Available online at www.sciencedirect.com
SCIENCE DIRECT®
Bioorganic & Medicinal Chemistry 14 (2006) 611–621
Bioorganic & Medicinal Chemistry

Constructing plasma protein binding model based on a combination of cluster analysis and 4D-fingerprint molecular similarity analyses

Jianzhong Liu,[a,*] Liu Yang,[a] Yi Li,[b] Dahua Pan[b] and Anton J. Hopfinger[b]

[a]*Department of Chemistry and Biochemistry, University of Delaware, Newark, DE 19716, USA*
[b]*Laboratory of Molecular Modeling and Design (M/C 781), College of Pharmacy, The University of Illinois at Chicago, 833 South Wood Street, Chicago, IL 60612-7231, USA*

Received 12 July 2005; revised 22 August 2005; accepted 22 August 2005
Available online 7 October 2005

Abstract—Based on 2D-connectivity molecular similarity and cluster analyses, a dataset for HSA binding is divided into the training set and the test set. 4D-fingerprint similarity measures were applied to this dataset. Four different predictive schemes (SM, SA, SR, and SC) were applied to the test set based on the similarity measures of each compound to the compounds in the training set. The first algorithmic scheme (SM), which only takes the most similar compound in the training set into consideration, predicts the binding affinity of a test compound. This scheme has relatively poor predictivity based on 4D-fingerprint similarity analyses. The other three algorithmic schemes (SM, SR, and SC), which assign a weighting coefficient to each of the top-ten most similar training set compounds, have reasonable predictivity of a test set. The algorithmic scheme which categorizes the most similar compounds into different weighted clusters predicts the test set best. The 4D-fingerprints provide 36 different individual IPE/IPE type molecular similarity measures. Further investigation shows that the NP/HA, HS/HA, and HA/HA IPE/IPE type measures predict the test set well. Moreover, these three IPE/IPE type similarity measures are very similar to one another for the particular training and test sets investigated. The 4D-fingerprints have relatively high predictivity for this particular dataset.

1. Introduction

Cluster analysis is the process in which groups are detected within a set of objects where members in a group are 'similar' to one another with respect to some attribute while 'dissimilar' to members in other groups. Cluster analysis was first used in taxonomy for species classification but has been gradually adopted by other disciplines where information about grouping is necessary. Before a cluster method can be applied, measures representing attributes of objects have to be derived as the basis for comparison among different objects. Criteria to determine what measures should be used to reflect the essential properties of an object mainly depend on the purpose of the study. Thus, selection of attributes is a subjective process. Furthermore, choosing an appropriate clustering set of criteria (method) for the task at hand also relies on the attributes of the dataset and the goals one hopes to achieve. Essentially, there is no one 'correct' set of clusters for a particular set of objects.[1] Cluster analysis has been adopted by scientists working in computer-aided drug design owing to the 'principle' that molecules having similar structural and physiochemical properties are likely to behave similarly in a biological system. Thus, the rational classification of a large compound library may be useful for identifying drug-likeness hit compounds.

Human serum albumin (HSA),[2] the most abundant protein in blood plasma (MR 66 kDa, concentration 0.53–0.75 mM), has multiple hydrophobic binding sites (a total of eight for fatty acids, endogenous ligands of HSA) and is known to bind a diverse set of drugs, especially neutral and negatively charged hydrophobic compounds. The structure of HSA and its multiple binding sites are shown in Figure 1. For drug-like compounds, two high affinity binding sites have been proposed in subdomains IIA and IIIA of HSA, which are known as Sudlow's sites I and II[3,4], and are highly elongated

Keywords: Molecular similarity; 4D-fingerprint similarity; HSA; Cluster analysis.
* Corresponding author. Tel.: +1 302 831 3522; fax: +1 302 831 633; e-mail: zhong@udel.edu

0968-0896/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2005.08.035

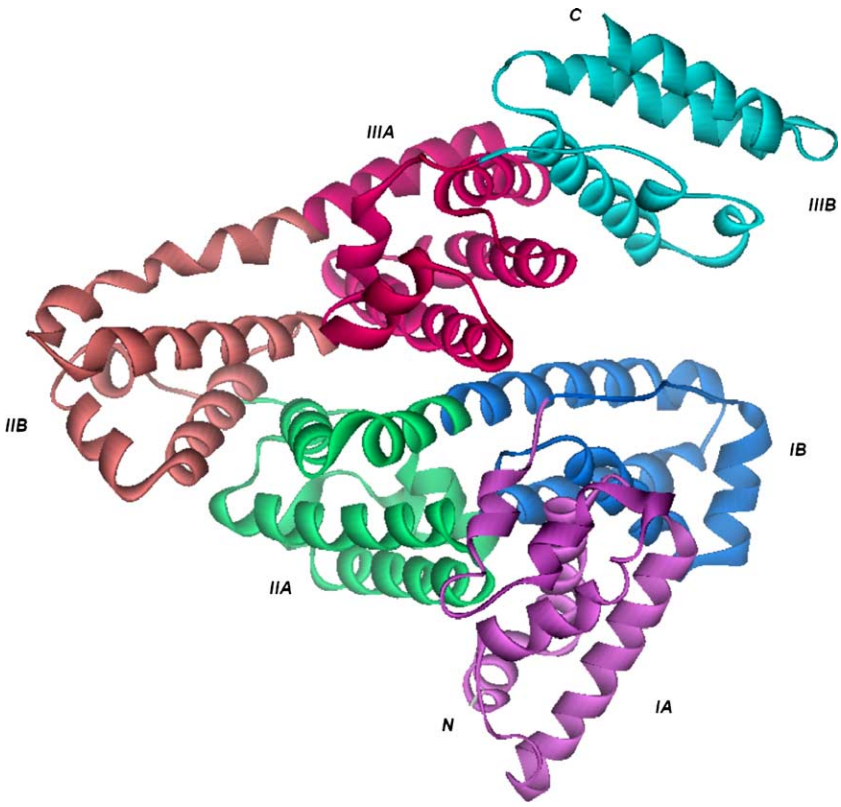

Figure 1. A schematic representation of the native structure of HSA and its six known binding sites.

hydrophobic pockets with charged lysine and arginine residues near the surface which can function as attachment points for polar groups of ligands.[5] High resolution X-ray crystal structures of HSA complexes with two anesthetics, propofol and halothane[6], and with the widely used anticoagulant warfarin, have been reported.[7] These crystal structure studies have also found that by increasing the ligand concentration, the ligands can occupy all known HSA binding sites, each with a different affinity and a different pharmacological relevance. A detailed understanding of the HSA binding modes of most drugs is, however, still missing. Drug binding to HSA is widely assumed to be simply dominated by the lipophilicity of a ligand.[8] Recent QSAR analysis demonstrates that overall lipophilicity is rather poorly correlated to HSA binding for a diverse set of molecules. This finding is in contrast to studies on congeneric series where lipophilicity is often found to be the dominant factor, suggesting that specific molecular recognition elements beside physicochemical parameters are essential.[9]

The recognition of the critical role that ligand molecular shape, and more generally, steric aspects play in ligand–receptor binding has stimulated an interest in developing methods for comparing shapes of molecules as a means of predicting binding. However, while similar molecules are expected to exert similar activities, there is no rigorous or unambiguous method for defining and calculating their similarity. It may be expected that, in certain cases, overall molecular similarity will produce similar activity, whereas, in other cases, only the molecular similarity of certain (active) regions of the molecules will give rise to similar activities. Moreover, there is more than one similarity metric, and there are different techniques with which to evaluate molecular similarity and to model properties which depend on spatial and shape features of molecule.

