

## Four-Dimensional Quantitative Structure–Activity Relationship Analysis of a Series of Interphenylene 7-Oxabicycloheptane Oxazole Thromboxane A<sub>2</sub> Receptor Antagonists

M. G. Albuquerque,<sup>†,§,△</sup> A. J. Hopfinger,<sup>\*,†,‡</sup> E. J. Barreiro,<sup>§</sup> and R. B. de Alencastro<sup>△</sup>

Laboratory of Molecular Modeling and Design, M/C 781, College of Pharmacy, The University of Illinois at Chicago, 833 South Wood Street, Chicago, Illinois 60612-7231, The Chem21 Group, Inc., 1780 Wilson Drive, Lake Forest, Illinois 60045, LASSBio, Depto. de Fármacos, Faculdade de Farmácia, UFRJ, Centro de Ciências da Saúde, Bloco B (sub-solo), Ilha do Fundão, Rio de Janeiro, RJ 21944-910, Brasil, and LFQO, Depto. de Química Orgânica, Instituto de Química, UFRJ, Centro de Tecnologia, Bloco A, Lab. 609, Ilha do Fundão, Rio de Janeiro, RJ 21944-000, Brasil

Received May 26, 1998

A series of 39 (a training set of 29 and a test set of 10) interphenylene 7-oxabicyclo[2.2.1]heptane oxazole thromboxane A<sub>2</sub> (TXA<sub>2</sub>) receptor antagonists were studied using four-dimensional quantitative structure–activity relationship (4D-QSAR) analysis. Two thousand conformations of each analogue were sampled to generate a conformational energy profile (CEP) from a molecular dynamic simulation (MDS) of 100 000 trajectory states. Each conformation was placed in a grid cell lattice for each of six trial alignments. Cubic grid cell sizes of 1 and 2 Å were considered. The frequency of occupation of each grid cell was computed for each of seven types of pharmacophoric group classes of atoms of each compound. These grid cell occupancy descriptors (GCODs) were then used as independent variables in constructing three-dimensional (3D)-QSAR models after data reduction. The types of data reduction included doing no reducing, reduction based on individual GCOD correlation with activity, and reduction from minimum variance constraints over the GCOD population. The 3D-QSAR models were generated and evaluated by a scheme that combines a genetic algorithm (GA) optimization with partial least squares (PLS) regression. The 3D-QSAR models were evaluated by cross-validation using the leave-one-out technique. The cross-validated correlation coefficient,  $Q^2$ , ranged from 0.27 to 0.86. The models are not from chance correlation because a scrambled data set was generated and evaluated ( $Q^2 = 0.25–0.37$ ). A composite 3D-QSAR model was constructed using the best models derived from GCODs of both 1 and 2 Å grid cell size lattices. The 3D-QSAR models provide detailed 3D pharmacophore requirements in terms of atom types and corresponding locations needed for high TXA<sub>2</sub> inhibition activity. Specific sites in space that should not be occupied by an active inhibitor are also specified. The GCOD measures for the compounds in the training set permit reference points regarding which pharmacophore sites can provide the largest boosts in inhibition activity relative to the existing analogues.

### INTRODUCTION

Thromboxane A<sub>2</sub> (TXA<sub>2</sub>) is an extremely potent, short-lived endogenous mediator that stimulates platelet activation and aggregation, as well as smooth muscle contraction.<sup>1</sup> TXA<sub>2</sub> has been implicated as a potential contributor in the pathogenesis of thrombotic, renal, and vasospastic diseases.<sup>2</sup> Preclinical studies employing stable, selective TXA<sub>2</sub> agonists and antagonists have suggested that TXA<sub>2</sub> antagonists may be useful in the treatment of these disorders.<sup>3</sup>

To exert its effects, TXA<sub>2</sub> appears to act on a specific receptor known as the TP receptor. Hirata and co-workers<sup>4</sup> have identified and characterized a (human) TP receptor as a protein of 343 amino acids with seven putative transmembrane domains, consistent with the receptor being in the G-protein-linked class. Yamamoto et al.<sup>5</sup> proposed a three-dimensional (3D) model for the TP receptor using molecular modeling techniques. However, the tertiary structure of the

receptor is not yet available from crystallographic and/or nuclear magnetic resonance (NMR) studies.

There are some reported structure–activity molecular modeling studies of TP antagonists.<sup>6–9</sup> In the work reported in this paper we have developed 3D quantitative structure–activity relationship (3D-QSAR) models for a series of interphenylene 7-oxabicyclo [2.2.1] heptane oxazoles, synthesized as described by Misra et al.,<sup>10</sup> which are highly potent, selective, and long-acting TXA<sub>2</sub> receptor antagonists. A new method called four dimensional QSAR (4D-QSAR) analysis<sup>11</sup> has been used to develop the 3D-QSAR models. The method is capable of exploring large degrees of both conformational and alignment freedoms in the search for the active conformation and binding mode, respectively, of each compound studied. As such, 4D-QSAR analysis is well suited to study flexible molecules with multiple candidate pharmacophore sites such as the TXA<sub>2</sub> receptor antagonists investigated in this study. Indeed, 4D-QSAR analysis was successfully applied to a highly flexible set of PGF<sub>2</sub>α prostaglandins expressing antinatory activity as part of the initial set of applications of this new method.<sup>11</sup>

\* Author to whom correspondence should be addressed.

† Laboratory of Molecular Modeling and Design.

‡ The Chem21 Group, Inc.

§ LASSBio.

△ LFQO.

**Table 1.** Structures and Biological Activities of the Interphenylene 7-Oxabicyclo[2.2.1]heptane Oxazoles Series<sup>a</sup>

compound	structure	p <i>K</i> <sub>d</sub>	compound	structure	p <i>K</i> <sub>d</sub>	compound	structure	p <i>K</i> <sub>d</sub>	compound	structure	p <i>K</i> <sub>d</sub>
1		8.92	2		7.72	21		8.89	22		9.52
3		10.00	4		8.49	23		8.41	24		9.70
5		7.44	6		7.35	25		9.52	26		7.66
7		8.00	8		8.40	27		7.97	28		9.00
9		9.70	10		8.24	29		10.30	30		7.74
11		8.33	12		8.32	31		7.79	32		9.07
13		7.70	14		9.10	33		6.95	34		8.80
15		7.59	16		7.72	35		7.96	36		9.15
17		8.54	18		9.22	37		9.30	38		8.54
19		8.35	20		9.15	39		7.66			

<sup>a</sup> Activity is measured as the ability to inhibit the specific binding of [<sup>3</sup>H]SQ-29548 (HSQ) to TXA<sub>2</sub> receptors in human platelet membranes and given as p*K*<sub>d</sub>, (see ref 10) compounds **1**–**29** are the training set and **30**–**39** are the test set.

In addition to developing highly significant 3D-QSAR models for a set of TXA<sub>2</sub> antagonists, a detailed exploration of the statistical procedures used to derive the models was carried out. The reasons for this exploration are (a) to better understand the properties of the 4D-QSAR descriptors, (b) to derive optimum data reduction procedures for 4D-QSAR analyses, (c) to determine how to identify the manifold of multiple high-quality 3D-QSAR models that can result from a 4D-QSAR study, and (d) to develop procedures to test whether members of the manifold of 3D-QSAR models can be combined to create a single best and/or most comprehensive 3D-QSAR model.

## METHODS

**1. Biological Data.** A series of interphenylene 7-oxabicyclo[2.2.1]heptane oxazoles were synthesized and evaluated as TXA<sub>2</sub> receptor antagonists. In addition, some aspects of the corresponding structure–activity relationship (SAR) were explored.<sup>10</sup> The biological activity of each these compounds was evaluated as the ability to inhibit the aggregation of human-platelet-rich plasma in response to the exogenous TXA<sub>2</sub> mimetic U-46619 (UIPA), and to inhibit the specific binding of [<sup>3</sup>H]SQ-29,548 to TXA<sub>2</sub> receptors in human platelet membranes. The biological activities from the binding assays were compiled from the original reference<sup>10</sup> as equilibrium affinity constants (*K*<sub>d</sub>, nM), converted to molar

**Table 2.** The Ten Operational Steps in Performing an (RI) 4D-QSAR Analysis

step number	description of the step operation
1	Generate the reference grid and initial 3D models for all compounds in the training set.
2	Select the trial set of interaction pharmacophores elements (IPEs).
3	Perform a conformational ensemble sampling of each compound to generate its conformational ensemble profile (CEP).
4	Select a trial alignment.
5	Place each conformation of each compound in the reference grid cell space according to the alignment, and record the grid cell occupancy profile (GCOP) for each IPE and choice in occupancy measure. The resulting composite set of grid cell properties constitutes the set of grid cell occupancy descriptors, GCODs.
6	Perform a PLS data reduction of the entire set of GCODs against the biological activity measures.
7	Use the most highly weighted PLS GCODs, and any other user-selected descriptors, for the initial descriptor basis set in a GA analysis.
8	Return to step 4 and repeat step 4–7 unless all trial alignments have been included in the analysis.
9	Select the optimum set of 3D-QSAR models with respect to alignment and any of the methodology parameters.
10	Adopt the lowest-energy conformer state, from the set sampled for each compound, which predicts the maximum activity using the optimum 3D-QSAR model, as the “active” conformation (shape).

units and then expressed in negative logarithmic units,  $pK_d$  ( $-\log K_d$ ). The  $pK_d$  values are given in Table 1 and comprise the set of dependent variables in the 4D-QSAR analyses. The range in activity for the analogues in Table 1 is about three (from 7.35 to 10.30)  $pK_d$  units. The structures of 29 compounds (**1–29**) are also given in Table 1 and, together with the  $pK_d$  values, constitute the SAR training set. An additional 10 compounds (**30–39**) were selected as an external validation set. That is, these 10 analogues represent a test set of compounds not included in developing the 3D-QSAR models. The structures and  $pK_d$  values of the additional 10 compounds (**30–39**) are also listed in Table 1.

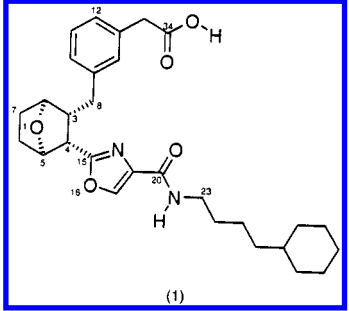
**2. 4D-QSAR Analysis.** The current methodology formulation of 4D-QSAR analysis consists of 10 operational steps that are given in Table 2. Implementation of 4D-QSAR analysis for the analogues listed in Table 1 is described later.

#### a. Model Building and Conformational Sampling.

Three-dimensional structures of each of the 39 analogues reported in Table 1 in their neutral forms were constructed using the Hyperchem 5.0 software.<sup>12</sup> Each structure was energy-minimized using the Hyperchem 5.0 MM+ force field without any restriction. Partial atomic charges were computed using the AM1<sup>13</sup> semiempirical method (MOPAC 6.0).<sup>14</sup> The Hyperchem geometry was used as the initial structure in each AM1 calculation, and this geometry was optimized for each compound as part of AM1-based partial atomic charge estimation.

