Prediction of Biological Targets for Compounds Using Multiple-Category Bayesian Models Trained on Chemogenomics Databases

Nidhi,[†] Meir Glick,[‡] John W. Davies,[‡] and Jeremy L. Jenkins*[,‡]

Indiana University−Purdue University Indianapolis, School of Informatics, Indianapolis, Indiana 46202, and Novartis Institutes for Biomedical Research Inc., Lead Discovery Center, 250 Massachusetts Ave., Cambridge, Massachusetts 02139

Received January 3, 2006

Target identification is a critical step following the discovery of small molecules that elicit a biological phenotype. The present work seeks to provide an in silico correlate of experimental target fishing technologies in order to rapidly fish out potential targets for compounds on the basis of chemical structure alone. A multiple-category Laplacian-modified naïve Bayesian model was trained on extended-connectivity fingerprints of compounds from 964 target classes in the WOMBAT (World Of Molecular BioAcTivity) chemogenomics database. The model was employed to predict the top three most likely protein targets for all MDDR (MDL Drug Database Report) database compounds. On average, the correct target was found 77% of the time for compounds from 10 MDDR activity classes with known targets. For MDDR compounds annotated with only therapeutic or generic activities such as "antineoplastic", "kinase inhibitor", or "anti-inflammatory", the model was able to systematically deconvolute the generic activities to specific targets associated with the therapeutic effect. Examples of successful deconvolution are given, demonstrating the usefulness of the tool for improving knowledge in chemogenomics databases and for predicting new targets for orphan compounds.

INTRODUCTION

Chemical genetic screens are increasingly identifying small-molecule probes that modulate biological processes[1−3] yet remain "orphan ligands" in the sense that their macromolecular targets may remain unknown. After the chemical-genetic screening phase, identifying the biological target of a probe can be a complex task requiring novel technologies.[4] At the same time, decades of published, patented, and proprietary research on protein−ligand interactions contain a wealth of biological compound activities (IC50/EC50 data). This historical information can provide clues as to the targets of new compounds, given a methodology to classify new compound structures in the proper context of the collective structure−activity relationship. By combining data-mining methods and diverse chemogenomics libraries, it is possible to map known "chemistry space" onto known "biological activity space" in the form of models that enable a priori prediction of the targets, pathways, or therapeutic relevance of phenotypically interesting compounds given only their chemical structures. Such an approach to in silico chemogenomics[5] could help speed up target identification by prioritizing targets or target families according to the strength of their established associations with the structural motifs in small-molecule probes.

The relationship of chemical classes to protein or activity classes in chemogenomics databases has been explored in recent years. In an experimental setting, annotated compound libraries were screened in cellular assays to generate hypotheses about biological mechanisms corresponding to cellular phenotypes.[6] Computationally, Schuffenhauer et al. introduced a "homology-based similarity searching" approach[7] that demonstrated chemical structural information can be used to find inhibitors of not just targets but also target families in MDDR (MDL Drug Database Report; Molecular Design Ltd., San Leandro, CA). Another method that used MDDR for validation was developed to identify chemical substructures that occur in more than one activity class.[8] Further, the similarity between MDDR activity classes has been quantified by measuring the similarity of compound structures and substructures associated with those biological activities.[9] Most pertinent to the current study, the PASS algorithm pioneered data mining in annotated chemical libraries for new activity predictions by using "multilevel neighborhoods of atoms" descriptors to provide probability scores for a list of pharmacological effects.[10,11]

The ability to predict the specific protein targets of small molecules is, in principle, becoming easier because of databases developed in recent years. These databases derive from commercial sources, such as WOMBAT (World Of Molecular BioAcTivity),[12] MDDR (MDL Drug Database Report),[13] stARLITe,[14] Jubilant BioSys databases,[15] and GVK Biosciences;[16] public sources, such as the Harvard *ChemBank*[17] and the NIH Roadmap Molecular Libraries Initiative;[18] and proprietary data (however, see the Discussion for the importance of database format). In tandem, novel 2D and 3D chemical descriptors[19−24] have been developed that can be deployed with robust statistical models tailored specifically for structure−activity model generation. In particular, Laplacian-modified naïve Bayesian classifiers[25]

* Corresponding author phone: 617-871-7155; fax: 617-871-4088; e-mail: jeremy.jenkins@novartis.com.
[†] Indiana University−Purdue University Indianapolis.
[‡] Novartis Institutes for Biomedical Research Inc.

10.1021/ci060003g CCC: $33.50
Published on Web 03/22/2006

have proven capable of handling the stochastic noise generated by high-throughput screening data[26−28] and are amenable to training on large-scale chemogenomics databases because computing time scales linearly with the number of data points and significant tuning by the user is not required. Laplacian-modified naïve Bayes classifiers in combination with chemical descriptors have been used in recent studies to identify antagonists of a specific target class,[25] to enrich docking data,[29] and to analyze high-throughput screening (HTS) data.[26,27]

In the present study, we apply a multiple-category Laplacian-modified naïve Bayesian model trained on the compound−target pairings in the WOMBAT database to predict the most likely activities for all compounds from the MDDR. In many cases, the specific protein targets are annotated in MDDR, enabling us to perform a retrospective analysis of target prediction accuracy for compounds from 10 activity classes: ACE (angiotensin-converting enzyme), AChE (acetylcholinesterase), AT1 (angiotensin 1), AT2 (angiotensin 2), COX-1 (cyclooxygenase 1), COX-2 (cyclooxygenase 2), H+/K+/-ATPase, HIV-1P (HIV-1 protease), PDE4 (phosphodiesterase 4), and HIV-1RT (HIV-1 reverse transcriptase). In other cases, a therapeutic activity rather than a specific target is listed in the compound MDDR activity index. We also show that for five selected therapeutic/generic classes—antineoplastic, kinase inhibitors, antiarthritic, antihypertensive, and anti-inflammatory—the model is able to provide WOMBAT protein target predictions that align well with the generic or therapeutic MDDR annotations.

MATERIALS AND METHODS

Databases. WOMBAT is a target-annotated chemogenomics database available from Sunset Molecular Discovery (Santa Fe, New Mexico). WOMBAT 2005.1 contains 117 007 compounds with 104 230 unique SMILES, which target 1021 unique proteins published in 4773 articles in medicinal chemistry journals between 1975 and 2004. The database follows a hierarchical scheme that includes information about target families. For the purpose of this study, only compounds with activity ($IC_{50}/EC_{50}/K_i/K_b/K_d$/MIC/$ED_{50}$) $< 30\ \mu$m were selected because this cutoff ensures reasonable potency while not resulting in a substantial loss of target information. Thus, a reduced data set of 103 735 compounds with 964 biological targets was obtained. We made no attempt to distinguish between agonists and antagonists—all compounds with binding capability were considered jointly. Also, WOMBAT compounds may have more than one target listed, so each ligand−target pairing was considered during model training. Other automated curation steps were carried out using custom scripts, such as the reassignment of WOMBAT property names ACT.LIST.ENZ.FAMILY, ACT.LIST.PRO.FAMILY, and ACT.LIST.REC.FAMILY to a single property called FAMILY.

