A Novel Structure-Based Multimode QSAR Method Affords Predictive Models for Phosphodiesterase Inhibitors

Xialan Dong,[†] Jerry O. Ebalunode,[†] Sung Jin Cho,[‡] and Weifan Zheng*[,†]

Department of Pharmaceutical Sciences, BRITE Institute, North Carolina Central University, 1801 Fayetteville Street, Durham, North Carolina 27707 and CHDI Foundation, Inc., 6080 Center Drive, Suite 100, Los Angeles, California 90045

Received August 1, 2009

Quantitative structure−activity relationship (QSAR) methods aim to build quantitatively predictive models for the discovery of new molecules. It has been widely used in medicinal chemistry for drug discovery. Many QSAR techniques have been developed since Hansch's seminal work, and more are still being developed. Motivated by Hopfinger's receptor-dependent QSAR (RD-QSAR) formalism and the Lukacova−Balaz scheme to treat multimode issues, we have initiated studies that focus on a structure-based multimode QSAR (SBMM QSAR) method, where the structure of the target protein is used in characterizing the ligand, and the multimode issue of ligand binding is systematically treated with a modified Lukacova−Balaz scheme. All ligand molecules are first docked to the target binding pocket to obtain a set of aligned ligand poses. A structure-based pharmacophore concept is adopted to characterize the binding pocket. Specifically, we represent the binding pocket as a geometric grid labeled by pharmacophoric features. Each pose of the ligand is also represented as a labeled grid, where each grid point is labeled according to the atom types of nearby ligand atoms. These labeled grids or three-dimensional (3D) maps (both the receptor map (R-map) and the ligand map (L-map)) are compared to each other to derive descriptors for each pose of the ligand, resulting in a multimode structure−activity relationship (SAR) table. Iterative partial least-squares (PLS) is employed to build the QSAR models. When we applied this method to analyze PDE-4 inhibitors, predictive models have been developed, obtaining models with excellent training correlation (r^2 = 0.65−0.66), as well as test correlation (R^2 = 0.64−0.65). A comparative analysis with 4 other QSAR techniques demonstrates that this new method affords better models, in terms of the prediction power for the test set.

INTRODUCTION

Since its introduction by Hansch et al. in 1963, the quantitative structure−activity relationship (QSAR) method[1] has evolved from the initial form of a linear free-energy relationship to a wide variety of methods based on graph theoretic indices[2] and comparative molecular field analysis (CoMFA).[3] In the early phase of development, molecular features have been described by various types of electronic, steric, and hydrophobic descriptors.[4] Later, both two-dimensional (2D) topological and three-dimensional (3D) field descriptors have been developed. The 2D descriptors are calculated from molecular graphs as graph theoretic indices.[2,5−7] Although these descriptors represent different features of molecular structures, their physicochemical meaning is often unclear. The 3D descriptors were then developed to address some of the problems of the 2D techniques (e.g., their inability to distinguish stereoisomers). These 3D methods were exemplified by Hopfinger's work on molecular shape analysis (MSA)[8] and Crippen's work on the Voronoi method, respectively.[9] Perhaps, the best-known 3D method is CoMFA, which was developed by Cramer et al.[3] and has been widely used in medicinal chemistry projects.

The above methods are, by nature, ligand-based, where the descriptors used in the QSAR analyses are derived from the molecular structures of the ligands only. No structural information of the biological target is included in calculating the molecular descriptors. Although these descriptors are universally calculable for any organic molecule, the relevance of the calculated descriptors to a particular target is not directly encoded into the descriptors; rather, it is often revealed via statistical and machine learning analysis, often coupled with different variable selection methods. The optimal selection of variables is achieved by combining stochastic search techniques with different regression methods such as multiple linear regression (MLR), partial least-squares (PLS), and artificial neural networks (ANNs).[10−15]

One of the main challenges—and, in our view, opportunities—is how to utilize the increasing amount of X-ray structural information in constructing QSAR models. For example, the X-ray structures of important drug targets of the major gene families (NHR, kinase, protease, and PDE) have been known. In 2009, the total number of 3D structures of proteins reached >55 000 in the Protein Data Bank (www.rcsb.org/pdb/home/home.do). However, there is a lack of systematic approaches to utilize the abundant 3D structural information in QSAR modeling. Some analyses were at-

* To whom all correspondence should be addressed. Tel.: 919 530 6752. E-mail: wzheng@nccu.edu.
[†] BRITE Institute, NC.
[‡] CHDI Foundation, Inc. CA.

10.1021/ci900283j
Published on Web 01/22/2010

tempted, mostly in *ad hoc* fashions. For example, Oprea and Waller analyzed HIV protease inhibitors using CoMFA analysis, where the target structure was used in creating the alignment rules for CoMFA analysis.[16] Cho and Tropsha analyzed a set of AChE inhibitors where the protein structure of AChE was used to calculate the field properties for CoMFA analysis.[17] Martin et al. recently published two papers[18,19] that combined structure-based pharmacophore descriptors and PLS analysis to generate target-specific docking scoring functions. However, their methods were not designed for QSAR modeling *per se*. To our best knowledge, the first systematic development of a protein structure-based QSAR method was performed by Hopfinger's group in a series of studies,[20−22] where they described their methods as "receptor-dependent" QSAR (RD-QSAR). They use MD (molecular dynamics) simulations to generate multiple conformations and binding modes, followed by spatial occupancy analysis, leading to descriptors such as Grid Cell Occupancy Descriptor (GCOD)[23] and other descriptors. Even though multiple conformations were considered in Hopfinger's RD-QSAR approach, the effect of multiple binding modes was implicitly encoded in the GCOD-like descriptors, as opposed to being taken into account in an explicit fashion.

We have been systematically developing structure-based QSAR methods to address the issue of how to best encode the structural information of the binding site into the QSAR models. In a recent study, we have demonstrated that the predictiveness of a structure-based QSAR (SB-PPK) model was superior to other models that had been developed using more-traditional, ligand-based QSAR techniques.[24] This may be due to the fact that the SB-PPK descriptors are generated based on how the inhibitors match the pharmacophore features of the target binding site, and, thus, they are target-specific, whereas traditional QSAR methods are ligand-based where no target information is used to calculate the descriptors. Hence, it seems that target-specific descriptors can afford more-predictive models than universal ligand-based descriptors. However, issues that were not addressed in our previous work are conformational flexibility and multiple potential binding modes of ligand molecules. Instead, it allowed only one conformation/mode per inhibitor, similar to most current 3D QSAR methods. In our current work, we intend to include multiple conformations and multiple poses of each inhibitor in the QSAR analysis, which should account for unexplained variances in the biological activity by other single-mode QSAR approaches. To solve the multiconformation, multipose problem, a robust mathematical scheme is greatly needed.

Recently, Lukacova and Balaz systematically addressed the multimode QSAR topic in the context of ligand-based CoMFA.[25] They reported the thermodynamic principle behind the multimode methodology, as well as a mathematical formalism that solves otherwise nonlinear equations. This scheme is the first attempt to handle the problem of multiple-mode issue of 3D-QSAR in an explicit fashion. An iterative PLS procedure has been proposed to solve the resultant linearized equations. They have implemented their multimode QSAR formalism in the ligand-based CoMFA technique.[25]

Thus, we have initiated studies that focus on a protein structure-based multimode QSAR (SBMM QSAR) method, in which the structure of the target protein is used to characterize a small molecule ligand, and the issues of multiple conformations and multiple poses of the ligand are systematically treated with a slightly modified Lukacova−Balaz scheme. Because of our previous success in using structure-based pharmacophore key descriptors,[24] we have adopted the structure-based pharmacophore concept[26] to characterize the structural features of both the binding pocket of the target protein and the small molecule ligand. Specifically, we represent the binding pocket as a geometric grid that is labeled by various types of pharmacophoric features. Small molecule ligands are also represented as labeled grids, where each grid point is labeled according to the atom types of the nearby ligand atoms.

