

Prediction of Ligand–Receptor Binding Free Energy by 4D-QSAR Analysis: Application to a Set of Glucose Analogue Inhibitors of Glycogen Phosphorylase

Prabha Venkatarangan[†] and Anton J. Hopfinger*

Laboratory of Molecular Modeling and Design (M/C-781), College of Pharmacy, University of Illinois at Chicago, 833 South Wood Street, Chicago, Illinois 60612-7231

Received April 5, 1999

A set of 47 glucose analogue inhibitors of glycogen phosphorylase was investigated using 4D-QSAR analysis. A single significant 4D-QSAR model, having no outliers, was found as a function of alignment and conformational sampling. This 4D-QSAR model consists of six grid cell occupancy descriptors which defines the thermodynamic averaged spatial pharmacophore for this set of analogues. The 4D-QSAR model was validated by aligning it upon the crystal structures of glycogen phosphorylase with some bound inhibitors of the training set. Validation of the 4D-QSAR model was realized by establishing the consistency of the types and locations of the grid cell occupancy descriptors relative to the binding interaction sites of the crystal complex. The loss in binding free energy, due to loss in inhibitor conformational entropy upon binding to the enzyme, was computed and found to be in the 0–2 kcal/mol range for these inhibitors. The “active” conformation of each analogue was also postulated from the 4D-QSAR model.

INTRODUCTION

In this paper we report the four-dimensional quantitative structure–activity relationship, 4D-QSAR, analysis of a set of 47 glucose analogue inhibitors of glycogen phosphorylase *b* (GP_b). Glycogen is the carbohydrate reserve of most metabolically active cells in mammals. GP catalyzes the first step in the phosphorolysis of glycogen to glucose-1-phosphate. Glycolysis of glucose-1-phosphate in muscle provides energy to sustain muscle contraction, while in the liver the production of glucose supplies an energy source for extrahepatic tissue. GP exists in two interconvertible states through reversible phosphorylation, the inactive *b* form (predominantly *T* state) and the active *a* form (predominantly *R* state). Hepatic glycogen metabolism is regulated by glucose through promotion of inactivation of GP_a.¹ Glucose inactivates GP_a by competitive inhibition of glucose-1-phosphate and stabilizes the inactive *T* state.

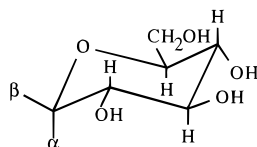
Diabetes Mellitus (DM) is characterized by chronic elevated blood glucose levels. Hyperglycemia in non-insulin-dependent (NIDDM) patients is a result of diminished insulin release and/or insulin resistance that leads to impaired tissue uptake and impaired suppression of hepatic glucose production. In response to insulin release after a meal, muscle glycogen synthesis is the principal pathway of glucose disposal, and defects in muscle glycogen synthesis have a dominant role in the insulin resistance that occurs in persons with NIDDM. At basal insulin levels, impaired suppression of hepatic output of glucose arising from gluconeogenesis and glycogen breakdown is the principal cause of high glucose concentration in NIDDM patients. In an animal model, the genetically diabetic *db/db* mouse, the activity of hepatic phosphorylase is increased compared with nondiabetic mice.² Weak inhibitors of glycogen phosphorylase are weakly hypoglycemic.³ These observations suggest that inhibitors

of glycogen phosphorylase may help shift the balance between glycogen synthesis and glycogen degradation in favor of glycogen synthesis in both muscle and liver, and such inhibitors may be useful therapeutic agents for treatment of diabetes. Thus, glucose analogue inhibitors of glycogen phosphorylase may be of clinical interest in the regulation of glycogen metabolism in diabetes.

4D-QSAR analysis⁴ is a relatively new molecular modeling method to develop QSAR models. The method is capable of exploring and sampling 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 having multiple candidate pharmacophore patterns as demonstrated in a study of a set of interphenylene 7-oxabicycloheptane oxazole TXA₂ receptor antagonists.⁵ The capacity to sample and explore geometric and pharmacophore-group degrees of freedom crucial to building a comprehensive QSAR model is the fourth “dimension” considered in 4D-QSAR analysis. Thus, a working definition of 4D-QSAR analysis is a set of procedures which allows the construction of optimized dynamic spatial QSAR models, in the form of 3D pharmacophores, which are dependent on conformation, alignment, and pharmacophore-grouping.

