Pharmacophore Fingerprinting. 1. Application to QSAR and Focused Library Design

Malcolm J. McGregor and Steven M. Muskal*

Affymax Research Institute, 3410 Central Expressway, Santa Clara, California 95051

Received November 16, 1998

A new method of rapid pharmacophore fingerprinting (PharmPrint method) has been developed. A basis set of 10 549 three-point pharmacophores has been constructed by enumerating several distance ranges and pharmacophoric features. Software has been developed to assign pharmacophoric types to atoms in chemical structures, generate multiple conformations, and construct the binary fingerprint according to the pharmacophores that result. The fingerprint is used as a descriptor for developing a quantitative structure−activity relationship (QSAR) model using partial least squares. An example is given using sets of ligands for the estrogen receptor (ER). The result is compared with previously published results on the same data to show the superiority of a full 3D, conformationally flexible approach. The QSAR model can be readily interpreted in structural/chemical terms. Further examples are given using binary activity data and some of our novel in-house compounds, which show the value of the model when crossing compound classes.

INTRODUCTION

Pharmacophore based screening has become commonplace in the field of computer aided drug design. The pharmacophore concept is based on the kinds of interactions observed in molecular recognition: hydrogen bonding, charge−charge, and hydrophobic interactions. A pharmacophore is a set of functional group types in a spatial arrangement that represents the interactions made in common by a set of small molecule ligands with a protein receptor. Some advantages of the methodology are as follows: (i) it can be used where the only data available are a set of known "hits" without a knowledge of the structure of the receptor; (ii) the pharmacophore specification is sufficiently general that it can be applied across different classes of ligands, i.e., a pharmacophore can be derived from one class and used to predict activity in another; and (iii) the methodology can be applied to large data sets in high throughput screening (HTS) applications. Thus the method has wide utility.

In the usual application of the methodology a single pharmacophore hypothesis, or a small number of them, is derived from a set of ligands with known activity. (The term "hypothesis" is usually used at this stage to indicate that this is a computational result, as opposed to an empirical one arrived at by experiment or observation of structure complexes.) The hypothesis is then computationally screened across a database of compounds to narrow the selection for biological screening. A measure of the increase in hit rate achieved is desirable. Several software systems are commercially available that support pharmacophoric development and use. Widely used examples are Catalyst[1−3] by Molecular Simulations Inc.[4] and the ChemDiverse module of Chem-X by Chemical Design Ltd.[5] However, a major limitation is that compounds must be registered into the proprietary, closed database system provided by the respective vendors.

Pharmacophore fingerprinting is an extension of this approach whereby a basis set of pharmacophores is generated by enumerating all pharmacophoric types with a set of distance ranges. The concept has been described previously,[6−9] and applications to structure−activity relationships have been explored with atom pair descriptors.[10] In the present study a basis set of three-point pharmacophores has been generated. Tools have been developed to fingerprint large libraries of compounds with the entire set and incorporated into a program called PharmPrint. The efficiency of the code is such that it can be applied to high-throughput electronic ("in silico") screening applications, in our case involving combinatorial libraries.

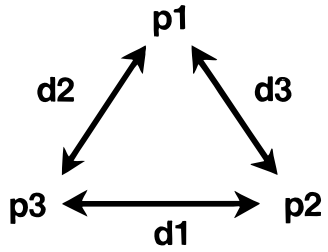

Figure 1. Schematic representation of pharmacophoric types (p) and distances (d).

The question then arises as to how to derive a function that relates the fingerprints to biological activity. Methods that suggest themselves are neural networks, genetic algorithms, and regression techniques. We chose one of the latter methods, namely, partial least squares (PLS). As a natural outcome of this methodology, we are able to address the issues of chemical diversity and coverage of drug space (in preparation) in addition to activity prediction.

METHODOLOGY

Fingerprint Generation. Figure 1 shows the general definition of the three-point pharmacophore. The pharmacophoric groups used were the six commonly used in this type of analysis: hydrogen bond acceptor (A) and donor (D), groups with formal negative (N) and positive (P) charges, and hydrophobic (H) and aromatic (R). In addition a seventh

* To whom correspondence should be addressed.