In the following sections, we describe the SBMM QSAR methodology, followed by the results of a SBMM QSAR analysis of a set of PDE4 (phosphodiesterase 4) inhibitors. These inhibitors have potentials in treating neurodegenerative diseases and cognitive disorders. Preclinical studies indicate that PDE-4 inhibitors can counteract deficits in long-term memory caused by pharmacological agents, aging, or over-expression of mutant forms of human amyloid precursor proteins.[27] Several QSAR models have been published based on the same set of inhibitors used in this study, and a comparative analysis has been made in this work. Based on the prediction statistics for both the training set and the test set, our new models are more robust and predictive than those obtained by traditional ligand-based QSAR techniques, as well as that obtained with the SBPPK method reported in our previous work.[24] All the QSAR models developed in this work have been rigorously validated, following a previously published QSAR workflow.[24,28,29]

MATERIALS AND METHODS

Overall Workflow of SBMM QSAR. The overall workflow of the SBMM (structure-based multiple mode) QSAR scheme is shown in Figure 1. It involves several major steps: (1) preparation of the multimode dataset, where binding poses for each ligand in the dataset are generated by structure-based docking of the ligand to the biological target; (2) generation of grid representation of both the binding pocket and each ligand pose; (3) calculation of structure-based descriptors for each docking pose of the ligand; (4) deriving SBMM QSAR models using an iterative PLS (partial least-squares) procedure; and (5) model validation, where each model is tested using a test set. We detail each of the major steps as follows.

Generation of Multiple Binding Poses for Each Ligand. A dataset that consists of 35 PDE-4 inhibitors[30] was used in this analysis. The molecular identifications (IDs) and the biological activities against PDE-4 are given in Table 1. The structures are given as SMILES in Table S1 of the Supporting Information. The compounds contain an indole moiety, which replaces the "rolipram-like" 3-methoxy-4-cyclopentoxy motif. The *in vivo* activities were determined from measurement of serum TNF-α levels in LPS challenged mice.[31] All the molecules were built using the builder module of the MOE package (Chemical Computing Group, Montreal, Canada). OMEGA Version 2.0 (OMEGA, OE Scientific, NM, USA) was employed to generate multiple conformers for each of the ligand molecules (MAX_CONF is set to be 2000). The 3D conformational database of the PDE4 inhibitors was used in docking to the binding pocket of a

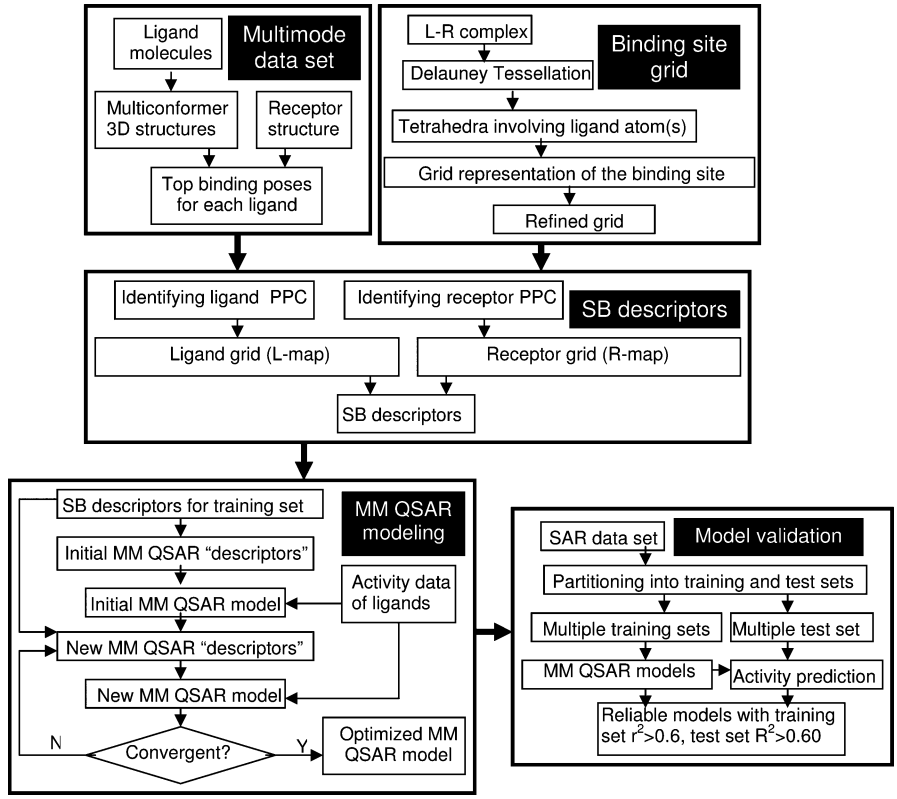

Figure 1. Overall workflow of the SBMM QSAR method.

PDE4 structure using the FRED program (FRED version 2.2.4, OE Scientific). The X-ray structure of the PED4-inhibitor complex (Protein Data Bank accession code 1xon) was used to create the receptor file for FRED docking. The top poses of each ligand ranked by FRED/Chemgauss3 scoring function were selected as the binding poses for that ligand (MAX_POSE is set to be 100). The choice of the Chemgauss scoring function is because we are in need of a scoring function that can generate high-quality docking poses, as opposed to rank different ligands in terms of their binding affinities. After each ligand was docked to the 3D protein structure (1xon), all the ligand poses were aligned inside the binding pocket in a unified coordinate frame.

Grid Representation of the Binding Pocket. Based on the X-ray structure of the receptor−ligand complex (1xon), Delauney Tessellation (DT) implemented in the QHULL program,[32] which was employed to capture the neighboring relationship among the atoms of the bound ligand and the binding site of the protein. Figure 2 schematically shows this process in two dimensions, where the ligand atoms (in open circles) and the binding pocket atoms (in filled circles) form Delauney simplices (Figure 2a, note that the simplices are triangles in 2D space, and tetrahedra in 3D space). Here, the space occupied by the Delauney simplices that contain at least one ligand atom represents the space of the binding pocket (Figure 2b). Once the binding pocket space is identified, as described previously, we then construct a regular geometric grid to approximate the entire binding pocket space (Figure 2c). A geometric grid of arbitrary spacing can be built resulting in an approximation of the space at different resolutions. Based on our experience, we have adopted a grid spacing of 0.8 Å to achieve the optimal

Table 1. Molecular Identifications (IDs) and Biological Activity

Mol_id	pIC_{50} ($-\log M$)
mol1	7.82
mol2	7
mol3	8.49
mol4	7.6
mol5	7.74
mol6	8.28
mol7	7.15
mol8	7.6
mol9	7.34
mol10	7.1
mol11	7.22
mol12	7.92
mol13	6.73
mol14	7.38
mol15	6.7
mol16	6.52
mol17	7.28
mol18	7.35
mol19	7.72
mol20	7.52
mol21	8.39
mol22	7.8
mol23	7.7
mol24	8.1
mol25	7.62
mol26	7.85
mol27	8.15
mol28	7.34
mol29	7.66
mol30	7.42
mol31	6.59
mol32	7.92
mol33	6.3
mol34	5.85
mol35	7.1