For a set of N molecules, $N \times N$ similarity matrices, in which each molecule is compared with all the other molecules under study, have been constructed. The chemical information inherent to the similarity matrix was correlated to the biological activity measures of the N compounds through a combined partial least-squares,[10,11] genetic algorithm,[12,13] and artificial neural network algorithm[14,15] and QSARs have been reported using this approach.

The theoretical basis of molecular similarity in QSAR analysis is different from the traditional 2D-QSAR paradigm. The molecular similarity measure is not like conventional parameters (e.g., σ, π, and MR), as it does not encode physicochemical property measures that are specific to molecular substituents. The molecular similarity measure is the resemblance between a pair of molecules based on some composite molecular feature like spatial or electrostatic attributes, or a combination of the two.[12] The use of molecular similarity offers a new descriptor dimension to QSAR studies. Instead of a correlation between substituent properties and activities, a molecular similarity based QSAR establishes an association between global structural properties and activity variation among a series of molecules. The implicit assumption in

a molecular similarity QSAR analysis is that globally similar compounds have similar activities.[16] Different notions of molecular similarity have been suggested and used based on molecular formula, molecular graphs, molecular skeletons, atom types, and positions, conformations, van der Waals surfaces, and/or molecular fields.[17] In this study, we describe a new, generic prediction tool applied to predict the binding of drugs to HSA. And the molecular similarity calculations are built on the 4D pharmacophore similarity concept, first discussed by Duca and Hopfinger.[18]

2. Results

The details of the 4D-fingerprints in molecular similarity analyses are given by Duca and Hopfinger.[18] There are a total of 36 different IPE pair types of molecular similarity matrices, which are defined in Table 5. For each of them, the four schemes mentioned in Section 4 have been applied. The r^2_{pre} of each scheme for each IPE/IPE molecular similarity matrix are shown in Table 1. The three IPEs yielding the best predictions for each method are shown in bold. In general, nearly all of the IPE pair type molecular similarity matrices do not predict the test set well using the SM scheme (Eq. 2). Thus, the most similar compound's HSA binding affinity cannot be used to represent the test compound's binding affinity when molecular similarity is based on a single individual compound and IPE type similarity. Since 4D-fingerprint molecular similarity is based on sampling many conformations of each compound, the 4D-fingerprint molecular similarity value contains the Boltzmann average spatial properties of each pair of compounds. This situation is not like 2D-connectivity similarity where only the chemical structure is captured in a similarity comparison. The poor r^2_{pre} results using the SM scheme and each of the individual IPE types suggest that a single IPE type and a single compound, even when employing the compound's 3D spatial structure, cannot adequately represent the compound with respect to HSA binding. The most similar training set compound for a single IPE type and the test compound may bind at different binding sites, or in different binding modes, to HSA.

But the other three schemes, SA, SR, and SC, can better predict a test compound's HSA binding affinity. This general result again suggests that using a reasonable algorithm which summarizes all of the pertinent similar training set compounds to the test compound is a preferable predictive approach to the use of only the most similar training set compound based on a particular similarity measure. The last scheme, SC, which first clusters similar compounds and then assigns them different weighting coefficients, see Eqs. 4 and 5, predicts the test set better than the other schemes. Of the 36 IPE/IPE molecular similarity measures, it is interesting to note that the NP/HA IPE/IPE measure yields the best predictions based on the r^2_{pre} value. This type of IPE/IPE molecular similarity reflects the spatial relationship of the nonpolar groups and the hydrogen bond acceptors over a molecule. Under this IPE/IPE measure, if the hydrogen bond acceptor and nonpolar features in a reference molecule are jointly similar to those of a test compound, the two molecules are measured as being very similar. An isoproperty surface of salicylic acid is shown in Figure 2a. The red region within the circle represents the hydrogen bond acceptor zone. The most similar compound for NP/HA, fentiazac, is shown in Figure 2b, which has a similar hydrogen bond acceptor region to salicylic acid, even though these two molecules are different in molecular size and shape. This case is an example that 4D-fingerprint molecular similarity can separate out molecular space similarity according to IPE types. The successful predictions using the NP/HA IPE/IPE type similarity measure suggest that hydrogen bonding and nonpolar interactions are very important for drug-like compounds binding to HSA.

Table 1. The r^2_{pre} values using the four schemes described in Section 4 and applied using the 36 different IPE types of molecular similarity measures

IPE/IPE	r^2_{pre}				IPE/IPE	r^2_{pre}			
	SM	SA	SR	SC		SM	SA	SR	SC
Sim00	**0.258**	0.169	0.259	0.240	Sim33	−0.904	0.032	0.188	0.260
Sim01	−1.622	−0.221	0.017	−0.041	Sim34	−0.424	0.129	0.161	0.107
Sim02	−2.061	0.101	0.146	0.141	Sim35	−1.740	−0.124	0.061	−0.029
Sim03	−1.041	0.175	0.358	0.215	Sim36	−5.882	−0.299	0.198	−0.737
Sim04	−0.949	**0.334**	**0.376**	0.359	Sim44	−0.636	0.282	0.286	**0.413**
Sim05	−1.278	0.260	0.218	0.272	Sim45	−2.411	0.052	0.184	0.121
Sim06	−4.632	0.116	0.240	−0.427	Sim46	−5.214	−0.013	−0.033	−0.511
Sim11	**0.251**	0.011	**0.368**	0.348	Sim55	−1.831	−0.483	−0.167	−0.308
Sim12	−1.580	0.112	0.294	0.169	Sim56	−6.345	−1.112	−0.856	−1.970
Sim13	**−0.067**	0.198	0.336	0.346	Sim66	−4.956	−0.165	−0.122	−0.779
Sim14	−0.501	**0.339**	**0.372**	**0.471**	Sim70	−0.241	−0.074	0.160	0.127
Sim15	−1.937	0.256	0.296	0.228	Sim71	−0.414	−0.085	0.138	−0.121
Sim16	−4.657	0.054	0.060	−0.495	Sim72	−0.262	0.107	0.157	0.178
Sim22	−1.975	−0.406	−0.116	−0.392	Sim73	−0.710	0.157	0.191	0.187
Sim23	−0.914	−0.281	0.143	−0.155	Sim74	−1.497	**0.368**	0.331	**0.446**
Sim24	−1.476	−0.185	0.138	−0.104	Sim75	−2.255	0.100	0.119	0.173
Sim25	−1.257	0.106	0.192	0.178	Sim76	−4.709	0.132	0.108	−0.447
Sim26	−5.830	−0.490	−0.390	−1.098	Sim77	−0.118	0.167	0.234	0.277

Sim of Sim*ij* in the IPE/IPE column refers to the IPE type 4D-fingerprint similarity measure, and the *i* and *j* identify the IPE types as defined in Table 5. The values in **bold** are the three highest for each scheme.