10.1021/ci980159j CCC: $18.00
Published on Web 03/19/1999

type was added for any atom that is not labeled with any of the first six types (X). Six distance ranges were used: 2.0−4.5, 4.5−7.0, 7.0−10.0, 10.0−14.0, 14.0−19.0, and 19.0−24.0 Å. These have been used previously, where distributions of hits were analyzed,[7] and smaller distance bins were assigned to the more populated regions. The three-point pharmacophores were enumerated for all combinations of types and distances and then subjected to two additional constraints: (i) triangle rule, i.e., the length of each side of a triangle cannot exceed the sum of the lengths of the other two sides, otherwise this would produce a geometrically impossible object; (ii) elimination of redundant pharmacophores related by symmetry. The resulting number is 10 549.

Most of the software used for this analysis was written in-house. The exception was the Corina program for 3D structure generation.[11] This was chosen because it runs in batch mode, accepts a variety of standard molecule formats, and has been shown to generate good quality structures.[12] The output from Corina is used as input to the PharmPrint program which was developed in-house. Only heavy atoms are used in computations; if hydrogens are present, they are ignored. The function of the PharmPrint program is 3-fold.

(1) It assigns the pharmacophoric groups to atoms. This is done using a substructure search for the relevant fragments, using an atom-by-atom mapping algorithm.[13] The fragments are chosen using heuristics about which substructures contain the pharmacophoric groups. For the most part this is fairly straightforward; e.g., a carboxylate has a negative charge, an aliphatic amine has a positive charge, a hydroxyl is a hydrogen bond donor and acceptor, etc. The most problematic assignment was that of hydrophobic. The following was tried and appears to work well: an atom of type C, Cl, Br, or I is considered hydrophobic if it is further than two bonds removed in 2D connectivity from any atom of type N, O, P, or S−H.

(2) It rotates about bonds to generate multiple conformations. This utilizes the quaternion rotation algorithm.[14] Heuristics are used to determine which bonds are rotatable and the angles they can assume. There is a limit to the number of conformations generated (set at 1000); rotations that affect the largest number of atoms are performed first, so that if the limit is reached, then as much of the conformational space is covered as possible. A simple energy function is implemented to eliminate conformations with severe steric overlaps. Bonds in rings are assigned to be nonrotatable and the multiple ring conformation option in Corina is switched on.

(3) It builds the pharmacophore fingerprint by measuring distances between pharmacophoric groups. The output from the program is a fingerprint for each structure in the original SD file (SD format is the MDL ASCII molecule format), with an identifier derived from a specified data field.

As a binary descriptor, the fingerprint is efficient to deal with computationally. It can be represented in computer memory using one integer for 32 bits in the bitstring. This can be unpacked into one integer or floating point number per bit, but for some calculations it does not need to be unpacked. An example is the Tanimoto coefficient, a measure of bitstring (and therefore molecular) similarity, which can be calculated using bitwise operators in a programming language like C.

The computational speed of fingerprinting is dependent on the parameters set and the nature of the structures, in particular the number of total atoms and rotatable bonds. With the MDDR9104 set (below), and with the parameters used, the speed was of the order of 10 structures/min on a Silicon Graphics R10000 processor.

Preparation of the MDDR Subset. The MDDR (MDL Drug Data Report)[15] was used as a reference for drug like compounds. It is a database of biologically active compounds with associated data, including activity classes. Version 98.1 contains 92 604 entries. A subset was prepared using the following criteria.

(1) Structures that have a molecular weight in the drug like range of 200−700 are used. For this and subsequent steps it is important that salts are removed from the original database structure. To this end a program "stripsalt" was used to remove small-disconnected fragments in SD files.

(2) Only structures which consist entirely of atoms from the following list are included: C, N, O, H, S, P, F, Cl, Br, I. This preserves only the types of structures which reflect the chemistry carried out in our laboratories and removes unusual structures such as metal complexes.

(3) The compound activity class, as given in the activ_class and activ_index fields in the MDDR, indicates a well-defined target (i.e., individual enzyme or receptor), as opposed to a broad therapeutic class. The file activty.txt, provided by MDL, lists the classes. This file was manually inspected to extract all such classes.

(4) The 2D chemical similarity between any structure and all other structures in the final list is below a certain threshold. This eliminates close analogues that might bias the analysis. The measure chosen was the Tanimoto coefficient with the MDL 166 user keys, and the threshold was 0.8. The keys are 2D fragment-based descriptors which are calculated automatically in MDL ISIS databases.[16]

(5) Classes that had less than eight members, and compounds that belonged only to those classes, were eliminated.