The MDDR version used in the present study contains 159 967 compounds and 761 biological activities derived mainly from patents and journal articles. All records lacking useful structural information were removed, resulting in 156 873 compounds with 659 biological activities. Most of the biological activities are not protein-specific (~70%); for example, antineoplastic and antihypertensive encompass therapeutic activities that derive from the modulation of a number of very different proteins. Further, the annotated classes do not necessarily preclude activity in other classes, and some activity classes are highly related or overlapping.[9]

Descriptors. Extended Connectivity Fingerprint (ECFP) descriptors developed by SciTegic are circular substructural fingerprints that are based on the Morgan algorithm.[30] An ECFP feature represents an exact structure with limited and specified attachment points. ECFPs are generated in an iterative fashion. Initially, each atom is assigned a code that derives from the number of atomic connections, element type, charge, and mass within an atomic neighborhood size of "0". In the first iteration, information about each atom's immediate neighbors is collected and a new code representing the atom and its immediate neighbors is generated. In each iteration, the neighborhood size becomes larger and the updated codes of the atoms from previous iterations are used for assigning new codes. When the desired neighborhood size is reached, the set of all features is returned as a fingerprint. ECFPs with a neighborhood size of six were used as structural descriptors to train the multiple-category Laplacian-modified naïve Bayesian classifiers. ECFPs were deemed sufficient for the purpose of the present study because of their strong performance in other reports.[20,26,27,29] ECFP performance was not compared with other chemical descriptors in the present study.

Multiple-Category Models (MCM). Naïve Bayes is a statistical classification method based on the Bayes rule of conditional probability, which states that, given two events A and B, the probability of event A occurring, given that B has already occurred, $P(A|B)$, is given by

$$P(A|B) = P(B|A)\,P(A)/P(B)$$

where $P(A)$ and $P(B)$ are the probabilities of events A and B, respectively. The Bayesian classifier is called naïve because it naïvely assumes the features are independent. From this assumption, it is valid to multiply probabilities of the individual events. As described below, the "naïve" assumption will be employed: the probabilities of the individual events will be multiplied; however, probabilities themselves will not be calculated by the above equation but instead using a Laplacian-corrected estimator.[25,31] The estimator is used to adjust the uncorrected probability estimate of a feature to account for the different sampling frequencies of different features. In the context of this paper, the Laplacian-corrected estimator for a compound being active given a feature Fi is calculated according to the following equations (Dave Rogers, SciTegic, personal communication), where A compounds are active in T total compounds and feature Fi is contained in T_{Fi} samples and A_{Fi} samples containing feature Fi are active.

We start with the baseline probability of a compound being active:

$$P(\text{active}) = A/T \tag{1}$$

If the molecule contains feature Fi, the *uncorrected* estimate of activity should be

$$P(\text{active}|\text{Fi}) = A_{Fi}/T_{Fi} \tag{2}$$

But if the number of compounds in the data set containing feature Fi is small, this estimate may be overconfident. More

sampling of a feature is desirable to increase confidence. Typically, if we sample a feature K times, we would expect the number of active compounds to be KA_{Fi}/T_{Fi}.

We can correct every Fi encountered by adding virtual samples:

$$[A_{Fi} + (A/T)K]/(T_{Fi} + K) \quad (3)$$

If we have few samplings of the feature, the probability of P(active|Fi) should approach P(active). The Laplacian correction substitutes $K = 1/P$(active) or T/A:

$$(A_{Fi} + 1)/(T_{Fi} + T/A) \quad (4)$$

Rearrange by multiplying the numerator and denominator by A/T to yield

$$[(A_{Fi} + 1)(A/T)]/[T_{Fi}(A/T) + 1] \quad (5)$$

To get the *relative* estimator, we divide eq 5 by P(active) or A/T:

$$P_{final}(A|Fi) = (A_{Fi} + 1)/[T_{Fi}(A/T) + 1] \quad (6)$$

This is the same as eq 8 in the appendix of Hert et al.[31]

Because several features are normally required to characterize a compound, the multiple features Fi in a sample have to be combined. To allow for an easier combining of multiple features into a Bayesian score, the SciTegic implementation uses the sum of the log values of the individual feature probabilities. Given n features for a compound, the combined estimation, $P_{combined}$, is calculated as follows:

$$P_{combined} = \log[P_{final}(\text{active}|F1)] + \log[P_{final}(\text{active}|F2)] \ldots + \log[P_{final}(\text{active}|Fn)]. \quad (7)$$

In the present case involving a chemogenomics database, the objective is to build a Laplacian-modified naïve Bayesian model for *each* target class in the database. Because the number of targets is on the order of hundreds or even thousands, an automation of the process is desirable. The Learn Molecular Categories component in Pipeline Pilot automatically builds multiple Laplacian-modified Bayesian models. The user specifies a "CategoryProperty"—for example, the WOMBAT target name property—which is subsequently used to create a model for each of the categories. In the creation of the model, the compounds in the other activity categories are defined as the inactive set.

To predict the target of a test compound, it is passed through each Laplacian-modified naïve Bayesian model of each target class. The relative estimator score for each of the target classes is calculated. The target with the highest score is assumed to be the most probable target for that compound. Similarly, the next highest score for a target can be assigned as the second most likely target and so on (see the Discussion).

MCM Testing. Two separate MCMs were created for the purpose of the present study.

(i) A multiple-category model was created using 85% of the WOMBAT compound records as the training set, so that the remaining 15% of the database could be used as a test set for model validation, ensuring no overlapping structures in the 85/15 partition. The WOMBAT target name was used as the CategoryProperty for model training.

(ii) An MCM was trained on 100% of the WOMBAT database. When building multiple-category models, the "inactives" for training any single category derive from the "actives" for every other category. Pipeline Pilot version 4.5.2 was employed to build the MCMs. The computation was carried out on an Intel Xeon 2.7 GHz dual processor server with 4 GB of RAM. The training of the MCM on the current data set took nearly 6 h. The subsequent target prediction rate is 1000 compounds per minute, including the generation of ECFP_6 descriptors and all scripts to handle data postprocessing. All 156 873 MDDR records were passed through the WOMBAT MCM. Notably, 11% of the MDDR structures are exact matches with WOMBAT structures; these compounds were not removed from testing to serve as positive controls in retrospective analyses. MDDR activity classes associated with 10 specific targets were selected to help assess MCM predictivity: ACE (angiotensin-converting enzyme), AChE (acetylcholinesterase), AT1 (angiotensin 1), AT2 (angiotensin 2), COX-1 (cyclooxygenase 1), COX-2 (cyclooxygenase 2), H+/K+/-ATPase, HIV-1P (HIV-1 protease), PDE4 (phosphodiesterase 4), and HIV-1RT (HIV-1 reverse transcriptase). Additionally, five therapeutic activity classes without specifically associated targets were assessed. Table 1 summarizes the number of compounds in the respective test data sets.

Table 1. MDDR Activity Classes

MDDR activity class	number of compounds
Phosphodiesterase IV Inhibitor	2000
Cyclooxygenase-1 Inhibitor	88
Cyclooxygenase-2 Inhibitor	1055
Acetylcholinesterase Inhibitor	810
Angiotensin II AT1 Antagonist	2185
Angiotensin II AT2 Antagonist	53
ACE Inhibitor	570
Reverse Transcriptase Inhibitor	819
HIV-1 Protease Inhibitor	1027
H+/K+-ATPase Inhibitor	751
Antineoplastic	19 621
Kinase Inhibitor	1858
Antiarthritic	11 147
Antihypertensive	11 124
Anti-inflammatory, Intestinal	29

A scaffold analysis between MDDR and WOMBAT was carried out using the "Murcko Assemblies" parameter option in the Generate Fragments component in Pipeline Pilot. Murcko assemblies consist of the core ring structure of the compound and any chains contained in the core structure minus any "side-chain" substituents.[32] Tanimoto similarity values between compounds were determined using ECFP_6 fingerprints.

