Bridging Chemical and Biological Space: "Target Fishing" Using 2D and 3D Molecular Descriptors

James H. Nettles,* Jeremy L. Jenkins, Andreas Bender, Zhan Deng, John W. Davies, and Meir Glick

Lead Discovery Informatics, Lead Discovery Center, Novartis Institutes for BioMedical Research Inc., 250 Massachusetts Avenue, Cambridge, Massachusetts 02139

Received July 28, 2006

Bridging chemical and biological space is the key to drug discovery and development. Typically, cheminformatics methods operate under the assumption that similar chemicals have similar biological activity. Ideally then, one could predict a drug's biological function(s) given only its chemical structure by similarity searching in libraries of compounds with known activities. In practice, effectively choosing a similarity metric is case dependent. This work compares both 2D and 3D chemical descriptors as tools for predicting the biological targets of ligand probes, on the basis of their similarity to reference molecules in a 46 000 compound, biologically annotated chemical database. Overall, we found that the 2D methods employed here outperform the 3D (88% vs 67% success) in correct target prediction. However, the 3D descriptors proved superior in cases of probes with low structural similarity to other compounds in the database (singletons). Additionally, the 3D method (FEPOPS) shows promise for providing pharmacophoric alignment of the small molecules' chemical features consistent with those seen in experimental ligand/ receptor complexes. These results suggest that querying annotated chemical databases with a systematic combination of both 2D and 3D descriptors will prove more effective than employing single methods.

1. Introduction

The relationship between chemical structure and biological function is the basis of modern drug discovery and development.[1,2] Current methods of cellular screening and pharmacogenetic profiling can rapidly reveal phenotypic responses to drugs but do not immediately pinpoint their molecular target.[3,4] Affected pathways may require broad arrays of secondary assays to define the specific binding partner(s) before reaching the goal of "rationally" optimizing the hit into a potent and selective lead.[5] Streamlining chemogenomics-based discovery,[6] through the assistance of *in silico* target fishing, is one of the long-range goals of this study. Many computational studies that have been performed to evaluate chemical/ biological relations have done so from the perspective of "virtual screening", searching a small molecule library for compounds with similar activity to a single, known, biological target.[7,8] Our computational study is designed toward a different experimental goal: identifying the molecular target for a single chemical entity, or "target fishing", based on similarity of a new compound to structures where activities against a broad panel of targets is already known (i.e., an annotated compound library).

Ongoing efforts to combine and curate databases relating large numbers of diverse chemical structures with their biological activities hold the promise of revealing patterns that provide new insight into the molecular features responsible for such activities. Examples of those databases are manifold, such as the StARLITe database,[9] BioPrint,[10] and WOMBAT.[11] Preliminary success has recently been reported from combining large data sets.[12] However, effective use of such databases depends on having reliable methods of relating the chemical structure of the query compound to the reference compounds in the database. Many methods of assigning and quantifying molecular similarity exist and have been recently reviewed elsewhere.[13] To be effective in a large pharmaceutical environment, an optimal method needs to be fast and sufficiently robust to process millions of similarity calculations. A common technique for high throughput cheminformatics analysis is the reduction of a given molecular structure into a set of "descriptors" that can be rapidly compared and evaluated numerically.

Existing 2D methods to describe molecules, which are based solely upon the topological connectivity of a molecular structure, are very fast given that the "problem" of exploring the conformational space is ignored. It is expected that such methods would be useful for clustering similar compounds,[14] or selecting diverse subsets from large libraries,[15] but these methods are also surprisingly effective for "virtual screening" of actives from large compound sets using the structures of a small number of known actives as probes.[2,16,17] Although 3D methods are computationally more expensive due to a need to consider conformers, tautomers, charge distribution, alignment, etc., they have arguably performed less well in comparative virtual screening exercises,[18,19] though there are exceptions.[20−24] It seems intuitive that 3D methods would include greater information since the binding between a ligand and receptor is a 3D event. It should be noted that early 2D methods such as Daylight or Unity fingerprints were initially designed to find "more of the same" while 3D methods can enrich scaffold diversity. Recent work has shown the cost/benefit appears to be case dependent.[25] One goal of this work, is to examine methods for predicting a chemical probe's biological target based upon similarity to a reference compound (*in silico* "target fishing"). A second goal is to examine the relation of ligand-based 3D alignments to actual 3D receptor binding. We are particularly interested in identifying relations between novel chemical classes, through "scaffold hops".[26,27] In this study, we have tested two different 2D similarity methods (Scitegic's Extended Connectivity Fingerprints, and MDL's Public Keys)[28,29] and one 3D similarity method (FEature POint PharmacophorS)[30] using identical data to assess both overall accuracy, and versatility. Performance of the methods is specifically examined for natural products and compounds without structural neighbors, single-

* Corresponding author. E-mail: James.Nettles@novartis.com. Phone: (617) 871-3904. Fax: (617) 871-4088.

10.1021/jm060902w CCC: $33.50
Published on Web 10/19/2006

tons. Singletons are inherently challenging for similarity methods. Specifically, we have examined potential for complementarity between different descriptor types, by varying relevant information (that is, similar molecules) present in the database. One hypothesis, supported by the results, is that 2D methods are favored in case of close analogues, but that 3D methods offer advantages below a certain similarity threshold.

2. Methods

(A) Database, Bioactivity Data, and Database Preparation. WOMBAT 2005.1 (WOrld of Molecular BioAcTivity) database.,[11] containing over 100 000 unique structures described as SMILES and over 240 000 biological activities in separate ISIS databases, was merged into a single delimited text file for manipulation and analysis in PipelinePilot 5.0.[28] To reduce artifacts from nonspecific binding, only those compounds having measured IC_{50} activity <30 μm were used in our current analysis. The final groomed database contained 47 505 unique chemical structures associated with 544 biological targets. The frequency of molecules affecting specific targets ranged from 1281 to 1. A target class with only 1 member cannot be located with our procedure because "self" is removed from the comparison. However, they were left in to serve as decoys during the search.

(B) 2D Target Fishing: MDL Structural Keys and ECFP_6 Fingerprints. I. All WOMBAT: The 2D similarity protocols, using each descriptor type, were run against the groomed database using the complete set of 47 505 chemical structures as probes.

II. 5%WOMBAT: A smaller probe set of 2351 molecules reflecting a randomly chosen 5% of the entire chemical library was generated using the Random Percent Filter in Pipeline Pilot. Second runs against the full database were performed with the 2351 molecules set using both 2D descriptor types.

2D descriptors for all compounds were computed using both MDL public keys and SciTegic's Extended Connectivity Fingerprints (ECFP_6) in Pipeline Pilot. Pairwise similarity comparisons between the compounds were done with SciTegic's Tanimoto component. Both the first and second "nearest neighbor" for each probe, with Tanimoto similarity <0.99, were flagged for subsequent evaluation. Only the first nearest neighbor was used as the reference compound for target identification and numerical scoring.