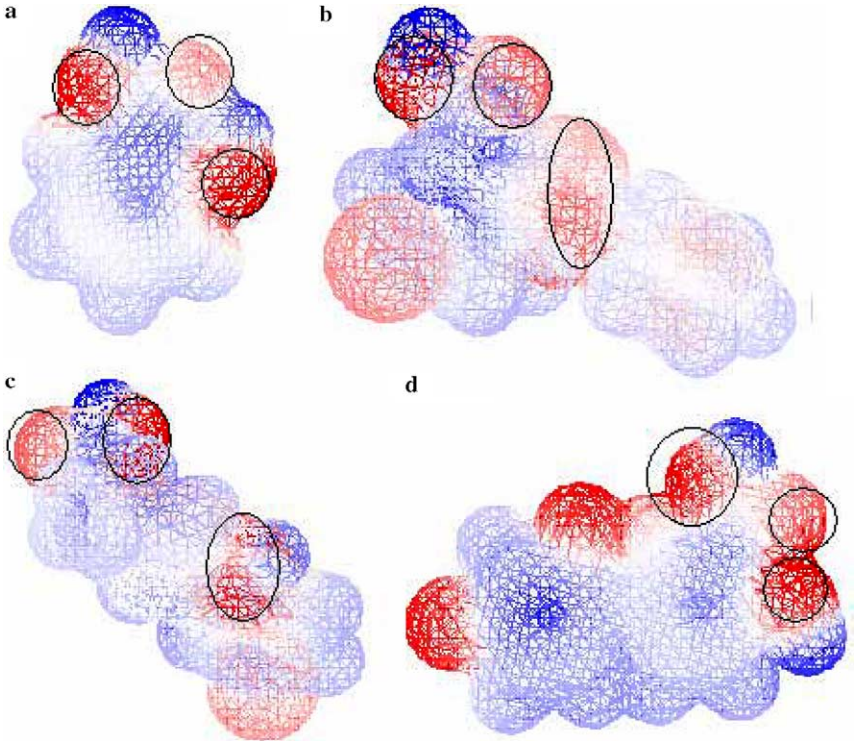

Figure 2. The isoproperty surface of salicylic acid (a), fentiazac (b), carprofen (c), and diflunisal (d) in wire mesh contours. Fentiazac, carprofen, and diflunisal are the most similar compounds to salicylic acid according to NP/HA, HS/HA, and HA/HA IPE/IPE types. The red and pink contours in the circles represent the regions of hydrogen bond acceptors.

It is not surprising that the HS/HA IPE/IPE similarity type predicts HSA binding behavior second best based upon r^2_{pre}, which is 0.446. In this dataset, there are much fewer polar atoms than there are non-polar atoms in most compounds. Thus, the HS/HA molecular similarity measures are very close to NP/HA molecular similarity measures. Using salicylic acid as an example, Table 2 contains the top-ten most similar compounds according to NP/HA and HS/HA IPE/IPE types. Seven compounds of the two IPE/IPE measures are the same, but the similarity order to salicylic acid is different. Figure 3 shows the molecular similarity measures of the compounds in the training set relative to salicylic acid according to the HS/HA, and NP/HA IPE/IPE types. The trends across these IPE/IPE types are very similar, which suggests that the HS/HA and NP/HA IPE/IPE similarity measures are, indeed, highly similar (Table 3).

The HA/HA IPE/IPE similarity type, which is based solely on the spatial distribution of the hydrogen bonding acceptors in a molecule, also predicts the test set well. Table 2 includes a listing of the top-ten most similar compounds to salicylic acid according to the HA/HA IPE/IPE molecular similarity type. Six and seven, compounds are the same as those found using the NP/HA and HS/HA IPE/IPE type measures, respectively, but the ranking by similarity to salicylic acid is different. Figure 2 also shows the isoproperty surfaces of carprofen (c) and diflunisal (d), which are the most similar pair of compounds, according to the HS/HA and HA/HA IPE/IPE based molecular similarity calculations. It is clear that all compounds in Figure 2 have highly similar isoproperty surfaces. These results demonstrate that the NP/HA, HS/HA, and HA/HA IPE/IPE types yield very similar results for this data set, and all three IPE/IPE

Table 2. The top-ten most similar training set compounds to salicylic acid using the NP/HA, HA/HA, and HS/HA IPE/IPE types

Sim14(NP/HA IPE/IPE type)		Sim74(HS/HA IPE/IPE type)		Sim44(HA/HA IPE/IPE type)	
Similarity	Compounds	Similarity	Compounds	Similarity	Compounds
0.881	Fenbufen	0.904	Pirprofen	0.829	Cicletanine
0.906	Cicletanine	0.909	Hydroxycoumarin	0.867	Carbamazepine
0.915	Suprofen	0.941	Fenbufen	0.871	Fentiazac
0.924	Clofibric acid	0.945	Cicletanine	0.896	Fenbufen
0.928	Carbamazepine	0.946	Fentiazac	0.919	Hydroxycoumarin
0.929	Carprofen	0.955	Clofibric acid	0.951	Clofibric acid
0.932	Hydroxycoumarin	0.969	Diflunisal	0.964	Mepivacaine
0.946	Pirprofen	0.971	Itanoxone	0.965	Itanoxone
0.958	Sulindac	0.978	Naproxen	0.965	Lidocaine
0.98	Fentiazac	0.996	Carprofen	0.998	Diflunisal