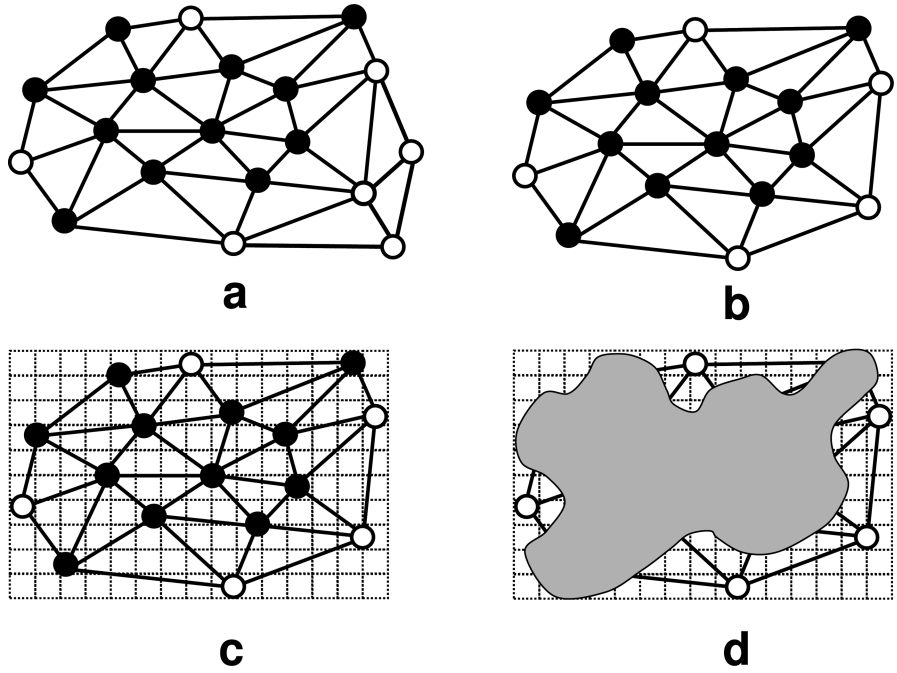

Figure 2. Identification and representation of a binding pocket: (a) Delauney tessellation of ligand atoms and binding pocket atoms (Delauney simplices are formed among both the ligand atoms and the protein atoms); (b) identification of all the Delauney simplices that have one or more ligand atoms; (c) grid approximation of the space occupied by the Delauney simplices identified in panel b; and (d) all grid points that are inside one of the above Delauney simplices are kept (an overall contour of the pocket is indicated).

result. In the final refinement step, the grid points that are too close to the protein atoms are removed, avoiding any clashes that may occur between the grid points and the protein atoms. The cutoff distance in the above refinement is set to be 2.8 Å. Figure 2d shows the contour of the resulting pocket shape, approximated by the geometric grid. Analysis of a real ligand−receptor complex would result in a 3D geometric grid that approximates the binding pocket of the receptor. This geometric grid is the common reference frame for the generation of the receptor map (R-map, representing pharmacophoric characteristics of the binding pocket), and the ligand map of a given ligand pose (L-map) (representing the pharmacophoric features of a docked ligand).

Generation of L-maps for the Ligands. Ligand pharmacophoric centers (ligand PPC) are first identified for each of the ligand molecules in the dataset. Each pharmacophore center is labeled with one of the five pharmacophoric types: positively charged (P), negatively charged (N), hydrophobic (L), hydrogen bond donor (D), and hydrogen bond acceptor (A). The Patty program (OE Scientific, NM, USA) was employed to type the ligand atoms. Once the pharmacophoric centers are defined for each ligand, one can characterize each binding pose of a ligand based on how it is placed inside the receptor's binding pocket, in terms of how it occupies the geometric grid of the binding site. The grid points close to the ligand pharmacophore centers are labeled as one or more of the following types: positively charged (P), negatively charged (N), hydrophobic (L), hydrogen bond donor (D), hydrogen bond acceptor (A), and unoccupied (U). The radius of nearness is set to be 1.5 Å to a particular pharmacophore center. For example, if a grid point is near a type "P" ligand center within 1.5 Å, that grid point is labeled as "P". All other grid points are labeled in a similar fashion, and the fully labeled geometric grid characterizes how a particular ligand pose occupies the binding pocket, and is therefore called the ligand map (or L-map) for that particular ligand pose. Since multiple docking poses of a ligand are included in this study, each ligand will have as many L-maps as the number of selected docking poses. These L-maps will be used to compare with the reference R-map (see below) to generate descriptors for each pose of a ligand.

Generation of the R-map for the Binding Pocket. The grid points of the binding site geometric grid can be labeled according to their locations, with respect to binding site protein atoms (or receptor PPC). A grid point can be labeled as one or more of the following pharmacophoric types: positively charged (P), negatively charged (N), hydrophobic (L), hydrogen bond donor (D), hydrogen bond acceptor (A), other (O), and unoccupied (U). The typing of the grid points is based on the properties of nearby amino acid atoms, and the radius of nearness is defined as being in the range of 2.5−3.6 Å, depending on what centers are being defined. For example, 3.6 Å is used to define hydrophobic centers. Since each grid point is labeled with a pharmacophoric type, the entire grid defines the pharmacophoric characteristics of the binding pocket. Thus, we named this labeled grid the receptor map (or R-map), because it fully characterizes the receptor's binding pocket. The R-map will be used as the reference frame to generate the structure-based descriptors for each docking pose of a ligand.

Generation of Structure-Based Descriptors. All ligand poses can be characterized according to how they fit to the binding pocket, in terms of pharmacophoric matches. Because each pose of a ligand is characterized by its L-map, and the receptor binding pocket is characterized by its R-map, the fit between a ligand pose and the binding pocket is simply a comparison of the properties of the grid points on both the R-map and the L-map of a given docking pose. Note that both the L-map and the R-map share the same reference frame, which is the geometric grid generated in a previous step (see the Grid Represen-

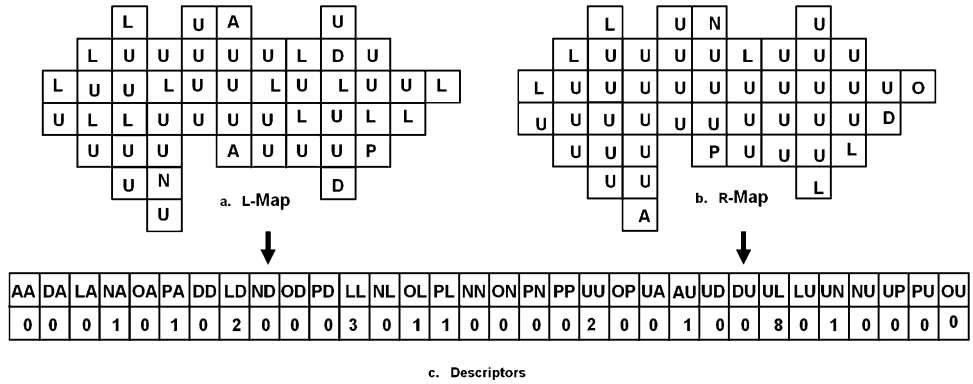

Figure 3. Generation of structure-based descriptors for each pose of a ligand based on its L-map and the binding pocket R-map: (a) labelling of all the grid points of an L-map with one of the following types: A, D, P, N, U, L and U; (b) labelling of the grid points of the receptor's R-map with one of the following types: A, D, L, P, N, O, U; and (c) generation of the structure-based descriptors for each pose of a ligand. (In this work, 32 descriptors are generated for each ligand pose.)