RESULTS

Database Scaffold Comparison. The MDDR and WOMBAT databases are both comprised of historical activity data with some degree of overlapping information. How extensively the databases overlap should be understood in order to provide a fair evaluation of the WOMBAT-trained MCM on MDDR test compounds. Unfortunately, the degree of annotated targets in common is difficult to assess outright because of the incongruity in database design. However, in

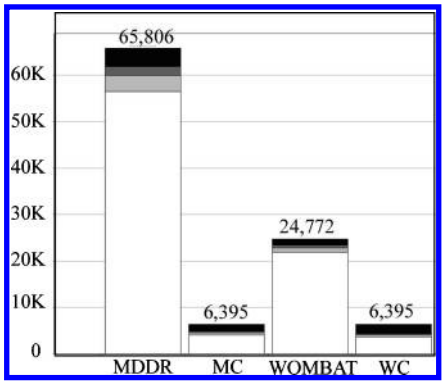

Figure 1. Frequency distribution of "Murcko assembly" scaffolds in MDDR and WOMBAT. MDDR, scaffolds found only in MDDR; MC, scaffolds from MDDR found in common with WOMBAT. WOMBAT, scaffolds found only in WOMBAT; WC, scaffolds from WOMBAT found in common with MDDR. The frequency of scaffold occurrence is denoted as follows: one occurrence (white), i.e., "singletons", two occurrences (light gray), three occurrences (dark gray), and four or more occurrences (black).

terms of compound overlap, we observed that only ~11% of MDDR canonical smiles identically matched compounds in WOMBAT. To further understand the structural relatedness of the databases, a comparative scaffold diversity analysis was carried out. We found 65 806 Murcko assemblies (i.e., scaffolds) unique to MDDR (Figure 1) and 24 772 Murcko assemblies unique to WOMBAT; 6395 Murcko assemblies were found in common to both data sets. The scaffold analysis revealed that MDDR is more diverse than WOMBAT in terms of the ratio of total scaffolds to total compounds. Although there is some scaffold overlap between both data sets, there are a large number of scaffolds unique to each database, suggesting that WOMBAT and MDDR are sufficiently different for the current purpose of serving as training and test sets.

MCM Validation. An MCM trained on 85% of the WOMBAT compounds was first used to predict the targets for the remaining 15% of the database. For each compound, the MCM component generates a score for all possible targets. However, for pragmatic reasons, only the top three target predictions were evaluated to reflect a real-life situation where only a small number of target predictions could be tested experimentally. The 85/15 model ranks the correct target highest among all 964 targets for 82% of the compounds (Figure 2a). The target is correctly predicted on the second guess for 8% of the compounds and correctly predicted on the third guess for 2% of the compounds. In total, 92% of the compounds are correctly assigned to their known targets within three guesses. Because WOMBAT records are also annotated with target family names (e.g., GPCRs, voltage-gated ion channel, oxidoreductases), a further analysis was done to examine the accuracy of target family prediction. The 85/15 model trained on family categories predicts the correct target family for 89% of the compounds on the first guess (Figure 2b) and the correct target family for 95% of the compounds in the first three guesses.

Prediction of WOMBAT Targets for MDDR Compounds. All MDDR compounds were passed through the MCM trained on the full WOMBAT database, and the top three target predictions were saved. To retrospectively assess the model, we examined the target predictions made for 10

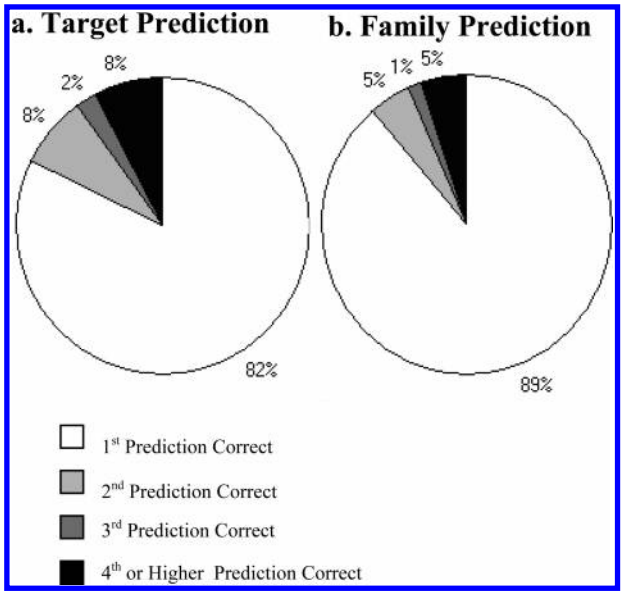

Figure 2. a. Accuracy of target prediction by the WOMBAT 85% MCM trained on Target categories. Shown are the percentages of test compounds (WOMBAT 15%) with targets predicted correctly on the first guess (white), second guess (light gray), third guess (dark gray), and all others (black). b. Accuracy of family prediction by the WOMBAT 85% MCM.

different MDDR activity classes with specific protein targets. Figure 3a shows the percentage of compounds for which the correct target is predicted correctly within the top three guesses. Correspondingly, Figure 3b depicts the percentage of compounds for which the correct target family is predicted. On average for the 10 test cases, the target is predicted correctly in the top three guesses for 77% of the compounds and the family is predicted correctly in the top three guesses for 79% of the compounds. Using only the top guess, the MCM predicts the correct target 53% of the time and the correct family 68% of the time on average. These results illustrate that the family is much easier to predict than the specific target on the first guess because of the "noise" created by homologous family member proteins. In other words, the presence of closely related activity classes[9] can cause infidelity in target prediction because related proteins often share related chemical inhibitors.[7,8] However, this effect is abrogated after three guesses as the accuracy of target prediction becomes similar to that of family prediction.

Machine learning algorithms extrapolate information from their training data sets. The structural diversity among underlying training and test set compounds will therefore influence the success of such algorithms. For COX-1, only 27% of the MDDR COX-1 inhibitors are correctly assigned. To investigate the basis for the low prediction success, a 2D fingerprint similarity analysis between the test set (COX-1 compound sets from MDDR) and the training set (COX-1 compounds from WOMBAT) was carried out. Strikingly, we found that nearly 95% of the MDDR COX-1 inhibitors have a Tanimoto similarity less than 0.4 with WOMBAT COX-1 compounds (using ECFP_6 descriptors), indicating that there is a limit to how far the MCM can extend outside of the chemical space in the training set to classify compounds in the test set.