This procedure resulted in 9104 compounds and 152 classes.

Partial Least Squares. Partial least squares (PLS)[17−19] was applied to derive a function that relates the fingerprints to the activity values using code written in-house. The algorithm used is based on the NIPALS algorithm for principle component analysis (PCA).[18] PLS has been previously applied to the analysis of chemical structure, most notably in the CoMFA methodology.[20] The data were mean centered but not variance scaled. The data for the independent variables were entered as 1.0 for a pharmacophore hit or 0.0 for no hit, corresponding to the binary fingerprint. The data for the dependent variables were either the log of the relative binding affinity (RBA) for data sets 1−3 or 1.0/0.0 to represent active or inactive in data set 4.

Data Sets. We chose data sets for the estrogen receptor because of the recent therapeutic interest in this class of targets and the fact that there have been several quantitative structure−activity relationship (QSAR) models developed for ER ligands.[21−28] The crystal structure for ER-α has also recently been reported.[29] The data sets are summarized in Table 1 and described in detail below.

Data Sets 1 and 2. The first two data sets used for the QSAR analysis were a set of 31 ER ligands with activity

Table 1

set	training	testing
1	31 literature compounds[30] (RBA for human ER-α: 0.001−468)	leave-one-out cross-validation on training set
2	31 literature compounds[30] (RBA for rat ER-β: 0.001−404)	leave-one-out cross-validation on training set
3	set 1 + 17 proprietary heterocycles (RBA for human ER-α: 0.002−5.5)	18 heterocycles (distinct from training set) (RBA for human ER-α: 0.017−9.4)
4	15 from set 1 (RBA ≥10.0) + 750 MDDR "not estrogen"	86 proprietary compounds (<1 μM for ER-α) + 250 MDDR "estrogen" compounds + 8290 MDDR "not estrogen" compounds

values for binding to human ER-α (set 1) and rat ER-β (set 2).[30] These are given as relative binding activity (RBA) compared to the activity of the natural ligand, estradiol (E2), which is given a value of 100.0. Thus the higher the value, the more active the compound. The structures are illustrated in Figure 2. It can be seen that they are reasonably diverse, spanning several structural classes; some of the structures are quite rigid while others are somewhat flexible. There are two published crystal structures of ER-α with different ligands bound,[29] which cast light on the nature of the protein−ligand interactions. One crystal structure contains in the active site the natural ligand and agonist E2. The other has the antagonist raloxifene. Raloxifene is not part of data sets 1 and 2 but is structurally analogous to compounds 23, 24, 25, and 26 (Figure 2). Raloxifene makes additional interactions compared to E2, involving a positively charged group on a flexible side chain on the ligand. This is accommodated by a conformational shift in the protein that is the structural basis for antagonism in the ER. This indicates that this training set contains ligands with at least two different binding modes.

Three different QSAR methods have been previously applied to these data sets[28] that we use for comparison of results. The methods apply PLS to different molecular descriptors: (1) comparative molecular field analysis (CoMFA), a widely used method based on the calculation of a steric and electrostatic field on a grid around each ligand; (2) The CoDESSA program, which calculates descriptors for 2D and 3D structures and quantum-mechanical properties; and (3) hologram QSAR (HQSAR), which uses as a descriptor a molecular hologram constructed from counts of substructural molecular fragments (2D only).

The results for these data sets are presented as r^2 and q^2, comparing the predicted and actual activity values. The q^2 calculation (cross-validated r^2) uses the leave-one-out (LOO) procedure.

Data Set 3. This makes use of 35 heterocyclic compounds that have shown activity in our ER-α assay. The activity values have been expressed as RBA to make them equivalent to set 1; 17 of these were added to set 1 to give a training set of 48. The remaining 18 compounds were used for testing using the model derived from the training set, and the result is given for these compounds only.