The structures from the AM1 calculations were used as the initial structures in each molecular dynamics simulation (MDS)<sup>15</sup> used to construct the conformational ensemble profile (CEP) of each compound. The Molsim software (Molsim 3.0)<sup>16</sup> was used to perform the MDS and to generate the trajectories for, in turn, deriving the CEP. The MDS protocol employed 100 000 steps for each compound, the step size was 0.001 ps and the simulation temperature was 298 K. An output trajectory file was saved every 50 steps to generate a CEP consisting of 2000 conformations. The 4D-QSAR analysis does not use a single conformation in 3D-QSAR model building, but rather the intrinsic conformational flexibility of each compound is taken in account through its CEP. Others workers have also considered conformational flexibility in QSAR studies.<sup>17,18</sup>

**b. Independent Variable Generation.** The CEP consisting of the 2000 conformations generated by MDS sampling for each compound was overlayed onto a cubic lattice of a selected grid cell size. Six different 3-atom alignments were

**Table 3.** Atom Numbers and Three-Atom Sequences Defining the Six Alignments<sup>a</sup>


alignment	1 <sup>st</sup> atom	2 <sup>nd</sup> atom	3 <sup>rd</sup> atom	GCOD	$Q^2$
1	3	5	7	5245	0.73
2	3	34	12	6826	0.62
3	1	16	23	5304	0.60
4	1	8	15	4674	0.67
5	4	20	34	3607	0.55
6	3	34	20	4584	0.76

<sup>a</sup> The trial number of GCODs and the  $Q^2$  for the best model with three independent variables for each alignment are also given; compound **1** is used to define the atom numbering.

selected to define the lattice overlays. The atom numbers and corresponding sequence for each alignment are listed in Table 3. The set of alignments was chosen to span the major pathways of the bonding topology for this class of compounds.

Two sizes of grid cells were explored, 1 and 2 Å. The 2 Å grid cell size studies was used to select the best alignment, and the 1 Å grid cell size studies used the best alignment with the intention to refine the 2 Å grid cell 3D-QSAR models. The atom occupancy of each grid cell is a descriptor in 4D-QSAR analysis. These grid cell occupancy descriptors (GCODs) were computed for each of seven types of atomic groups (IPEs):

- (1) occupancy by any type of atom – *all* (0)
- (2) nonpolar atoms – *np* (1)
- (3) polar atoms of positive charge – *polar plus* (2)
- (4) polar atoms of negative charge – *polar minus* (3)
- (5) hydrogen bond acceptor – *ha* (4)
- (6) hydrogen bond donor – *hd* (5)
- (7) aromatic carbons and hydrogen – *aromatic* (6)

No reference standard compound<sup>11</sup> was used to compute the GCODs. Thus, the normalized absolute occupancy of a grid cell, defined as the number of times a cell was occupied by an atom type over the MDS divided by the size of the CEP (2000) was used to define the GCODs. This set of

GCODs constituted the complete set of independent variables used in the construction of 3D-QSAR models.

**c. Data Reduction.** Three serial levels of data reduction were considered. The first level of data reduction uses one of two filtering criteria. One filter is to eliminate GCODs for which their variance (self-variance) over the set of analogues is less than a prechosen fraction or percentage. The other filter is to eliminate GCODs that individually have less than a prechosen correlation coefficient,  $R$ , with activity. A second level of data reduction employs partial least squares (PLS) regression to identify the  $X$  most highly weighted GCODs from the population of available GCODs for a PLS model optimized as a function of the number of principal components. The  $X$  most highly weighted PLS GCODs are then used in the final data reduction step. The final step consists of constructing 3D-QSAR models using a genetic algorithm optimization. In this study, the Genetic Function Approximation (GFA)<sup>19</sup> using the Wolf 6.2<sup>20</sup> software, which is implemented with PLS regression,<sup>21,22</sup> was employed in 3D-QSAR model building and optimization. Others workers have also used a genetic algorithm (GA)/PLS combined analysis in chemometric applications.<sup>23</sup>

Three strategies of using the three types data reduction were tried. First, the entire set of GCODs (the "parent data set") was used as "input" to the second level of data reduction, the PLS variable ranking, without the elimination of GCODs either by their self-variance or individual correlation to activity. Second, a variance-filtering constraint was applied to the parent set of GCODs prior to applying the other two steps. Different self-variance cutoffs were explored. Finally, the individual GCOD correlation to activity constraint was used as an initial filter. Individual GCOD correlations against the activity data set (the independent variables) with a value of  $R < |\alpha|$  were eliminated from the independent variable set. Different value for  $\alpha$  were considered. In this initial reduction scheme, all but one of the GCODs of the same grid cell, but of different atom types, were also eliminated if their  $R$  values to one another were  $>0.95$ .

The GFA optimizations were initiated using 300 randomly generated 3D-QSAR models. Mutation probability over the crossover, optimization cycle was set at 33%, and a smoothing factor (the variable that controls the number of independent variables in the models) ranged from 0.1 to 3. The number of crossover operations was either 6000 or 12 000. The 12 000 GFA runs were developed by combining two 6000 GFA runs with smoothing factors of 0.1 and 3, with the goal of generating highly scored models with only two to six independent variables (GCODs).

**3. Independent Variable Cross-Correlation.** The cross-correlation matrix of the GCODs from the 3D-QSAR models was calculated to determine if two or more highly correlated GCODs appear in the same 3D-QSAR model. Models with highly correlated combinations of GCODs were eliminated from further consideration.

**4. Independent 3D-QSAR Models.** The correlation coefficient of the residuals (observed activity less calculated) between each pair of 3D-QSAR models were computed as a diagnostic measure<sup>11,24</sup> to select the subset of independent 3D-QSAR models from the entire set of highly scored models. The idea behind this calculation is that nearly equivalent models will have nearly the same distribution of

residuals (and an  $R$  near 1); whereas independent models will have nearly uncorrelated residuals ( $R$  near 0). The resulting set of independent, top-scored 3D-QSAR models is referred to as the *manifold* of 3D-QSAR models for the training set.

**5. Model Validation. a. Cross-Validation.** The 10 best 3D-QSAR models as scored by the lack-of-fit (LOF) measure<sup>19</sup> from the GFA analysis were evaluated by leave-one-out cross-validation. The cross-validated models were then decreasingly ranked by a combination of increasing number of independent variables used in each correlation equation, decreasing value of cross-validated  $R^2$ , and by increasing number of outliers.

**b. External Validation.** The manifold of 3D-QSAR models were individually applied to the test set of 10 compounds that were *not* included in the training set used in the 4D-QSAR analysis. The measures of fit realized in constructing the manifold of 3D-QSAR models were used as reference baselines to evaluate the measures of fit determined for the test set.

**c. Randomized Activity Data Set.** To explore the possibility of chance correlations in constructing the 3D-QSAR models, as well as to test the robustness of the models, the  $pK_d$  values for the training set were randomized ("scrambled") with respect to compound chemical structure and corresponding GCODs. An attempt was then made to construct good 3D-QSAR models using the scrambled data set. If low  $R^2$  and  $Q^2$  values are obtained for 3D-QSAR models from a scrambled data set, the probability of a chance correlation of a highly scored 3D-QSAR model from the actual training set is low.

## RESULTS

**1. CEP Convergence.** Convergence of each of the CEPs was established in terms of the behavior of the corresponding MDS trajectories. The total potential energy (PE) versus simulation time was monitored. The realization of PE fluctuations that average to a constant value over simulation time for each of the compounds was used as a measure of convergence. For each compound, 2000 conformations were recorded and used to generate the CEP of the compound. No predominant conformation was observed over the MDS trajectory of any compound in the training set. The population of conformational states, with respect to PE, fit a Boltzmann distribution for each of the compounds in the training set. The realization of Boltzmann distributions of conformational states was taken as another diagnostic of achieving CEP convergence.

**2. Alignment Evaluation for 2 Å Grid Cells.** Six 3-atom alignments were selected to determine the preferred alignment using a grid cell size of 2 Å. As can be seen in Tables 3 and 4, the best alignment is number 6, which leads to a cross-validated  $R^2$  ( $Q^2$ ) = 0.76 and a standard error (SE) = 0.40 for a 3D-QSAR model with three GCODs (independent variables). A plot of  $Q^2$  versus the number of independent variables (see Figure 1) indicates that the 3D-QSAR models with three to five independent variables have the highest values of  $Q^2$ . The parameter  $Q^2$  decreases when two or six independent variables are used. This behavior in  $Q^2$  is independent of alignment. The same type of behavior is observed for the SE of the cross-validated models (see



**Table 4.** Statistical Measures of Fit for Alignments 1 to 6 (grid cell size of 2 Å) for Each of the Best Models with 2, 3, 4, 5, and 6 GCODs

alignment	GCOD	$R^2$	$Q^2$	SE	outliers
1	2	0.61	0.48	0.58	1
	3	0.79	0.73	0.41	3
	4	0.82	0.76	0.39	4
	5	0.84	0.77	0.38	4
	6	0.84	0.38	0.61	3
2	2	0.68	0.56	0.52	2
	3	0.70	0.62	0.49	2
	4	0.79	0.69	0.44	4
	5	0.82	0.71	0.42	4
	6	0.85	0.63	0.48	2
3	2	0.62	0.45	0.59	1
	3	0.67	0.60	0.50	2
	4	0.79	0.68	0.45	2
	5	0.77	0.57	0.52	3
	6	0.79	0.56	0.53	5
4	2	0.62	0.50	0.56	1
	3	0.69	0.67	0.45	2
	4	0.77	0.71	0.43	2
	5	0.80	0.71	0.42	3
	6	0.78	0.60	0.50	2
5	2	0.51	0.27	0.68	1
	3	0.65	0.55	0.53	0
	4	0.72	0.62	0.49	2
	5	0.74	0.60	0.50	1
	6	0.75	0.58	0.52	0
6	2	0.67	0.61	0.50	1
	3	0.82	0.76	0.40	2
	4	0.87	0.81	0.35	3
	5	0.89	0.83	0.33	4
	6	0.89	0.77	0.38	4

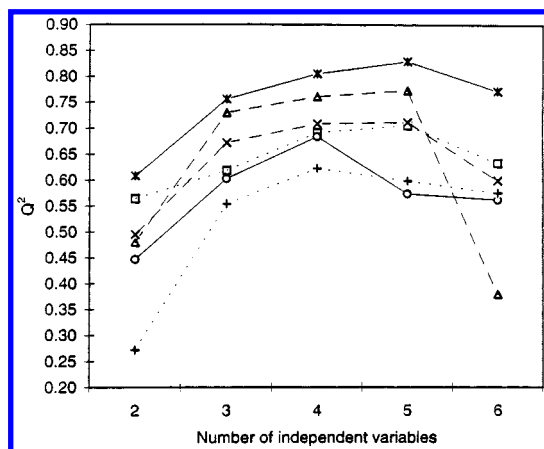
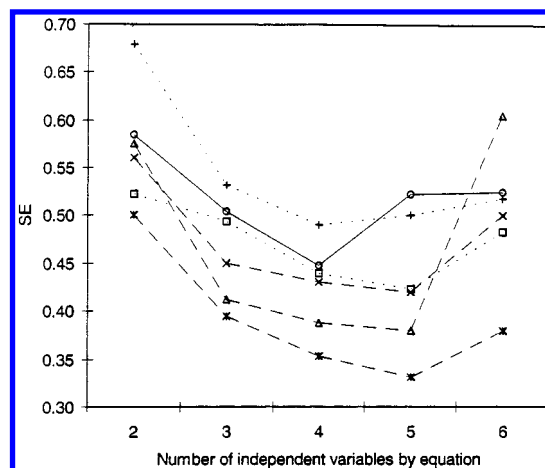
**Figure 1.** The parameter  $Q^2$  versus number of independent variables (GCODs) for alignments 1–6 using a grid cell size of 2 Å. Key to alignments: ( $\Delta$ ) 1; ( $\square$ ) 2; ( $\circ$ ) 3; ( $\times$ ) 4; (+) 5; (\*) 6.