Prediction of WOMBAT Targets for MDDR Therapeutic Activities. Compounds with therapeutic rather than

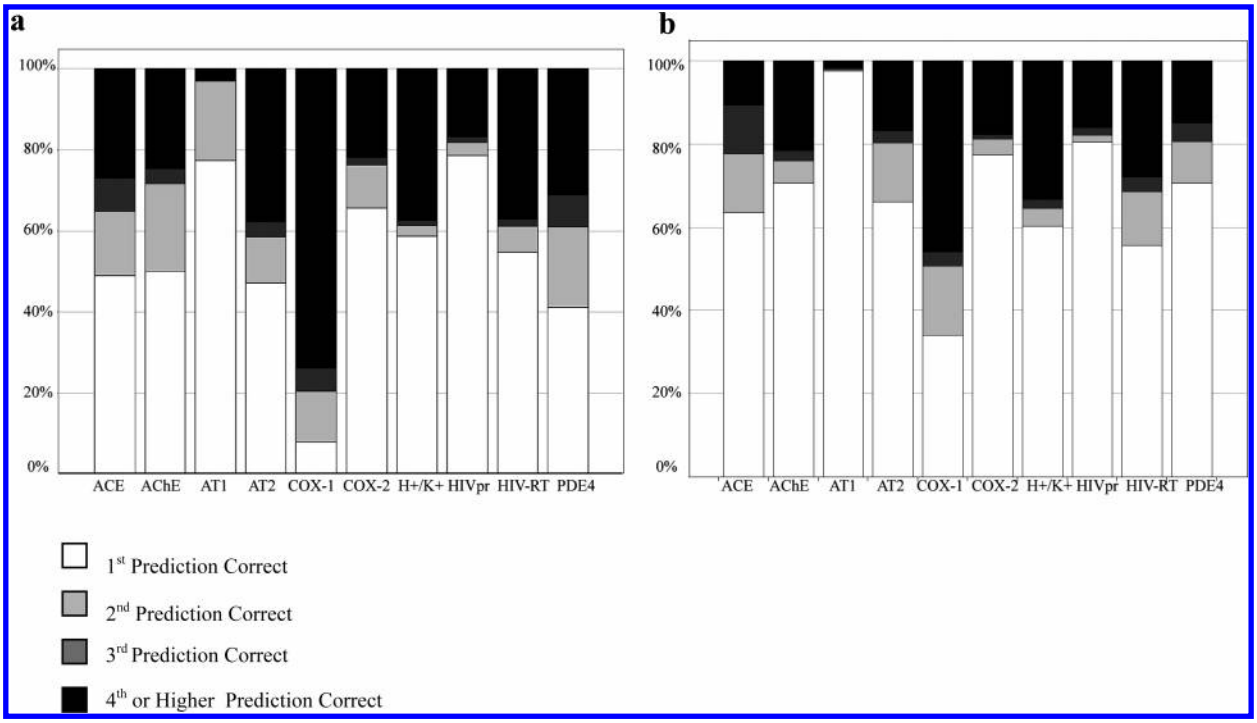

Figure 3. Prediction accuracy of the WOMBAT MCM on test compounds from 10 MDDR activity classes. a. Target prediction. b. Family prediction. The number of predictions required to correctly find the target or family is shown.

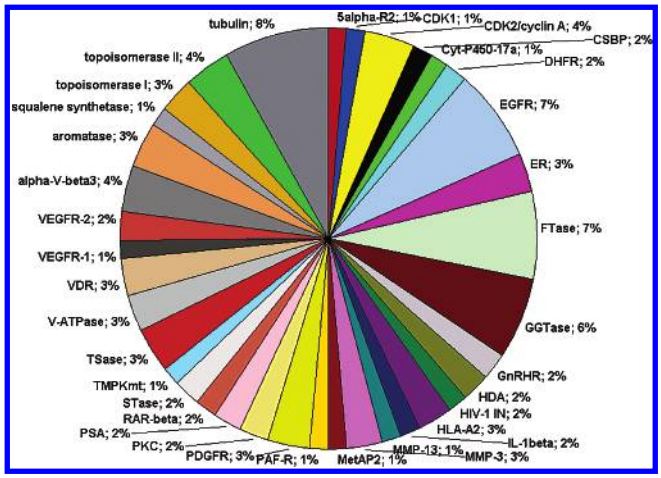

Figure 4. Distribution of the top targets predicted for generic MDDR activity class Antineoplastic. Only targets accounting for more than 1% of all predictions are shown.

target-specific annotations were assessed to see if the WOMBAT targets predicted were known to be associated with the MDDR therapeutic activity. The exercise is more qualitative than quantitative in nature because the exact target is not known or not annotated for a large portion of the compounds. The top targets predicted for five therapeutic or generic activity classes were examined: Antineoplastics (inhibitors of cell growth); Kinase Inhibitors (generic); Antiarthritic; Antihypertensive; and Anti-inflammatory, Intestinal. Overall, we found that the most frequently predicted targets were well-established modulators of the corresponding generic activities.

Antineoplastics. In the case of antineoplastics (Figure 4), tubulin is the most frequently predicted target (8%). (See Appendix 1 in the Supporting Information for the full names of the abbreviated targets.) Inhibition of tubulin assembly and motility is clearly linked to inhibition of the cell cycle and growth. Other highly represented targets are growth factor receptors (EGFR, FGFR, VEGF-R, and PDGF-R), cell cycle and cell signaling kinases [PKC, PKA, CDK2, CDK4, Tie-2, adenosine kinase (AK), c-Src, Flt-1, Lck, TMPKmt, and CSBP/p38], and other anticancer targets such as farnesyltransferase (FTase), geranylgeranyltransferase (GGTase), matrix metalloproteinases (MMP), topoisomerases, aromatase, aldose reductase (AR), and neutral sphingomyelinase (n-SMase).

In a substantial number of cases, the predicted target could be verified by examining the more specific text provided in the MDDR's patent and literature fields; however, if target information exists, it is not always translated to the ACTIV_CLASS field in MDDR (which complicates automated model building using MDDR). In other cases where no specific text was given, the predictions were potentially interesting in a prospective sense, such as the assignment of histone deacetylase (HDAC) to trichostatin D. HDACs accounted for 2% of all target predictions for the antineoplastics despite the fact that HDACs are not an ACTIV_CLASS in MDDR. HDACs catalyze the removal of acetyl groups from histone lysine residues, a process that regulates gene expression and affects transformed cell growth and survival.[33] A well-established paninhibitor of HDACs is the antifungal antibiotic trichostatin A (TSA; Figure 5a) produced by *Bacillus megaterium*, whose hydroxamic acid moiety is known to chelate zinc in the HDAC active site resulting in potent binding and inhibition.[34−36] Compound EXTREG 286017, or trichostatin D (Figure 5b), is a TSA analogue produced by *Streptomyces violaceusniger* listed as antineoplastic on the basis of a publication describing its ability to induce phenotypic reversions in transformed cells.[37] Trichostatin D is otherwise identical to TSA except for an attached sugar ring that partially "masks" the critical hydroxamic acid group. It is conceivable that the sugar ring will be cleaved off in vivo to reveal the hydroxamic acid

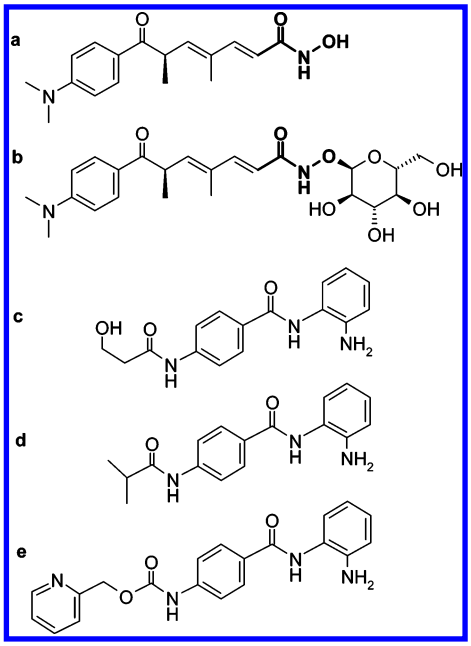

Figure 5. a. Trichostatin A, a known HDAC inhibitor. b. Trichostatin D. MDDR EXTREG: 286017. c. EXTREG 142057 [N-(2-aminophenyl)-4-(3-hydroxypropanamido) benzamide], a substructure in many compounds with anticancer activity and predicted as an HDAC inhibitor. d. EXTREG 142055 [N-(2-aminophenyl)-4-isobutyramidobenzmide], an analogue of EXTREG 142057. e. MS-27-275 benzamide-type HDAC inhibitor.

