of protein targets of a small molecule. *Proteins: Struct., Funct., Bioinf.* **2001**, *43*, 217–226.

(18) Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R. Comparative Assessment of Scoring Functions on a Diverse Test Set. *J. Chem. Inf. Model.* **2009**, *49*, 1079–1093.

(19) Li, X.; Li, Y.; Cheng, T.; Liu, Z.; Wang, R. Evaluation of the performance of four molecular docking programs on a diverse set of protein-ligand complexes. *J. Comput. Chem.* **2010**, *31*, 2109–2125.

(20) Nigsch, F.; Bender, A.; Jenkins, J. L.; Mitchell, J. B. O. Ligand-Target Prediction Using Winnow and Naive Bayesian Algorithms and the Implications of Overall Performance Statistics. *J. Chem. Inf. Model.* **2008**, *48*, 2313–2325.

(21) 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.

(22) Wale, N.; Karypis, G. Target Fishing for Chemical Compounds Using Target-Ligand Activity Data and Ranking Based Methods. *J. Chem. Inf. Model.* **2009**, *49*, 2190–2201.

(23) Cheng, T.; Zhao, Y.; Li, X.; Lin, F.; Xu, Y.; Zhang, X.; Li, Y.; Wang, R.; Lai, L. Computation of Octanol−Water Partition Coefficients by Guiding an Additive Model with Knowledge. *J. Chem. Inf. Model.* **2007**, *47*, 2140–2148.

(24) Willett, P. Evaluation of Molecular Similarity and Molecular Diversity Methods Using Biological Activity Data. *Methods Mol. Biol.* **2004**, *275*, 51–64.

(25) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **2002**, *45*, 4350–4358.

(26) Nettles, J. H.; Jenkins, J. L.; Bender, A.; Deng, Z.; Davies, J. W.; Glick, M. Bridging Chemical and Biological Space:□ "Target Fishing" Using 2D and 3D Molecular Descriptors. *J. Med. Chem.* **2006**, *49*, 6802–6810.

(27) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(28) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **2009**, *37*, W623–633.

(29) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **2008**, *36*, D901–906.

(30) 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*, D668–672.

(31) Zhu, F.; Han, B.; Kumar, P.; Liu, X.; Ma, X.; Wei, X.; Huang, L.; Guo, Y.; Han, L.; Zheng, C.; Chen, Y. Update of TTD: Therapeutic Target Database. *Nucleic Acids Res.* **2010**, *38*, D787–D791.

(32) Chen, X.; Ji, Z. L.; Chen, Y. Z. TTD: Therapeutic Target Database. *Nucleic Acids Res.* **2002**, *30*, 412–415.

(33) Paull, K. D.; Shoemaker, R. H.; Hodes, L.; Monks, A.; Scudiero, D. A.; Rubinstein, L.; Plowman, J.; Boyd, M. R. Display and Analysis of Patterns of Differential Activity of Drugs Against Human Tumor Cell Lines: Development of Mean Graph and COMPARE Algorithm. *J. Natl. Cancer Inst.* **1989**, *81*, 1088–1092.

(34) Shoemaker, R. H. The NCI60 human tumour cell line anti-cancer drug screen. *Nat. Rev. Cancer* **2006**, *6*, 813–823.

(35) Weinstein, J. N.; Myers, T. G.; O'Connor, P. M.; Friend, S. H.; Fornace, A. J., Jr.; Kohn, K. W.; Fojo, T.; Bates, S. E.; Rubinstein, L. V.; Anderson, N. L.; Buolamwini, J. K.; van Osdol, W. W.; Monks, A. P.; Scudiero, D. A.; Sausville, E. A.; Zaharevitz, D. W.; Bunow, B.; Viswanadhan, V. N.; Johnson, G. S.; Wittes, R. E.; Paull, K. D. An Information-Intensive Approach to the Molecular Pharmacology of Cancer. *Science* **1997**, *275*, 343–349.