Table 2. Multimode QSAR Table

			Mol_1/A_1			
1	n_{111}	n_{112}	⋯	n_{11h}	⋯	n_{11H}
2	n_{121}	n_{122}	⋯	n_{12h}	⋯	n_{12H}
⋮	⋮	⋮	⋯	⋮	⋯	⋮
j	n_{1j1}	n_{1j2}		n_{1jh}		n_{1jH}
⋮	⋮	⋮		⋮		⋮
J	n_{1J1}	n_{1J2}	⋯	n_{1Jh}	⋯	n_{1JH}
			Mol_2/A_2			
1	n_{211}	n_{212}	⋯	n_{21h}	⋯	n_{21H}
2	n_{221}	n_{222}	⋯	n_{22h}	⋯	n_{22H}
⋮	⋮	⋮	⋯	⋮	⋯	⋮
j	n_{2j1}	n_{2j2}		n_{2jh}		n_{2jH}
⋮	⋮	⋮		⋮		⋮
J	n_{2J1}	n_{2J2}	⋯	n_{2Jh}	⋯	n_{2JH}
⋮	⋮	⋮		⋮		⋮
			Mol_i/A_i			
1	n_{i11}	n_{i12}	⋯	n_{i1h}	⋯	n_{i1H}
2	n_{i21}	n_{i22}	⋯	n_{i2h}	⋯	n_{i2H}
⋮	⋮	⋮	⋯	⋮	⋯	⋮
j	n_{ij1}	n_{ij2}		n_{ijh}		n_{ijH}
⋮	⋮	⋮		⋮		⋮
J	n_{iJ1}	n_{iJ2}		n_{iJh}	⋯	n_{iJH}
⋮	⋮	⋮		⋮		⋮
			Mol_I/A_I			
1	n_{I11}	n_{I12}	⋯	n_{I1h}	⋯	n_{I1H}
2	n_{I21}	n_{I22}	⋯	n_{I2h}	⋯	n_{I2H}
⋮	⋮	⋮	⋯	⋮	⋯	⋮
j	n_{Ij1}	n_{Ij2}		n_{Ijh}		n_{IjH}
⋮	⋮	⋮		⋮		⋮
J	n_{IJ1}	n_{IJ2}	⋯	n_{IJh}	⋯	n_{IJH}

tation of the Binding Pocket section). Thus, if a grid point on the L-map is of the type "P", and the corresponding grid point on the R-map is of the type "P", a "PP" descriptor is observed. Similarly, if a grid point on the L-map is of the type "N", and the corresponding grid point on the R-map is "P", then a "NP" descriptor is observed. This process of grid-point-to-grid-point matching between the L-map and the R-map is performed over the entire geometric grid. At the end of this process, the value of a descriptor (such as "PP") is the number of observed occurrences of that descriptor. Figure 3 schematically shows this process. Thus, a multimode structure−activity relationship (SAR) table is generated, as shown in Table 2, where i, j, and h are the indices for the Ith ligand, the jth pose, and the hth descriptor, respectively. I, J, and H represent the number of ligands, number of poses for each ligand, and the number of descriptors for each pose, respectively. Note that this table is different from a traditional SAR table in that the traditional SAR table would have one row per molecule, but the SAR table here has J rows per molecule (J is the number of poses per ligand), because of the multimode nature of this method.

Generation of Multimode QSAR Models. Once the multimode descriptor table is generated as described previously, QSAR models can be developed using a nontraditional PLS (partial least-squares) procedure called iterative PLS (or iPLS). This idea was first proposed in the work of Lukacova and Balaz.[25] We detail the theory in the context of our novel descriptors as follows.

The total association constant (K_i) of a ligand bound to a receptor in multiple modes can be expressed as the sum of the partial association constants K_{ij} (for $j = 1, 2, ..., J$), as in eq 1.

$$K_i = \sum_{j=1}^{J} K_{ij} \tag{1}$$

Similar to Lukacova and Balaz,[25] we hypothesized that the partial association constant (K_{ij}) for ligand i in its pose j is correlated to the structure-based descriptors n_{ijh} (for $h = 1, 2, ..., H$), through eq 2,

$$K_{ij} = \exp(c_0 + \sum_{h=1}^{H} c_h n_{ijh}) \tag{2}$$

where c_0 is the intercept and c_h ($h = 1, 2,. .. H$) are the regression coefficients. In this work, H remains constant ($H = 32$).

According to eqs 1 and 2, we have the following nonlinear equation, relating the value of K_i of a ligand with its structure-based descriptors n_{ijh}.

$$K_i = \sum_{j=1}^{J} K_{ij} = \sum_{j=1}^{J} \exp(c_0 + \sum_{h=1}^{H} c_h n_{ijh}) \tag{3}$$

The task of developing a QSAR model is to solve the above equation for the coefficients (c_0 to c_h), given a set of known K_i data and the corresponding descriptors, n_{ijh}, for each pose of the ligand. According to Lukacova and Balaz, the aforementioned equation can be transformed into a linear format, as shown in eq 4. The derivation of this equation is

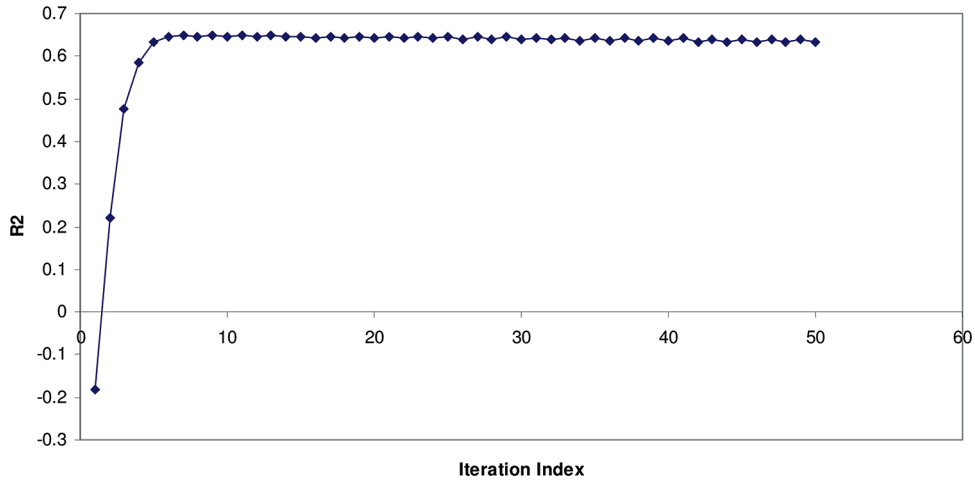

Figure 4. Convergence of the iterative partial least-squares (iPLS) process to solve eq 4.

detailed in Appendix S2 in the Supporting Information. We note that this equation is slightly different from that reported by Lukacova and Balaz in that we put the summation of K_{ij} as the denominators on both sides of eq 4.

$$\frac{K_i}{\sum_{j=1}^{J} K_{ij}} - 1 + \frac{\sum_{j=1}^{J} K_{ij} \ln(K_{ij})}{\sum_{j=1}^{J} K_{ij}} = c_0 + \frac{\sum_{h=1}^{H} c_h \sum_{j=1}^{J} K_{ij} n_{ijh}}{\sum_{j=1}^{J} K_{ij}} \tag{4}$$

This is a better fit to the iPLS procedure.

To solve the above equation for the coefficients (i.e., building a QSAR model), an iPLS procedure is employed. The left-hand side of eq 4 is the new dependent variable for the PLS analysis; and the independent variables for the iPLS procedure are not the original descriptors, n_{ijh}; rather, they are functions of partial association constants (K_{ij}) and the original descriptors (n_{ijh}), as shown in eq 4. In fact, the PLS variables also are dependent on the regression coefficients c_h (where $h = 1, 2, ..., H$) in each iteration. Once the regression coefficients are set in the first iteration, new K_{ij} values are calculated from eq 2, and then used to update the variables of eq 4. This procedure is repeated until the resultant models converge. The initial PLS variables for eq 4 are set as the average values of descriptors over all the binding poses for each of the ligands. After the iPLS procedure converges, the final coefficients (c_0 to c_h) define a QSAR model, and eq 3 is used to predict the activity values for unknown or test set molecules from their multimode descriptors (n_{ijh}).