Data Set 4. This set was designed so that the method could be tested using binary activity values. The PLS algorithm was run with activity values of either 1.0 or 0.0. Separate training and testing sets, each with actives and inactives, were established as follows. The training set actives were the 15 compounds from set 1 which have a RBA of ≥10.0 (numbers 2−8, 20−22, 24−27, and 29 in Figure 2). The testing set actives were the compounds in the complete MDDR which have the string "ESTROGEN" in the activity_class field. This set was pruned to eliminate obvious prodrugs and to exclude compounds in the training set, the resulting number being 250. The MDDR9104 subset was pruned to exclude compounds that had the string "ESTROGEN" in the activity_class field. This gave a set of 9040 from which 750 compounds were randomly chosen and included in the training set with activity values of 0.0, i.e., assumed inactive. The remaining 8290 were used for testing. To distinguish these compounds from ones which can be properly claimed to be inactive (as actually tested in the biological assay), we term these the "background" sets as opposed to "inactives". At the training stage the active compounds were duplicated 50 times to give as much overall weight to the active as the background compounds. An additional testing set was established of 86 compounds from our corporate database which have an activity for ER-α of better then 1 μM (the "ARI actives"). These were derived from combinatorial libraries, with scaffolds of three different chemical classes, none of which are represented in the training set, and which include most of the heterocyclics in data set 3.

Table 2

dataset	statistic	CoMFA	HQSAR	CODESSA	PharmPrint
1	q^2	0.70	0.67	0.46	0.75
	r^2	0.95	0.88	0.79	0.92
	PCs	4	4	2	4
2	q^2	0.60	0.68	0.61	0.71
	r^2	0.95	0.91	0.92	0.93
	PCs	4	5	4	5
3	q^2	N/A	N/A	N/A	0.88
	PCs	N/A	N/A	N/A	6

RESULTS AND DISCUSSION

Table 2 presents the results for the PharmPrint/PLS QSAR on data sets 1−3 as described in Methodology. For comparison, the same statistics are reproduced for the three other QSAR methods previously applied[28] to data sets 1 and 2: comparative molecular field analysis (CoMFA), classical QSAR using the CoDESSA program, and hologram QSAR (HQSAR). It can be seen that for both data sets the r^2 (non-cross-validated result) for the PharmPrint/PLS is comparable to the other three methods. With the cross-validated statistic, q^2, this value is higher for the Pharmprint/PLS than for any of the other methods. With set 3, the q^2 is higher still, at 0.88. This suggests that the phamacophore fingerprints are more successfully generalizing from the data. This may not be surprising since the methodology uses not only 3D features but conformational flexibility as well.

As a point of interest, our method initially used only the six pharmacophoric types A, D, H, N, P, R; unlabeled atoms