Figure 2). The smallest SEs are from the 3D-QSAR models with three, four, or five independent variables, and SE increases using two or six independent variables. Thus, 3D-QSAR models with three, four, or five GCODs have higher  $Q^2$ , and smaller SE values than models with fewer or greater numbers of independent variables.

**3. Evaluation of Alignment 6 and a 2 Å Grid Cell Lattice as a Function of Data Reduction.** The 3D-QSAR models derived from the parent data set are characterized by no first level GCOD data reduction. These 3D-QSAR models are reported as “a” type models and a number after “a” has been added that defines the number of independent variables used in the model. Thus, the coding “a2” defines a 3D-QSAR model derived directly from the parent data set and containing two independent variables (GCODs).

**Figure 2.** Standard error (SE) from the cross-validated models versus number of independent variables (GCODs) for alignments 1–6 using a grid cell size of 2 Å. Key to alignments: ( $\Delta$ ) 1; ( $\square$ ) 2; ( $\circ$ ) 3; ( $\times$ ) 4; (+) 5; (\*) 6.**Table 5.** Statistical Descriptions of the Five “Best” Models with Two to Six Selected Independent Variables (SIV) using Alignment 6 and a Grid Cell Size of 2 Å<sup>a</sup>

IDS	FIV	SIV	$R$	$R^2$	$F$	$Q$	$Q^2$	SE	outliers
4584	71	2	0.82	0.67	26.27	0.78	0.61	0.50	1
4584	136	3	0.91	0.82	38.67	0.87	0.76	0.40	2
4584	136	4	0.93	0.87	39.90	0.90	0.81	0.35	3
4584	136	5	0.95	0.89	38.21	0.91	0.83	0.33	4
4584	129	6	0.95	0.89	31.13	0.88	0.77	0.38	4
111	69	2	0.82	0.67	26.61	0.78	0.60	0.51	1
111	70	3	0.91	0.82	38.67	0.87	0.76	0.40	2
111	85	4	0.94	0.88	42.20	0.91	0.83	0.33	2
111	85	5	0.95	0.90	42.72	0.92	0.84	0.32	3
111	74	6	0.95	0.91	35.36	0.91	0.82	0.34	3

<sup>a</sup> The results are from two input data sets (IDS): (a) the parent data set (4584 gcods) and (b) the reduced data set (111 GCODs); final independent variables (fiv) is the number of gcods remaining at the end of the GFA crossover operation.

Only one of the first level data reduction filtering schemes was used in this study involving GCODs from a 2 Å grid cell lattice. The exclusion of each independent variable with an individual  $R$  against activity of less than  $|0.34|$ , and/or all but one of the GCODs of the same grid cell, the difference in the GCODs only being atom type, with a  $R > 0.95$ , was employed in the initial data reduction. The 3D-QSAR models obtained from this initial data reduction filtering are coded as “b” type models. Using this initial data reduction filtering, the parent data set of 4584 independent variables was reduced to only 111 independent variables (see Table 5). In other words, only 2.42% of the GCODs from the parent data set survived this first level of data reduction.

An inspection of Table 5 indicates that  $Q^2$  does not significantly increase by simply increasing the number of independent variables in the best 3D-QSAR models derived from either the parent data set or the reduced data set. Rather,  $Q^2$  has about the same high value for the models with two and three independent variables as compared with the 3D-QSAR models with four, five or six independent variables. There is a decrease in the number of outliers for the 3D-QSAR models derived from the reduced data set in which the numbers of independent variables are larger than three. Models derived from the reduced data set do not exhibit a quantitative improvement in fitting after outlier

**Table 6.** The Top Five 3D-QSAR Models with Two to Six Independent Variables (GCODs) using Alignment 6 and a Grid Cell Size of 2 Å Taken from (a) The Parent Data Set and (b) The Preprocessed Data Set<sup>a</sup>

data set	number of independent variables				
	2	3	4	5	6
a	$pK_d$ (HSQ) = 8.51 + 147.97 (2,-3,1,6) - 5.40 (3,1,2,2)	$pK_d$ (HSQ) = 7.13 + 96.14 (2,-3,1,6) + 10.72 (5,-3,2,1) + 6.18 (5,-1,0,0)	$pK_d$ (HSQ) = 7.08 + 97.86 (2,-3,1,6) + 6.69 (2,-1,5,6) + 10.72 (5,-3,2,1) + 5.91 (5,-1,0,0)	$pK_d$ (HSQ) = 7.14 + 95.08 (2,-3,1,6) + 6.48 (2,-1,5,6) - 90.24 (3,-3,4,4) + 10.98 (5,-3,2,1) + 5.69 (5,-1,0,0)	$pK_d$ (HSQ) = 7.19 + 94.27 (2,-3,1,6) + 3.82 (3,1,1,1) + 4.86 (5,-1,0,1) - 131.63 (5,0,-6,0) + 5.27 (6,-2,2,1) + 221.60 (10,1,-2,1)
b	$pK_d$ (HSQ) = 8.72 + 129.61 (2,-3,1,6) - 2.71 (3,0,2,3)	$pK_d$ (HSQ) = 7.13 + 96.14 (2,-3,1,6) + 10.72 (5,-3,2,1) + 6.18 (5,-1,0,0/1)	$pK_d$ (HSQ) = 7.70 + 109.38 (2,-3,1,6) - 1.38 (3,0,2,3) + 8.66 (5,-3,2,1) + 4.86 (5,-1,0,0/1)	$pK_d$ (HSQ) = 7.69 + 97.48 (2,-3,1,6) - 1.41 (3,0,2,3) + 4.48 (5,-3,2,1) + 2.53 (5,-2,3,1) + 4.53 (5,-1,0,0/1)	$pK_d$ (HSQ) = 7.75 + 100.99 (2,-3,1,6) - 1.56 (3,0,2,3) + 4.68 (5,-3,2,1) + 1.93 (5,-2,3,1) + 4.28 (5,-1,0,0/1) + 0.72 (7,-1,2,0)

<sup>a</sup> The first three numbers inside the parenthesis defining the GCODs are the *x*, *y*, and *z* spatial coordinates. the fourth number defines the atom type occupancy (0, all types; 1, nonpolar; 2, polar plus; 3, polar minus; 4, hydrogen bond acceptor; 5, hydrogen bond donor; 6, aromatic); the symbol "/" means "or".

**Table 7.** Linear Cross-Correlation Matrix of the Residuals from the Five Top 3D-QSAR Models with Two to Six GCODs from (a) the Parent Data Set, and (b) the Preprocessed Data Set, Using a Grid Cell Size of 2 Å and Alignment 6

	a2	a3	a4	a5	a6	b2	b3	b4	b5	b6
a2	1.00									
a3	0.42	1.00								
a4	0.35	0.86	1.00							
a5	0.38	0.78	0.91	1.00						
a6	0.39	0.71	0.49	0.43	1.00					
b2	0.77	0.30	0.22	0.22	0.39	1.00				
b3	0.42	1.00	0.86	0.78	0.71	0.30	1.00			
b4	0.55	0.84	0.69	0.64	0.72	0.62	0.84	1.00		
b5	0.46	0.74	0.57	0.54	0.77	0.54	0.74	0.88	1.00	
b6	0.44	0.73	0.58	0.53	0.78	0.54	0.73	0.87	0.98	1.00

removal, but the number of remaining outliers does decrease.

Table 6 lists the top five 3D-QSAR models with two to six independent variables constructed from alignment 6 and a grid cell size of 2 Å using (a) the parent data set and (b) the reduced data set. The GCOD (2,-3,1,6) is present in all of the top 3D-QSAR models. The first three numbers in this descriptor coding are the *x*, *y*, *z* Cartesian coordinates defining the grid cell location, and the last number is the atom type (already described and defined in the heading of Table 6) whose occupancy is recorded. The individual *R* of this GCOD with activity is 0.63. The GFA/PLS 3D-QSAR model building was able to identify this important GCOD with, or without, initial data reduction.

The top 3D-QSAR models are (see Table 6) more "additive" than "diverse" with respect to increasing number of GCODs in the models. This "additive" nature can be seen from models a3, a4, and a5 and models b3, b4, b5, and b6, where one more GCOD is added to each 3D-QSAR model of increasing number of independent variables without a change in the existing set of GCODS. Table 7 reports the linear cross-correlation matrix between the residuals (observed activity less calculated activity) of the top five models (with two to six independent variables) for (a) the parent data set and (b) the reduced data set. Rogers<sup>11,24</sup> proposed this type of error analysis to establish how much one model is different from, or similar to, another model. The 3D-QSAR models generated from the parent data set and from

the reduced data set have relatively high correlations among themselves and among models of the other data set. In fact, models a3 and b3 are exactly the same.

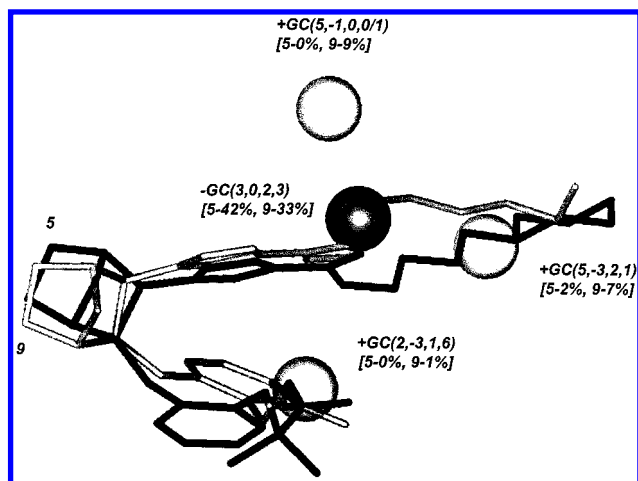
The 3D-QSAR models generated from the reduced data set have less "spurious" GCODs than those derived from the parent data set. For example, the GCOD (2, -1, 5, 6) that appears in both models a4 and a5 has an individual *R* with the activity of 0.23 and is found to involve only six compounds in the training set. GCOD (3, -3, 4, 4) that appears in model a5 has an individual *R* = -0.23 with activity, and is found to occur for only two compounds. In model a6, there are two spurious GCODs, (5, 0, -6, 0) and (10, 1, -2, 1), both with *R* < 0.11 with activity and each occurring for only two compounds. The tendency seems to be that 3D-QSAR models with small numbers of independent variables have less spurious variables than 3D-QSARs with larger numbers of GCODs.