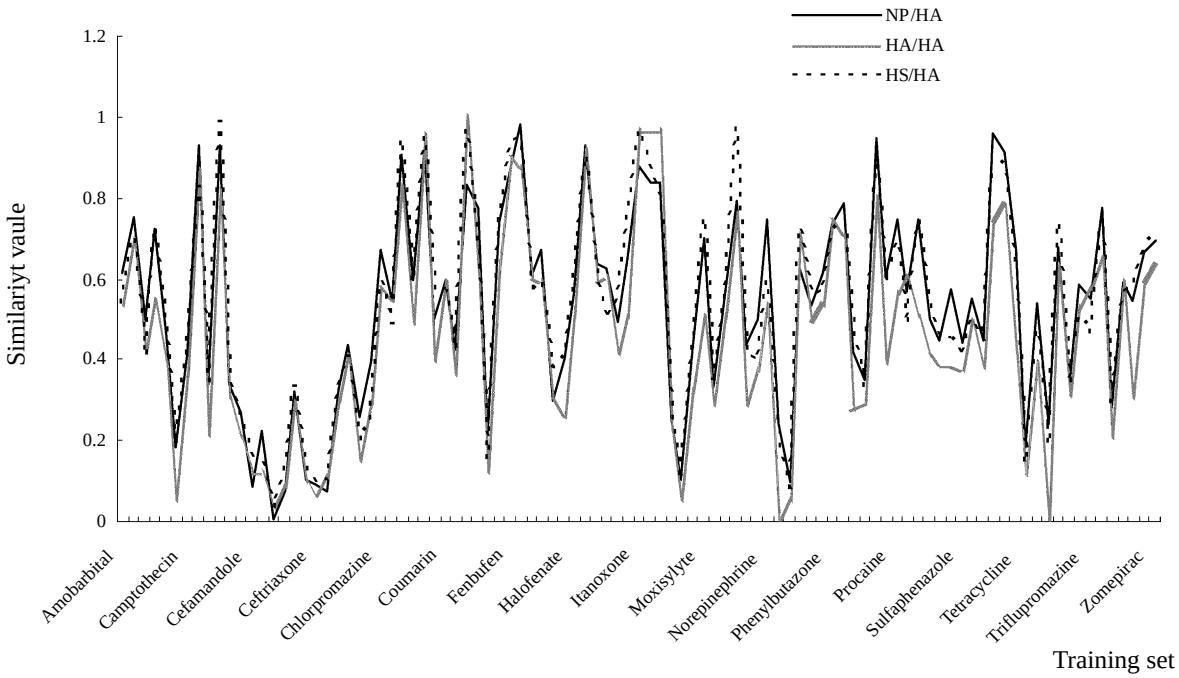

Figure 3. The molecular similarity values of the compounds in the training set relative to the test compound, salicylic acid, according to NP/HA, HS/HA, and HA/HA IPE/IPE types. The three similarity measures are very similar which indicates that the NP/HA, HS/HA, and HA/HA IPE/IPE types collect much of the same information.

types predict the test set well. According to Lipinski's 'rule of five,'[19] most drug-like compounds have less than 10 hydrogen-bond acceptors, which means hydrogen-bond acceptors generally compose only a limited part of a drug molecule. But this specific molecular feature among the molecules is identified as being highly the indicative of binding behavior to HSA.

Based on these results, it is clear that the three algorithmic schemes, SA, SR, and SC better predict HSA binding than the SM scheme. In turn, it would appear that the 4D-fingerprint measures of the top-ten most similar compounds provide better information than simply using the most similar compound 4D-fingerprint information. Overall, of the 36 IPE/IPE molecular similarity measures the NP/HA, HS/HA, and HA/HA IPE/IPE types yield similar results for this dataset, and all of them predict the test set well using the SC algorithmic scheme.

3. Discussion

A number of different schemes have been designed and used in the application of molecular similarity calculations for the creation of QSAR models. The easiest way to correlate molecular similarity measures with biological data is to use simple regression analysis. Many examples of such studies can be found in the literature. Seri-Levy and Richards[20,21] exploited molecular similarity data to construct QSARs for ligand enantiomer eudismic ratios. Burt et al.[22] correlated the activity of a set of nitromethylene insecticides using their molecular electrostatic potential (MEP) similarity to the most active molecule in the series. It is clear from these studies that molecular similarity-based-regression equations can produce good QSARs. Nevertheless, it is unlikely that a single molecule will contain all, or most, of the structural information inherent to a given ligand dataset, especially for the current study of ligand binding to HSA which has multiple binding sites. We applied regression analysis to the training set in Table 4 using 2D-connectivity similarity measure and built a QSAR model. However, the predictivity of this model is not good ($q^2 = 0.106$). This result is quite different from recent QSAR modelings of binding to human serum proteins with similar measurements, which produced a five variable model ($q^2 = 0.76$) QSAR model for penicillin and an eight variable model ($q^2 = 0.78$) QSAR model for penicillin.[23] The possible reason is that our data consist of the binding affinity constant from diverse drugs, but theirs is only for penicillin or its derivatives. Therefore, they have the same binding site, which makes the topological descriptor an important character to describe HSA binding of these compounds. A combination of E-state and molecular connectivity χ indices was introduced to get a six variable QSAR model ($q^2 = 0.77$).[24] Four of them are E-state-related descriptors. This may explain why we get so bad predictivity of our model.

Four new algorithmic schemes were developed to predict the HSA–ligand binding based on the assumption that similar compounds have similar binding modes and binding affinities. The first algorithmic scheme predicts the binding affinity of a test compound using only the most similar training set compound's binding affinity. This scheme has relatively poor predictivity, based on 4D-fingerprints. The other three algorithmic schemes, which assign a weighting coefficient to each of the top-ten most similar training set compounds, have reasonable predictivity of a test set. Finally, the algorithmic

Table 3. The correlation coefficients among the P^-, HA, and HD IPE type 4D-fingerprint molecular similarity measures