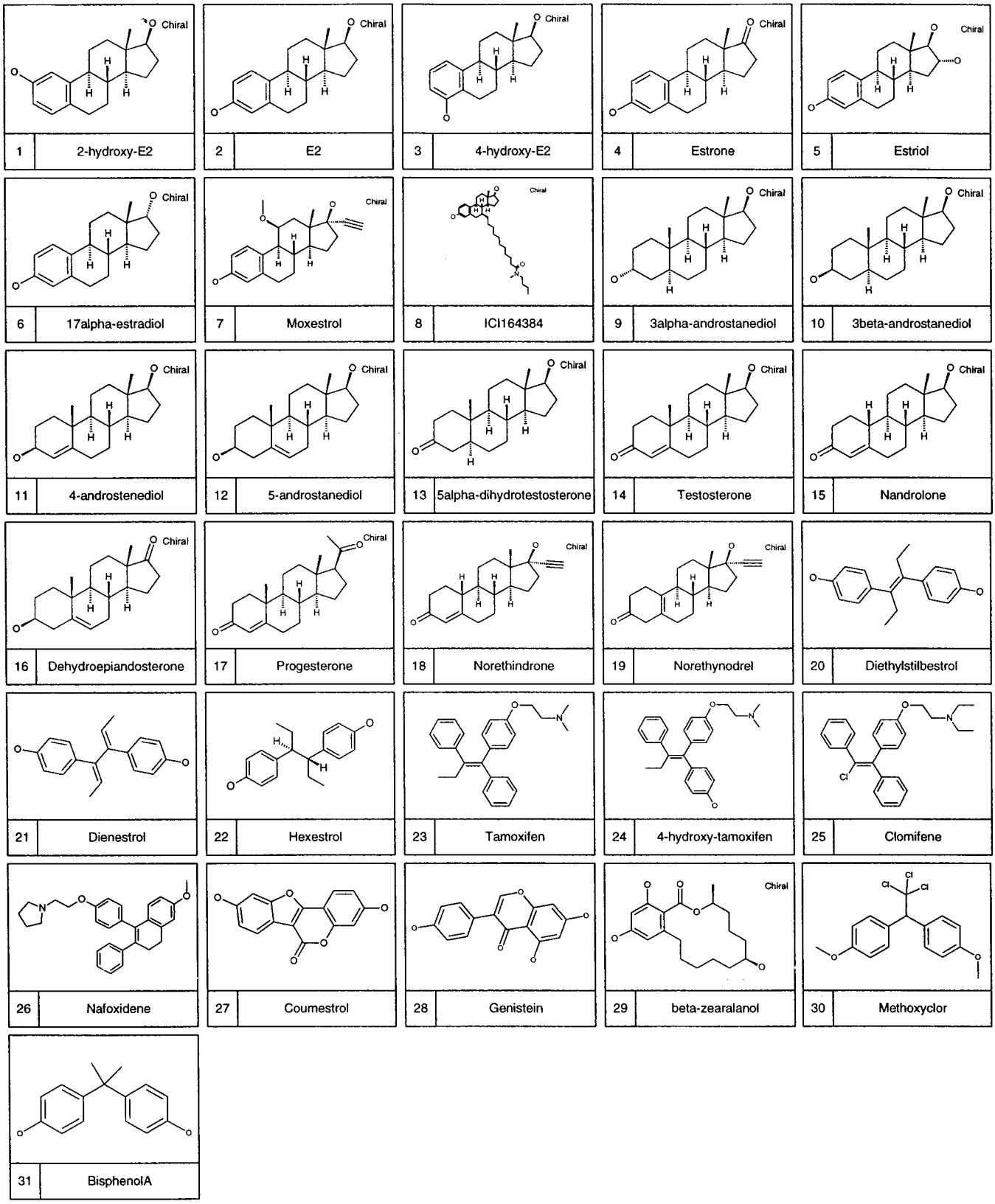

Figure 2. The 31 structures in data sets 1 and 2.

were ignored, giving a bitstring of length 6726. With this fingerprint it was very difficult to approach the accuracy given in Table 2, the q^2 statistic remaining around 0.60 or below (data not shown). Therefore it is clear that the previously unlabeled atoms, now given a default label X and included in the fingerprint of length 10 549, contain important information, probably related to molecular volume. The X type accounts for 49.7% of all atoms in molecules of the MDDR9104 set.

In QSAR studies of this type it is considered important to be able to deal with different kinds of activity data. One situation that is common, especially with the results of an

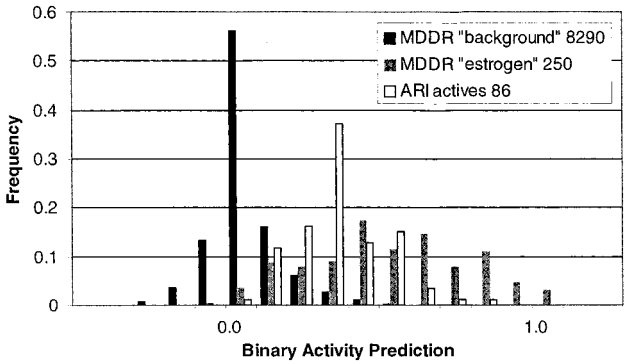

Figure 3. Distribution of binary activity prediction values for testing set 4.

Table 3

	mean	sd	% correct
MDDR background 8290	0.03	0.14	89.7
MDDR estrogen 250	0.53	0.26	87.4
ARI actives 86	0.37	0.15	87.2

initial screening of primary libraries, is that data are available as to which compounds are active or inactive, but reliable IC50 or EC50 data have not been established. Thus training set 4 was established consisting of the 15 compounds from set 1 which have RBA values of ≥10.0. These were used with activity values of 1.0, ignoring the actual affinity values. It was considered desirable to include as many inactive compounds as possible, as the determinants of activity are not just the pharmacophores that occur more frequently in the active compounds but also ones which occur less frequently compared to a background hit rate. These pharmacophores may have a negative impact on activity when they are present. Since the PharmPrint method is very efficient, it is possible at the training stage to include many or all of the compounds in the MDDR set, the great majority of which can be assumed to be inactive for a particular target, which we term the background compounds.