A drawback to the 3D-QSAR models derived from the reduced data set is an increased chance of high correlation between independent variables in the same model. For example, in model b5, the spatially adjacent GCODs (5, -3, 2, 1) and (5, -2, 3, 1) have a cross correlation of 0.69, in model b6 the same pair of GCODs occur and, in addition, there is a *R* of 0.79 between the GCODs (5, -2, 3, 1) and (7, -1, 2, 0).

The tendency also seems to be that 3D-QSAR models with smaller numbers of independent variables have GCODs more statistically independent of one another. Thus, 3D-QSAR models with three, four, and no more than five independent variables are judged to be of optimum size for this particular SAR data set.

3D-QSAR model b4 is identified as the "best" model for a grid cell size of 2 Å because it is an "economical" model (four GCOD terms), has an *R*<sup>2</sup> = 0.88, and does not possess spurious, or interrelated GCODs. This 3D-QSAR model will be referred to as Model II (II for 2 Å), is shown in Figure 3, and is given as

$$pK_d = 7.70 + 109.38 (2, -3, 1, 6) - 1.38 (3, 0, 2, 3) + 8.66 (5, -3, 2, 1) + 4.86 (5, -1, 0, 0/1)$$



**Figure 3.** Graphical representation of Model II (2 Å grid cell) composed of four GCODs as defined by eq 1. The lowest-energy CEP conformations of compounds **5** (low-activity of 7.44) and **9** (high-activity of 9.70) are shown for the best alignment, which is alignment 6. Compound **9** is displayed in "gray" scale color, whereas compound **5** is shown in "black" for reference. The grid cells are represented as spheres whose radii are equal to the sides of the grid cells. GC(*x, y, z, IPE*) defines the location of the grid cell in space and its relevant IPE type. The "+" and "-" in front of the GCs define whether occupancy enhances (+) or diminishes (-) activity. In brackets, for each grid cell, are compound numbers **5** and **9** and the corresponding percent occupancy of the grid cell for the CEP of each compound. IPE refers to interaction pharmacophore element which is a generalization of atom type.

$$N = 29 \quad R^2 = 0.88 \quad SE = 0.31 \quad F = 42.2 \quad (1)$$

**4. Evaluation of Alignment 6 and a 1 Å Grid Cell Lattice as a Function of Data Reduction.** In this section the "a" models are, again, those derived from the parent data set. In this specific study, two first level data reduction filtering schemes were employed. The filtering constraint based on variance of the independent variables was first applied. The "b" models are derived from a minimum variance cutoff value of 4.5% (50.80% of the GCODs from the parent data set survive using this cutoff value, see Table 8). The "c" models are derived from a minimum variance cutoff value of 21.3% (19.95% of the GCODs survive), and the "d" models are constructed from a minimum variance cutoff value of 37.6% (9.13% remain).

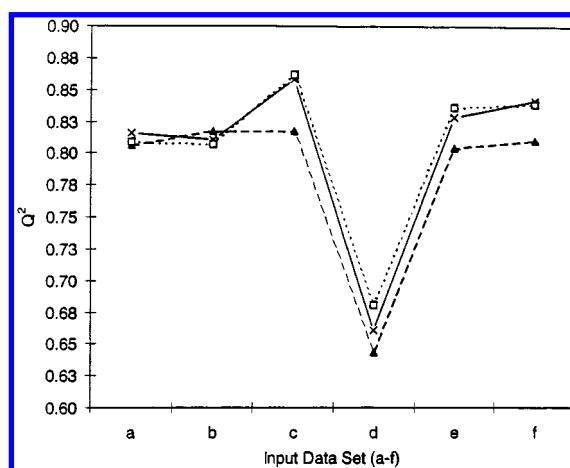
The other first level data reduction filter used in this particular study is the same as that employed for the 4D-QSAR analysis of GCODs from the 2 Å grid cell lattice; namely, a minimal acceptable level of *R* between the GCOD and activity. The "e" models are derived from this form of first level data reduction and only 3.95% of the GCODs from the parent data set remain after data reduction. The "f" models are derived from this same first level data reduction scheme with the additional constraint of excluding GCODs generated by <12 compounds. Thus, the resulting reduced data set is similar in derivation to the "b" data set for the 2 Å grid cell lattice. Only 2.17% of the GCODs from the parent data set survive this first level data reduction scheme of filtering.

The parameter  $Q^2$  does not vary in a significant way for the best 3D-QSAR models generated using the data sets from the "a", "b", "c", "e", and "f" first level of data reduction filtering schemes, see Figure 4 and Table 7. For the data set generated from the "d" filtering scheme, it appears the

**Table 8.** Statistical Descriptions of the Three "Best" Models with Three to Five Selected Independent Variables (SIV) Using Alignment 6 and a Grid Cell Size of 1 Å<sup>a</sup>

data set	IDS	FIV	SIV	$R^2$	F	$Q^2$	SE	outliers
a	25019	138	3	0.85	46.74	0.81	0.35	1
	(100%)	138	4	0.87	39.89	0.82	0.34	2
		138	5	0.88	32.35	0.81	0.35	3
b	12709	157	3	0.87	54.64	0.82	0.34	2
	(50.80%)	157	4	0.87	39.55	0.81	0.35	3
		157	5	0.87	33.19	0.81	0.35	3
c	4991	176	3	0.87	53.99	0.82	0.34	2
	(19.95%)	149	4	0.89	48.10	0.86	0.30	3
		149	5	0.92	50.33	0.86	0.30	1
d	2290	69	3	0.72	21.34	0.64	0.48	1
	(9.13%)	107	4	0.78	21.77	0.66	0.47	3
		175	5	0.83	22.45	0.68	0.45	3
e	988	68	3	0.86	50.33	0.81	0.35	2
	(3.95%)	68	4	0.88	43.56	0.83	0.33	1
		152	5	0.89	46.24	0.84	0.32	1
f	544	64	3	0.85	46.98	0.81	0.35	1
	(2.17%)	113	4	0.89	47.62	0.84	0.32	2
		108	5	0.89	44.21	0.84	0.32	3

<sup>a</sup> The results are from six different input data sets (IDS): (a) the parent data set and the b–f reduced data sets (see the text for the schemes of data reduction); final independent variable (FIV) is the number of GCODs remaining at the end of the GFA crossover operation.



**Figure 4.** The  $Q^2$  from 3D-QSAR models with three, four, and five selected independent variables (SIV) constructed from different reduced data (a–f) sets using alignment 6 and a grid cell size of 1 Å. Key: (—▲—) 3 SIV; (---×---) 4 SIV; (···□···) 5 SIV.

cutoff value was excessive. No 3D-QSAR model has a  $Q^2$  value >0.69. The cutoff for the "d" first level data reduction filtering scheme excludes important GCODs from the resultant data set. An example of an excluded, but important, GCOD is (11, -4, 5, 1) (see Table 8), which is present in all top 3D-QSAR models except those constructed from the "d" data set. This GCOD has and  $R = 0.74$  with activity.

Tables 8 (statistical fit properties) and 9 (regression equations) list each of the top three 3D-QSAR models that have three to five independent variables using alignment 6 and a grid cell lattice of 1 Å, and are derived from the "a"–"f" data sets. Again, many of the 3D-QSAR models are "additive" in the same fashion as models a3, a4, and a5 and models b4 and b5 of Table 5. Also, some of the 3D-QSARs are nearly equivalent based on the correlation of their residuals, see Table 10. Two examples of nearly equivalent 3D-QSAR models that can be seen in Table 10 are b3 and c3 ( $R = 0.99$ ) and e3 and f4 ( $R = 0.89$ ). One 3D-QSAR that is largely independent of all other models is d5 (Table



**Table 9.** The Top Three 3D-QSAR Models with 3 to 5 Independent Variables (GCODs) Using Alignment 6 and a Grid Cell Size of 1 Å from the a–f Reduced Data Sets

data set	number of independent variables		
	3	4	5
a	pK <sub>d</sub> (HSQ) = 7.68 + 20.65 (10, -2, 1, 1) + 53.32 (11, -4, 5, 1) - 41.88 (14, 1, 4, 1)	pK <sub>d</sub> (HSQ) = 7.35 + 2.70 (-1, 0, 6, 0) + 19.74 (10, -2, 1, 1) + 51.10 (11, -4, 5, 1) - 33.41 (14, 1, 4, 1)	pK <sub>d</sub> (HSQ) = 7.18 + 3.50 (-1, 0, 6, 0) + 15.75 (10, -2, 1, 1) + 3.89 (10, -1, 2, 1) + 50.77 (11, -4, 5, 1) - 33.29 (14, 1, 4, 1)
b	pK <sub>d</sub> (HSQ) = 7.57 - 41.00 (8, -5, 9, 0) + 21.47 (9, -3, 3, 0) + 58.52 (11, -4, 5, 1)	pK <sub>d</sub> (HSQ) = 7.59 - 28.82 (9, -4, 12, 0) + 29.67 (9, -3, 3, 0) + 45.63 (11, -4, 5, 1) - 15.05 (11, 1, 5, 0)	pK <sub>d</sub> (HSQ) = 7.56 + 1.20 (2, -3, 13, 0) - 22.60 (9, -4, 12, 0) + 33.85 (9, -3, 3, 0) + 44.18 (11, -4, 5, 1) - 20.23 (11, 1, 5, 0)
c	pK <sub>d</sub> (HSQ) = 7.55 - 40.64 (8, -5, 9, 1) + 21.54 (9, -3, 3, 0) + 59.09 (11, -4, 5, 1)	pK <sub>d</sub> (HSQ) = 7.39 + 36.90 (9, -3, 3, 0) - 14.96 (10, -3, 8, 1) + 54.88 (11, -4, 5, 1) - 27.26 (12, 1, 2, 1)	pK <sub>d</sub> (HSQ) = 7.47 - 13.82 (9, -4, 8, 0) + 35.78 (9, -3, 3, 0) + 49.29 (11, -4, 5, 1) - 32.67 (12, 1, 2, 1) + 17.12 (13, -1, 4, 6)
c	pK <sub>d</sub> (HSQ) = 7.26 + 13.17 (3, -5, 2, 1) + 14.16 (7, 0, 2, 1) + 28.42 (12, -2, 4, 0)	pK <sub>d</sub> (HSQ) = 7.95 - 11.67 (1, -2, 4, 1) + 18.89 (5, -4, 8, 1) + 12.05 (9, -3, 4, 1) + 19.50 (11, -2, 6, 1)	pK <sub>d</sub> (HSQ) = 6.88 + 14.61 (3, -5, 2, 6) + 8.53 (3, -1, 6, 0) + 10.20 (6, 0, 3, 1) + 14.01 (10, -1, 1, 0) + 30.11 (12, -2, 4, 0)
e	pK <sub>d</sub> (HSQ) = 6.72 + 3.63 (6, 0, 1, 0) + 34.20 (9, -3, 2, 0/1) + 46.25 (11, -4, 5, 1)	pK <sub>d</sub> (HSQ) = 7.33 - 304.26 (1, -4, 13, 0) + 6.50 (3, -5, 2, 6) + 29.54 (9, -3, 2, 0/1) + 44.73 (11, -4, 5, 1)	pK <sub>d</sub> (HSQ) = 7.28 - 4.45 (-2, 2, -1, 0/1) + 19.28 (3, -5, 2, 0/1) - 27.60 (8, -5, 8, 1) + 19.97 (10, -2, 1, 0/1) + 54.72 (11, -4, 5, 1)
f	pK <sub>d</sub> (HSQ) = 7.58 + 61.25 (9, -3, 4, 0/1) - 61.15 (9, -3, 5, 0/1) + 41.50 (11, -4, 5, 1)	pK <sub>d</sub> (HSQ) = 6.68 + 4.22 (6, 0, 1, 0) + 32.76 (9, -3, 2, 0/1) + 54.96 (11, -4, 5, 1) - 21.92 (12, 2, 7, 1)	pK <sub>d</sub> (HSQ) = 6.67 + 4.00 (6, 0, 1, 0) + 34.02 (9, -3, 2, 0/1) + 52.78 (11, -4, 5, 1) - 24.88 (13, -3, 5, 1) + 7.34 (13, -2, 5, 1)