Crystal structures of some glucose analogue inhibitors bound to GP_b are available.⁶ In fact, this structure information has been used to construct free energy force field (FEFF) 3D-QSAR ligand–receptor binding models.⁷ Thus, in addition to developing a receptor-independent (RI) 4D-QSAR model, the work reported here explores how much structural and thermodynamic information of the ligand–receptor complex is captured in the RI 4D-QSAR model. The exploration of structural information within the RI 4D-QSAR model is realized by aligning the grid cell occupancy descriptors (GCODs) of the model [see refs 4 and 5, and

[†] Current Address: Pharmacoceia Inc., CN 5350, Princeton, NJ 08543.

Table 1. Structure–Activity Data for the Glucose Analog Inhibitors of Glycogen Phosphorylase Used in the 4D-QSAR Training Set

compd no.	α	β	K_i (mM)	ΔG_{303} (kcal/mol)
1	H	NHC(=O)CH ₃	0.032	6.23
2	H	NHC(=O)CH ₂ CH ₃	0.039	6.11
3	H	NHC(=O)CH ₂ Br	0.044	6.04
4	H	NHC(=O)CH ₂ Cl	0.045	6.03
5	H	NHC(=O)C ₆ H ₅	0.081	5.67
6	H	NHC(=O)CH ₂ CH ₂ CH ₃	0.094	5.58
7	H	NHC(=O)NH ₂	0.14	5.34
8	H	C(=O)NHCH ₃	0.16	5.26
9	H	NHC(=O)CH ₂ NH ₂	0.37	4.76
10	C(=O)NH ₂	H	0.37	4.76
11	H	C(=O)NH ₂	0.44	4.65
12	H	C(=O)NHNH ₂	0.40	4.17
13	H	SH	1.00	4.16
14	CH ₂ OH	H	1.50	3.92
15	OH	H	1.70	3.84
16	H	C(=O)NHC ₆ H ₅	5.40	3.14
17	H	OH	7.40	2.95
18	H	CH ₂ CN	9.00	2.84
19	OH	CH ₂ OH	15.80	2.50
20	H	OCH ₃	24.70	2.23
21	CH ₂ NH ₂	H	34.50	2.03
22	C(=O)NHCH ₃	H	36.70	1.99
23	CH ₃	H	53.10	1.77
24	C(=O)NH ₂	NHCOOCH ₃	0.016	6.65
25	H	NHCOOCH ₂ Ph	0.35	4.79
26	H	NHC(=O)CH ₂ NHCOCH ₃	0.99	4.17
27	H	C(=O)NHNHCH ₃	1.8	3.81
28 ^a	OH	H	2	3.74
29	H	C(=O)NHCH ₂ CH ₂ OH	2.6	3.58
30	H	COOCH ₃	2.8	3.54
31	C(=O)NHNH ₂	H	3.0	3.50
32	H	SCH ₂ C(=O)NHPh	3.6	3.39
33	H	C(=O)NH-4-OHPh	4.4	3.27
34	H	CH ₂ CH ₂ NH ₂	4.5	3.25
35	C(=O)NH-4-OHPh	H	5.6	3.12
36	OH	CH ₂ N ₃	7.4	2.95
37	OH	CH ₂ CN	7.6	2.94
38	H	C(=O)NHCH ₂ CF ₃	8.1	2.90
39	C(=O)NHPh	H	12.6	2.63
40	COOH	H	15.2	2.52
41	H	CH ₂ NH ₂	16.8	2.46
42	C(=O)NHCH ₂ CH ₂ OH	H	16.9	2.46
43	H	SCH ₂ C(=O)NH-2,4-F ₂ Ph	18.9	2.39
44	H	SCH ₂ C(=O)NH ₂	21.1	2.32
45	CH ₂ N ₃	H	22.4	2.29
46	COOCH ₃	H	24.2	2.24
47	C(=O)NHCH ₂ -2,4-F ₂ Ph	H	27.2	2.17

^a The ring O replaced by S.

part b of the Methods] within the inhibitor binding site of Gpb to see if the 4D-QSAR model sterically fits and provides complementary ligand–receptor interaction sites. Thermodynamic binding information inherent to the RI 4D-QSAR model can be evaluated by comparing the energy terms of the FEFF 3D-QSAR model to the grid cell occupancy descriptors of the RI 4D-QSAR model.