(36) Zaharevitz, D. W.; Holbeck, S. L.; Bowerman, C.; Svetlik, P. A. COMPARE: a web accessible tool for investigating mechanisms of cell growth inhibition. *J. Mol. Graphics Modell.* **2002**, *20*, 297–303.

(37) Cheng, T.; Wang, Y.; Bryant, S. H. Investigating the Correlations among the Chemical Structures, Bioactivity Profiles, and Molecular Targets of Small Molecules. *Bioinformatics* **2010**, *26*, 2881–2888.

(38) Wang, Y.; Bolton, E.; Dracheva, S.; Karapetyan, K.; Shoemaker, B. A.; Suzek, T. O.; Wang, J.; Xiao, J.; Zhang, J.; Bryant, S. H. An overview of the PubChem BioAssay resource. *Nucleic Acids Res.* **2010**, *38*, D255–266.

(39) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402.

(40) Gigant, B.; Wang, C.; Ravelli, R. B. G.; Roussi, F.; Steinmetz, M. O.; Curmi, P. A.; Sobel, A.; Knossow, M. Structural basis for the regulation of tubulin by vinblastine. *Nature* **2005**, *435*, 519–522.

(41) Ravelli, R. B. G.; Gigant, B.; Curmi, P. A.; Jourdain, I.; Lachkar, S.; Sobel, A.; Knossow, M. Insight into tubulin regulation from a complex with colchicine and a stathmin-like domain. *Nature* **2004**, *428*, 198–202.

(42) Nogales, E.; Wolf, S. G.; Downing, K. H. Structure of the $\alpha\beta$ tubulin dimer by electron crystallography. *Nature* **1998**, *391*, 199–203.

(43) Mullard, A. 2010 FDA drug approvals. *Nat. Rev. Drug Discovery* **2011**, *10*, 82–85.

(44) Jordan, M. A.; Wilson, L. Microtubules as a target for anticancer drugs. *Nat. Rev. Cancer* **2004**, *4*, 253–265.

(45) Willett, P. Similarity-based approaches to virtual screening. *Biochem. Soc. Trans.* **2003**, *31*, 603–606.

(46) Cheng, T.; Liu, Z.; Wang, R. A knowledge-guided strategy for improving the accuracy of scoring functions in binding affinity prediction. *BMC Bioinf.* **2010**, *11*, 193.

(47) Wallqvist, A.; Huang, R.; Covell, D. G. Chemoinformatic Analysis of NCI Preclinical Tumor Data: Evaluating Compound Efficacy from Mouse Xenograft Data, NCI-60 Screening Data, and Compound Descriptors. *J. Chem. Inf. Model.* **2007**, *47*, 1414–1427.

(48) Gan, P. P.; Kavallaris, M. Tubulin-Targeted Drug Action: Functional Significance of Class II and Class IVb $\beta$-Tubulin in Vinca Alkaloid Sensitivity. *Cancer Res.* **2008**, *68*, 9817–9824.

(49) Lobert, S.; Vulevic, B.; Correia, J. J. Interaction of Vinca Alkaloids with Tubulin: A Comparison of Vinblastine, Vincristine, and Vinorelbine. *Biochemistry* **1996**, *35*, 6806–6814.

(50) Hastie, S. B. Spectroscopic and kinetic features of allocolchicine binding to tubulin. *Biochemistry* **1989**, *28*, 7753–7760.

(51) Zhao, H. Scaffold selection and scaffold hopping in lead generation: a medicinal chemistry perspective. *Drug Discovery Today* **2007**, *12*, 149–155.

(52) Böhm, H.-J.; Flohr, A.; Stahl, M. Scaffold hopping. *Drug Discovery Today: Technol.* **2004**, *1*, 217–224.

(53) McKie, J. H.; Douglas, K. T.; Chan, C.; Roser, S. A.; Yates, R.; Read, M.; Hyde, J. E.; Dascombe, M. J.; Yuthavong, Y.; Sirawaraporn, W. Rational Drug Design Approach for Overcoming Drug Resistance: Application to Pyrimethamine Resistance in Malaria. *J. Med. Chem.* **1998**, *41*, 1367–1370.