9). Despite this unique model, all 3D-QSAR models derived from the data set based on the “d” first level of data reduction were dismissed because of their low  $Q^2$  values.

Again, spurious independent variables are found in the 3D-QSAR models derived from both the parent data set, and the 3D-QSAR models derived from the various first level data reduction data sets using the 1 Å grid cell lattice. The GCOD, (14, 1, 4, 1), which is present in the “a” models, has an individual  $R$  with activity of only 0.01, the GCOD (8, -5, 9, 0) in model b3 has  $R = 0.10$ . The GCOD (11, 1, 5, 0), that is in both models b4 and b5 has  $R = 0.07$ , GCOD (8, -5, 9, 1) in model c3 has  $R = 0.12$ , and GCOD (12, 1, 2, 1) in both models c4 and c5 has  $R = 0.09$ . Only one 3D-QSAR derived from the reduced data set generated by the filter of individual GCOD cross-correlation against activity has a spurious independent variable. In this model (e5), GCOD (1, -4, 13, 0) is found for only five compounds.

Interrelated independent variables occur in both the same 3D-QSAR model and in different models, which is not a surprise because many 3D-QSARs are highly correlated with one another. Examples of this behavior include model c5, where the spatially adjacent GCODs (10, -2, 1, 1) and (10, -1, 2, 1) have  $R = 0.67$ ; model b5, where GCODs (2, -3, 13, 0) and (9, -4, 12, 0) have  $R = 0.77$ ; model c4, where GCODs (11, -4, 5, 1) and (10, -3, 8, 1) have  $R = 0.67$ ; models c4 and c5, where GCODs (9, -3, 3, 0) and (12, 1, 2, 1) have  $R = 0.65$ ; model e5, where GCODs (11, -4, 5,

1) and (8, -5, 8, 1) have  $R = 0.81$ , as well as GCODs (3, -5, 2, 0/1) and (-2, 2, -1, 0/1) that have  $R = 0.89$ ; model f3, where adjacent in space GCODs (9, -3, 4, 0/1) and (9, -3, 5, 0/1) have  $R = 0.94$ ; model f4, where GCODs (11, -4, 5, 1) and (12, 2, 7, 1) have  $R = 0.69$ ; and model f5, where spatially adjacent GCODs (13, -3, 5, 1) and (13, -2, 5, 1) have  $R = 0.91$ . Again, the tendency seems to be that 3D-QSARs with a smaller number of independent variables have less chance of having interrelated GCODs. Almost all 3D-QSAR models with five independent variables have a pair of interrelated GCODs.

The 3D-QSAR model e3 has been chosen as the best model for the grid cell size of 1 Å because this is a “economical” model (3 GCODs), it has an  $R^2 = 0.86$ , and it does not contain spurious or interrelated variables. This model will be referred to as Model I (1 for 1 Å), and is composed of GCOD (6, 0, 1, 0), which is an all-atom type occupancy near the amide group on the  $\omega$  chain, GCOD (9, -3, 2, 0/1) which is an all-atom type, or nonpolar atom occupancy at the “middle” of the  $\omega$  chain and GCOD (11, -4, 5, 1) which is a nonpolar atom occupancy at the “end” of the  $\omega$  chain. This 3D-QSAR model is shown in Figure 5 and reported below as eq 2

$$pK_d = 6.72 + 3.63 (6, 0, 1, 0) + 34.20 (9, -3, 2, 0/1) + 46.25 (11, -4, 5, 1)$$

$$N = 29 \quad R^2 = 0.86 \quad SE = 0.33 \quad F = 50.3 \quad (2)$$

**5. Evaluation of Model I and Model II and Construction of a Composite Model, Model III.** To evaluate the two best 3D-QSARs, Models I and II, the residuals of fit in activity from each of these models have been determined. Table 11 reports the residuals of fit, and Figure 6 contains the plots of these residuals for Models I and II. Compounds **1**, **4**, **7–11**, **16**, **18**, **20**, **22**, **25**, and **29** have residual values of opposite sign for Model I and Model II. Compounds **2**, **5–6**, **12**, **14**, **17**, **23**, and **24** have larger residuals from Model I as compared with Model II, whereas compounds **3**, **13**, **15**, **19**, **21**, and **26–28** have larger residuals for Model II relative to Model I.

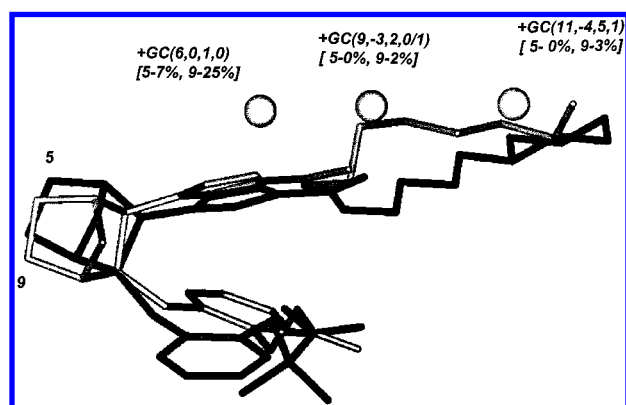
Using these observations, a composite 3D-QSAR model was built by combining noninterrelated GCODs from both Models I and II. One GCOD from Model I, (6, 0, 1, 0), with an individual correlation of  $R = 0.53$  with GCOD (2, -3, 1, 6) from Model II, was deleted in constructing the composite model. In addition, two GCODs from Model II were deleted; they are, GCOD (5, -1, 0, 0/1) with a correlation of  $R = 0.73$  with GCOD (9, -3, 2, 0/1) from Model I and (5, -3, -2, 1) with a correlation of  $R = 0.74$  with the GCOD (11, -4, 5, 1) from Model I. The remaining GCODs of Models I and II were used to construct the composite model, called Model III. The GFA regression fit using these GCODs gives Model III, which is represented by eq 3 and Figure 7.

$$pK_d = 7.84 + 26.45 (9, -3, 2, 0/1, 1) + 31.54 (11, -4, 5, 1, 1) + 79.79 (2, -3, 1, 6, 2) - 1.63 (3, 0, 2, 3, 2) \\ N = 29 \quad R^2 = 0.93 \quad SE = 0.23 \quad F = 85.5 \quad (3)$$



**Table 10.** The Linear Cross-Correlation Matrix of the Residuals of Fit from the Top Three 3D-QSAR Models with Three to Five GCODs for the a–f Reduced Data Sets for a Grid Cell Size of 1 Å and Alignment 6

	a3	a4	a5	b3	b4	b5	c3	c4	c5	d3	d4	d5	e3	e4	e5	f3	f4	f5
a3	1.00																	
a4	0.93	1.00																
a5	0.91	0.98	1.00															
b3	0.71	0.65	0.65	1.00														
b4	0.54	0.59	0.62	0.69	1.00													
b5	0.49	0.57	0.60	0.69	0.96	1.00												
c3	0.69	0.64	0.64	0.99	0.69	0.70	1.00											
c4	0.54	0.60	0.59	0.60	0.69	0.74	0.63	1.00										
c5	0.52	0.49	0.45	0.60	0.63	0.62	0.59	0.74	1.00									
d3	0.41	0.48	0.47	0.41	0.49	0.52	0.44	0.46	0.20	1.00								
d4	0.47	0.51	0.48	0.57	0.40	0.47	0.59	0.51	0.21	0.62	1.00							
d5	0.27	0.27	0.34	0.16	0.28	0.30	0.18	0.09	−0.10	0.72	0.29	1.00						
e3	0.54	0.58	0.61	0.57	0.59	0.59	0.58	0.66	0.48	0.40	0.59	0.20	1.00					
e4	0.30	0.41	0.37	0.55	0.68	0.65	0.56	0.54	0.43	0.45	0.55	0.12	0.70	1.00				
e5	0.48	0.56	0.53	0.64	0.42	0.48	0.64	0.42	0.34	0.53	0.52	0.32	0.50	0.55	1.00			
f3	0.57	0.51	0.50	0.61	0.42	0.45	0.61	0.53	0.40	0.48	0.58	0.23	0.45	0.29	0.41	1.00		
f4	0.54	0.59	0.61	0.60	0.56	0.57	0.63	0.66	0.40	0.52	0.59	0.36	0.89	0.70	0.56	0.47	1.00	
f5	0.48	0.49	0.49	0.55	0.40	0.40	0.57	0.48	0.26	0.43	0.61	0.22	0.81	0.61	0.53	0.29	0.84	1.00

**Figure 5.** Same as Figure 3, but for Model I (1 Å grid cell).

The last, that is the fifth, entry in the GCOD notation for Model III defines the grid cell size (1 or 2 Å) of the GCOD. The calculated activities and corresponding residuals in activity using Model III are also given in Table 11 for the training set. The residuals plot for Model III is also given in Figure 6. In general, the residuals are smaller for Model III than the residuals of both Models I and II.

**6. External Data Set Validation.** The 10 compounds (test set) that were not included in 3D-QSAR model construction were used to make “real” predictions using Models I–III. Table 12 lists both the observed activities and the predicted activities from Models I–III for the test set. The activity residuals for each model are also given in Table 12 and these residuals are also plotted in Figure 8. Smaller residuals are found, in general, for Model III than for Models I or II, again indicating that the composite model captures more structure–activity information than either individual model.