METHODS

Training Set of Glucose Analogue Inhibitors. The training set of 47 glucose analogue inhibitors of glycogen phosphorylase, Gpb, is given in Table 1. These structures, and the

corresponding inhibitory binding constants (K_i), are reported in refs 8–11. The change in the free energy of binding, ΔG , was estimated from the K_i values using eq 1 where T is temperature and R is the gas constant. The ΔG values for the inhibitors are also reported as part of Table 1.

$$\Delta G_i = -RT \ln K_i \quad (1)$$

The 10 Operational Steps in 4D-QSAR Analyses. The 10 steps involved in the current 4D-QSAR formalism are listed in Table 2.

Step 1 is analogous to initiating a CoMFA 3D-QSAR study¹² in that a reference grid cell space and a 3D structure

Table 2. The 10 Operational Steps of the Receptor-Independent (RI) 4D-QSAR Formalism

step number	description of 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 pharmacophore elements, IPEs
3	perform a MDS conformational ensemble sampling of each compound and generate its conformational ensemble profile, CEP
4	select a trial alignment
5	place each conformation of each compound in the reference grid space according to the alignment, and record the grid cell occupancy, GCOD, for each IPE and choice in occupancy measure
6	perform a partial least-squares (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 genetic algorithm (GA) analysis
8	return to step 4 and repeat steps 4–7 unless all trial alignments have been included in the analysis
9	select the optimum QSAR with respect to alignment and any of the methodological parameters
10	select 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)

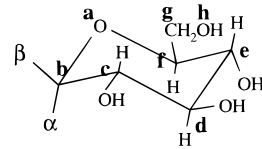
for each compound in the training set are specified. In this 4D-QSAR application space is divided into a cubic grid lattice of 40 Å on each side containing grid cells with a resolution of 1.0 Å. A 3D structure of each molecule in the training set is generated. The reference molecular structure to generate the analogue geometries is the bound glucose conformation obtained from the crystal structure of the glucose-glycogen phosphorylase complex.⁶ The Chemlab-II molecular modeling program¹³ was used to add the substituents to the bound glucose ring conformation. Substituent geometries were optimized by fixed valence geometry conformational analysis. Partial atomic charges were assigned to each of the ligands on the basis of CNDO calculations using Chemlab-II.¹³

The second step in 4D-QSAR analysis is the selection of the trial set of interaction pharmacophore elements, IPEs. In essence, the atoms of a molecule are partitioned into seven classes: polar-positive partial charge (p+), polar-negative partial charge (p-), nonpolar (np), hydrogen bond donor (hbd), hydrogen bond acceptor (hba), aromatic (ar) and unrestricted atom type occupancy (any).

The third step in the formalism, performing a conformational ensemble sampling of each compound in the training set, addresses the active conformation issue. The objective of this step is to Boltzmann sample the complete set of conformational states available to each molecule to generate its conformational ensemble profile, CEP. Clearly, complete conformational sampling may not be achieved for an arbitrary molecule, and can be considered a source of possible error in the construction of a 4D-QSAR model.

Molecular dynamics simulation, MDS, is used to generate the CEP of each molecule in the training set. The MDSs are done using the MOLSIM package¹⁴ with an extended MM2 force field.^{15,16} The temperature for the MDS is set at 300 K with a simulation sampling time of 100 ps with intervals of 0.001 ps for a total sampling of 100 000 conformations of each glucose analogue. The atomic coordinates of each conformation and its intramolecular energy sampled during the MDS are recorded every 0.01 ps for a total of 1000 "frames", or steps, in the CEP of each compound.