(54) Horton, J. R.; Sawada, K.; Nishibori, M.; Cheng, X. Structural Basis for Inhibition of Histamine N-Methyltransferase by Diverse Drugs. *J. Mol. Biol.* **2005**, *353*, 334–344.

(55) Birdsall, B.; Tendler, S. J. B.; Arnold, J. R. P.; Feeney, J.; Griffin, R. J.; Carr, M. D.; Thomas, J. A.; Roberts, G. C. K.; Stevens, M. F. G. NMR studies of multiple conformations in complexes of Lactobacillus casei dihydrofolate reductase with analogs of pyrimethamine. *Biochemistry* **1990**, *29*, 9660–9667.

(56) Denny, B. J.; Ringan, N. S.; Schwalbe, C. H.; Lambert, P. A.; Meek, M. A.; Griffin, R. J.; Stevens, M. F. Structural studies on bio-active compounds. 20. Molecular modeling and crystallographic studies on methylbenzoprim, a potent inhibitor of dihydrofolate reductase. *J. Med. Chem.* **1992**, *35*, 2315–2320.

(57) Robson, C.; Meek, M. A.; Grunwaldt, J.-D.; Lambert, P. A.; Queener, S. F.; Schmidt, D.; Griffin, R. J. Nonclassical 2,4-Diamino-5-aryl-6-ethylpyrimidine Antifolates: Activity as Inhibitors of Dihydrofolate Reductase from Pneumocystis carinii and Toxoplasma gondii and as Antitumor Agents. *J. Med. Chem.* **1997**, *40*, 3040–3048.

(58) Bram, E.; Ifergan, I.; Shafran, A.; Berman, B.; Jansen, G.; Assaraf, Y. Mutant Gly482 and Thr482 ABCG2 mediate high-level resistance to lipophilic antifolates. *Cancer Chemother. Pharmacol.* **2006**, *58*, 826–834.

(59) Rosowsky, A.; Queener, S. F.; Cody, V. Inhibition of dihydrofolate reductases from Toxoplasma gondii, Pneumocystis carinii, and rat liver by rotationally restricted analogues of pyrimethamine and metoprine. *Drug Des. Discovery* **1999**, *16*, 25–40.

(60) Galivan, J.; Nimec, Z.; Rhee, M.; Boschelli, D.; Oronsky, A. L.; Kerwar, S. S. Antifolate Drug Interactions: Enhancement of Growth Inhibition Due to the Antipurine 5,10-Dideazatetrahydrofolic Acid by the Lipophilic Dihydrofolate Reductase Inhibitors Metoprine and Trimetrexate. *Cancer Res.* **1988**, *48*, 2421–2425.

(61) Milletti, F.; Vulpetti, A. Predicting Polypharmacology by Binding Site Similarity: From Kinases to the Protein Universe. *J. Chem. Inf. Model.* **2010**, *50*, 1418–1431.

(62) Ertl, P.; Schuffenhauer, A.; Renner, S. The Scaffold Tree: An Efficient Navigation in the Scaffold Universe. *Methods Mol. Biol.* **2011**, *672*, 245–260.

(63) Whittle, M.; Gillet, V. J.; Willett, P.; Alex, A.; Loesel, J. Enhancing the Effectiveness of Virtual Screening by Fusing Nearest Neighbor Lists: A Comparison of Similarity Coefficients. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1840–1848.

(64) Fliri, A. F.; Loging, W. T.; Thadeio, P. F.; Volkmann, R. A. Biospectra Analysis: Model Proteome Characterizations for Linking Molecular Structure and Biological Response. *J. Med. Chem.* **2005**, *48*, 6918–6925.

(65) Fliri, A. F.; Loging, W. T.; Thadeio, P. F.; Volkmann, R. A. Biological spectra analysis: Linking biological activity profiles to molecular structure. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 261–266.