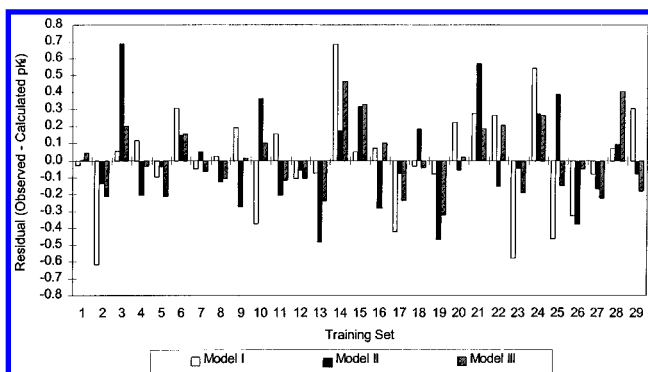
Figure 9 is a plot of the observed activity versus the predicted activity of both the training set and the test set (all 39 compounds) using Model III. To establish the outliers of Model III, the standard deviations (SD) of the residuals for both the 29 analogues of the training set, and also all 39 compounds in the total data set, were computed. Outliers are defined as compounds whose residuals are more than twice the SD of the residual of fit (see Tables 12 and 13). An inspection of Tables 11–13 and Figure 8 reveals three outliers for Model III for all 39 compounds. The outliers

**Table 11.** Residuals of Fit (Observed – Calculated  $pK_d$ ) from Models I, II, and III<sup>a</sup>

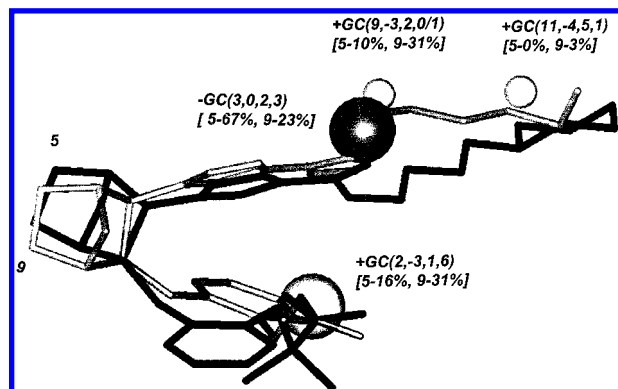
training set	model I		model II		model III	
	calculated	residual	calculated	residual	calculated	residual
1	8.95	−0.03	8.92	0.00	8.88	0.04
2	8.34	−0.62	7.86	−0.14	7.93	−0.21
3	9.94	0.06	9.31	<b>0.69</b>	9.80	0.20
4	8.37	0.12	8.70	−0.21	8.52	−0.03
5	7.54	−0.10	7.48	−0.04	7.65	−0.21
6	7.05	0.30	7.20	0.15	7.19	0.16
7	8.05	−0.05	7.95	0.05	8.07	−0.07
8	8.38	0.02	8.53	−0.13	8.51	−0.11
9	9.51	0.19	9.98	−0.28	9.69	0.01
10	8.61	−0.37	7.87	0.37	8.14	0.10
11	8.17	0.16	8.53	−0.20	8.45	−0.12
12	8.43	−0.11	8.38	−0.06	8.43	−0.11
13	7.78	−0.08	8.18	−0.48	7.94	−0.24
14	8.42	<b>0.68</b>	8.92	0.18	8.64	<b>0.46</b>
15	7.54	0.05	7.27	0.32	7.26	0.33
16	7.65	0.07	8.00	−0.28	7.62	0.10
17	8.96	−0.42	8.62	−0.08	8.78	−0.24
18	9.26	−0.04	9.03	0.19	9.27	−0.05
19	8.43	−0.08	8.82	−0.47	8.67	−0.32
20	8.93	0.22	9.21	−0.06	9.13	0.02
21	8.62	0.27	8.32	<b>0.57</b>	8.70	0.19
22	9.25	0.27	9.67	−0.15	9.31	0.21
23	8.99	−0.58	8.46	−0.05	8.60	−0.19
24	9.16	0.54	9.42	0.28	9.43	0.27
25	9.99	−0.47	9.13	0.39	9.67	−0.15
26	7.99	−0.33	8.04	−0.38	7.71	−0.05
27	8.05	−0.08	8.14	−0.17	8.19	−0.22
28	8.93	0.07	8.90	0.10	8.59	0.41
29	9.99	0.31	10.38	−0.08	10.48	−0.18

<sup>a</sup> The values in bold are outliers (greater than twice the standard deviation).

are compounds **33**, **36**, and **37**, all of which are from the test set. Compound **33** is predicted to be more active (predicted  $pK_d$  = 8.27) than is observed ( $pK_d$  = 6.95). The probable reason compound **33** is an outlier is that the training set does not contain any compounds without a carboxylic acid function in the  $\alpha$  chain (compound **33** has a hydroxyl function, see Table 1). The carboxyl carbon of the compounds in the training set and the carbon bonded to the hydroxyl in compound **33** are the second atom in alignment 6. Thus, the GCODs associated with these atoms are constant for all compounds. Compounds **36** and **37** are each predicted to be less active than is observed. No good explanation(s) for the predicted loss in activity is obvious.



**Figure 6.** The residuals of fit (observed — calculated  $pK_a$ ) plot of the training set of compounds (1–29) for Models I, II, and III.



**Figure 7.** Same as Figure 3, but for Model III (combined 1 and 2 Å grid cells).

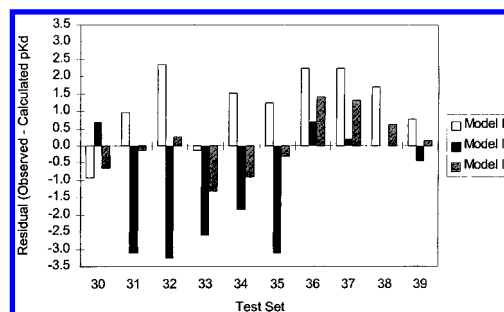
**Table 12.** Observed and Calculated  $pK_a$ , and the Corresponding Residuals of Fit (Observed — Calculated  $pK_a$ ), for the Test Set of Compounds using Models I, II, and III

test set	observed $pK_a$	model I		model II		model III	
		calculated	residual	calculated	residual	calculated	residual
30	7.74	8.66	-0.92	7.05	0.69	8.38	-0.64
31	7.79	6.84	0.95	10.91	-3.12	7.91	-0.12
32	9.07	6.72	<b>2.35</b>	12.34	-3.27	8.82	0.25
33	6.95	7.08	-0.13	9.55	-2.60	8.27	-1.32
34	8.80	7.29	1.51	10.65	-1.85	9.70	-0.90
35	7.96	6.72	1.24	11.08	-3.12	8.28	-0.32
36	9.15	6.91	<b>2.24</b>	8.46	0.69	7.74	<b>1.41</b>
37	9.30	7.06	<b>2.24</b>	9.09	0.21	7.98	<b>1.32</b>
38	8.54	6.83	<b>1.71</b>	8.54	0.00	7.92	0.62
39	7.66	6.89	0.77	8.09	-0.43	7.49	0.17

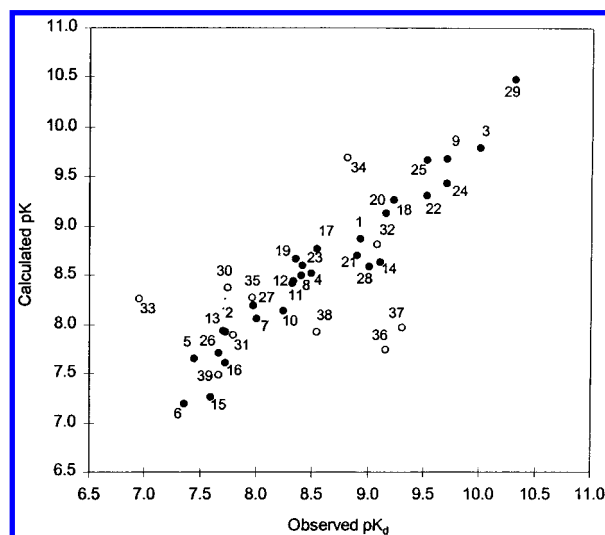
<sup>a</sup> The values in bold are the outliers (greater than twice the standard deviation of fit using all 39 compounds).

Both of these compound have similar structures to compounds in the training set. Compound 36 is the saturated derivative of compound 20, which has the same observed activity as compound 36 ( $pK_a = 9.15$ ). Compound 37 is the *O*-demethylated derivative of compound 25, which also has a similar observed activity.

**7. Graphical Representations and Interpretation of Models I–III.** Graphical representations of the 3D-QSAR Models I–III are shown, respectively, in Figures 3, 5, and 7. In each of these figures, the lowest-energy CEP conformation of compound 5 (low activity of 7.44) and compound 9 (high activity of 9.70) are shown superimposed according to alignment 6. Compound 9 is displayed in “gray” scale color whereas compound 5 is shown in “black”. The grid cells are shown as spheres whose diameters are equal to the sides of the grid cells. Next to each grid cell “sphere”



**Figure 8.** The residuals of fit (observed — calculated  $pK_a$ ) plot for test set of compounds (30–39) using Models I, II, and III.



**Figure 9.** Observed versus calculated  $pK_a$  values for the training set (compounds 1–29, black circles) and for the test set (compounds 30–39, white circles) using Model III.

**Table 13.** Standard Deviation of Fit of the Residuals (Observed — Calculated  $pK_a$ ) for the Test Set (29 Compounds) and for All the Compounds (39 compounds) for Models I, II, and III

number of compounds	standard deviation of residuals		
	model I	model II	model III
29	0.31	0.29	0.21
39	0.79	1.02	0.47

representation is its relative location in space, ( $x, y, z$ ) and IPE type (see Figure 3), “GC( $x, y, z$ , IPE)”. The “+” or “–” in front of each GC( $x, y, z$ , IPE) defines whether activity is enhanced (+) or diminished (–) as a function of increasing grid cell occupancy. In brackets below each GC( $x, y, z$ , IPE) are the compound number, 5 and 9, and the corresponding percent occupancy of the grid cell for the CEP of each of the two compounds.

Model I (Figure 5) is composed of (i) GCOD (6, 0, 1, 0), which is most frequently occupied by carbon and oxygen atoms of the amide function of the  $\omega$  chain, (ii) GCOD (9, –3, 2, 0/1), which is most often occupied by alkyl carbon and hydrogen atoms at the middle position of the  $\omega$  chain, and (iii) GCOD (11, –4, 5, 1), which is most frequently occupied by alkyl and aromatic carbons and hydrogen atoms at the end of the  $\omega$  chain.

The regression coefficients of each of the GCODs of Model I are positive. Hence, the greater the occupancy of each of these grid cells by the appropriate atom type of an analogue, the greater the biological activity of the analogue.