Validation of the SBMM QSAR Models. Model validation is a critical step in a QSAR study. As shown in Figure 1, we have employed the standard workflow reported first by Golbraikh and Tropsha, and then adopted by us[24,28,29] to build predictive QSAR models. Splitting of the dataset into training and test sets, as well as obtaining the statistics for both the training set and test set, is critical in this model validation protocol. Thus, a given data set is split into multiple pairs of training and test sets by a clustering procedure based on the Adaptive Resonance Theory (ART), specifically the ART-2a algorithm.[33] The theory was developed by Grossberg and Carpenter on aspects of how the brain processes information. It is a category of neural network models. ART 2-A is an algorithm for rapid category learning and recognition, and it was used here as the clustering method for obtaining training and test sets in a rationale fashion to ensure that the training set is well-designed to cover molecules in the test set. A vigilance parameter governs how many potential clusters will be generated. The more stringent it is, the more clusters are obtained. We adjust this parameter so that a proper number of clusters are obtained. Specifically, the entire set of molecules were first clustered into multiple groups. The vigilant parameter was adjusted to obtain seven (7) multimember clusters, and one molecule from each cluster was randomly selected into a test set, resulting in seven-member test sets. After molecules were selected into a test set, the remaining molecules in the dataset were put into the corresponding training set. Thirty pairs of such training and test sets were generated, with the training set having 28 molecules and the test sets having 7 molecules. For each pair of training and test sets, the iPLS procedure was used to train the models on the training set to build a QSAR model, which was then used to predict the test set molecules. Model quality was calculated as r^2 for the training set and R^2 for the test set. While the former (r^2) reflects the model quality as measured on the training set, the latter (R^2) is a more-robust indicator of the predictive power of the model. In the end, only those model(s) with both indicators (r^2 and R^2) greater than a preset threshold (0.60 in this work) were retained as the final models for future use. Although the exact value of the threshold is somewhat arbitrary, the choice of 0.60 was based on most QSAR literature. This issue has also been addressed extensively by Golbraikh et al.[28,29] They have explored various ways of validating QSAR models, and a standard QSAR workflow has been proposed, which has been employed in this work. Thus, the use of 0.60, although an arbitrary threshold, does allow us to identify good-quality QSAR models that are comparable to the QSAR literature.

RESULTS AND DISCUSSION

Convergence of the Iterative PLS Procedure. Figure 4 shows that the iterative PLS (iPLS) procedure converges fairly fast. In this experiment, all 35 molecules were used for analysis with the iPLS procedure. The squared correlation coefficient (r^2) calculated from the actual and predicted activities of the 35 molecules reached 0.65 within just five

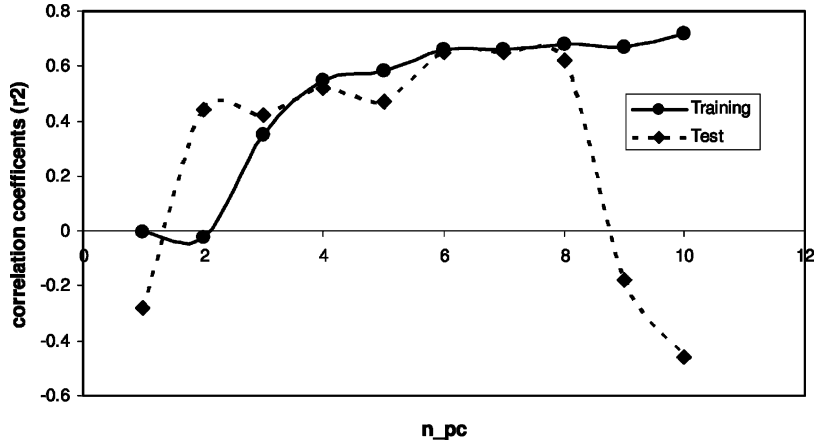

Figure 5. Effect of the number of principal components (n_PC) on model quality measured in terms of the training set (r^2) and ther test set (R^2).

iterations. This indicates that stable solutions of the coefficients (c_0, ..., c_h) for eq 2 have been obtained with relatively low computational costs. The same convergence behavior was also observed when this procedure was employed to analyze multiple training sets as a result of the dataset splitting. Thus, this procedure for solving the multimode problem is extremely efficient, compared to other methods, which are explained as follows.

The challenge in treating the multimode issue using traditional 3D QSAR techniques is the enormous combinatorial space that must be searched.[18] Because the true binding pose is often not at the very top of the list of docking poses, one must consider more candidate poses for each ligand. In a hypothetical case, if top 10 docking poses are used for each ligand in a 10-molecule dataset, one would need to explore 10^{10} combinations of 3D alignments for exhaustive 3D QSAR analyses to determine the best models. This is computationally prohibitive. The real case scenarios are even worse than this hypothetical example. Most methods try to overcome this problem by exploring the combinatorial space with a stochastic search technique, such as the genetic algorithms (GAs), to minimize the search effort. Martin et al.[18] highlighted the inefficiencies of different such approaches and developed an iterative technique to gradually select the right poses for building target-specific docking scoring functions. Although their method is not for QSAR modeling *per se*, the methodology is worth noting. However, even their method does not treat multimode issues in an explicit manner. At the end, only one pose per ligand was allowed for building the models. Our method, following the Lukacova−Balaz scheme, offers fast convergence (as demonstrated in Figure 4) and handles multiple poses per ligand explicitly, based on thermodynamic principles and a sound mathematical solution to the problem.

Effect of Number of Principal Components on Model Quality. We have examined the effect of number of principal components used in PLS analysis on the quality of the derived QSAR models. Figure 5 shows one example of how the correlation for both the training and test sets vary with the number of principal components used. Two different patterns have been observed. For the training set, the r^2 value increases with the number of principal components. The R^2 value for the test set increases only when the number of principal components is between 1 and 6, but it reaches a plateau when the number of principal components falls between 6 and 8, and then it degenerates very quickly. We chose to use six principal components for PLS analysis in the rest of this work, because it is the minimum number of principal components that affords the optimal results. This is designed to minimize overfitting during the training process.

Effect of Number of Poses Used on Model Quality. It is conceivable that the number of docking poses used for the SBMM QSAR analysis has a significant effect on the model quality. To demonstrate this, we have conducted experiments that examine the effect of the number of poses used on the correlation coefficients obtained for both the training and test sets. Figures 6a and 6b shows two of the examples. Both cases seem to indicate that 10−11 poses may afford the most balanced model quality, in terms of r^2 and R^2 for the training and test sets, respectively. When a small number of docking poses were used in building the SBMM QSAR models, the model quality was poor. This was probably because the conformational space was not sufficiently sampled, and the true (or nearly true) binding poses were not observed at the top. On the other hand, when too many poses were used, more noise might have been introduced into the process, because of the inclusion of too many irrelevant binding poses, resulting in less qualified models for the test sets; however, the model quality for training sets is less sensitive to the high number of poses used in the analysis. Thus, it is critical to examine the effect of the number of poses on the quality of SBMM QSAR models, so that an optimal number of poses can be chosen for model development in a specific case.

Because multiple poses are explicitly incorporated in our analysis, our method offers more tolerance on the requirement of having just the correct pose at the very top of a docking experiment. Methods such as those published by Cho et al.[17] are strongly dependent on the selection of a right pose for each ligand, while our method requires only several true or nearly true binding poses for each ligand. This relaxed requirement works well with state-of-the-art docking tools. It is conceivable that these models can be used in conjunction with docking tools for virtual screening of molecular databases. In such cases, traditional modeling methods (one pose per ligand) would require extremely high accuracy of docking before QSAR prediction, while our method can tolerate less-accurate docking results.