such that trichostatin D could behave as a natural prodrug of TSA. Although HDAC is not mentioned in the patent and literature fields of the MDDR record for trichostatin D, it is extremely likely that HDAC inhibition is the source of phenotypic reversions in transformed cells. Despite the obvious similarity to TSA, this validation of HDAC as a target of trichostatin D could not be found explicitly in MDDR or any other database that we searched. As a second example of deconvolution from therapeutic activity to a specific target, antineoplastic compounds EXTREG 142057 and 142055 (Figure 5c,d) were also predicted to be HDAC inhibitors, although their MDDR records deriving from a 1985 publication[38] do not describe this activity. However, these benzamide chemotypes are close analogues of compound MS-27-275[39] that was later patented as an HDAC inhibitor and used in clinical trials (Figure 5e). Importantly, the two MDDR compounds are not members of the WOMBAT database used to train the MCM. These findings strongly underscore the potential use of automated target prediction, both for updating chemogenomics database information and for discovering new targets for "orphan" compounds.

Kinases Inhibitors. MDDR contains the generic ACTIV-_CLASS Tyrosine-Specific Protein Kinase Inhibitor, as well as the somewhat more specific classes of Thymidine Kinase Inhibitor, Protein Kinase C Inhibitor, Adenosine Kinase Inhibitor, Nucleoside Diphosphate Kinase Inhibitor, and Phosphatidylinositol Kinase Inhibitor. We predicted the specific protein targets for this collective set of kinase activity classes (Figure 6). As expected, the assignment of specific targets yields mostly kinase targets predictions. EGFR (27%) and PKC (16%) account for the largest portions. Less than 10% of Figure 6 is comprised of nonkinase targets. The relationship of kinases to the "false positive" or nonkinase

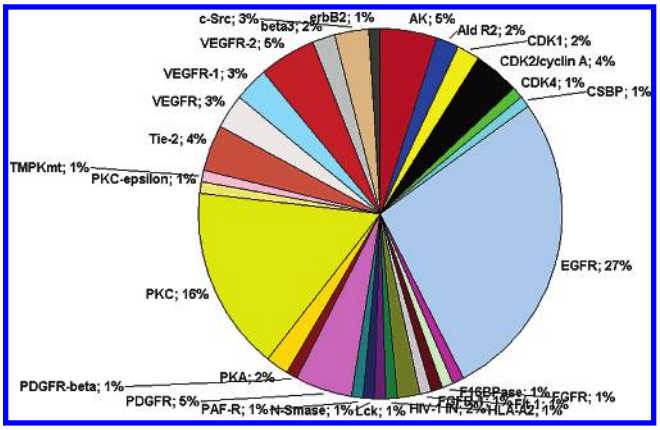

Figure 6. Distribution of the top targets predicted for kinase MDDR activity classes (thymidine kinase inhibitor, protein kinase C inhibitor, tyrosine-specific protein kinase inhibitor, adenosine kinase inhibitor, nucleoside diphosphate kinase inhibitor, and phosphatidylinositol kinase inhibitor). Only targets accounting for more than 1% of all predictions are shown.

targets is of interest for further exploration in terms of the relatedness of the associated ligands. For example, although HIV integrase (2% prediction frequency) is not a kinase family member, it is clearly associated with nucleotide-like inhibitors which block its natural mode of action; nucleotide-based inhibitors are structurally related to many ATP-competitive kinase inhibitors. Thus, the infidelity of target prediction reveals a ligand-centric relationship between the correct target and the false-positive target, which could be mined further.

One important issue to address is the dependency of model accuracy on the chemical similarity between training set and test set structures. In other words, can the model correctly assign a compound's target when the compound has a low topological similarity to the training set compounds? The answer is not entirely intuitive because Bayesian models place weights on certain substructures according to their relevance to target activity. We explored this issue by binning the MDDR kinases according to their Tanimoto similarity to the WOMBAT kinases and plotting prediction accuracy (Figure 7). The number of guesses required to identify Kinase as the correct WOMBAT family is indicated by bar coloration. The compound binning approach revealed a correlation between Tanimoto coefficient (T_c) similarity and target prediction accuracy. For test compounds with $T_c > 0.80$ to any training set compound from its family, the initial prediction was nearly always correct. Interestingly, the predictions remain reasonably good for $T_c = 0.30-0.40$ but drop off precipitously when $T_c < 0.30$ (using ECFP_6). Thus, there is an asymptotic limit to target/family predictions that derives from the overlap of training set and test set chemical space. However, for compounds in the $T_c = 0.30-0.50$ range, the MCM predictions are surprisingly reliable.

Anti-Inflammatory, Intestinal. The distribution of targets predicted for MDDR compounds with therapeutic activity for treating intestinal inflammation are shown in Figure 8. NK1 (neurokinin 1) is the most frequently predicted target, followed by NEP (neutral endopeptidase) and LTD4 (leukotriene D4 receptor). NK1 is a member of the tachykinin receptor class known to play a role in regulating intestinal motor function, secretion, and inflammation.[40] NEP down-