The GCODs are normalized to values between 0 and 1 with respect to occupancy for each analogue. Thus, the GCOD regression coefficients are direct measures of the relative importance of the GCODs on biological activity. However, it must be kept in mind that occupancies of the GCODs can be highly coupled to one another through conformation and alignment. Hence, an increase in occupancy of one grid cell by an appropriate atom type may come at the expense of a loss of useful grid cell occupancy by an atom type somewhere else.

Model I expresses biological activity by three GCODs that characterize the shape of the  $\omega$  chain. Only *all-atom* and *nonpolar* atom types are expressed in Model I. The role of the  $\alpha$  chain is only indirectly expressed through its conformational and alignment coupling to the  $\omega$  chain.

In contrast, Model II (Figure 3) is composed of GCOD (2, -3, 1, 6) that is most frequently occupied by aromatic hydrogen atoms of the phenyl ring on the  $\alpha$  chain. GCOD (3, 0, 2, 3) of Model II is most often occupied by oxygen and nitrogen atoms of the amide function on the  $\omega$  chain. This GCOD has a negative regression coefficient in the equations representing Models II and III. In other words, occupation of grid cell (3, 0, 2) decreases activity, and, in fact, for the least active compound (number 6 of Table 1), this is the grid cell most frequently occupied. The other GCODs of Model II are (5, -3, 2, 1), which is most frequently occupied by alquil and aromatic carbons and hydrogen atoms at the end of the  $\omega$  chain, and (5, -1, 0, 0/1), which is most frequently occupied by alquil carbon and hydrogen atoms near the middle of the  $\omega$  chain. It is not clear from Figure 3 that (5, -1, 0) can be occupied, but the CEPs of some compounds have conformations in which the  $\omega$  chain adopts a "kinked" geometry where the apex of the kink occurs near this grid cell.

Model II is similar to Model I in terms of both the number of independent variables (four for Model I and three for Model II) and statistical significance. However, the GCODs of the two models are somewhat different. GCODs (6, 0, 1, 0) [Model I] and (5, -1, 0, 0/1) [Model II] are adjacent grid cells in space and involve, largely, "all atom" types. However, the other GCODs of these two models are different in both spatial location and atom type. Model II explicitly incorporates features from both the  $\alpha$  and  $\omega$  chains, whereas Model I incorporates GCODs from only the  $\omega$  chain. Model III is composed of GCODs (9, -3, 2, 0/1, 1), (11, -4, 5, 1, 1), (2, -3, 1, 6, 2) and (3, 0, 2, 3, 2) from both Models I and II. Model III is a better 3D-QSAR than both Models I and II by all statistical measures of fit and robustness. The need for a 2 Å, as opposed to a 1 Å grid cell in the region described by (2, -3, 1, 6, 2) may be due to this space being associated with the end of a long, flexible chain; that is, there is considerable conformational latitude/uncertainty in defining spatial needs from the available SAR data (see Figure 7).

#### 8. Scrambling the Activity Measures of the Data Set.

As a final validation test of the 4D-QSAR analysis, the activity measures (dependent variables) of the 29 compounds of the training set were randomly "scrambled" with respect to compound chemical structure and an attempt was made to derive new 4D-QSAR models. The same data reduction schemes used in the original study were employed for the scrambled activities training set. The scrambled activity data

set has a correlation with the original data set of  $R = -0.18$ . None of the 3D-QSAR models from the scrambled activity training set were significant. The  $Q^2$  values obtained are in the range of 0.25 to 0.37. These results indicate that the  $Q^2$  of the 3D-QSAR models obtained from the original training set data set are not due to chance correlation.

Overall, the results from this study are in agreement with a study from Clark and Cramer<sup>25</sup> who found that the frequency of chance correlation using PLS decreases indefinitely as the number of descriptors becomes increasingly larger than the number of observations (compounds), as is usually the case in CoMFA applications.

## DISCUSSION

Significant 3D-QSAR models (Models I–III) were constructed for a set of highly flexible molecules using 4D-QSAR analysis. A variety of data reduction schemes were used in model construction. The quality of a resultant 3D-QSAR model is sensitive to the method of data reduction. It would seem that different data reduction schemes need to be considered in optimizing a 3D-QSAR model using 4D-QSAR analysis.

A very surprising, but highly significant, finding from this study is that "pieces" of good 3D-QSAR models can be combined to generate even better models. Moreover, these pieces can come from models arrived at using not only different data reduction schemes, but concurrently, or independently, from 3D-QSARs derived using *different grid cell sizes* to compute the GCODs. In this study, the sets of 3D-QSAR models employed as sources of the "pieces" were arrived at by trying different data reduction schemes for a fixed grid cell size. However, there is no reason the generation of an optimized 3D-QSAR from pieces of other good models cannot be generalized. A series of data reduction schemes can be explored for a set of GCODs derived from *multiple grid cell sizes*. GCODs from different size grid cell lattices can be very rapidly generated from the CEPs and alignments, and easily incorporated into the GA optimization by expanding the (x, y, z, IPE) representation of a GCOD to a (x, y, z, IPE, grid cell size) linear string. This representation of GCODs is, in fact, used for Model III. Moreover, GA mutation operators can be constructed to test for the significance of spatially overlapping GCODs. This strategy is being tested in our current 4D-QSAR analysis software.

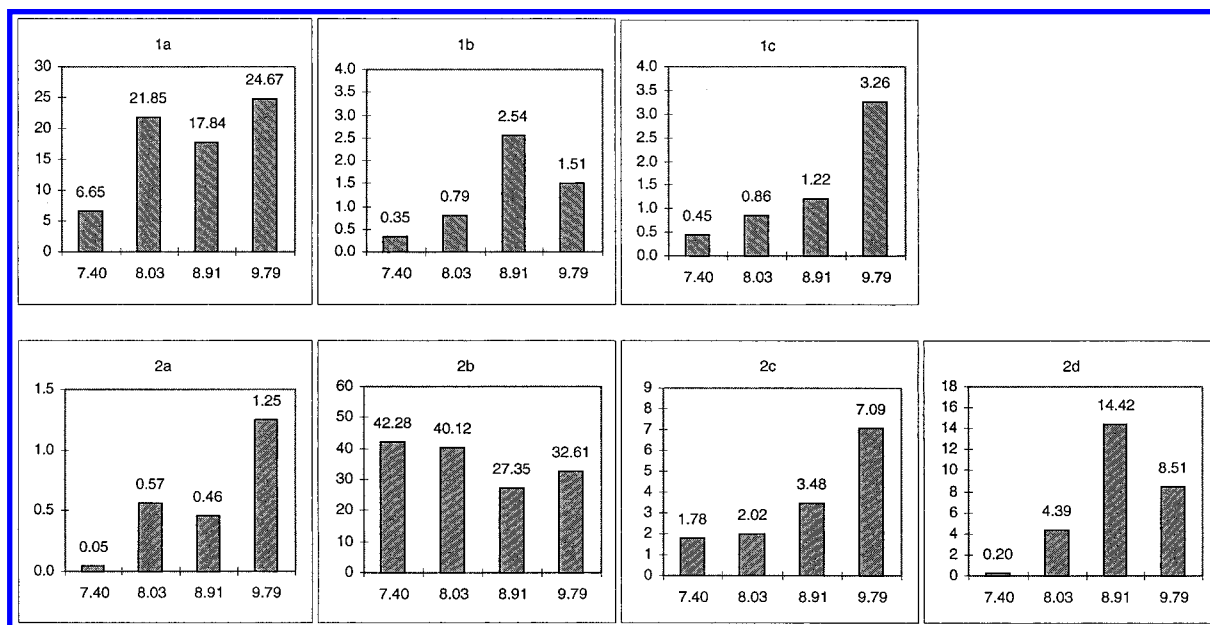
Some plausible reasons that large/small grid cell size GCODs may be selected in preference to GCODs from smaller/larger grid cell lattices, and/or "all-atom type" IPEs may be selected in deference to specific atom type IPEs, in the optimum 3D-QSAR model are

**(1) Overcoming "Signal-to-Noise Problems." (a) Adoption of Large Grid Cells.** The training set may contain only a few compounds, and/or low-energy conformations, that explore a critical region of space with respect to activity. A larger grid cell size may permit the statistical identification of this space by providing larger population GCODs of the region than can be realized for the corresponding GCODs derived from smaller size lattices. The GCOD (2, -3, 6, 2) of Models II and III may result from this type of behavior in the training set. **(b) Adoption of smaller grid cells.** The argument for larger grid cell GCODs just presented implicitly

**Table 14.** Four Compound Subsets of the Training Set Consisting of Less Active to the Most Active Compounds

compound subset <sup>b</sup>	average activity	1 Å grid cell (% average occupancy)			2 Å grid cell (% average occupancy)			
		1a	1b	1c	2a	2b	2c	2d
(A) 5–6	7.40	6.65	0.35	0.45	0.05	42.28	1.78	0.20
(B) 2; 7–8; 10–13; 15–16; 19; 23; 26–27	8.03	21.85	0.79	0.86	0.57	40.12	2.02	4.39
(C) 1; 4; 14; 17–18; 20–21; 28	8.91	17.84	2.54	1.22	0.46	27.35	3.48	14.42
(D) 3; 9; 22; 24–25; 29	9.79	24.67	1.51	3.26	1.25	32.61	7.09	8.51

<sup>a</sup> The average activity and the average occupancy (given in percent, %) of the 1 Å (1a, 1b, and 1c) and 2 Å (2a, 2b, 2c, and 2d) grid cells found in Models I–III are reported: 1a = (6,0,1,0); 1b = (9,−3,2,0/1); 1c = (11,−4,5,1); 2a = (2,−3,1,6); 2b = (3,0,2,3); 2c = (5,−3,2,1), and 2d = (5,−1,0,0/1). <sup>b</sup> Compound number is from Table 1.



**Figure 10.** The histogram representation of the average occupancy (%) of the 1 Å and 2 Å grid cells (Y-axis) of Models I–III for the compound subsets of average activity 7.40, 8.03, 8.91, and 9.79 (X-axis). See Table 14 for the specific compounds in the subsets and definitions of the GCODs.

assumes the distribution of occupancy in the large grid cell is random. However, if the distribution of occupancy is largely localized to a specific subspace of the grid cell, then a smaller grid cell encompassing that subspace may provide a more resolved GCOD than the larger grid cell. Moreover, in some cases multiple small grid cell size GCODs may provide more information to the QSAR than the single large grid cell GCOD that includes all of these smaller grid cells.

**(2) Characterizing “Hydrophobic Pockets.”** Hydrophobic pockets of a receptor binding site are usually larger than the size of an individual grid cell. Different parts and amounts of a hydrophobic pocket are usually filled by each of the ligands. Thus, the probability that the population contribution from any individual GCOD of a small grid cell exploring the hydrophobic pocket will be statistically meaningful is small. However, a larger grid cell size should capture a greater atomic population of this region by the ligands, and, therefore, more likely lead to a statistically meaningful GCOD in the 4D-QSAR analysis. In a sense, this is also a signal-to-noise issue, but arises from a particular type of geometric and energetic interaction. The GCODs involving nonpolar atom types should be strong candidates for larger grid cell sizes using this reasoning.