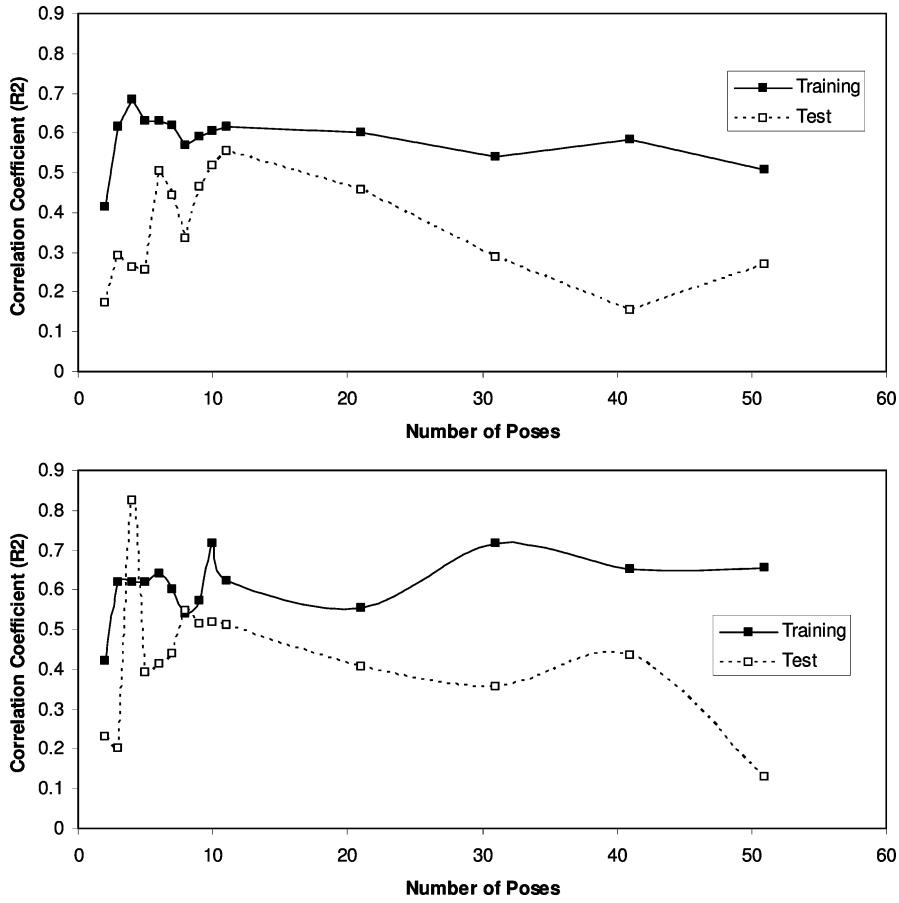

Figure 6. Effect of number of poses on model quality for two pairs of training and test sets.

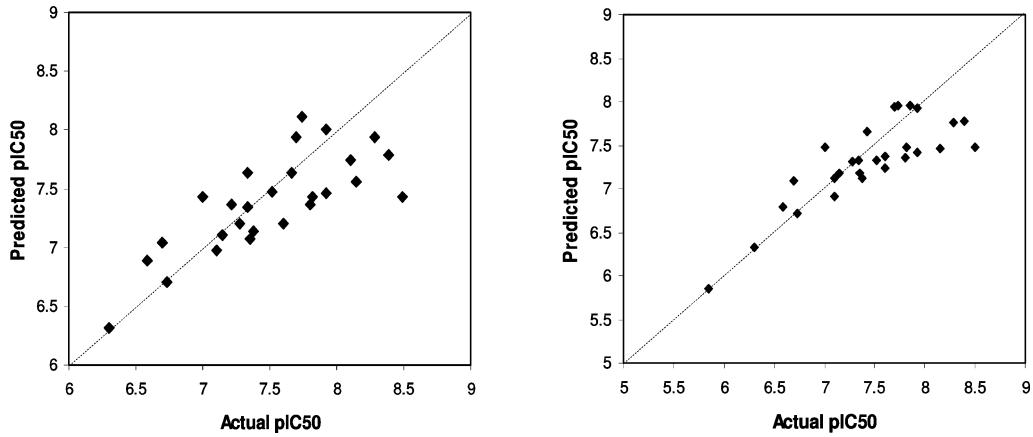

Figure 7. (Left) Correlation between predicted and actual activity for training set 1, $r^2 = 0.65$. (Right) Correlation between predicted and actual activity for training set 2, $r^2 = 0.66$.

Predictive Models Have Been Derived from Rigorous Model Validation. As advocated by Golbraikh et al.,[28,29] multiple splits of training and test sets were generated, and they were all used to build and validate the SBMM QSAR models. For each of the training sets, the iPLS procedure was employed to derive the corresponding QSAR model. Accordingly, the r^2 value was calculated for the training set, and the R^2 value for corresponding test sets was also obtained. From all the models developed, only those that passed the criteria ($r^2 > 0.60$ and $R^2 > 0.60$) were selected as the final models for future use.

To demonstrate the predictiveness of the final models, the scatter plots of actual against predicted activities are shown in Figures 7 (of training) and 8 (of testing) for two final models that satisfy our validation criteria ($r^2 > 0.60$ and $R^2 > 0.60$). Excellent correlations between the actual and predicted activities can be observed for the training and the test sets in both cases. The training set r^2values are 0.66 and 0.65 for case 1 and case 2, respectively. The test set R^2 values are 0.65 and 0.64 for case 1 and case 2, respectively. In both cases, the test set size is 7 and the training set size is 28. The number of principal components used is 6, and the number of poses used is 11. The absolute values of the prediction errors in both cases range from 0.06−0.45 log units and 0.05−0.45 log units, respectively. In relative terms, the prediction errors range from 0.6% to 5.5%. The details of the prediction errors are shown as bar graphs in Figure 9. Overall, the predictive qualities of the SBMM QSAR models

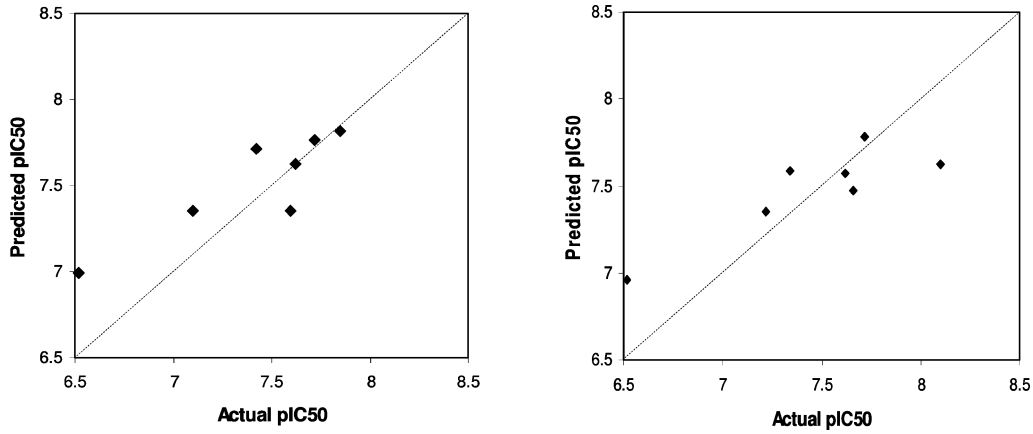

Figure 8. (Left) Correlation between predicted and actual activity for test set 1, $R^2 = 0.64$. (Right) Correlation between predicted and actual activity for test set 2, $R^2 = 0.65$.