Selection of the trial alignments is step 4 in a 4D-QSAR analysis. To be clear, 4D-QSAR analysis does not "solve" the alignment problem. Rather, the 4D-QSAR scheme permits a rapid evaluation of individual trial alignments. Consequently, the alignment problem can be treated as a search and sample operation analogous to conformational profiling. The ability to rapidly evaluate alignments in the

Table 3. Set of Trial Alignments Used in Constructing the 4D-QSAR Models


alignment no.	atom 1	atom 2	atom 3
1	a	b	c
2	g	f	a
3	f	a	b
4	c	d	e
5	f	g	h

4D-QSAR algorithm is due to the complete decoupling of conformational analysis from alignment analysis and rapid descriptor estimation for each alignment.

A single CEP of each compound in the training set is used to evaluate the five trial alignments defined in Table 3. A three-atom-pair match is currently used to define an alignment. Each conformation of each compound is aligned relative to the reference grid through the invariant coordinates of the three alignment atoms. The alignments are chosen to explore the set of potential binding sites on the glucose analogues. The strategy was to pick alignments close to and far from the α and β substituent sites, alignments spanning the ring, and an alignment involving the CH₂OH substituent.

Each conformation from the CEP of each compound is placed in the reference grid cell space according to the trial alignment under consideration as part of step 5. The grid cell occupancy descriptors, GCODs, for each of the chosen IPEs are then computed and used as the basis set of trial 4D-QSAR descriptors. Figure 1 illustrates a general glucose analogue inhibitor in grid cell space under some alignment.

Step 6 is the data reduction step identical to that done in a CoMFA. 4D-QSAR analysis, like CoMFA, generates an enormous number of trial QSAR descriptors because of the large number of grid cells and seven possible IPEs. Unlike CoMFA, however, the 4D-QSAR scheme defines the *complete set* of grid cells of the training set to include in an analysis. The composite set of grid cells occupied at least once in constructing the CEP for the *any atom* IPE of each member of the training set defines the complete set of grid cells. CoMFA uses a distance cutoff in the evaluation of field potential to limit the set of descriptor grid cells.

Partial least squares (PLS) regression analysis¹⁷ is used to perform the data reduction fit between the observed

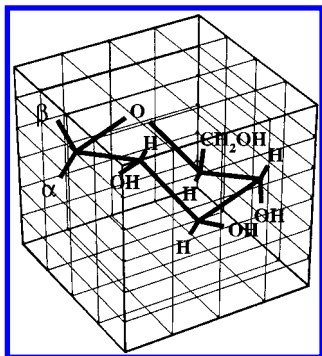


Figure 1. Each conformation of each compound aligned in the reference grid cell space.

dependent variable (ΔG) measures and the corresponding GCOD values. Before PLS is applied, two different types of data filtering can be employed. First, all grid cells not occupied by any atom of any compound for all the conformations can be omitted. Second, the GCODs resulting from occupancies of less than two atoms over the entire MDS-CEP can also be excluded prior to PLS being performed. The most highly weighted PLS descriptors are used to form the trial basis set for deriving the final optimum 4D-QSAR models.

The compact 4D-QSAR models are actually generated in step 7 as part of model building, optimization, comparison, and evaluation using a genetic algorithm, GA. The M most highly weighted PLS descriptors are used to form the trial basis set for the GA analysis. Currently, $M = 200$, and only linear terms are used in multiple linear regression, MLR, fits in the GA optimization. Other descriptors, not derived from 4D-QSAR analysis, can be added to the trial basis set at the start of this step by the user. The specific genetic algorithm currently used in the 4D-QSAR program¹⁸ is a modification of the genetic function approximation, GFA.^{19,20} Smoothing factors of 0.2 and 0.5 are used to determine the optimal number of descriptors in the 3D-QSAR model²¹ on the basis of Friedman's lack-of-fit, LOF,²² as the optimization metric.