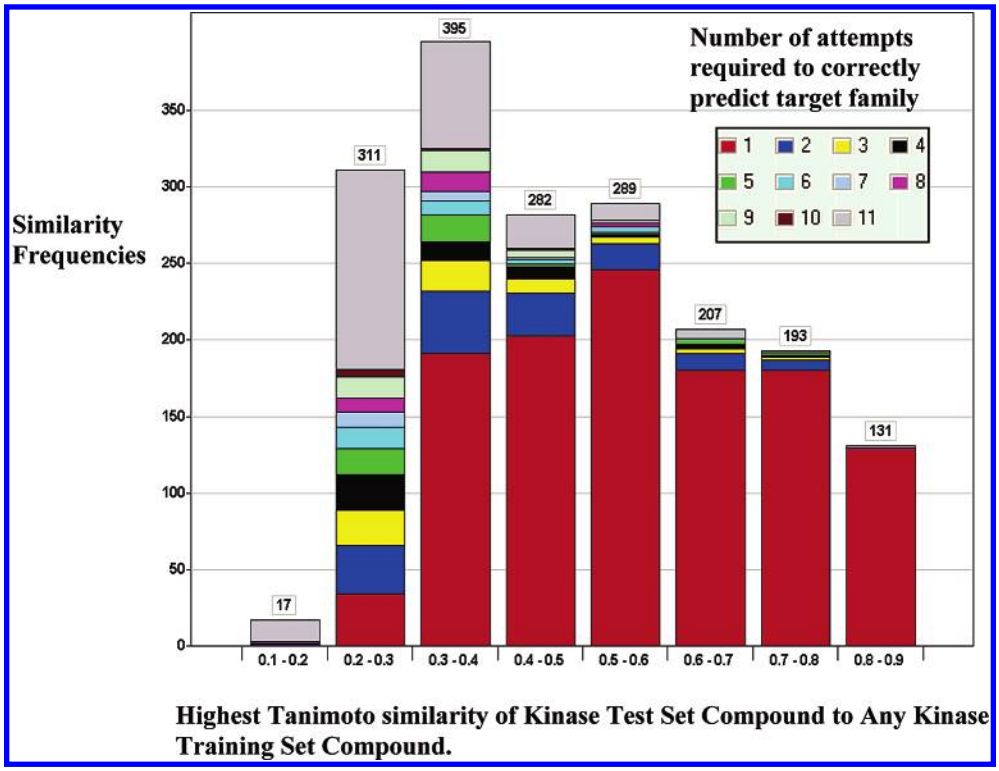

Figure 7. The influence of chemical similarity on prediction accuracy. The 2D similarity between all MDDR test set kinase inhibitors to all WOMBAT training set kinase inhibitors was determined. MDDR test set compounds were then binned according to their highest Tanimoto similarity to any WOMBAT kinase inhibitor in the training set. The ability of the MCM to accurately predict kinase as the target family is denoted by color: red (kinase was the first family predicted), blue (kinase was the second family predicted), and so on.

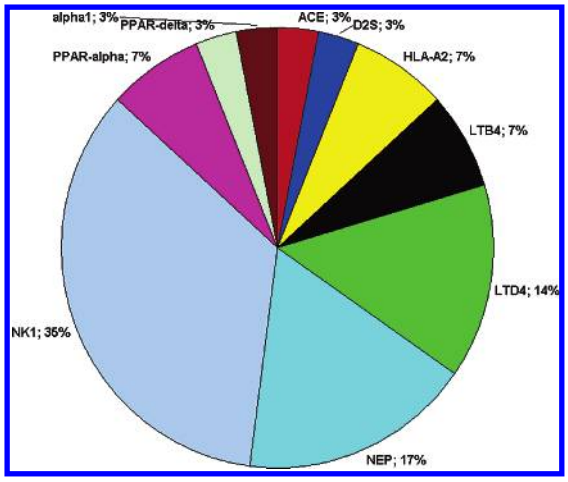

Figure 8. All targets predicted for MDDR activity class Anti-inflammatory, Intestinal.

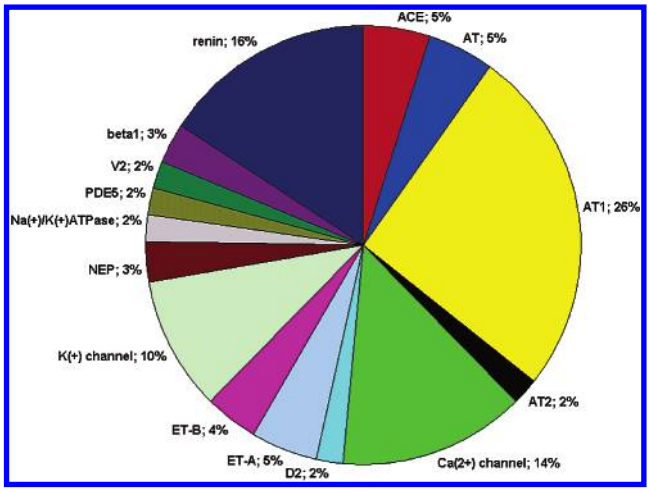

Figure 9. Distribution of the top targets predicted for MDDR activity class Antihypertensive.

regulates intestinal inflammation[41] and LTD4 and LTB4 are proinflammatory mediators implicated in inflammatory bowel disease.[42] Peroxisome proliferator-activated receptors (PPARs) are also suggested as targets; PPAR modulators have been investigated for the treatment of ulcerative colitis[43] (see the Discussion).

Antihypertensive. Many of the well-established players in the renin-angiotensin system are predicted WOMBAT targets for the MDDR antihypertensive compounds (Figure 9), including AT (general), AT1, AT2, ACE, and renin—all targets of antihypertension drugs approved or in clinical trials.[44,45] Additionally, Ca^+ and K^+ channels which are known to closely regulate vascular reactivity and membrane potential[46] are implicated.

Antiarithritic. As would be expected, COX-2, COX-1, and MMPs are frequently predicted as antiarthritic targets, along with TNFα and proteins involved in its biogenesis—TNFα-converting enzyme (TACE) and CSBP kinase (Figure 10).

DISCUSSION

Naïve Bayesian modeling used in conjunction with Extended Connectivity Fingerprints has previously been shown to perform well in identifying actives in comparison to other substructure-based searching methods.[26,28,29] In the current study, we applied multiple-category naïve Bayesian models to predict targets for compounds with various levels of annotation. When further information was available, it

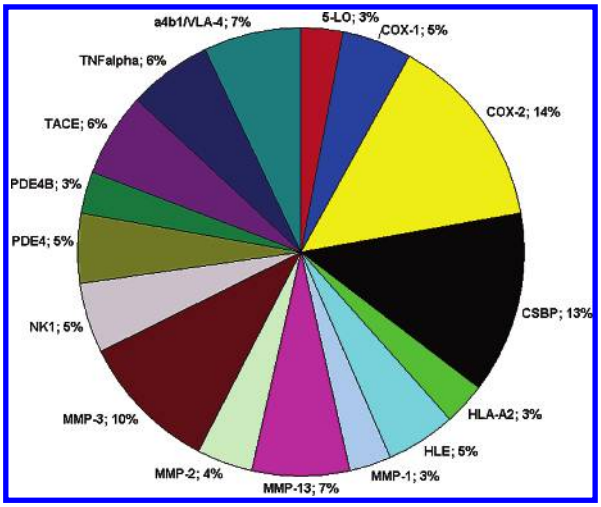

Figure 10. Distribution of the top targets predicted for MDDR activity class Antiarithritic.

was typical to find that generic activities could be deconvoluted to specific target activities in a way that agreed with surrounding patent and published literature. This information could of course be curated manually; however, MCMs offer a systematic and automated way of rapidly adding new, predictive information to large databases of compounds. For MCM predictions, the target guess is accompanied by Bayes scores that help to quantify the degree of belief in the prediction. With enough empirical data, rules applying to Bayes score thresholds could be generated by the user of the model. It is worth noting that not every compound will have a strong target match nor does every compound necessarily have an in vivo "target" at all.

A typical usage of the Laplacian-modified naïve Bayesian classifier is to analyze data from a HTS campaign where a compound collection is screened using one assay for one target. These data can be treated as one data set of active and inactive compounds where the objective of the analysis is to classify *compounds* and specifically identify false negatives. In contrast, the objective in this study is to classify *targets* rather than compounds. Each target class is a different data set, where the number of active compounds is different and a small fraction of the inactives is different as well. In the HTS analysis, compounds are prioritized with the same model, and therefore, a comparison between compounds or prioritization is straightforward. Here, the models were trained on targets with different numbers of active compounds. Because the Laplacian-modified naïve Bayesian model employs two corrections based on the number of active compounds, one has to bear in mind that comparing scores between "relative" target models is not entirely valid in a statistical sense. Although an "absolute" model is available in Pipeline Pilot 4.5.2 to offset this problem, in our hands, the component did not improve target prediction performance (data not shown), and so, it was not used.