	Sim03	Sim04	Sim05	Sim13	Sim14	Sim15	Sim23	Sim24	Sim25	Sim33	Sim34	Sim35	Sim44	Sim45	Sim55	Sim73	Sim74	Sim75
Sim03	1.000	0.707	0.046	0.945	0.684	0.051	0.154	0.160	0.052	0.920	0.599	0.047	0.623	0.050	0.046	0.968	0.684	0.043
Sim04	0.707	1.000	0.076	0.664	0.943	0.082	0.208	0.222	0.082	0.630	0.901	0.078	0.912	0.083	0.064	0.689	0.970	0.073
Sim05	0.046	0.076	1.000	0.032	0.057	0.982	0.307	0.323	0.891	0.035	0.061	0.949	0.059	0.949	0.353	0.048	0.076	0.984
Sim13	0.945	0.664	0.032	1.000	0.702	0.037	0.131	0.140	0.040	0.906	0.576	0.035	0.604	0.036	0.039	0.924	0.643	0.030
Sim14	0.684	0.943	0.057	0.702	1.000	0.065	0.184	0.198	0.068	0.632	0.880	0.060	0.895	0.064	0.055	0.666	0.922	0.054
Sim15	0.051	0.082	0.982	0.037	0.065	1.000	0.320	0.336	0.891	0.040	0.067	0.935	0.065	0.939	0.300	0.053	0.083	0.978
Sim23	0.154	0.208	0.307	0.131	0.184	0.320	1.000	0.943	0.274	0.141	0.201	0.310	0.188	0.315	0.116	0.154	0.206	0.304
Sim24	0.160	0.222	0.323	0.140	0.198	0.336	0.943	1.000	0.300	0.144	0.211	0.327	0.203	0.338	0.126	0.159	0.218	0.320
Sim25	0.052	0.082	0.891	0.040	0.068	0.891	0.274	0.300	1.000	0.039	0.065	0.878	0.063	0.885	0.329	0.053	0.082	0.882
Sim33	0.920	0.630	0.035	0.906	0.632	0.040	0.141	0.144	0.039	1.000	0.612	0.037	0.645	0.038	0.032	0.914	0.624	0.033
Sim34	0.599	0.901	0.061	0.576	0.880	0.067	0.201	0.211	0.065	0.612	1.000	0.063	0.941	0.066	0.050	0.594	0.901	0.057
Sim35	0.047	0.078	0.949	0.035	0.060	0.935	0.310	0.327	0.878	0.037	0.063	1.000	0.060	0.982	0.326	0.049	0.077	0.945
Sim44	0.623	0.912	0.059	0.604	0.895	0.065	0.188	0.203	0.063	0.645	0.941	0.060	1.000	0.064	0.042	0.615	0.906	0.056
Sim45	0.050	0.083	0.949	0.036	0.064	0.939	0.315	0.338	0.885	0.038	0.066	0.982	0.064	1.000	0.319	0.052	0.082	0.949
Sim55	0.046	0.064	0.353	0.039	0.055	0.300	0.116	0.126	0.329	0.032	0.050	0.326	0.042	0.319	1.000	0.047	0.064	0.303
Sim73	0.968	0.689	0.048	0.924	0.666	0.053	0.154	0.159	0.053	0.914	0.594	0.049	0.615	0.052	0.047	1.000	0.694	0.046
Sim74	0.684	0.970	0.076	0.643	0.922	0.083	0.206	0.218	0.082	0.624	0.901	0.077	0.906	0.082	0.064	0.694	1.000	0.074
Sim75	0.043	0.073	0.984	0.030	0.054	0.978	0.304	0.320	0.882	0.033	0.057	0.945	0.056	0.949	0.303	0.046	0.074	1.000

Sim of Sim*ij* in the first column/row refers to the IPE type 4D-fingerprint similarity measure, and *i* and *j* denote the IPE types defined in Table 5.

scheme which categorizes the most similar compounds into different weighted clusters predicts the ligand–protein binding affinity of the test set best.

The 4D-fingerprints provide 36 different individual IPE/IPE type molecular similarity measures. The NP/HA, HS/HA, and HA/HA IPE/IPE type measures predict the test set well. Moreover, these three IPE/IPE type similarity measures are very similar to one another for the particular training and test sets investigated. Using the similarity measures of salicylic acid to each compound in the training set as an example, the correlation coefficients between each of the three IPE/IPE types are 0.895 (NP/HA:HS/HA), 0.906 (NP/HA:HA/HA), and 0.922 (HS/HA:HA/HA).

To minimize chance correlation from using only salicylic acid as an example, the correlation coefficients between each IPE/IPE type of molecular similarity are calculated using the similarity measures of all compounds in the test set to each compound in the training set, which are 0.943 (NP/HA:HS/HA), 0.895 (NP/HA:HA/HA), and 0.922 (HS/HA:HA/HA). This finding again supports that the NP/HA, HS/HA, and HA/HA IPE/IPE types of similarity measures are, in turn, highly similar to one another for this dataset. In fact, among the 36 different IPE/IPE type molecular similarity measures, many measures are similar for this dataset. For example, the correlation coefficient between the A/HD and NP/HD, P^+/HD, P^-/HD, HA/HD, and HS/HD IPE/IPE type similarity measures are 0.982, 0.891, 0.949, 0.949, and 0.984, respectively. Therefore, these IPE pair type similarities have nearly identical predictivities, where the r^2_{pre} are 0.272, 0.228, 0.178, 0.029, 0.121, and 0.173, respectively, using the SC scheme. However, the correlations between these IPE pair type measures, and the NP/HA, HS/HA, and HA/HA IPE/IPE type similarities are low. For example, the similarity values between the A/HD IPE/IPE type, and the NP/HA, HS/HA, and HA/HA IPE/IPE types are 0.065, 0.065, and 0.083, respectively. These results also suggest that hydrogen bond donor features in a molecule are very important determinants to explore the HSA binding structure–activity relationship.

Correlation coefficients between the A/P^- IPE pair type similarity, and the NP/P^-, P^-/P^-, and HS/P^- IPE/IPE type similarity measures are also very high, 0.945, 0.920, and 0.968, respectively. These IPE/IPE type similarity measures also have a relatively high predictivity using the SC scheme. IPE pair types involving P^- focus on the molecular similarity measures of the polar atoms with negative partial charges in a molecule. Most hydrogen bond acceptors in an organic molecule are nitrogen or oxygen atoms, and are also polar atoms with negative partial atomic charges. Therefore, P^- IPE types of similarity measures should show a correspondence to the HA or HD IPE type similarity measures. Table contains the similarity values of the P^-, HA, and HD IPE types of similarity measures. Most of the P^- IPE type similarity measures each have a relatively high correspondence to the HA IPE type measures, but have a low correspondence to the HD IPE type of similarity measures. These results suggest that 4D-fingerprints can

separate molecular space properties with respect to the IPE types of atoms composing a molecule. All of the 4D-fingerprint similarity measures can be used to predict a test set. In this case, the P^- and HD IPE similarity measures are not better than the HA IPE type similarity measure. In addition, the 4D-fingerprints seem to search out the most important types of interactions between a ligand and its receptor by using sets of pairwise IPEs.

As described in Section 4, the test set was selected based on the 2D-connectivity molecular similarity measures by defining the chemical space of the total dataset. But the same test set was used in evaluating 4D-fingerprint molecular similarity to predict HSA binding. Ideally, the comparison of 2D-connectivity and 4D-fingerprint HSA binding affinity predictions should be done under the condition that the training set and the test set are the same, and have the same chemical space and biological activity distribution for both the 4D-fingerprint and the 2D-connectivity dendrograms. The dendrogram distribution of the combined training set and test set, according to the NP/HA 4D-fingerprint similarity measure, is shown in Figure 4. Obviously, the test set distribution is not as uniform as that found using 2D-connectivity similarity and shown in Figure 5. But even under this restricted condition, the NP/HA 4D-fingerprint molecular similarity measures still have high predictivity.

The data points to a conclusion that may be summarized in simple term: the 4D-fingerprints provide a set of molecular similarity measures that are meaningful and comprehensive. That is, 4D-fingerprints provide 36 different IPE pair types of molecular similarity measures to separate molecular space properties and provide multiple pathways to interpret structure-activity data. Moreover, the algorithmic schemes, which assign a weighting coefficient to each of the top-ten most similar training set compounds, have reasonable predictivity of a test set and the algorithmic scheme which categorizes the most similar compounds into different weighted clusters predicts the ligand–protein binding affinity of the test set best.