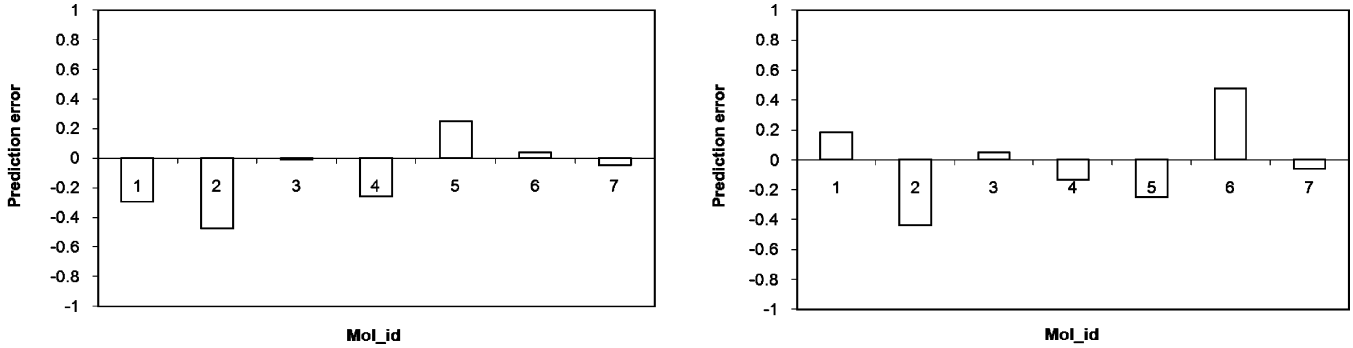

Figure 9. Prediction errors for two test sets of molecules.

are comparable or superior to the qualities of published results in terms of the R^2 values for the test sets.

SBMM Models Are More Predictive Than Ligand-Based and Single-Mode Models. To demonstrate the effectiveness of this new structure-based multimode modeling method, we have collected the modeling statistics obtained using six different QSAR techniques; these are shown in Figures 10a (for the training set) and Figure 10b (for the test set). The SBPPK and MOE-2D models refer to the best models developed in our previous work,[24] where SBPPK is a structure-based single conformational pharmacophore key model, while MOE-2D was developed based on the standard 2D descriptors in MOE and a PLS regression method. The CoMFA and CoMSIA models refer to the best models developed by Chakraborti et al.[30] The SBMM QSAR and SBSM QSAR refer to the models developed using our method in two different operating options: multipose option and single-pose option, respectively. For the single pose method, we have selected the best docking pose of each inhibitor, calculated the pharmacophore descriptors of each inhibitor in the same way as the multimode method. We then used PLS (partial least-squares) regression to build the QSAR models. The r^2 values are for the training set used by a method, and the R^2 values are for the test set used by that method.

For the training set, CoMFA and CoMSIA models reported the highest r^2 values. One may draw a conclusion that these two methods afforded the best models. However, the R^2 values for the corresponding test set are very poor. This issue has been the subject of Golbraikh et al.,[28,29] who developed their QSAR workflow to enable the development of rigorously validated models. According to them, these two models must have been a result of overtraining and underperforming. The best models should be those that afford the most

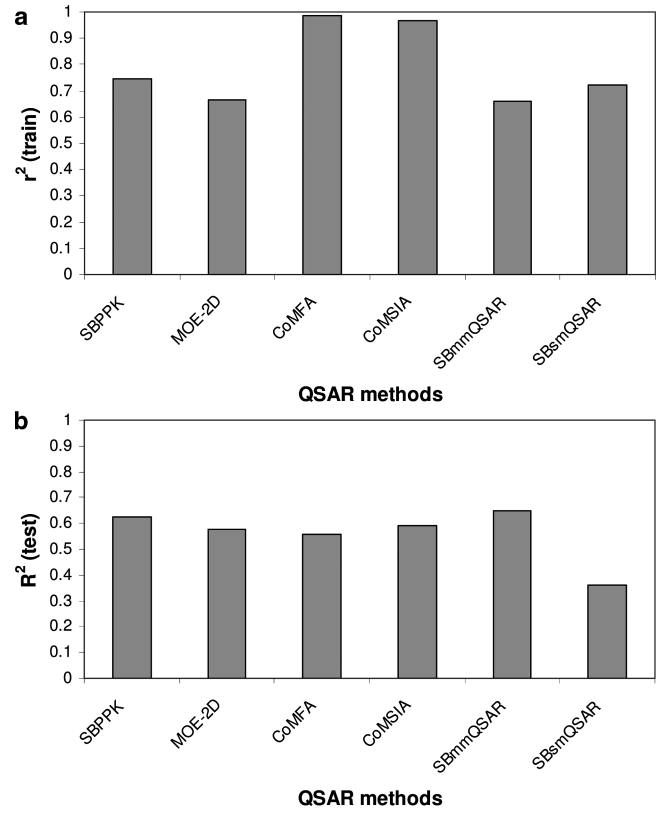

Figure 10. (a) Model performance of six different QSAR methods for the training set. (b) Model performance of six different QSAR methods for the test set.

balanced statistics. Thus, SBPPK and SBMM QSAR models are the best. It is interesting to note that SBSM model performs very well in training but fails during the test set prediction. We attribute this observation to the instability of the single-mode QSAR approach, which is typical for other single-mode QSAR techniques as well. In contrast, the structure-based multimode technique seems to be able to overcome the instability issue.

Taken together, the structure-based, multimode method (SBMM QSAR) can afford better models than ligand-based methods (MOE-2D, CoMFA, and CoMSIA), as well as the structure-based single-mode methods (SB-PPK and SBSM models).

CONCLUSIONS AND COMMENTS

We have developed a new structure-based multimode QSAR (SBMM QSAR) method that combines the concept of RD-QSAR[20,21] and the scheme for treating multimode issues by Lukacova.[25] We describe the top docking poses of each ligand as L-maps, and the pharmacophoric features of the binding pocket as an R-map. The descriptors for each docking pose of a ligand are generated based on how the L-map matches the R-map. This target-specific description of ligand molecules is critical in receptor-dependent QSAR (RD-QSAR) formalisms. An iterative PLS procedure has been developed to solve the linearized equations for the coefficients of the SBMM QSAR models. The results obtained from analyzing a set of PDE4 inhibitors have demonstrated that predictive QSAR models can be developed with this new SBMM QSAR method. Further comparative analysis indicates that this new method affords better models than models developed with five other QSAR techniques built based on the same dataset.

It is important to note that this new approach is most applicable to QSAR analysis when the receptor's three-dimensional (3D) structure is available. This is increasingly the case, because of the structural genomics effort. For example, the X-ray structures of many important drug targets of the major gene families (NHR, kinase, protease, PDE) have been solved. As mentioned earlier, the total number of 3D structures of proteins reaches over 55 000 in the Protein Data Bank. Thus, this new method should find increasingly more applications in the future for drug discovery projects where the target 3D structure is known.

One of the often asked questions about a QSAR approach is its interpretability (i.e., can one use the model to help propose new molecules). Ideally, a QSAR method should generate both predictive and interpretable models for a given problem. However, this has been very difficult for most known methods. The interpretability, although desirable, has not been the main focus of this work. The fact that multiple conformations are being used simultaneously in the model tends to complicate the interpretation of the resultant models. However, it is known that multiple conformations/poses contribute to the binding affinity of a ligand, and the overall biological activity is a result of the combined contribution from different poses and conformations. Thus, we argue that good predictability, if achieved by combining multiple conformations and poses, is worth having, even if more-complicated and less-interpretable models are generated. Moreover, our QSAR models are intended for virtual screening projects where searching large database of existing molecules is the task at hand. Thus, the main application domain of this QSAR method is in the area of virtual screening or combinatorial library design as opposed to aiding chemists to conduct De Novo design.