Machine learning algorithms are largely dependent on the training data sets. The quality of curation of the underlying chemogenomics database is vital to the success of the computational model. The target listed for a given compound should be well-established in the citation source, requiring careful reading of the primary literature on the part of the database curator. Consistency in the target naming schema should also be considered for automated methods to work properly. Consistent naming is particularly problematic with many of the current commercial databases. For example, if COX-2-related records are called both "COX-2" and "cyclooxygenase-2", separate MCM models will be constructed, causing a loss of information to each model. A one-name-for-one-target classification system is imperative.

The additional annotation of target family names in WOMBAT was found to be useful to "widen the net" during target fishing. This could be important in prospective cases where the test compound target does not exist in the training database but other members of the same family are present. Further annotations such as protein structural domains or pathway information would be valuable. For example, many of the WOMBAT targets assigned to MDDR activity classes "Anti-inflammatory, Intestinal" and "Antiarthritic" are members of the arachidonic acid metabolism pathways (e.g., COX-2, 5-LO, and LTB4); pathway annotation would raise the confidence level in predictions as multiple targets from the same pathway are fished out. On the other hand, even incorrect target predictions contain useful information because an incorrect target prediction implies a level of chemical relatedness in the chemotypes that bind to the known target and the incorrectly predicted target. On closer inspection of MDDR Anti-inflammatory, Intestinal compounds EXTREG 167764 and 167767, we found them to be characterized as 5-lipoxygenase (5-LO) inhibitors. Although our model predicted PPAR rather than 5-LO as the most likely target, known endogenous ligands of 5-LO (arachidonic acid) and PPAR (eicosanoids such as leukotriene B4) are clearly closely related.[47] If the endogenous ligands are chemically related, small-molecule modulators of the two targets may look chemically similar. Consistent with this discussion, the COX-2 inhibitor celecoxib was found recently to be an agonist of PPARα.[48]

Successful predictions are fostered by structural diversity among compounds in the chemogenomics database. To compensate for a lack of topological structural diversity, it may be of interest to also incorporate 3D chemical descriptors that encode pharmacophoric information in an effort to encourage scaffold hopping.[23,49]

The automated nature of multiple-category naïve Bayesian models trained on the knowledge in chemogenomics databases lends itself to the creation of very large predicted chemogenomics databases that would enable additional data mining, such as searching for common bioisosteres that link targets, families, or therapeutic indications. However, MCMs could be used for a number of different applications. For example, target-focused or family-focused libraries could be created from large vendor or corporate collections by selecting only compounds with the highest naïve Bayes scores for a relevant target. Conversely, if one wanted to avoid finding bioactive compounds that resemble known chemotypes,[50] lead compounds could be prioritized according to those with the lowest Bayes scores for relevant targets to encourage novelty. The MCM could also be used to predict or explain off-target effects, especially if trained on a database of targets relevant to preclinical profiling. Importantly, in silico target prediction is highly complementary to experimental target fishing methods that are vital to the emerging field of chemical genetics.

ACKNOWLEDGMENT

The authors thank Drs. Mahesh Merchant and Douglas Perry (IUPUI School of Informatics) for their support and the following people for their comments: Drs. Edgar Jacoby, Jim Nettles, John Tallarico, Dmitri Mikhailov, Ansgar Schuffenhauer, Kamal Azzaoui, Markus Dobler, Dejan Bojanic, Christian Parker, Zhan Deng, and Andreas Bender of Novartis and Dave Rogers and the SciTegic staff.