The significance of 4D-fingerprint molecular similarity analyses lies at the high throughout screening before experiment data. In this study, we can build a model to predict plasma protein binding model based on the algorithmic scheme which categorizes the most similar compounds into different weighted clusters by using 44D-fingerprint molecular similarity measures. The practical application depends on the dataset. Generally, the more dataset, the wider value of data, the more reliable the model.

4. Materials and methods

4.1. Dataset

A 115 compound HSA ligand binding dataset was constructed from literature data. The data percentage of HSA plasma protein binding comes from the work by Kratochwil and co-workers,[9] and from the textbooks of Goodman and Gilman[25] and of Dollery,[26] where, in general, f_b of a drug refers to the fraction bound to HSA plasma protein. The percentage bound values were converted into equivalent binding affinities, k_b, using Eq. 1, which is derived from the law of mass action. k_b is the drug binding affinity to HSA under the assumption that binding occurs exclusively to HSA, a binary complex is formed, and an excess of HSA (concentration 0.6 mM, [HSA]) is present as compared to the concentration of the drug. Table 4 contains the complete set of data for both the training and test sets

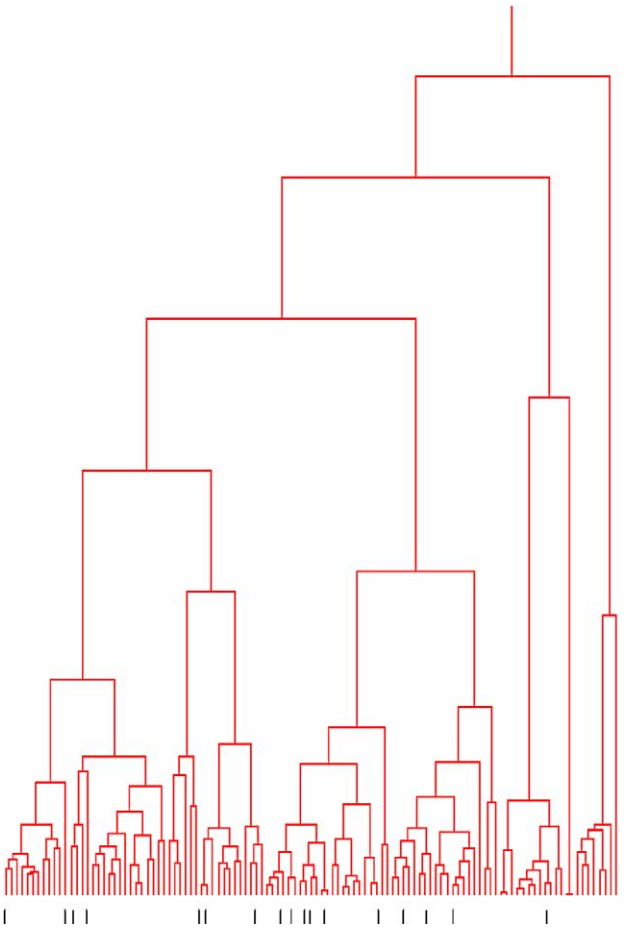

Figure 4. The dendrogram distribution of both the training set and test set according to the NP/HA IPE/IPE 4D-fingerprint molecular similarity measure. The short and black lines along the '*x*-axis' represent the positions of the test set compounds in the dendrogram.

$$\log k_b = \log \frac{[f_b]}{1-[f_b]} - \log[\mathrm{HSA}]. \quad (1)$$

The lowest binding compound in the training set is chlorpromazine and the highest is bilirubin. The obvious structural dissimilarity of compounds in this dataset suggests that these compounds bind to HSA in multiple ways. Thus, this dataset presents a good opportunity to explore the relationship between molecular similarity and HSA binding constant. At the same time, molecular similarity differences are represented using 4D-fingerprint measures, which produce relatively high prediction on test set. And the 4D-fingerprints also provide additional information which is helpful to interpret the binding mechanisms of drug-like compounds to HSA,

Table 4. The training and test set compounds for the HSA binding analyses

Compound	$\log k_b$
Training set	
Amobarbital	3.66
Aspirin	4.37
Azapropazone	5.88
Benoxaprofen	6.28
Benzylpenicillin	3.04
Bilirubin	7.79
Camptothecin	6.56
Carbamazepine	3.13
Carbenoxolone	7.7
Carprofen	6.64
Cefaclor	2.11
Cefadroxil	3.28
Cefamandole	3.57
Cefazolin	4.36
Cefoperazone	4.54
Cefotaxime	3.1
Cefradine	2.54
Cefsulodin	2.58
Ceftazidime	2.68
Ceftriaxone	3.14
Cefuroxime	3.63
Cephalexin	2.5
Cephaloridine	3.61
Cephapirin	3.31
Chlorothiazide	4.61
Chlorpromazine	1.98
Chlorpropamide	4.83
Cicletanine	4.88
Cimoxatone	4.61
Clofibric acid	5.52
Clometacin	4.44
Coumarin	3.89
Dicoumarol	6.1
Diflunisal	6.2
Disopyramide	3.66
Doxycycline	4.42
Etodolac	5.03
Fenbufen	5.62
Fentiazac	5.57
Fluindione	5.69
Flurbiprofen	5.95
Furosemide	5.27
Fusidine	4.89
Halofenate	5.2
Hydroxycoumarin	5.64
Ibuprofen	5.52
Imipramine	4.38
Indomethacin	5.71
Indoprofen	5.27
Itanoxone	5.29
Lidocaine	3.74
Mepivacaine	5.4
Methicillin	2.96
Methotrexate	3.45
Methylorange	5.57
Moxisylyte	2.83
Nafcillin	4.08
Nalidixic acid	4.34
Naproxen	6.2
Nicergoline	3.79
Nimesulide	5.69
Norepinephrine	7
Nortriptyline	3.51
Novobiocin	5.74
Oxazepam	4.56
Oxyphenbutazone	5.29
Phenobarbital	3
Phenylbutazone	5.54
Phenytoin	4.07
Pipotiazine	3.27
Piretanide	5.13
Pirprofen	5.75
Pregnenolone	4.63
Procaine	3.49
Promazine	4.93
Quinine	3.88
Sotalol	3.3
Sulfaethidole	5.18
Sulfamethoxazole	3.7
Sulfaphenazole	5.3
Sulfathiazole	4.4
Sulfisoxazole	4.34
Sulindac	5.4
Suprofen	5.18
Testosterone	4.47
Tetracycline	3.64
Thyroxine	5.77
Ticlopidine	3.97
Tinoridine	4.38
Tolazamide	4.94
Tolbutamide	6.52
Triflupromazine	4.74
Tryptophane	4.59
Urapidil	3.34
Valproic acid	4.76
Verapamil	3.43
Warfarin	5.33
Zomepirac	4.28
Test set	
Acenocoumarin	5.32
Acetylcoumarin	4.39
n-Butyl-*p*-aminobenzoate	4.45
Binedaline	4.48
Bupivacaine	3.88
Befoxitin	3.05
Beftriaxone	4.72
Betiedil	3.99
bhlorpropamide	4.98
Diclofenac	5.90
Digitoxingenin	4.53
Fenoprofen	5.67
Ketoprofen	6.16
Pentobarbital	3.08
Practolol	2.42
Salicylic acid	4.93
Tetracycline	4.64

and is also very insightful in the design of compounds which bind to HSA in a given affinity range.