The results are presented graphically in Figure 3 and statistically in Table 3. The 8290 MDDR "background" compounds in the testing set are clustered close to zero. The 250 MDDR "estrogen" testing compounds and 86 ARI estrogen compounds are distributed between 0.0 and 1.0. In Figure 3, it is clear that both have a distribution that is clearly distinct from the background compounds. The ARI compounds have a distribution that is somewhat to the left of the MDDR estrogen compounds. This can be interpreted by considering that the MDDR estrogen compounds are generally of the same class as the training set. The ARI compounds, however, are derived from our combinatorial libraries and are of three distinct classes, none of which are represented in the training set. This gives some measure of the predictive ability across different classes of molecules. Table 3 gives the percentage of correctly classified compounds, assuming the MDDR background set is inactive and the MDDR estrogen and ARI compounds are active, and taking an arbitrary discrimination cutoff of 0.2. The results are 89.7, 87.4, and 87.2%, respectively.

Given good results, it is also desirable to be able to interpret them in terms of structural features. With some computational methods this may be difficult. However, with the fingerprints, we can look at the weights produced by the

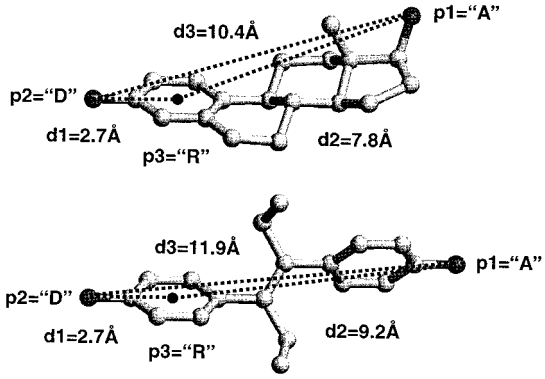

Figure 4. Illustration to show a mapping of pharmacophore 1624 from Table 4 onto (i) estradiol (top), the natural ligand, and (ii) diethylstilbestrol (bottom), the most potent compound in data set 1. The hydroxyl groups are both hydrogen bond donors and acceptors (D, A), and the centroid of the aromatic ring is shown.

Table 4

			distances			types		
rank	pharm no.	weight	1	2	3	1	2	3
1	1624	0.0832	2−4.5	7−10	10−14	A	D	R
2	1673	0.0805	2−4.5	7−10	10−14	D	D	R
3	1670	0.0778	2−4.5	7−10	10−14	D	D	H
4	1667	0.0770	2−4.5	7−10	10−14	D	D	X
5	1663	0.0755	2−4.5	7−10	10−14	D	A	H
6	840	0.0724	2−4.5	4.5−7	7−10	H	D	R
7	3889	0.0724	4.5−7	4.5−7	10−14	D	D	H
8	1666	0.0719	2−4.5	7−10	10−14	D	A	R
9	1618	0.0680	2−4.5	7−10	10−14	A	D	X
10	1660	0.0680	2−4.5	7−10	10−14	D	A	X
10 540	3523	−0.0638	4.5−7	4.5−7	4.5−7	X	A	D
10 541	728	−0.0649	2−4.5	4.5−7	7−10	A	X	R
10 542	365	−0.0657	2−4.5	2−4.5	7−10	A	R	X
10 543	696	−0.0662	2−4.5	4.5−7	7−10	X	H	D
10 544	484	−0.0672	2−4.5	4.5−7	4.5−7	X	A	A
10 545	3522	−0.0672	4.5−7	4.5−7	4.5−7	X	A	A
10 546	1443	−0.0674	2−4.5	7−10	7−10	H	X	X
10 547	681	−0.0716	2−4.5	4.5−7	7−10	X	A	A
10 548	288	−0.0719	2−4.5	2−4.5	7−10	X	A	X
10 549	485	−0.0754	2−4.5	4.5−7	4.5−7	X	A	D