**(3) Identifying Local Molecular Flexibility.** Some regions of a ligand may need to be highly flexible in order for other parts of the ligand to simultaneously achieve tight receptor binding. Such local flexible structures of a ligand

are likely to be located in the binding site such that there is space around them to accommodate their movements. If this type of local conformational flexibility is distributed over a ligand training set and crucial to activity, then larger grid cell sizes are needed for the corresponding GCODs to capture this needed flexibility in terms of variable spatial locations of the corresponding atoms. This type of reasoning would argue the GCODs of “all-atom” type of occupancy should reflect this type of behavior.

The predictions for the test set of 10 compounds are not, in composite, as good as those for the training set using any of the three models (see Tables 11 and 12). This result can be indicative of

- (1) structure–activity diversity in the test set is not captured in the training set;
- (2) overfitting of the 3D-QSAR models to the training set; and
- (3) slightly different multiple modes of ligand–receptor binding some of which for the test set are different from those used to construct the models in the analysis of the training set.

Schemes can be developed to explore the division of a SAR data set into a training set and a test set so that the predictions for the test set are as good as for the training set. However, such schemes are particular forms of cross-validation, and may provide no additional information than



can be achieved by simple combining of the test and training sets into a larger training set and doing an analysis.

In this study, the training set was designed to mimic the case where a 3D-QSAR model had been constructed, new analogues subsequently suggested and a test made to see how well, in fact, their activities could be predicted. In light of this type of "real word" application, a better approach to dealing with the SAR data of a training set is to take advantage of the evolutionary capabilities of genetic algorithms and devise a scheme to evolve a 3D-QSAR model from its original training set to include new compounds and their activities as they become available.

The spatial pharmacophore (refer to Figures 3, 5, and 7) that emerges for the TXA<sub>2</sub> antagonists in the training set, based on the GCODs of 3D-QSAR models I–III, has the following features,

(1) a specific shape (conformation), defined in terms of "all atom" and/or "nonpolar" atom types, is required for the  $\omega$  chain; the GCODs (6,0,1,0), (9,−3,2,0/1), (5,−3,2,1), (5,−1,0,0/1), and (11,−4,5,1) define this shape.

(2) occupancy of the space "between" the  $\alpha$  and  $\omega$  chains, near the amide function of the  $\omega$  chain by polar atoms with negative charge densities is detrimental to activity according to GCOD (3,0,2,3);

(3) specific local shape/conformation of the  $\alpha$  chain near the aromatic ring such that grid cell (2,−3,1) is occupied by aromatic atoms of the phenyl ring is required.

Overall, a relatively extensive "recipe" for predicting the pK<sub>d</sub> activity of new TXA<sub>2</sub> antagonist analogues is given by Models I–III and Figures 3, 5, and 7. Any new candidate analogue can be explored and its activities predicted using Models I–III by constructing the requisite CEP and applying alignment 6. However, additional insight into selecting new trial analogues prior to their computational evaluation can be realized by knowing which grid cells of Models I–III are most highly (lowly) populated for the most active analogues, as compared with less active analogues, in the training set. Table 14 and Figure 10 report the grid cell occupancy of each of the GCODs of Models I–III for four subsets of the training set that have been partitioned by activity, pK<sub>d</sub>. Some definite trends in the grid cell occupancy behavior of the analogues in the training set can be gleaned from Table 14 and Figure 10. Grid cells (6, 0, 1) [1a] of Model I and (3, 0, 2) [2b] of Models II and III are more highly occupied than the other grid cells of Models I–III. A decrease in the occupancy of (3, 0, 2) by polar atoms with a negative charge density would be desirable because occupancy of this space by such atom types decreases activity. Increases in the occupancy of (11, −4, 5) by any type of atom, (2, −3, 1) by aromatic atoms, and (5, −3, 2) by any atom type each correlate with increasing activity. All three of these grid cells are, on average, occupied <10% for the CEPs of the compounds in the training set. Thus, new analogues that increase the occupancy of any of these sites, even modestly, should be active.

The discussion just presented highlights the fact that one cannot think of a rigid template structure of a molecule when discussing SARs from a 4D-QSAR analysis. The probability that parts of a molecule are found in interrelated regions of space becomes a component of the basis for describing the SAR. Medicinal chemists generally accept this conceptual

"picture" of a ligand–receptor SAR, but may now have to learn to actually work with concrete representations of the concept.

## ACKNOWLEDGMENT

This work was supported in part by an NSF SBIR Phase I Grant (No. DMI-9560439) to The Chem21 Group, Inc. Resources of The Laboratory of Molecular Modeling and Design were used to perform the reported work. MA is grateful to the CNPq of Brazil for fellowship support and to the Department of Medicinal Chemistry and Pharmacognosy of UIC. MA acknowledges W. Dunn III for use of the HyperChem software and J. Tokarski, B. Jin, S. Wang, Prabha V., Prakash M., R. Wood, C. Catana, and C. Klein, all of UIC, for useful discussions over the course of this study.

## REFERENCES AND NOTES

- (1) Samulsson, B.; Goldyne, M.; Granstrom, E.; Hamberg, M. Hammarstrom, B.; Malmsten, C. Prostaglandins and thromboxanes. *Ann. Rev. Biochem.* **1978**, *47*, 997–1029.
- (2) Armstrong, R. A.; Wilson, N. H. Aspects of the thromboxane receptor system. (Review) *Gen. Pharmacol.* **1995**, *26* (3), 436–472.
- (3) Hall, S. E. Thromboxane A<sub>2</sub> receptor antagonists. *Med. Res. Rev.* **1991**, *11* (5), 503–579.
- (4) Hirata, M.; Hayashi, Y.; Ushikubi, F.; Yokota, Y.; Kageyama, R.; Nakanishi, S.; Narumiya, S. Cloning and expression of cDNA for a human thromboxane A<sub>2</sub> receptor. *Nature* **1991**, *349*, 617–620.
- (5) Yamamoto, Y.; Kamiya, K.; Terao, S. Modeling of a human thromboxane A<sub>2</sub> receptor and analysis of the receptor–ligand interaction. *J. Med. Chem.* **1993**, *36* (7), 820–825.
- (6) Ezumi, K.; Yamakawa, M.; Narisada, M. Computer-aided molecular modeling of a thromboxane receptor antagonist S-145 and its related compounds. *J. Med. Chem.* **1990**, *33* (4), 1117–1122.
- (7) Fukumoto, S.; Shiraishi, M.; Terashita, Z.; Ashida, Y.; Inada, Y. Synthesis and thromboxane A<sub>2</sub>/prostaglandine H<sub>2</sub> receptor antagonistic activity of phenol derivatives. *J. Med. Chem.* **1992**, *35* (12), 2202–2209.
- (8) Jin, B.; Hopfinger, A. J. A proposed common spatial pharmacophore and the corresponding active conformation of some TXA<sub>2</sub> receptor antagonists. *J. Chem. Inf. Comput. Sci.* **1994**, *34* (4), 1014–1021.
- (9) Albuquerque, M. G.; Rodrigues, C. R.; Alencastro, R. B.; Barreiro, E. J. Design of new potential 5-lipoxygenase inhibitors, dual thromboxane synthase inhibitors, and thromboxane A<sub>2</sub> receptor antagonists by AM1. *Int. J. Quantum Chem., Quantum Biol. Symp.* **1995**, *22*, 181–190.
- (10) Misra, R. N.; Brown, B. R.; Sher, P. M.; Patel, M. M.; Hall, S. E.; Han, W. C.; Barrish, J. C.; Kocy, O.; Harris, D. N.; Goldenberg, H. J.; Michel, I. M.; Schumacher, W. A.; Webb, M. L.; Monshizadegan, H.; Ogletree, M. L. Interphenylene 7-oxabicyclo[2,2,1]heptane oxazoles. Highly potent, selective, and long-acting thromboxane A<sub>2</sub> receptor antagonists. *J. Med. Chem.* **1993**, *36* (10), 1401–1417.
- (11) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M. G.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509–10524.
- (12) *HyperChem Program Release 4.5 for Windows*; Molecular Modeling Systems; 1995.
- (13) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107* (13), 3902–3909.
- (14) Stewart, J. J. P. *Mopac Manual v.6.0*; Frank J. Seiler Research Laboratory, United States Air Force Academy, Colorado Springs, CO, 1990.
- (15) van Gunsteren, W. F.; Berendsen, H. J. C. Computer simulation of molecular dynamics: methodology, applications, and perspectives in chemistry. *Angew. Chem., Int. Ed. Engl.* **1990**, *29*, 992–1023.
- (16) *Molsim User's Guide v.3.0*, Molecular Mechanics and Dynamics Simulation Software; D. C. Doherty and The Chem21 Group, Inc., 1780 Wilson Drive, Lake Forest, IL 60645; 1994.
- (17) Joao, H. C.; De Vreese, K.; Pauwels, R.; De Clercq, E.; Henson, G. W.; Bridger, G. J. Quantitative structural activity relationship study of bis-tetraazacyclic compounds. A novel series of HIV-1 and HIV-2 inhibitors. *J. Med. Chem.* **1995**, *38* (19), 3865–3873.

- (18) Grigorov, M.; Weber, J.; Tronchet, J. M. J.; Jefford, C. W.; Milhous, W. K.; Maric, D. (1997) A QSAR study of the antimalarial activity of some synthetic 1,2,4-trioxanes. *J. Chem. Inf. Comput. Sci.* **1997**, 37 (1), 124–130.
- (19) 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.
- (20) *Wolf Manual v.6.2, Wolf Genetic Function Approximation*; D. Rogers and Molecular Simulation, Inc.; 1994.
- (21) Dunn, W. J., III; Rogers, D. Genetic Partial Least Squares in QSAR. In *Genetic Algorithms in Molecular Modeling*; Devillers, J., Ed.; Academic: London, 1996; pp 109–130.
- (22) Glen, W. G.; Dunn, W. J., III; Scott, D. R. Principal components analysis and partial least squares. *Tetrahedron Comput. Methodol.* **1989**, 2 (6), 349–376.
- (23) Hasegawa, K.; Miyashita, Y.; Funatsu, K. (1997) GA strategy for variable selection in QSAR studies: GA-based PLS analysis of calcium channel antagonists. *J. Chem. Inf. Comput. Sci.* **1997**, 37 (2), 306–310.
- (24) *4D-QSAR Manual v.1.0*, A. J. Hopfinger and The Chem21 Group, Inc., 1780 Wilson Drive, Lake Forest, IL 60045; 1997.
- (25) Clark, M.; Cramer, R. D., III The probability of chance correlation using partial least squares (PLS). *Quant. Struct.-Act. Relat.* **1993**, 12 (2), 137–145.

CI980093S