(C) 3D Target Fishing: FEPOPS. All 47 505 unique compounds were input as SMILES strings. FEature POint PharmacophoreS (FEPOPS) were calculated as described below and stored as text in a 3D descriptor database (3DDD) yielding 815 676 records with compound IDs and associated feature point information. Similarity was determined by calculating Pearson correlations between the four atomic feature points of each probe/reference pair.[31]

I. All WOMBAT: 3D analysis was not run on the full data due to computational expense.

II. 5%WOMBAT: The same 2351 chemical structures, used for the 2D comparison, were expanded with all FEPOPS conformer/tautomers, as described below (40 852 entries), and compared to the residual set of 774 824 descriptors. (Since we are search the database against itself, if probes are not removed the algorithm will find the probe itself as the closest reference structure.) The single compound having the maximal Pearson correlation of FEPOPS descriptors to the probe was selected as the reference for target assignment. Analysis of the $40\,852 \times 774\,824$ system takes approximately 17 h on our Pipeline Pilot server.

The FEPOPS descriptors were computed for each database entry according to the method of Jenkins et al.[30] Using this method, compounds are preprocessed to generate 3D structures, assign protonation states, enumerate tautomers, and calculate partial charges and atomic log *P* values. Multiple conformers are generated by systematic rotation of flexible bonds. Ligand atoms are partitioned into four *k*-mean clusters based upon their spatial coordinates. Centroids are defined from the atoms of each cluster. Partial charges, log *P*, and hydrogen bond donors and acceptors of the atoms belonging to each cluster are summed and encoded into the centroids to create "feature points". The distances between feature points are recorded after sorting on the basis of quadrupole directionality. *K*-medoids clustering of feature points is performed to find a smaller number of representative conformers. Up to seven conformations for each of five tautomers may be retained, making a maximum of 35 different 3D structures possible for each unique 2D chemical entry.

(D) Fishing for Chemical Diversity with 3D Descriptors. I. Nearest Neighbor Misses (NNmiss). A subset of the probes from the "All WOMBAT" dataset that did not achieve target matches using 2D similarity criteria (see Results) was selected for additional 3D searching. The rationale for working with this subset was 3-fold. First, we wanted to assess 3D performance in cases where 2D fails. Second, the smaller number of probes reduced computing times and allowed for multiple conditions to be examined as described in the next section. Third, using a randomly selected subset of the total nearest neighbor misses as probes left the remainder of the 2D misses in the 3DDD. Since certain target classes have representatives with low 2D similarity in the parent database, this criteria was important for reasonable evaluation of 3D performance (see Discussion).

II. Similarity Filters. To explore the added value of the 3D FEPOPS method for finding correct probe/reference pairings with low structural similarity, the NNmiss set of probes (see Results and Discussion) was run multiple times against the 3DDD with different filters. The initial test ran the NNmiss probe set against the full 3DDD with only the probes removed. Separate runs subsequently removed sets of compounds having similarity scores to the probes greater than ECFP 0.85, MDL 0.85, MDL 0.80, and MDL 0.60 from the 3DDD when searched.

III. Probe for related chemistry/biology: ATP WOMBAT. A typical prospective exercise for "target fishing" would be to use a single molecule with observed activity as a probe to identify possible binding partners. As an example, FEPOPS descriptors were calculated for ATP and used to identify potential targets based upon 3D similarity to other molecules in the WOMBAT 3DDD. The use of such a highly flexible probe molecule was intended to explore 3D's ability to located correct functional targets based upon difficult, low similarity, reference structures.

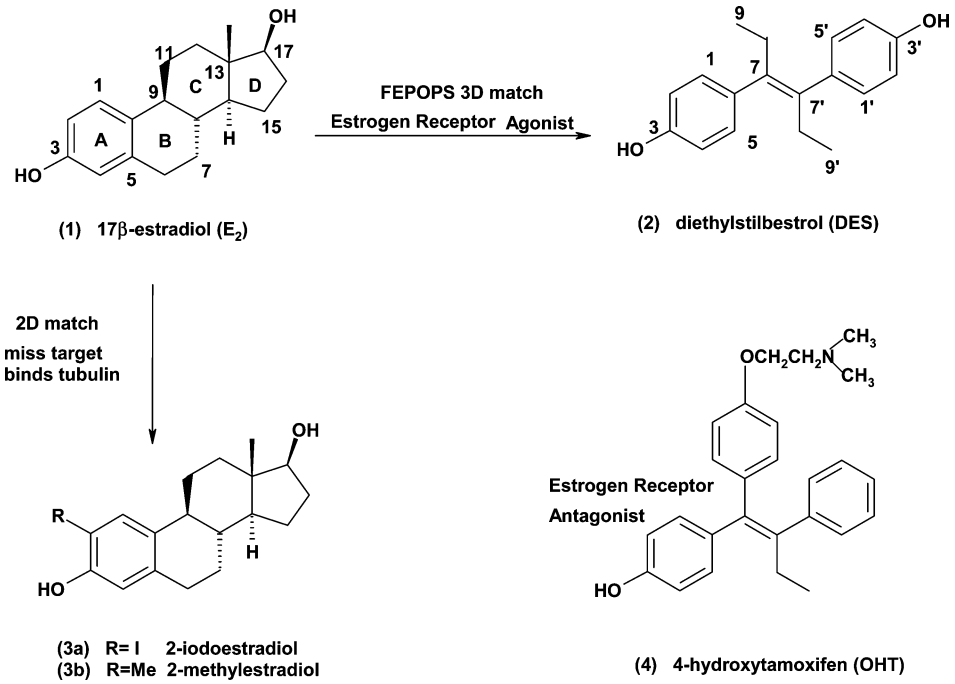
(E) Modeling of 3D Ligand Alignments in Biological Space. I. Analysis of Biostructural Correlation. Molecular pairs with correct target matches from FEPOPS 3D NNmiss comparisons were output from Pipeline Pilot as SD files and input to MOE 2005.06 for structural analysis.[32] The goal of this exercise was to examine the binding mode for probe/reference drug pairs having dissimilar chemical scaffolds, but similar biological activity. A custom SVL script was used to search the Protein Data Bank (PDB) for protein targets that contain bound ligands with high structural similarity to probe/reference pairs identified by FEPOPS.[33] PDB searching was performed using the prebuilt PDB_05_04.mdb as provided in MOE 2005.06.

II. Mapping of FEPOPS Centroids. Atom IDs associated with feature points used for alignments were output from FEPOPS as mol2 files. These atom IDs were used to define centroids associated with a crystallographic or modeled ligand structures using Unity in Sybyl7.1. For the experimental systems described below, a spatial constraint was projected from each centroid and color coded to match the FEPOPS assignment. (The current version of the FEPOPS algorithm returns a mol2 with the atom IDs for the fit conformer/tautomer but not the actual coordinates. A feature returning coordinates with IDs is under development.)