4.2. Molecular similarity calculations

Although it is possible to define molecular similarity within the context of global, topological, or even substituent-based parameters, just as done in 3D-QSAR models, a single 'default' conformation is generally used for each molecule. To overcome this limitation, 4D molecular similarity, 4D-MS, developed by Duca and

Hopfinger,[18] includes the thermodynamic distribution of conformation ensemble available to a molecule in constructing a set of similarity/diversity descriptors.[27–29] In order to fully investigate possible relationships between molecular similarity measures and HSA binding constants, 4D-fingerprint molecular similarity measures (4D-MS) are calculated.

The fourth dimension is exactly the time average conformation ensemble. The theory and corresponding methodology for constructing the main distance-dependent matrix, MDDM, matrices, and computing corresponding eigenvalues for each matrix, using 4D molecular similarities, are presented in detail by Duca and Hopfinger.[18] Absolute molecular similarity MDDMs for the pairs of atoms with the same IPE types are computed for each molecule in the dataset. Eigenvalues of the MDDM matrices are then employed as 4D-fingerprints to represent a molecule with respect to a particular IPE type. Thus, for each molecule, eight MDDMs can be constructed and eight sets of eigenvalues (4D-fingerprints) computed that correspond, individually, to the eight IPE types which are presented in Table 5. A threshold cutoff value for the eigenvalues is applied, and those normalized eigenvalues below the threshold cutoff value are disregarded. For this study, the threshold cutoff was set at 0.002.

4.3. The selection of the training set and the test set

In order to investigate how clustering a large dataset into smaller groups may influence the resultant predictions, the 2D molecular similarity matrix[30], the original dataset, was used as a preprocessor to cluster the dataset into training set and a corresponding test set. The underlying idea for this type of compound separation is that both the training and test sets should span the entire molecular similarity range, and the test set be distributed in similarity measures in the same way as the training set. In this process, a hierarchical clustering method was applied to classify these compounds into different clusters using the SAS 8.2 software.[31] The complete dataset is first partitioned into six groups, which contain 29, 24, 8, 17, 17, and 20 compounds, respectively. The purpose of this step is to accomplish the 'chemical' space separation of the complete dataset. Then compounds from each subset having high, medium, and low binding affinities to HSA are arbitrarily extracted and placed in the test set. This second step accomplishes the 'biological' activity separation. These two separation processes result in 98 compounds in the training set and 17 compounds in the test set. In this way, compounds in both the training set and the test set are distributed across the different clusters and also span the entire range in HSA binding affinity. The dendrogram distribution of the entire dataset leading to the training set and the test set is shown in Figure 5. The short and black lines along the '*x*-axis' at the bottom of the plot represent the positions of the test set compounds in the dendrogram. It is clear that these test set compounds are proportionally distributed across the entire original dataset.

Table 5. The set of interaction pharmacophore elements (IPEs)

IPE description	Symbol	Number code
Any type of atom	A	0
Nonpolar atom	NP	1
Polar atom of positive partial charge	P^+	2
Polar atom of negative partial charge	P^-	3
Hydrogen bond acceptor	HA	4
Hydrogen bond donor	HD	5
Aromatic atoms of molecule	Ar	6
Non-hydrogen atoms	HS	7

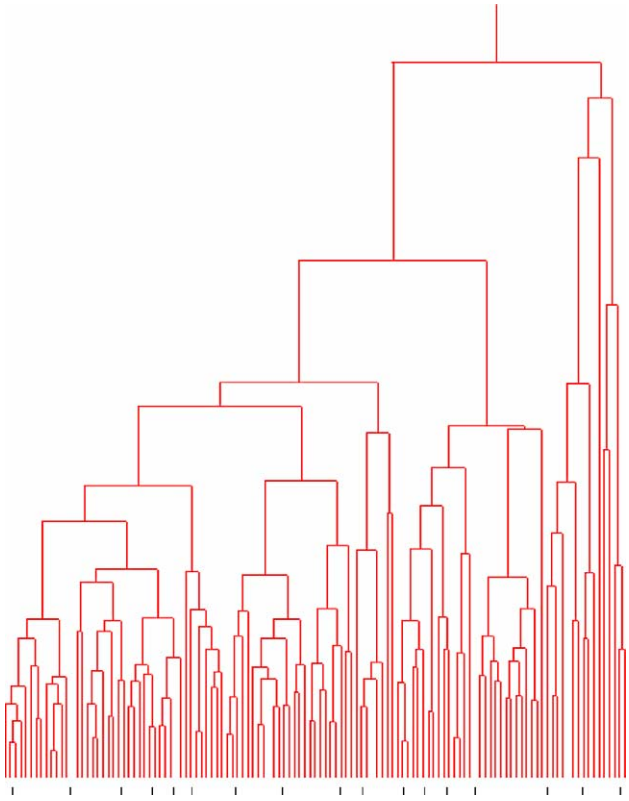

Figure 5. The dendrogram distribution of the training set and the test set compounds based upon the 2D-connectivity molecular similarity matrix. The short and black lines at the bottom along the '*x*-axis' represent the position of the test set compounds in the dendrogram.

4.4. HSA binding affinity prediction of test set based on molecular similarity

The most direct way to predict the test set is to simply build a QSAR model using regression analysis. However, this approach was not successful when applied to this dataset. We developed an alternate methodology to successfully predict the HSA binding affinity of the test set. As mentioned in the Introduction section, the implicit assumption behind our approach is that globally similar compounds exhibit similar activities. Based on this assumption, the following algorithm has been evolved:

(a) For each compound in the test set, the most similar compound in the training set was identified and its activity used as the predicted activity of the test compound. This method is abbreviated as SM when used later in the analysis.

$$\log k_{\text{predicted}} = \log k_{\text{max-similar}}, \quad (2)$$

where $k_{\text{predicted}}$ means the predicted activity of the test compound, whereas $k_{\text{max-similar}}$ is the activity of the most similar compound in the training set.