The diagnostic measures used to analyze the resultant 4D-QSAR models generated by the GFA include descriptor usage as a function of crossover operation, linear cross-correlation among descriptors and/or dependent variables (biological activity measures), number of significant models, and measures of model significance including the correlation coefficient, r^2 , leave-one-out cross-validation correlation coefficient, $xv-r^2$, and LOF.²² Analogues of the training set are considered outliers when the differences in predicted and observed energies exceed 2.0 standard deviations from the mean. Random scrambling of the dependent variable (biological activity measure) with respect to the independent variables (QSAR descriptors of each compound) is also carried out to test for chance correlations.

Steps 5–7 are performed for a fixed alignment. Step 8 is a decision/selection operation to consider additional trial alignments in 4D-QSAR model construction, or to proceed with an evaluation of the composite set of 4D-QSAR models generated by the repetitive application of steps 4–7. Currently, step 8 is carried out directly by the investigator, and treatment of alignment is by sampling and not by an optimization scheme.

Once a desired set of trial alignments has been included in the 4D-QSAR model construction portion of the algorithm, the inspection and evaluation of the entire population of models are made. This is step 9. The principal objective of step 9 is to identify the “best” 4D-QSAR model with respect to alignment. However, this objective can be generalized to permit exploration and optimization of 4D-QSAR models with respect to not only alignment but also conformational sampling, IPE, and grid cell sizes. Moreover, while a best 4D-QSAR model can be identified using one or more measures of significance of fit, added information comes from comparing a family of best models. In essence, the best and distinct set of 4D-QSAR models from GA analysis are used in composite to build a *manifold* 4D-QSAR model.

Each alignment tested will lead to a particular best 4D-QSAR model *for that specific alignment*. The alignment corresponding to the 4D-QSAR model with the *overall highest r^2 and $xv-r^2$* , for all alignments tested, is selected as the *best* alignment. For the best alignment, a cross-correlation matrix of the residuals in error (observed ΔG – predicted ΔG) between pairs of the top 10 4D-QSAR models, based on their $xv-r^2$, is built. This is done to determine if the top 10 4D-QSAR models are providing common, or distinct, structure–activity information. In other words, it is possible to identify the set of *unique* best 4D-QSAR models. Overall, the best alignment and the corresponding best (highest explanation of the variance in the dependent variable, ΔG) 4D-QSAR model are identified as part of this procedure. Also, the linear cross-correlation matrix of the GCODs for the best 4D-QSAR model for the best alignment is built to determine if these significant GCODs are correlated to one another. Outlier analysis of the 4D-QSAR model is also performed.

The final step, step 10, is to hypothesize the “*active*” *conformation* of each compound in the training set. This is achieved by first identifying all conformer states sampled for each compound, one at a time, that are within ΔE of the global minimum energy conformation of the CEP. Currently, ΔE is set at 2 kcal/mol. The resulting set of low-energy conformations are individually evaluated using the equation of the best 4D-QSAR model. Since only a *single* conformer state is considered, grid cell occupancy is either zero or one depending on conformation and alignment. The single conformation within ΔE , which predicts the highest “activity”, the largest value of ΔG in this case, is selected as the active conformation of the compound. The postulated active conformations can be used as structure design templates, which includes their deployment as the molecular geometries of each ligand in a structure-based ligand–receptor binding study.

RESULTS

Of the five alignments considered, which are defined in Table 3, alignment 1 provides the best 4D-QSAR models as defined by the highest cross-validated correlation coefficient. The optimum $xv-r^2$ values are reported in Table 4 for each test alignment.