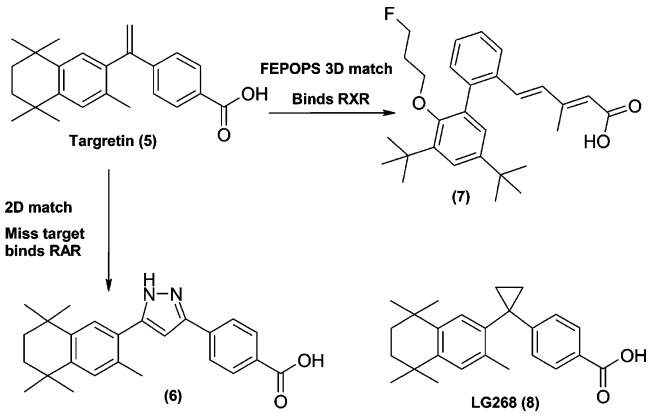
a. Modeling of Estrogen Receptor Ligands. Crystallographic coordinates of the estrogen receptor (ER) complexed with both the molecular probe 17-β-estradiol (Scheme 1), 1a52.pdb, and the reference compound diethylstilbestrol (2), 3erd.pdb, were downloaded from the PDB. Backbone atoms of residues Ser305-Phe461 (628 atoms) were aligned using the "Match" function in Syby 7.1 with an RMSD of 0.86A. FEPOPS centroids were mapped directly to the aligned experimental structures.

b. Modeling of Retinoid Receptors Ligands. A 3D model of the molecular probe targretin (Scheme 2, compound **5**) was generated

Scheme 1



Scheme 2



by modifying the cyclopropyl moiety of compound LG268 (**8**) as found in complex with the receptor in 1h9u.pdb[34] to the alkene of **5**. The new model was minimized using the Tripos force field to a gradient cutoff of 0.05 kcal/mol. The minimized model of **5** differed from the bound structure of **8** with a RMSD of 0.21A for all conserved atoms and was used to map FEPOP centroids.

The reference compound **7** was built in MOE and minimized using the MMFF94x to a gradient cutoff of 0.05 kcal/mol. A stochastic search of conformational space was performed in MOE with the following settings (bond rotation bias-30, Cartesian Perurbation Delta 0.001, Cartesian Minimization RMS Gradient 0.01, Energy Cutoff 7, Conformation Limit 1000, Failure Limit 20, Iteration Limit 1000, RMS Tolerance, MM Iteration Limit 200). Results were aligned, yielding two conformational clusters. FEPOPS centroids were mapped to the two lowest energy conformations representing both conformational clusters (relative energy difference 1.14 kcal/mol).

FEPOPS centroids were mapped to each conformer, and each was aligned to the centroids of **5** and evaluated as described in Results and Discussion**.**

c. Modeling of ATP/Balanol. Coordinates of protein complexes 1atp.pdb and 1bx6.pdb were downloaded from the PDB. Backbone atoms of residues Val15-Trp196 (728 atoms) were aligned using the "Match" function in Sybyl 7.1 with an RMSD of 0.92A. FEPOPS centroids were mapped directly to the aligned experimental structures.

Table 1. Results of "Target Fishing" Using 2D and 3D Chemical Descriptors

	total probes	total reference	match target	miss target	% success
All WOMBAT	47505	47505			
2DNN (ECFP-6)	47505	47505	42511	4994	89.5
2DNN (MDL)	47505	47505	41224	6281	86.8
3D FEPOPS	-	-	-	-	-
5% WOMBAT	2351	47505			
2DNN (ECFP-6)	2351	47505	2118	233	90.0
2DNN (MDL)	2351	47505	2046	305	87.0
3D FEPOPS	2351	45154	1588	768	67.5
NNmiss	339				
2DNN (ECFP-6)	339	47505	0	339	0
FEPOPS- self-filter	339	47166	109	230	32.1
FEPOPS < 0.85 ECFP filter	339	46933	91	248	26.8
FEPOPS < 0.85 MDL filter	339	43689	69	270	20.3
FEPOPS < 0.80 MDL filter	339	41103	61	278	18.0
FEPOPS < 0.60 MDL filter	339	5626	13	326	4.0

3. Results and Discussion

2D and 3D Target Fishing. The 2D Nearest Neighbor (2DNN) analysis was performed using MDL and ECFP descriptor keys as described in Methods and was run on both the full and 5% data sets. Raw numerical results are presented in Table 1. A control comparison illustrating compound distribution in the All WOMBAT and 5% WOMBAT dataset as a function of similarity are shown in Figure 1. 3D FEPOPS results using the 5%WOMBAT probe set and the 2DNN (ECFP-6) "Miss Target" with applied 2D filters are also given in Table 1.

Figure 1 shows the composition of both the full database and the 5% set, using both 2D descriptors. MDL and ECFP_6 keys were used to calculate Tanimoto similarity between each compound and its nearest neighbor in the full (blue) and 5% (red) WOMBAT data sets as described in methods. The compounds were placed into one of 10 bins based upon similarity scores between 0 and 1. The MDL keys cluster 80% of compounds in both the full and 5% database into the bin of 0.90 or greater similarity. The number of compounds with nearest neighbors less than 0.80 similarity represent less than 5% of the total databases. ECFP keys show that the greatest

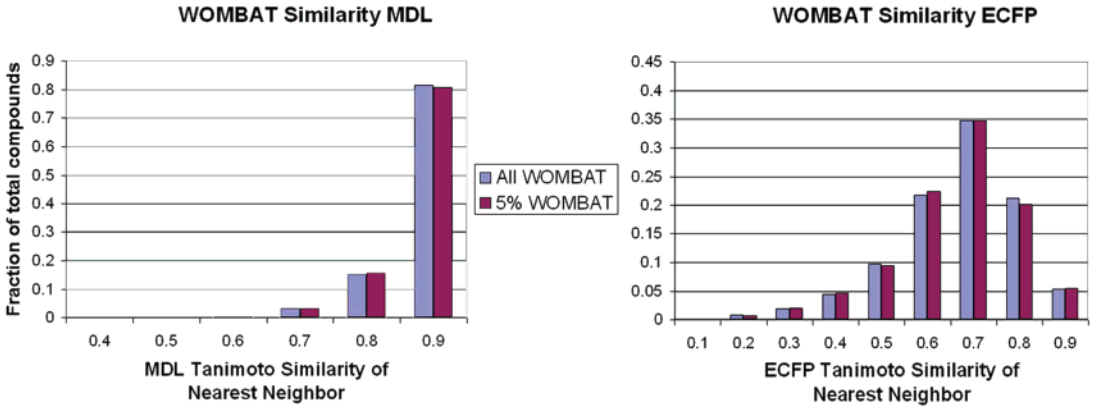


Figure 1. Comparison of data sets used in target exercise. Heights of blue bars indicate the fraction of molecules, in all of WOMBAT, having a nearest neighbor (NN) with similarity greater than or equal to the range represented by the *X*-axis value. Height of red bars displays the proportion of similar molecules in a randomly selected subset as compared to all of WOMBAT. The close correlation using both MDL and ECFP descriptors suggests that results from searches performed using the 5% subset can represent results from the whole database.

population of neighbors has similarity between 0.70 and 0.80 with a distribution on either side of that maximum. Like that shown with the MDL keys, there is a small population of compounds in the database that have no similar neighbor, the singletons (<0.40 ECFP, <0.80 MDL). The similar distribution of molecular pairs from this figure, along with the numerical results given in Table 1, suggests that the 5% set adequately represents the full dataset, for use in the 3-way comparison including the 3D method.