Future work will focus on improving the characterization of pharmacophore features for both the ligand molecules and the binding pocket. Other physicochemical properties may be included to label the grid points. New methods that focus on finding accurate docking poses must be developed to ensure that the true and/or nearly true binding poses for each ligand are ranked as high as possible on the list of docking poses. Other docking tools should be tested in the context of this SBMM QSAR approach. In fact, any docking tool (e.g., DOCK or AutoDock or other commercial packages) that can reliably generate high-quality docking poses is, in theory, applicable to this methodology. These aspects may be the subject of a future benchmark study comparing different docking methods on the quality of the resultant QSAR models.

ACKNOWLEDGMENT

We acknowledge the generous software support from the OpenEye Scientific, NM, USA. We would like to acknowledge the financial support by the Golden Leaf Foundation through the BRITE Center, North Carolina Central University. W.Z. would also like to acknowledge funding from NIH (National Institutes of Health) (1SC3GM086265).

Supporting Information Available: Table of molecular IDs, p_IC50, and SMILES representation of structures, and theoretical background (PDF). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Hansch, C.; Muir, R. M.; Fujita, T.; Maloney, P. P.; Geiger, F.; Streich, M. The Correlation of Biological Activity of Plant Growth Regulators and Chloromycetin Derivatives with Hammett Constants and Partition Coefficients. *J. Am. Chem. Soc.* **1963**, *85* (18), 2817–2824.
(2) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*. Academic Press: New York, 1976.
(3) Cramer, R. D.; 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* (18), 5959–5967.
(4) Leo, A.; Heller, S. R. *Fundamentals and Applications of Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.
(5) Hall, L. H.; Kier, L. B., The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure−Property Modeling. In *Reviews in Computational Chemistry*; Lipkowitz, K. B.; Boyd, D. B., Eds; VCH: Cambridge, U.K., 1991; Vol. 2, pp 367−422.
(6) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Research Studies Press: Chichester, England, 1986.
(7) Randic, M. Characterization of molecular branching. *J. Am. Chem. Soc.* **1975**, *97* (23), 6609–6615.
(8) Hopfinger, A. J. A QSAR investigation of dihydrofolate reductase inhibition by Baker triazines based upon molecular shape analysis. *J. Am. Chem. Soc.* **1980**, *102* (24), 7196–7206.
(9) Crippen, G. M. Distance geometry approach to rationalizing binding data. *J. Med. Chem.* **1979**, *22* (8), 988–997.
(10) Kubinyi, H. Variable Selection in Qsar Studies. I. An Evolutionary Algorithm. *Quant. Struct.−Act. Relat.* **1994**, *13* (3), 285–294.
(11) Kubinyi, H. Variable Selection in QSAR Studies. II. A Highly Efficient Combination of Systematic Search and Evolution. *Quant. Struct.−Act. Relat.* **1994**, *13*, 393–401.
(12) Luke, B. T. Evolutionary programming applied to the development of quantitative structure-activity relationships and quantitative structure−property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1279.
(13) Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure−Activity Relationships and Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34* (4), 854–866.

(14) So, S.-S.; Karplus, M. Evolutionary Optimization in Quantitative Structure−Activity Relationship: An Application of Genetic Neural Networks. *J. Med. Chem.* **1996**, *39* (7), 1521–1530.
(15) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated Descriptor Selection for Quantitative Structure−Activity Relationships Using Generalized Simulated Annealing. *J. Chem. Inf. Comput. Sci.* **1995**, *35* (1), 77–84.
(16) Oprea, T. I.; Waller, C. L.; Marshall, G. R. Three-dimensional quantitative structure-activity relationship of human immunodeficiency virus (I) protease inhibitors. 2. Predictive power using limited exploration of alternate binding modes. *J. Med. Chem.* **1994**, *37* (14), 2206–2215.
(17) Cho, S. J.; Garsia, M. L.; Bier, J.; Tropsha, A. Structure-based alignment and comparative molecular field analysis of acetylcholinesterase inhibitors. *J. Med. Chem.* **1996**, *39* (26), 5064–5071.
(18) Martin, E. J.; Sullivan, D. C. AutoShim: empirically corrected scoring functions for quantitative docking with a crystal structure and IC50 training data. *J. Chem. Inf. Model.* **2008**, *48* (4), 861–872.
(19) Martin, E. J.; Sullivan, D. C. Surrogate AutoShim: predocking into a universal ensemble kinase receptor for three dimensional activity prediction, very quickly, without a crystal structure. *J. Chem. Inf. Model.* **2008**, *48* (4), 873–881.
(20) Pan, D.; Liu, J.; Senese, C.; Hopfinger, A. J.; Tseng, Y. Characterization of a ligand−receptor binding event using receptor-dependent four-dimensional quantitative structure−activity relationship analysis. *J. Med. Chem.* **2004**, *47* (12), 3075–3088.
(21) Pan, D.; Tseng, Y.; Hopfinger, A. J. Quantitative structure-based design: formalism and application of receptor-dependent RD-4D-QSAR analysis to a set of glucose analogue inhibitors of glycogen phosphorylase. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (5), 1591–1607.
(22) Santos-Filho, O. A.; Hopfinger, A. J. Structure-Based QSAR Analysis of a Set of 4-Hydroxy-5,6-dihydropyrones as Inhibitors of HIV-1 Protease: An Application of the Receptor-Dependent (RD) 4D-QSAR Formalism. *J. Chem. Inf. Model.* **2006**, *46* (1), 345–354.
(23) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *J. Am. Chem. Soc.* **1997**, *119* (43), 10509–10524.
(24) Dong, X.; Zheng, W. A New Structure-Based QSAR Method Affords both Descriptive and Predictive Models for Phosphodiesterase-4 Inhibitors. *Curr. Chem. Genomics* **2008**, *2* (11), 29–39.
(25) Lukacova, V.; Balaz, S. Multimode ligand binding in receptor site modeling: implementation in CoMFA. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (6), 2093–2105.
(26) Wolber, G.; Langer, T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model.* **2005**, *45* (1), 160–169.
(27) Chen, R. W.; Williams, A. J.; Liao, Z.; Yao, C.; Tortella, F. C.; Dave, J. R. Broad spectrum neuroprotection profile of phosphodiesterase inhibitors as related to modulation of cell-cycle elements and caspase-3 activation. *Neurosci. Lett.* **2007**, *418* (2), 165–169.
(28) Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y. D.; Lee, K. H.; Tropsha, A. Rational selection of training and test sets for the development of validated QSAR models. *J. Comput.-Aided Mol. Des.* **2003**, *17* (2−4), 241–253.
(29) Golbraikh, A.; Tropsha, A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Mol. Divers.* **2002**, *5* (4), 231–243.
(30) Chakraborti, A. K.; Gopalakrishnan, B.; Sobhia, M. E.; Malde, A. 3D-QSAR studies of indole derivatives as phosphodiesterase IV inhibitors. *Eur. J. Med. Chem.* **2003**, *38* (11−12), 975–982.
(31) Hulme, C.; Moriarty, K.; Miller, B.; Mathew, R.; Ramanjulu, M.; Cox, P.; Souness, J.; Page, K. M.; Uhl, J.; Travis, J.; Huang, F. C.; Labaudiniere, R.; Djuric, S. W. The synthesis and biological evaluation of a novel series of indole PDE4 inhibitors I. *Bioorg. Med. Chem. Lett.* **1998**, *8* (14), 1867–1872.
(32) Barber, C. B.; Dobkin, D. P.; Huhdanpaa, H. T. The Quickhull algorithm for convex hulls. *ACM Trans. Math. Software* **1996**, *22* (4), 469–483.
(33) Carpenter, G.; Grossberg, S.; Rosen, D. Art 2-A: an adaptive resonance algorithm for rapid category learning and recognition. *Neural Networks* **1991**, *4* (4), 493–504.

CI900283J