PLS analysis. Table 4 presents the 10 highest and 10 lowest pharmacophores rank ordered by the magnitude of the weights for the first principle component from data set 4. Positive weights indicate pharmacophores that are common in the active compounds relative to the background compounds, and negative weights indicate the reverse. For example, the pharmacophore ranked highest, number 1624, can be seen to be a strong feature of the active compounds in the training set. It consists of an aromatic group (R) 2.0−4.5 Å from a hydrogen bond donor (D); this maps to the phenol group that is common to the most active compounds. There is a hydrogen bond acceptor atom (A) 7−10 Å from the ring centroid (R) and 10−14 Å from the (D) atom. This maps to another hydroxyl or carbonyl group further away. Figure 4 shows how pharmacophore 1624 maps to the molecular structures of estradiol, the natural ligand, and diethylstilbestrol, the most active compound in set 1. This illustrates that when one pharmacophore hits two molecules, it is because of a similar presentation of functionality, even though the compounds may be of different structural classes. Most of the rest of the top scoring pharmacophores are similar to the first, having the same distance ranges and varying only in some of the types. The pharmacophores with

the negative weights are more difficult to interpret in terms of the structures of the active compounds but appear to contribute roughly equally in magnitude to the model.

CONCLUSIONS

Pharmacophore fingerprinting is a promising approach to computer aided drug design. The PharmPrint fingerprint is a compact but information-rich descriptor. It is based on features observed to be important in ligand−receptor interactions, and it takes into account not only 3D structure but multiple conformations. We note the value of the X atom type that might help describe additional skeletal, support feature, and/or overall molecular volume.

The QSAR performs well with structures where there is some conformational flexibility and where there are multiple structural classes and binding modes. The QSAR does not require the structure of the receptor, and no assumptions are required about how molecules overlap at the binding site. However, one can envision a pharmacophore description being derived from the binding site of a known receptor structure and a function used to compare that to the molecule fingerprint. This is an area of current investigation for us.

The PharmPrint software runs in batch mode, requiring only a single 3D structure as input, and the output is a flat file. It does not require the construction of specialized databases. Thus it integrates well with our chemical information systems and can be triggered automatically as structures or libraries are added. Performance is such that tens of thousands of compounds can be processed each day. It is thus well-suited to our main application of building block selection for the design of focused or targeted combinatorial libraries using in silico screening. We are currently exploring ways to increase the efficiency of the calculation; these include parallelization strategies and ways to take advantage of the redundancy in enumerated combinatorial libraries where the same building block appears many times on the same scaffold. We will also be reporting on the use of this descriptor for what we consider the complementary problem to targeted design—primary library design—which brings up the questions of molecular diversity and coverage of drug space.

ACKNOWLEDGMENT

We would like to thank Charlie Peng for supplying the code for the rotation algorithm and Charles Hart and Ron Hale for the biology and chemistry, respectively.

REFERENCES AND NOTES

(1) Sprague, P. W. Automated Chemical Hypothesis Generation and Database Searching with Catalyst. *Perspectives in Drug Discovery and Design*; Müller, K., Ed.; *ESCOM Science Publishers B. V.*: Leiden, The Netherlands, 1995; Vol. 3, pp 1-20.

(2) Barnum, D.; Greene, J.; Smellie, A.; Sprague, P. Identification of Common Functional Configurations among Molecules. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 563−571.

(3) Greene, J.; Kahn, S.; Savoj, H.; Sprague, P.; Teig, S. Chemical Function Queries for 3D Database Search. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1297−1308.

(4) Molecular Simulations Inc., San Diego, CA.

(5) Chemical Design Ltd., Oxfordshire, U.K.

(6) Good, A. C.; Kuntz, I. D. Investigating the extension of pairwise distance pharmacophore measures to triplet-based descriptors. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 373.

(7) Pickett, S. D.; Mason, J. S.; McLay, I. M. Diversity Profiling and Design Using 3D Pharmacophores: Pharmacophore-Derived Queries (PDQ). *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1214−1223.

(8) Mason, J. S.; Pickett, S. D. Partition-based selection. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 85−114.

(9) Pickett, S. D.; Luttmann, C.; Guerin, V.; Laoui, A.; James, E. DIVSEL and COMPLIB−Strategies for the design and comparison of combinatorial libraries using pharmacophoric descriptors. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 144−150.

(10) Chen, X.; Rusinko, A.; Young, S. S. Recursive partitioning analysis of a large structure−activity data set using three-dimensional descriptors. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1054−1062.

(11) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic Generation of 3D-Atomic Coordinates for Organic Molecules. *Tetrahedron Comput. Methodol.* **1990**, *3*, 537−547.

(12) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000−1008.

(13) Gluck, D. J. A chemical structure storage and search system developed at Du Pont. *J. Chem. Doc.* **1965**, *5*, 43−51.