Figure 1 also illustrates that there is significant difference between the ECFP fingerprints and MDL public keys descriptors. The MDL public key method was one of the first and is still a widely used tool in cheminformatics.[29] The "Public" keys describe a chemical entity based upon the presence or absence of 166 substructural fragments from a predefined library. The Extended Connectivity FingerPrint (ECFP) is a representation assigned based upon each atom and its structural neighbors and belongs to the group of circular fingerprints.[28,35] The fingerprint is built iteratively by adding bits that represent larger and larger substructures to the features from all the previous steps. Accordingly, ECFP may assign thousands or millions of bits to describe a given molecule. The difference shown between Figure 1a and 1b reflects this variation of assignment method used. It should be noted that the lower Tanimoto similarity values for ECFP relative to MDL is a function of different scales. Tanimoto values should only be compared when computed using the same method. The granularity of separation between similar compounds is much greater for the ECFP method than the MDL method, but, probably due to the congeneric nature of our data, this does not significantly affect performance. [See Table 1: All WOMBAT Match Target (ECFP 42 511 correct, MDL 41 224 correct)]. Given a differently distributed set of probes, they may not perform equally. For example, in a study of similarity-based virtual screening using MDDR compounds, ECFP circular descriptors were found superior to structural key descriptors.[17, 35]

The performance of both types of 2D descriptors in our 5% WOMBAT analysis (success ECFP-6 90.0%, MDL 87.0%) may be elevated due to the composition of the reference database. Since WOMBAT is primarily comprised of congeneric series from chemistry programs reported in the literature, the high percentage of molecules with a close 2D neighbor and same target is to be expected. If our probes had not been pulled from the same congeneric sets, it is likely that there would be at least a small reduction in overall performance of the 2D *in silico* methods.

The 3D FEPOPS analysis was performed using the same 5% probe set as was used for the 2D analysis. A cursory examination of the raw percentage of successful results returned from our comparative exercise (ECFP 90%, MDL 87%, FEPOPS 68%) may tempt one to conclude that *in silico* methods for target fishing are highly effective, that 2D descriptors significantly outperform 3D, and that there is little difference between the 2D methods tested. *Deeper analysis reveals important qualifiers to those three assertions.*

Figure 2 highlights the relative performance of each of the 2D descriptors compared to the 3D method based upon the topological similarity between the probe and reference compound used to determine its target. Tanimoto similarity for 1588 correct 3D FEPOPS matches was computed using both ECFP-6 and MDL descriptors and plotted beside the results obtained using that 2D method. Light green and dark green bars illustrate the comparable performance of ECFP keys for prediction of targets in both the full WOMBAT database and the 5% random sample. The yellow bars associated with the 3DMS(FEPOPS) pairing show low correlation with 2D similarity. In fact, FEPOPS finds approximately 75% correct matches across the most populated chemistry space, 0.4−1.0 ECFP (see Figure 1 and discussion in text). The performance of 3D(FEPOPS) exceeds the 2D in correct matching for similarity pairings less than 0.4 ECFP. The blue bars, in Figure 2b, showing correct matching of target pairs using MDL descriptors marks dramatic decline in performance with decreasing similarity. The 3D(FEPOPS) method out performs the MDL 2D descriptors for values less than 0.80 MDL. The similarity dependence of 2D methods is also suggested by results from a recent study which built multiclass Laplacian-modified naive Bayesian models trained on WOMBAT using ECFP keys but probed with MDL Drug Data Report (MDDR) compounds.[36] That study reported a 77% success for prediction of targets, however, drawing a direct numerical correlate from that work is complicated by inherent differences of annotation between the two databases used and details of data preparation.

To test the potential of 3D as a compliment for 2D methods, we randomly selected a small subset of the compounds that failed to match using 2DNN (ECFP-6) (NNmiss) for further FEPOPS analysis. The raw numerical results for all cases are presented in Table 1. For the NNmiss set with only a "self" filter, 3D resulted in 32.1% success compared to 0% for 2D. A detailed examination of these results showed that the matches found by 3D were still quite similar in 2D chemical space. As one of our goals was to explore the 3D method's ability to bridge

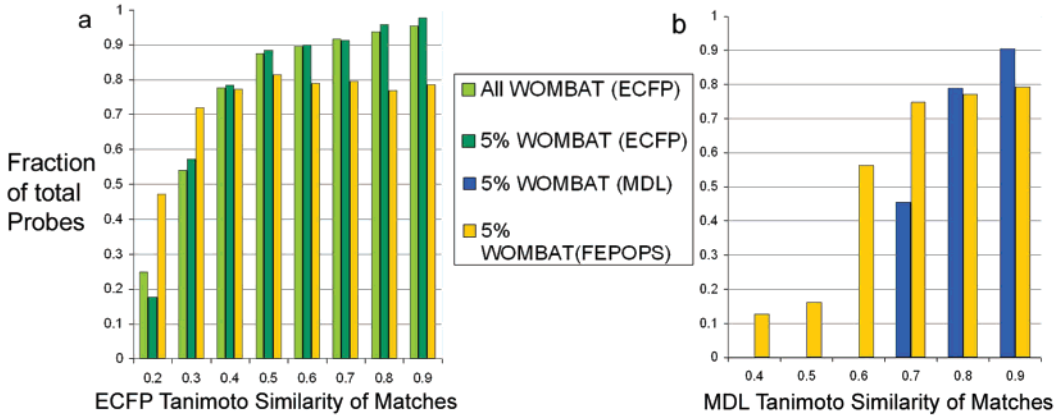


Figure 2. 2D and 3D success for target prediction as a function of 2D similarity. The similarity between probe and reference pairs used to successfully predict biological targets are compared measures of 2D similarity. The 2D similarity of matching pairs obtained using 3D FEPOPS were calculated using ECFP-6 (a) and MDL public keys (b). The 3D FEPOPS results are displayed as gold bars on the same graph with results from the particular 2D method.

distant chemical spaces that share biological function, we systematically removed molecules with 2D similarity to the probe from the reference database before searching. Consistent with the population distributions of similar compounds shown in Figures 1 and 2 and discussed earlier, we found that the results of filtering above the 0.85 level using either 2D descriptor did not return appreciably better structural diversity. There were a sufficient number of similar molecules left in the database for 3D to pick them as the closest match. When we used MDL filters that removed all molecules from the 3DDD with similarity less than 0.80 from reference set, the results changed significantly. The full results set with structural drawings and 2D Tanimoto indices for the 3D similar pairs in the 0.80 and 0.60 analysis are shown in the Supporting Information (SI).