(b) The molecular similarity matrix contains the global collection of the similarity information of all the training and test set compounds. Often the second or third most similar training set compound to a test compound may have a different binding mode. Thus, it is unlikely that only the most similar molecule will contain all, or even most, of the structural information inherent to a given ligand dataset. This is particularly the case for HSA, which has many binding sites and many drug binding modes are likely still unknown. In order to collect a larger amount of relevant molecular similarity information, the ten percent most similar training set compounds (in this study the top-ten similar compounds) were identified for a test compound, and their binding measures were averaged as the predicted HSA binding value of the test compound. This approach is abbreviated as SA

$$\log k_{\text{predicted}} = \frac{\sum_{j=1}^{n} \log k_{j\text{th-similar}}}{n}, \quad (3)$$

where n is the number of top-ten similar compounds, whereas $k_{j\text{th-similar}}$ is the activity of the jth most similar compound in the training set.

(c) Based on the SA method, it was next assumed that the difference in molecular similarity of a training set compound to a test compound should reflect the difference in their binding to HSA and be represented by a corresponding weighting coefficient for similarity in HSA binding activity. In order to explore the significance of this assumption, the 10% most similar training set compounds were used again along with a weighting coefficient for each training set compound.

$$\log k_{\text{predicted}} = \sum_{j=1}^{n} w_j \log k_{j\text{th-similar}}. \quad (4)$$

In this approach, two different ways were used to assign the weighting coefficient. In the first assignment, called SR, the weighting coefficient is the ratio of the molecular similarity value to the sum of the 10% most similar training set compounds' similarity values

$$w_j = \frac{s_j(t)}{\sum_{i=1}^{n} s_i(t)}, \quad (5)$$

where $s_i(t)$ defines the similarity value of the training set compound j to the test compound. In this assignment, the greater the similarity of the training set compound to the test compound, the larger the weighting coefficient.

Another way to define the weights is based on the cluster analysis of the training set compounds, which will be called SC. Each cluster has the same contribution to the test compound, but within each cluster the greater the similarity to the test compound of the members of the cluster, the greater the contribution to w_j

$$w_j = \frac{w'_j}{n}, \quad (6)$$

where w'_j is the weighting coefficient of the compound in a cluster and n is the number of clusters.

Using these schemes, the goal is to find a 'best' way to accurately predict the HSA binding potency of each test set compound. Each of these schemes, using the 4D-fingerprints, has been applied to predict HSA binding.

4.5. Evaluation of predictivity

The overall predictive quality of each of the schemes defined above was determined by applying the validated correlation factor, r^2_{pre}. For perfect prediction of the data, r^2_{pre} has a value of 1, and for predictions which are not better than random it has a value of 0 or even a negative value. The formula of r^2_{pre} is

$$r^2_{\text{pre}} = 1 - \frac{\sum_{i=1}^{n} (y_{i,\text{observed}} - y_{\text{predicted}})^2}{\sum_{i=1}^{n} (y_{i,\text{observed}} - \bar{y}_{i,\text{observed}})^2}. \quad (7)$$

References and notes

1. Barnard, J. M.; Downs, G. M. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 644–649.
2. Torres, P. J. *All about Albumin: Biochemistry Genetics and Medical Applications*; Academic Press: San Diego, 1996.
3. Sudlow, G.; Birkett, D. J.; Wade, D. N. *Mol. Pharm.* **1975**, *11*, 824–832.
4. Sudlow, G.; Birkett, D. J.; Wade, D. N. *Mol. pharm* **1976**, *12*, 1052–1061.
5. Watanbe, S.; Tanase, K. N.; Maruyama, T.; Kragh-Hansen, U.; Otagiri, M. *Biochem. J.* **2000**, *349*, 813–819.
6. Bhattacharya, A. A.; Curry, S.; Franks, N. P. *J. Biol. Chem.* **2000**, *275*, 38731–38738.
7. Petitpas, I.; Bhattacharya, A. A.; Twine, S.; East, M.; Curry, S. *J. Biol. Chem.* **2001**, *276*, 22804–22809.
8. Morris, J. J.; Bruneau, P. P. Prediction of Physicochemical Properties. In *Virtual Screening For Bioactive Molecules*; Böhm, H.-J., Schneider, G., Eds.; Wiley-VCH: Weinheim, 2000; pp 33–56.
9. Kratochwil, N. A.; Huber, W.; Müller, F.; Kansy, M.; Gerber, P. R. *Biochem. Pharmacol.* **2002**, *64*, 1355–1374.
10. Lesk, A. M. *Proteins: Structure, Function, Genetics* **1998**, *33*, 320–328.
11. Kubinyi, H.; Hamprecht, F. A.; Mietzner, T. *J. Med. Chem.* **1998**, *41*, 2553–2564.
12. So, S. S.; Karplus, M. *J. Med. Chem.* **1997**, *40*, 4347–4359.
13. So, S. S.; Karplus, M. *J. Med. Chem.* **1997**, *40*, 4360–4371.
14. Good, A. C.; So, S. S.; Richards, W. G. *J. Med. Chem.* **1993**, *36*, 433–438.
15. Cruz, R.; Lopez, N.; Quintero, M.; Rojas, G. *J. Math. Chem.* **1997**, *20*, 385–394.
16. Kubinyi, H. A General View on Similarity and QSAR Studies. In *Computer-Assisted Lead Finding and Optimization: Current Tools for Medicinal Chemistry*; van de Waterbeemd, H., Testa, B., Folkers, G., Eds.; VHCA, Wiley-VCH: Basel, Weinheim, 1997; pp 7–28.
17. Langer, T. *Perspect. Drug Discov. Des.* **1998**, *12*, 215–231.
18. Duca, J. S.; Hopfinger, A. J. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1367–1387.
19. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.

20. Seri-Levy, A.; Richards, W. G. *Tetrahedron: Asymmetry* **1993**, *4*, 1917–1921.
21. Seri-Levy, A.; West, S.; Richards, W. G. *J. Med. Chem.* **1994**, *37*, 1727–1732.
22. Burt, C.; Huxley, P.; Richard, W. G. *J. Comput. Chem.* **1990**, *11*, 1139–1146.
23. Hall, L. M.; Hall, L. H.; Kier, L. B. *J. Comput. Aided Mol. Des.* **2003**, *17*, 103–118.
24. Hall, L. M.; Hall, L. H.; Kier, L. B. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2120–2128.
25. Goodman, G. A. *The Pharmacological Basis of Therapeutics*, 9th ed.; McGraw-Hill: New York, 1996.
26. Dollery, C. *Therapeutic Drugs*, 2nd ed.; Churchill Livingstone: Edinburgh, 1999.
27. Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. *J. Am. Chem. Soc.* **1997**, *119*, 10509–10524.
28. Liu, J.; Pan, D.; Tseng, Y.; Hopfinger, A. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2170–2179.
29. Pan, Dahua; Liu, Jianzhong; senese, Craig; Tseng, Yufeng; Hopfinger, AJ *J. Med. Chem.* **2004**, *47*, 3075–3088.
30. Kier, L. B.; Hall, L. H. *J. Pharm. Sci.* **1981**, *70*, 583.
31. SAS, Version 8.1 for Windows, SAS Institute, 2001.