(14) Shoemake, K. Animating rotation with quaternion curves. *SIGGRAPH* **1985**, *19*, 245−254.

(15) MDL Information Systems, Inc., San Leandro, CA.

(16) McGregor, M. J.; Pallai, P. V. Clustering of Large Databases of Compounds: Using the MDL "Keys" as Structural Descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443−448.

(17) Lindberg, W.; Persson, J.-A.; Wold, S. Partial least-squares method for spectrofluorimetric analysis of mixtures of humic acid and lignin sulfonate. *Anal. Chem.* **1983**, *55*, 643−648.

(18) Geladi, P.; Kowalski, B. R. Partial least-squares regression: A tutorial. *Anal. Chim. Acta* **1986**, *185*, 1−17.

(19) Wold, S.; Sjöström, M.; Eriksson, L. Partial least squares projections to latent structures (PLS) in chemistry. *Encyclopedia of Computational Chemistry*; John Wiley & Sons: New York, 1998; pp 2006−2021.

(20) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(21) Williams, C. L.; Stancel, G. M. 1996 Estrogens and Progestins. In *Goodman and Gillman's The Pharmacological Basis of Therapeutics*, 9th ed.; *Hardman, J. G., Limbird, L. E., Eds.;* McGraw-Hill: New York, 1996; pp 1411−1440.

(22) Tong, W.; Perkins, R.; Strelitz, R.; Collantes, E. R.; Keenan, S.; Welsh, W. J.; Branham, W. S.; Sheehan, D. M. Quantitative structure−activity relationships (QSARs) for estrogen binding to the estrogen receptor: Predictions across species. *Environ. Health Perspect.* **1997**, *105*, 1116−1124.

(23) Tong, W.; Perkins, R.; Xing, L.; Welsh, W. J.; Sheehan, D. M. QSAR models for binding of estrogenic compounds to estrogen receptor α and β subtypes. *Endocrinology* **1997**, *138*, 4022−4025.

(24) Waller, C. L.; Minor, D. L.; Mckinney, J. D. Examination of the estrogen receptor binding affinities of polychlorinated hydroxybiphenyls using three-dimensional quantitative structure−activity relationships. *Environ. Health Perspect.* **1996**, *103*, 702−707.

(25) Bradbury, S. P.; Mekenyan, O. G.; Ankley, G. T. Quantitative structure−activity relationships for polychlorinated hydroxybiphenyl estrogen receptor binding affinity: An assessment of conformer flexibility. *Environ. Toxicol. Chem.* **1996**, *15*, 1945−1954.

(26) Gantchev, T. G.; Ali, H.; van Lier, J. E. Quantitative Structure−Activity Relationships/Comparative Molecular Field Analysis (QSAR/CoMFA) for Receptor-Binding Properties of Halogenated Estradiol Derivatives. *J. Med. Chem.* **1994**, *37*, 4164−4176.

(27) Waller, C. L.; Oprea, T. I.; Chae, K.; Park, H.-K.; Korach, K. S.; Laws, S. C.; Wiese, T. E.; Kelce, W. R.; Gray, L. E., Jr. Ligand-Based Identification of Environmental Estrogens. *Chem. Res. Toxicol.* **1996**, *9*, 1240−1248.

(28) Tong, W.; Lowis, D. R.; Perkins, R.; Chen, Y.; Welsh, W. J.; Goddette, D. W.; Heritage, T. W.; Sheehan, D. M. Evaluation of Quantitative Structure−Activity Relationship Methods for Large-Scale Prediction of Chemicals Binding to Estrogen Receptor. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 669−677.

(29) Brzozowski, A. M.; Pike, A. C. W.; Dauter, Z.; Hubbard, R. E.; Bonn, T.; Engstrom, O.; Ohman, L.; Greene, G. L.; Gustafsson, J.-A.; Carlquist, M. Molecular basis of agonism and antagonism in the estrogen receptor. *Nature (London)* **1997**, *389*, 753−758.

(30) Kuiper, G. G. J. M.; Carlsson, B.; Grandien, K.; Enmark, E.; Haeggblad, J.; Nilsson, S.; Gustafsson, J.-A. Comparison of the ligand binding specificity and transcript tissue distribution of estrogen receptors α and β. *Endocrinology* **1997**, *138*, 863−870.

CI980159J