The idea that a ligand centric 3D method may play a role in understanding the similar biological activity of structurally dissimilar drugs has precedent in the literature. Inverse docking protocols have been used to identify potential targets, using small molecule probes,[37] but this is limited by the availability of receptor structures for the targets of interest. Pharmacophore and 3D-QSAR-based methods have also shown promise for both identification of potential nontarget interactions and scaffold hopping within a given target but have a dependence upon the external knowledge used to train the models.[38,39] In the case of target fishing, one may not have external data, and the target information would need to be decoded from a single chemical probe. Although not explicitly attempting target identification, others have projected distance-based pharmacophoric features from the topology of a probe[21,26] or used explicit shape and electrostatics as descriptors for recovering actives.[24,40] FEPOPS are an inherently fuzzy description of a molecules potential shape(s) and chemical space. As described in Methods and earlier work by Jenkins et al., the feature points are representative of clusters of conformations available to different tautomers of the small molecule.[30] This approach is very fast compared to 3D-QSAR or structure-based methods and can compare all tautomer/conformer representations of a single probe to all 815 676 records of our 3DDD to produce a ranked alignment of features in approximately 2 min. In the current application of the method, each representation holds four spatially separated feature points that are associated with specific atoms of the small molecule. We hypothesized that the maximally overlapped feature points derived from exploring the chemistry space of molecules that affect the same target should correlate with the biological space of the actual target.

To test FEPOPS ability for predicting a biologically relevant pharmacophore from only the chemical description, we searched for examples having experimental crystal structure complexes that contained both our probe and reference structures.

Estrogen Receptor. Endogenous estrogens such as estradiol (E_2) (**1**, Scheme 1) exert their physiological effects by binding to estrogen receptors (ER), inducing nuclear translocation, and increasing transcription.[41] The synthetic nonsteroidal compound diethylstilbestrol (DES) (**2**) found by our 3D target fishing protocol also binds to ER with high affinity and similarly increases transcriptional events. By mapping FEPOPS-defined centroids onto the structural complexes of E_2 and DES bound to ERα, we can see the similar alignment of chemical features associated with the program's correct prediction of biological target, Figure 3a,b.[42,43] The coloring scheme (red, green, blue, violet) represents the algorithm's assignment of feature point (FP) number based upon the chemical properties of the underlying atoms. Atoms defined as FP 1 for each these compounds are found similarly located in a charged, polar environment of the receptor. The atoms assigned within FP 2 (green) present a hydrogen bond donating group within an otherwise hydrophobic portion of the site. Of course, the symmetry of DES makes distinction between atoms of FP 1 vs 2 and FP 3 vs 4 only possible relative to its alignment with the asymmetry of E_2. The significant differences of structural scaffolding presenting FP 3 (blue) and FP 4 (violet) in the two molecules highlights the utility of the 3D method. The 2D Tanimoto similarities between these molecules are 0.42-MDL and 0.13-ECFP while the 3D Pearson correlation of the feature points is 0.91.

We immediately questioned why the 2D similarity had erroneously predicted the target given such a heavily studied system as ER. Compounds **3a** and **3b** of Scheme 1 with similarities of 0.62, 0.60 (ECFP_6) 0.89, 0.97 (MDL) were the first and second nearest neighbors of E_2 **(1)**. It has been shown that the estrogenic properties of E_2 are diminished by O-methylation at the 2 position, and other substitutions, such as I or Me, likewise change the binding preference of the compounds from ER to tubulin.[44] The 3D-FEPOPS alignments of these molecules with Pearson coefficients of 0.89 and 0.83 reverses the orientation of the fused ring system 180° along the axis between the diols, suggesting that these molecules would not have the same binding mode. The large effect of small changes at the 2-position upon target specificity underscores the 3D nature of the binding event and an inherent limitation of 2D

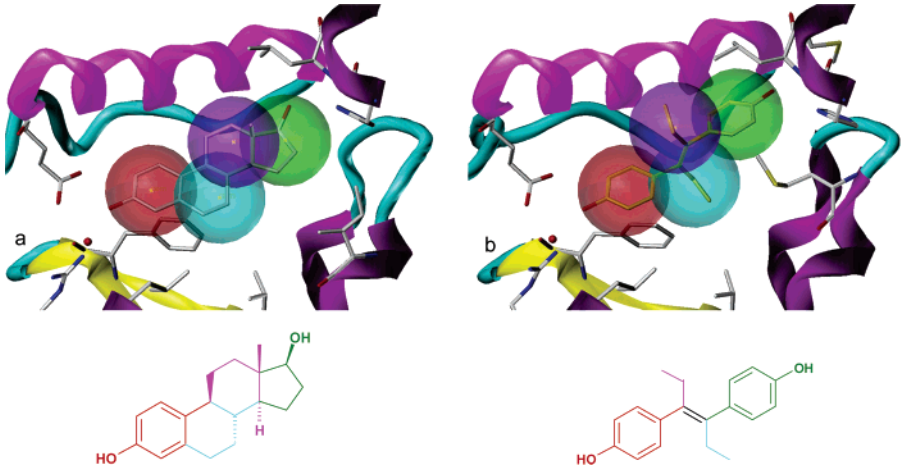


Figure 3. FEPOPS chemical space alignment in experimental biological space. Estrogen receptor (ER) agonists (a) β-estradiol as bound in 3erd.pdb and (b) diethylstilbestrol; DES, as bound in 1a52.pdb. Centroids, displayed as 2.5 Å diameter spheres, were computed from the crystallographic coordinates using the FEPOPS feature point atom definitions represented in the color coded structural drawings as described in the main text.