Supporting Information Available: A table of the full names and associated abbreviations for relevant targets. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Peterson, R. T.; Shaw, S. Y.; Peterson, T. A.; Milan, D. J.; Zhong, T. P.; Schreiber, S. L.; MacRae, C. A.; Fishman, M. C. Chemical Suppression of a Genetic Mutation in a Zebrafish Model of Aortic Coarctation. *Nature Biotech.* **2004**, *22*, 595−599.
(2) Stockwell, B. R. Chemical Genetics: Ligand-Based Discovery of Gene Function. *Nat. Rev. Genet.* **2000**, *1*, 116−125.
(3) Stockwell, B. R. Exploring Biology with Small Organic Molecules. *Nature* **2004**, *432*, 846−854.
(4) Szardenings, K.; Li, B.; Ma, L.; Wu, M. Fishing for Targets: Novel Approaches Using Small Molecule Baits. *Drug Discovery Today: Targets* **2004**, *1*, 9−15.
(5) Bredel, M.; Jacoby, E. Chemogenomics: An Emerging Strategy for Rapid Target and Drug Discovery. *Nat. Rev. Genet.* **2004**, *5*, 262−275.
(6) Root, D. E.; Flaherty, S. P.; Kelley, B. P.; Stockwell, B. R. Biological Mechanism Profiling Using an Annotated Compound Library. *Chem. Biol.* **2003**, *10*, 881−892.
(7) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity Metrics for Ligands Reflecting the Similarity of the Target Proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391−405.
(8) Sheridan, R. P. Finding Multiactivity Substructures by Mining Databases of Drug-Like Compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1037−1050.
(9) Sheridan, R. P.; Shpungin, J. Calculating Similarities between Biological Activities in the MDL Drug Data Report Database. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 727−740.
(10) Lagunin, A.; Stepanchikova, A.; Filimonov, D.; Poroikov, V. PASS: Prediction of Activity Spectra for Biologically Active Substances. *Bioinformatics* **2000**, *16*, 747−748.
(11) Poroikov, V.; Akimov, D.; Shabelnikova, E.; Filimonov, D. Top 200 Medicines: Can New Actions Be Discovered through Computer-Aided Prediction? *SAR QSAR Environ. Res.* **2001**, *12*, 327−344.
(12) Olah, M.; Mracec, M.; Ostopovici, L.; Rad, R.; Bora, A.; Hadaruga, N.; Olah, I.; Banda, M.; Simon, Z.; Mracec, M.; Oprea, T. I. WOMBAT: World of Molecular Bioactivity. In *Cheminformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: New York, 2004; pp 223−239.
(13) Elsevier MDL Home Page. http://www.mdli.com (accessed Dec 2005).
(14) Inpharmatica Index. http://www.inpharmatica.co.uk (accessed Dec 2005).
(15) Jubilant Biosys. http://jubilantbiosys.com (accessed Dec 2005).
(16) GVK Bio. http://www.gvkbio.com (accessed Dec 2005).
(17) Chembank − Initiative for Chemical Genetics. http://chembank.broad.harvard.edu (accessed Dec 2005).
(18) NIH Roadmap for Medical Research − Molecular Libraries and Imaging Overview. http://nihroadmap.nih.gov/molecularlibraries/index.asp (accessed Dec 2005).
(19) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177−1185.
(20) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Topological Descriptors for Similarity-Based Virtual Screening Using Multiple Bioactive Reference Structures. *Org. Biomol. Chem.* **2004**, *2*, 3256−3266.
(21) Bender, A.; Mussa, H. Y.; Gill, G. S.; Glen, R. C. Molecular Surface Point Environments for Virtual Screening and the Elucidation of Binding Patterns (MOLPRINT 3D). *J. Med. Chem.* **2004**, *47*, 6569−6583.
(22) Haggarty, S. J.; Clemons, P. A.; Schreiber, S. L. Chemical Genomic Profiling of Biological Networks Using Graph Theory and Combinations of Small Molecule Perturbations. *J. Am. Chem. Soc.* **2003**, *125*, 10543−10545.
(23) Jenkins, J. L.; Glick, M.; Davies, J. W. A 3D Similarity Method for Scaffold Hopping from Known Drugs or Natural Ligands to New Chemotypes. *J. Med. Chem.* **2004**, *47*, 6144−6159.
(24) Rush, T. S., III; Grant, J. A.; Mosyak, L.; Nicholls, A. A Shape-Based Scaffold Hopping Method and Its Application to a Bacterial Protein−Protein Interaction. *J. Med. Chem.* **2005**, *48*, 1489−1495.
(25) Xia, X.; Maliski, E. G.; Gallant, P.; Rogers, D. Classification of Kinase Inhibitors Using a Bayesian Model. *J. Med. Chem.* **2004**, *47*, 4463−4470.
(26) Glick, M.; Klon, A. E.; Acklin, P.; Davies, J. W. Enrichment of Extremely Noisy High-Throughput Screening Data Using a Naive Bayes Classifier. *J. Biomol. Screening* **2004**, *9*, 32−36.
(27) Rogers, D.; Brown, R. D.; Hahn, M. Using Extended-Connectivity Fingerprints with Laplacian-Modified Bayesian Analysis in High-Throughput Screening Follow-Up. *J. Biomol. Screening* **2005**, *10*, 682−686.
(28) Glick, M.; Jenkins, J. L.; Nettles, J. H.; Hitchings, H.; Davies, J. W. Enrichment of High-Throughput Screening Data with Increasing Levels of Noise Using Support Vector Machines, Recursive Partitioning, and Laplacian-Modified Naive Bayesian Classifiers. *J. Chem. Inf. Model* **2006**, *46*, 193−200.
(29) Klon, A. E.; Glick, M.; Thoma, M.; Acklin, P.; Davies, J. W. Finding More Needles in the Haystack: A Simple and Efficient Method for Improving High-Throughput Docking Results. *J. Med. Chem.* **2004**, *47*, 2743−2749.
(30) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures − A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107−113.
(31) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. New Methods for Ligand-Based Virtual Screening: Use of Data Fusion and Machine Learning to Enhance the Effectiveness of Similarity Searching. *J. Chem. Inf. Model.* **2006**, in press.
(32) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(33) Kouzarides, T. Histone Acetylases and Deacetylases in Cell Proliferation. *Curr. Opin. Genet. Dev.* **1999**, *9*, 40−48.
(34) Yoshida, M.; Kijima, M.; Akita, M.; Beppu, T. Potent and Specific Inhibition of Mammalian Histone Deacetylase both in Vivo and in Vitro by Trichostatin A. *J. Biol. Chem.* **1990**, *265*, 17174−17179.
(35) Vanhaecke, T.; Papeleu, P.; Elaut, G.; Rogiers, V. Trichostatin A-Like Hydroxamate Histone Deacetylase Inhibitors as Therapeutic Agents: Toxicological Point of View. *Curr. Med. Chem.* **2004**, *11*, 1629−1643.
(36) Meinke, P. T.; Liberator, P. Histone Deacetylase: A Target for Antiproliferative and Antiprotozoal Agents. *Curr. Med. Chem.* **2001**, *8*, 211−235.
(37) Hayakawa, Y.; Nakai, M.; Furihata, K.; Shin-ya, K.; Seto, H. Trichostatin D, a New Inducer of Phenotypic Reversion in Transformed Cells. *J. Antibiot.* **2000**, *53*, 179−183.
(38) Berger, M. R.; Bicshoff, H.; Fritschi, E.; Henne, T.; Herrmann, M.; Pool, B. L.; Satzinger, G.; Schmal, D.; Weiershausen, U. Synthesis, Toxicity, and Therapeutic Efficacy of 4-Amino-N-(2′-aminophenyl)-benzamide: A New Compound Preferentially Active in Slowly Gowing Tumors. *Cancer Treat. Rep.* **1985**, *69*, 1415−1424.
(39) Saito, A.; Yamashita, T.; Mariko, Y.; Nosaka, Y.; Tsuchiya, K.; Ando, T.; Suzuki, T.; Tsuruo, T.; Nakanishi, O. A Synthetic Inhibitor of Histone Deacetylase, MS-27−275, with Marked in Vivo Antitumor Activity Against Human Tumors. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 4592−4597.
(40) Lecci, A.; Capriati, A.; Maggi, C. A. Tachykinin NK2 Receptor Antagonists for the Treatment of Irritable Bowel Syndrome. *Br. J. Pharmacol.* **2004**, *141*, 1249−1263.
(41) Barbara, G.; De Giorgio R.; Stanghellini, V.; Corinaldesi, R.; Cremon, C.; Gerard, N.; Gerard, C.; Grady, E. F.; Bunnett, N. W.; Blenner-hassett, P. A.; Collins, S. M. Neutral Endopeptidase (EC 3.4.24.11) Downregulates the Onset of Intestinal Inflammation in the Nematode Infected Mouse. *Gut* **2003**, *52*, 1457−1464.
(42) Hyun, C. S.; Binder, H. J. Mechanism of Leukotriene D4 Stimulation of Cl^- Secretion in Rat Distal Colon in Vitro. *Am. J. Physiol.* **1993**, *265*, G467−G473.
(43) Ardizzone, S.; Porro, G. B. Biologic Therapy for Inflammatory Bowel Disease. *Drugs* **2005**, *65*, 2253−2286.
(44) Gradman, A. H.; Schmieder, R. E.; Lins, R. L.; Nussberger, J.; Chiang, Y.; Bedigian, M. P. Aliskiren, a Novel Orally Effective Renin Inhibitor, Provides Dose-Dependent Antihypertensive Efficacy and Placebo-Like Tolerability in Hypertensive Patients. *Circulation* **2005**, *111*, 1012−1018.
(45) Wood, J. M.; Schnell, C. R.; Cumin, F.; Menard, J.; Webb, R. L. Aliskiren, a Novel, Orally Effective Renin Inhibitor, Lowers Blood Pressure in Marmosets and Spontaneously Hypertensive Rats. *J. Hypertens.* **2005**, *23*, 417−426.

(46) Valenzuela, F.; Garcia-Saiso S.; Lemini, C.; Ramirez-Solares, R.; Vidria, H.; Mendoza-Fernandez, V. Metamizol Acts as an ATP Sensitive Potassium Channel Opener to Inhibit the Contracting Response Induced by Angiotensin II but not to Norepinephrine in Rat Thoracic Aorta Smooth Muscle. *Vasc. Pharmacol.* **2005**, *43*, 120−127.

(47) Funk, C. D. Prostaglandins and Leukotrienes: Advances in Eicosanoid Biology. *Science* **2001**, *294*, 1871−1875.

(48) Lopez-Parra, M.; Claria, J.; Titos, E.; Planaguma, A.; Parrizas, M.; Masferrer, J. L.; Jimenez, W.; Arroyo, V.; Rivera, F.; Rodes, J. The Selective Cyclooxygenase-2 Inhibitor Celecoxib Modulates the Formation of Vasoconstrictor Eicosanoids and Activates PPARgamma. Influence of Albumin. *J. Hepatol.* **2005**, *42*, 75−81.

(49) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed.* **1999**, *38*, 2894−2896.

(50) Oprea, T. I.; Bologa, C. G.; Edwards, B. S.; Prossnitz, E. R.; Sklar, L. A. Post-High-Throughput Screening Analysis: An Empirical Compound Prioritization Scheme. *J. Biomol. Screening* **2005**, *10*, 419−426.

CI060003G