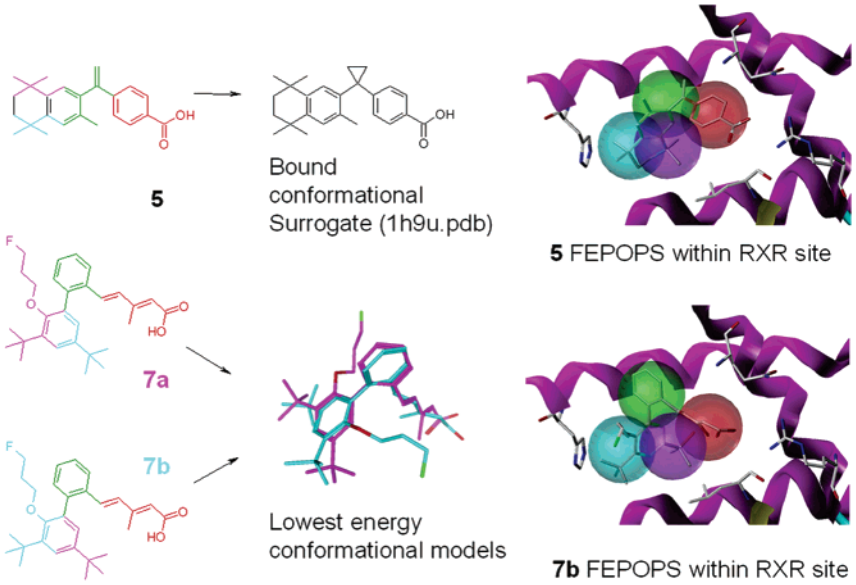


Figure 4. Chemical feature points in receptor space as models of biological function. FEPOPS centroids of the RXR selective agonist **5** were used to align the RXR selective antagonist **7** (see Scheme 2) in the context of the biological receptor. Retinoid receptor (RXR alpha) agonists probe (a) targretin and top FEPOPS hit; (b) diethylstilbestrol, DES, as modeled into structure 1h9u.pdb. Compound **5** was modeled into the protein structure, as described in Methods, using the bound ligand surrogate shown. Centroids, displayed as 2.5 Å diameter spheres, were computed from the model coordinates using the feature point atom definitions as represented by the color coded structural drawings. The FEPOPS centroids defined by the low energy conformations of **6** were used to position the molecule in the binding site.

similarity descriptions alone. Likewise, compound **4** is a variation of the clinical antiestrogen Tamoxifen that builds upon the structural framework of compound **2**. Although **4** can bind similarly to **2**, additional bulk at the 7 position induces positional change of a helix within the receptor such that it inhibits the binding of a coactivator peptide needed for downstream signaling.[43] This general idea of directed 3D effector regions is expanded in the next example.

RXR Receptor. The retinoid receptors, characterized by the subfamilies of retinoic acid receptors (RAR) and retinoid X receptors (RXR), serve as obligate binding partners in many cell signaling pathways affecting, cell differentiation, proliferation, and tissue homeostasis.[45] Two forms of retinoic acid are primarily responsible for the varied biological response. *All-trans*-retinoic acid (ATRA) binds exclusively to RAR, while 9-*cis*-retinoic acid (9-*cis*-RA) is primarily responsible for RXR signaling along with RAR activation. Improving selectivity of retinoid targeted drugs is pivotal for improving their medical use.[45,46]

Targretin (**5**), shown in Scheme 2, was one of the earlier molecules discovered with selectivity for the RXR family of receptors,[47] and was incorrectly assigned RAR activity based upon 2D similarity to the weak RAR antagonist **6**.[48] FEPOPS correctly assigned RXR binding to **5** through identification with compound **7**, also known to be RXR selective,[49] as the closest 3D match. Although Tanimoto scores of 2D similarity (0.53-MDL, 0.09-ECFP_6) are very low, the Pearson coefficient of distances between the aligned FEPOPS of **5** and **7** is >0.92. It is important to note that only 22 molecules annotated as RXR specific exist in over 41 000 compounds queried. The correct selection of this target has a less than 1:1800 chance of occurring randomly. The 3D method appears to be locating some significant chemical/biological relationships that are not obvious from the underlying 2D structure. Figure 4 illustrates the FEPOPS defined overlap in the context of the RXRβ receptor. Importantly, FP one, two (red and green) and three, four (blue and violet) align in a way consistent with the bent cis geometry of its endogenous ligand.[50] As described in Methods, our probe

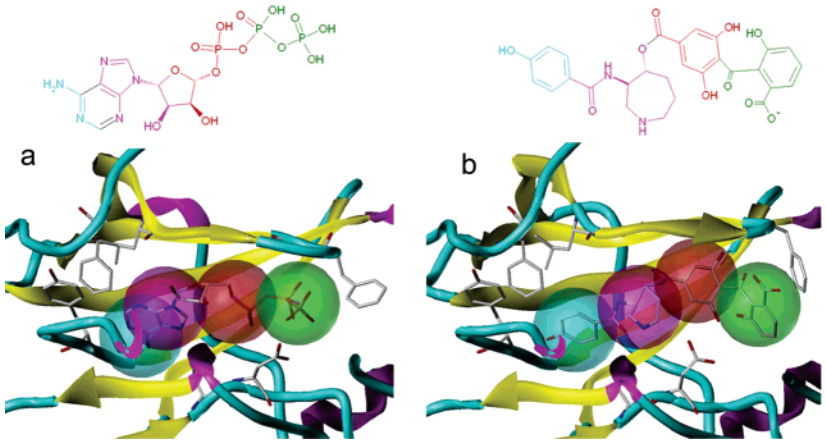


Figure 5. Cyclic AMP-dependent protein kinase (PKA) with bound ligands. (a) ATP as found in 1atp.pdb, and (b) balanol in 1bx6.pdb. The color coded chemical drawings of each molecule illustrate the atomic alignments derived during the 3D FEPOPS target fishing protocol. The corresponding colored spheres, of 3A diameter, illustrate the positioning of FEature POint PharmacoPhoreS when mapped into the crystallographic coordinates.

5 was modeled using the complex of structurally similar LG268 (**8**) with RXRβ;[34] however, the unusual scaffold of **7** was not available in the PDB. We used the two best scoring FEPOPS alignments to define centroids for the lowest energy conformations of **7**.[51] An internal energy difference of <2 kcal/mol between the conformers suggest that either might be viable candidates for binding. Figure 4 illustrates the two conformers (**7ab**) and the color coded atom alignments. Although atom assignments of FP 1 and 2 are identical in the 2 top scoring alignments, free rotation between the substituted phenyl rings of **7** allows two different, high scoring solutions for FP 3 and 4. The reversal of FP 4 and FP 3 suggested by the top two predicted alignments places the 3-fluoropropyl moiety of **7** either into or out of the binding pocket. Without an external reference, the FEPOPS results would best suggest the alignment of FP 1 and 2 but allow two options for FP 3 and 4. Given external knowledge gained from the ER systems, we can present a hypothesis. Compound **5** is a selective agonist for RXR while **7** is an antagonist, a situation analogous to DES, **2**, and OHT, **4**. The modeled binding mode of **7b** illustrated in Figure 4 projects the 3-fluoropropyl moiety out of the binding groove in the same receptor space relative to **5** as the side chain of **4** is to **2** in experimental solutions of the ER complexes.[43] That comparison of Shiau et al. suggests that antagonism in ER results from ligands that bind similarly to agonist, but with additional molecular features that disturb positioning of coactivating helix 12 responsible for downstream signaling. The tissue specific agonism/antagonism of various drugs is likely due to nonconserved protein side chain residues, outside the binding groove, that alter the position of the antagonizing feature relative to the ligand's core. Since RXR also requires alignment of its helix 12 outside the pocket for activation, a similar mechanism of a 3D-directed effect seems plausible.

PKA Receptor−ATP/Balanol. Given the initial success using 3D to fish WOMBAT for targets associated through dissimilar scaffolds, we cast a larger net using ATP as a probe and pulled back the top 30 targets predicted by FEPOPS. Analysis of scaffold diversity associated with correct matches revealed that three of the pairings, PKA, PKC-beta-1, and PKC-eta, were a variation of the same scaffold containing no phosphates, adenine, or ribose moieties, balanol. Figure 5 shows the FEPOPS alignment along with the crystallographic alignments of ATP and balanol bound to PKA (1atp.pdb, 1bx6.pdb). The natural product, balanol, is a potent inhibitor of serine and threonine kinases that competes with ATP for the same binding site. Detailed descriptions comparing the interactions of each ligand with complimentary residues of the protein have been given previously.[52] Although FP 2−4 have ligand/side-chain interactions that map directly across the two complexes, the negatively charged atoms of ATP's FP1 (red) accepts hydrogen bonds from donating side chains in the bottom of the pocket, while balanol's FP1 atoms accept hydrogen bonds from backbone amides of the flexible glycine rich loop above the ligands. This pharmacophoric overlap illustrates the utility of chemical descriptors that are not dependent upon the graph of the molecules. By searching flexible chemical space for 3-dimensional alignments, the method has found nonobvious ways of bridging the biological space of the target molecule. This appears particularly important for this system that involves significant spatial repositioning of the protein binding site. The "fuzzy" nature of the feature point alignment is a strength that allows it to accurately describe the general receptor environment of even flexible targets such as kinases. The enhanced performance of a different ligand-based 3D method compared to docking methods used for CDK2 supports the general usefulness of such an approach.[25]

4. Conclusions

We have used 2D and 3D methods of chemical description to search an annotated chemogenomics database for relationships to specific biological targets. We have shown that high 2D chemical similarity to a reference structure is a very good predictor of similar biological targets; however, there are limits to that effectiveness. This work shows that ECFP-6 descriptors sharply decline in predictive efficacy for neighbors with Tanimoto similarity indices less than 0.40 while MDL public keys decline at indices <0.80. The 3D FEPOPS method of description tested here was not able to achieve as high a percentage of correct prediction compared to 2D methods at similarity thresholds higher than those just described. However, the 3D method did outperform the 2D method below those thresholds, escaping clear correlation of performance to the underlying similarity of the chemical graph. These target fishing results have analogy to earlier experiments which show increased enrichment of chemotypes in 3D vs 2D virtual screening.[30,53,54] This is of particular importance when the chemotype of the query is missing from the reference database. The 3D method also demonstrated an ability to align chemical features in space similarly to those found in biological complexes. Though successful, a caution is raised by the RXR example that the whole molecule overlap of chemical space is not necessarily the best descriptor of biological function. A knowl-

edge-based distinction between a reference chemical binding and effecting regions may be required to separate functions such as antagonism and agonism. We have only tested one method of 3D description here. More extensive studies, comparing several 3D methods, may lead to overall improvements in performance. We believe additional work is merited, as our results support the idea that slower 3D methods can be effectively combined with 2D.

Although 90% of the tested compounds were correctly related to their biological target by at least one of the 2D methods, the analysis points out that the high performance was in a large part due to the congeneric nature of the data used. It also indicates that more than one neighbor should be considered for target identification in actual project work. Likewise the FEPOPS 3D method was able to correctly predict targets for a subset of those failing 2D comparison, but the overall lower fidelity indicates that its use should be reserved for second phase analysis in the case of very low neighborhood similarity. Clearly, no single method of chemical description is "best" in terms of functionally bridging a molecule with its biological activity. Fortunately, large compilations of chemical and biological data such as WOMBAT and others may provide sufficient test sets for developers to identify and correct specific deficiencies in new versions. We expect that combining different methods of 2D and 3D description with external learning frameworks such as Bayes Affinity Fingerprints[55] will be particularly useful for defining the particular chemical features responsible for biological activity and the spatial alignment necessary for their function.

Acknowledgment. We thank Chris Williams of Chemical Computing Group, and Andrei Caracoti of Scitegic, for assistance in custom scripting, and Greg Paris of Novartis for critical reading of the manuscript. J.N. and A.B. thank the Education Office of NIBR for postdoctoral fellowship support.

Supporting Information Available: Table of structures correctly matched by 3D methods from 2D NN misses. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Ehrhardt, P. W. Medicinal chemistry in the new millennium. A glance into the future. *Pure Appl. Chem.* **2002**, *74*, 703−785.
(2) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.
(3) Austen, M.; Dohrmann, C. Phenotype-first screening for the identification of novel drug targets. *Drug Discovery* Today **2005**, *10*, 275−282.
(4) Dobson, C. M. Chemical space and biology. *Nature* **2004**, *432*, 824−928.
(5) Stockwell, B. R. Exploring biology with small organic molecules. *Nature* **2004**, *432*, 846−854.
(6) Bredel, M.; Jacoby, E. Chemogenomics: An Emerging Strategy for Rapid Target and Drug Discovery. *Nature Rev. Genet.* **2004**, *5*, 262−275.
(7) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862−865.
(8) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual screening − an overview. *Drug Discovery Today* **1998**, *3*, 160−178.
(9) *StARLITe*; www.inpharmatica.co.uk: London, UK.
(10) BioPrint; www.cerep.com: Seattle, WA.
(11) Olah, M.; Mracec, M.; Ostopovici, L.; Rad, R.; Bora, A. et al. WOMBAT: World of Molecular Bioactivity. *Chemoinformatics in Drug Discovery*; Wiley-VCH: New York, 2004; pp 223−239.
(12) Paolini, G. V.; Shapland, R. H. B.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. Global mapping of pharmacological space. **2006**, *24*, 805−815.
(13) Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204−3218.
(14) Barnard, J. M.; Downs, G. M. Clustering of Chemical Structures on the Basis of Two-Dimensional Similarity Measures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 644−649.
(15) Matter, H.; Potter, T. Comparing 3D Pharmacophore Triplets and 2D Fingerprints for Selecting Diverse Compound Subsets. *J. Chem. Inf. Model.* **1999**, *39*, 1211−1225.
(16) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular similarity searching using atom environments, information-based feature selection, and a naive Bayesian classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170−178.
(17) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K. et al. Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Model.* **2004**, *44*, 1177−1185.
(18) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug* Discovery Today **2002**, *7*, 903−911.
(19) Brown, R. D.; Martin, Y. C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand−Receptor Binding. *J. Chem. Inf. Model.* **1997**, *37*, 1−9.
(20) Cleves, A. E.; Jain, A. N. Robust ligand-based modeling of the biological targets of known drugs. *J. Med. Chem.* **2006**, *49*, 2921−2938.
(21) Cramer, R. D.; Jilek, R. J.; Guessregen, S.; Clark, S. J.; Wendt, B. et al. "Lead Hopping". Validation of Topomer Similarity as a Superior Predictor of Similar Biological Activities. *J. Med. Chem.* **2004**, *47*, 6777−6791.
(22) Jain, A. N. Morphological similarity: a 3D molecular similarity method correlated with protein−ligand recognition. *J. Comput-Aided Mol. Des.* **2000**, *14*, 199−213.
(23) Makara, G. M. Measuring Molecular Similarity and Diversity: Total Pharmacophore Diversity. *J. Med. Chem.* **2001**, *44*, 3563−3571.
(24) Rush, T. S.; Grant, J. A.; Mosyak, L.; Nicholls, A. A Shape-Based 3-D Scaffold Hopping Method and Its Application to a Bacterial Protein−Protein Interaction. *J. Med. Chem.* **2005**, *48*, 1489−1495.
(25) Zhang, Q.; Muegge, I. Scaffold Hopping through Virtual Screening Using 2D and 3D Similarity Descriptors: Ranking, Voting, and Consensus Scoring. *J. Med. Chem.* **2006**, *49*, 1536−1548.
(26) 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.
(27) Renner, S.; Schneider, G. Scaffold-Hopping Potential of Ligand-Based Similarity Concepts. *ChemMedChem* **2006**, *1*, 181−185.
(28) Pipeline Pilot; SciTegic: San Diego, CA, 2005.
(29) *MDL Public Keys*; Elsevier MDL, Inc.: San Ramon, CA.
(30) 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.
(31) Pearson, K. Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. *Philos. Trans. R. Soc. London A Math. Phys. Eng. Sci.* **1896**, *187*, 253−318.
(32) *Molecular Operating Environment*; Computational Computing Group: Montreal, Canada, 2005.
(33) SVL script provided by Dr. Chris Williams at Chemical Computing Group.
(34) Love, J. D.; Gooch, J. T.; Benko, S.; Li, C.; Nagy, L. et al. The Structural Basis for the Specificity of Retinoid-X Receptor-selective Agonists: New Insights Into the Role of Helix H12. *J. Biol. Chem.* **2002**, 277, 11385−11391.
(35) Glen, R. C.; Bender, A.; Arnby, C. H.; Carlsson, L.; Boyer, S. et al. Circular fingerprints: Flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs* **2006**, *9*, 199−204.
(36) 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**, in press.
(37) 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* **2001**, *43*, 217−226.
(38) Ekins, S. Predicting undesirable drug interactions with promiscuous proteins in silico. *Drug Discovery Today* **2004**, *9*, 276−285.
(39) Marriott, D. P.; Dougall, I. G.; Meghani, P.; Liu, Y. J.; Flower, D. R. Lead generation using pharmacophore mapping and three-dimensional database searching: application to muscarinic M(3) receptor antagonists. *J. Med. Chem.* **1999**, *42*, 3210−3216.
(40) Nicholls, A.; MacCuish, N. E.; MacCuish, J. D. Variable selection and model validation of 2D and 3D molecular descriptors. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 451−474.

(41) Pearce, S. T.; Jordan, V. C. The biological role of estrogen receptors alpha and beta in cancer. *Crit. Rev. Oncol. Hematol.* **2004**, *50*, 3−22.
(42) Tanenbaum, D. M.; Wang, Y.; Williams, S. P.; Sigler, P. B. Crystallographic comparison of the estrogen and progesterone receptor's ligand binding domains. *Proc. Natl. Acad. Sci.* **1998**, *95*, 5998−6003.
(43) Shiau, A. K.; Barstad, D.; Loria, P. M.; Cheng, L.; Kushner, P. J. et al. The Structural Basis of Estrogen Receptor/Coactivator Recognition and the Antagonism of This Interaction by Tamoxifen. *Cell* **1998**, *95*, 927−937.
(44) Cushman, M.; He, H.-M.; Katzenellenbogen, J. A.; Lin, C. M.; Hamel, E. Synthesis, Antitubulin and Antimitotic Activity, and Cytotoxicity of Analogues of 2-Methoxyestradiol, an Endogenous Mammalian Metabolite of Estradiol That Inhibits Tubulin Polymerization by Binding to the Colchicine Binding Site. *J. Med. Chem.* **1995**, *38*, 2041−2049.
(45) Mark, M.; Ghyselinck, N. B.; Chambon, P. Function of retinoid nuclear receptors: lessons from genetic and pharmacological dissections of the retinoic acid signaling pathway during mouse embryogenesis. *Annu. Rev. Pharmacol. Toxicol.* **2006**, *46*, 451−480.
(46) Chambon, P. A decade of molecular biology of retinoic acid receptors. *FASEB J.* **1996**, *10*, 940−954.
(47) Boehm, M. F.; Zhang, L.; Badea, B. A.; White, S. K.; Mais, D. E. et al. Synthesis and structure−activity relationships of novel retinoid X receptor-selective retinoids. *J. Med. Chem.* **1994**, *37*, 2930−2941.
(48) Yoshimura, H.; Nagai, M.; Hibi, S.; Kikuchi, K.; Abe, S. et al. A Novel Type of Retinoic Acid Receptor Antagonist: Synthesis and Structure−Activity Relationships of Heterocyclic Ringcontaining Benzoic Acid Derivatives. *J. Med. Chem.* **1995**, *38*, 3163−3173.
(49) Michellys, P.-Y.; Ardecky, R. J.; Chen, J.-H.; D′Arrigo, J.; Grese, T. A. et al. Design, Synthesis, and Structure−Activity Relationship Studies of Novel 6,7-Locked-[7-(2-alkoxy-3,5-dialkylbenzene)-3-methylocta]-2,4,6-trienoic Acids. *J. Med. Chem.* **2003**, 46, 4087−4103.
(50) Egea, P. F.; Mitschler, A.; Moras, D. Molecular Recognition of Agonist Ligands by RXRs. *Mol. Endocrinol.* **2002**, *16*, 987−997.
(51) It is interesting to note that molecules which share the same binding site often produce more that one high scoring alignment with FEPOPS. This observation is subject of further investigation.
(52) Narayana, N.; Diller, T. C.; Koide, K.; Bunnage, M. E.; Nicolaou, K. C. et al. Crystal Structure of the Potent Natural Product Inhibitor Balanol in Complex with the Catalytic Subunit of cAMP-Dependent Protein Kinase. *Biochemistry* **1999**, *38*, 2367−2376.
(53) Good, A. C.; Hermsmeier, M. A.; Hindle, S. A. Measuring CAMD technique performance: a virtual screening case study in the design of validation experiments. *J. Comput-Aided Mol. Des.* **2004**, *18*, 529−536.
(54) Good, A. C.; Cho, S. J.; Mason, J. S. Descriptors you can count on? Normalized and filtered pharmacophore descriptors for virtual screening. *J Comput.-Aided Mol. Des.* **2004**, *18*, 523−527.
(55) Bender, A.; Jenkins, J. L.; Glick, M.; Deng, Z.; Nettles, J. H. et al. "Bayes Affinity Fingerprints" Improve Retrieval Rates in Virtual Screening and Define Orthogonal Bioactivity Space: When are Multi-Target Drugs a Feasible Concept? *J. Chem. Inf. Model.* **2006**, in press - DOI: 10.1021/ci600197y.

JM060902W