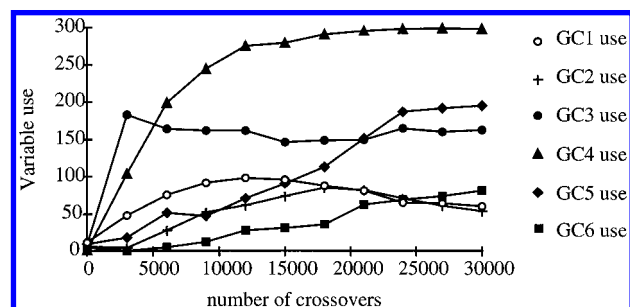
The top 10 4D-QSAR models of alignment 1 are summarized in Table 5. These models were obtained by using the GFA with 30 000 crossovers and a smoothing factor of 0.20. Descriptor usage as the 4D-QSAR models evolve and

Table 4. Cross-Validated Correlation Coefficient of the Best 4D-QSAR Model for Each Alignment

alignment no.	xv- r^2 (optimum)	alignment no.	xv- r^2 (optimum)
1	0.83	4	0.65
2	0.70	5	0.75
3	0.71		

Table 5. Summary of the Top 10 4D-QSAR Models Using Alignment 1 for the Glucose Analogue Inhibitors of Gp β

range in r^2 and (xv- r^2) in top 10 GFA models	0.86–0.87 (0.81–0.83)
number of unique grid cells in all top 10 GFA models	9
number of significant GFA descriptors	6
number of non-GCODs in all top 10 GFA models	none

**Figure 2.** GFA crossover plot of GCOD descriptor usage as the 4D-QSAR model evolves and is optimized.

are optimized, referred to as a crossover plot, is shown in Figure 2. The crossover plot demonstrates that the GFA model evolution has converged. Convergence is achieved when descriptor usage does not change as a function of the number of crossovers (a horizontal line).

The best 4D-QSAR model is

$$\Delta G = 5.04\text{GC1(hbd)} - 2.68\text{GC2(np)} + 11.22\text{GC3(p-)} + 4.87\text{GC4(any)} + 2.76\text{GC5(p+)} - 1.35\text{GC6(any)} + 2.89 \quad (2)$$

$$N = 47 \quad R^2 = 0.87 \quad \text{xv-}r^2 = 0.83$$

In eq 2 GCI(X) is the i th significant GCOD having the X type IPE as defined above in step 2 of the 4D-QSAR analysis process, and also in ref 4. No analogue in the data set of Table 1 was found to be an outlier. A variety of non-4D-QSAR descriptors including $\log P^{23}$ were included in the GFA model optimization. None of these non-4D-QSAR descriptors "survived" in the realization of the most significant 4D-QSAR models.

To determine if the top 10 4D-QSAR models are providing common, or distinct, structure-activity information, the correlation coefficients of the residuals in error between pairs of models were computed and are reported in Table 6. The idea in determining the residual-pair correlations is that equivalent models will have near-identical residuals, while distinct models should have noncorrelated residuals.²⁴ All of the top 10 4D-QSAR models have residuals of fit which are highly correlated to one another, indicating there is a single unique 4D-QSAR model which is selected as the model with the highest xv- r^2 value, namely, the 4D-QSAR model defined by eq 2.

The best 4D-QSAR model, as represented by eq 2, has been statistically evaluated. A linear cross-correlation matrix of the GCODs of eq 2 has been built (Table 7). Any pair of

Table 6. Linear Correlation Matrix of the Residuals of Error for the Top 10 4D-QSAR Models from GFA-MLR Optimization

model no.	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	1.00	1.00								
3	0.89	0.89	1.00							
4	0.91	0.89	0.85	1.00						
5	0.91	0.90	0.87	0.92	1.00					
6	0.85	0.90	0.94	0.91	0.93	1.00				
7	0.95	0.92	0.96	0.91	0.93	0.95	1.00			
8	0.91	0.89	0.98	0.94	0.92	0.95	0.96	1.00		
9	0.95	0.91	0.94	0.95	0.93	0.91	0.90	0.96	1.00	
10	0.95	0.96	0.93	0.90	0.89	0.91	0.95	0.93	0.92	1.00

Table 7. Linear Cross-Correlation Matrix of the GCODs Found in the Optimal 4D-QSAR Model (eq 2)

	GC1	GC2	GC3	GC4	GC5	GC6	BA
GC1	1.00						
GC2	-0.08	1.00					
GC3	0.02	-0.11	1.00				
GC4	-0.05	-0.17	0.58	1.00			
GC5	-0.08	-0.08	-0.13	-0.14	1.00		
GC6	-0.10	-0.10	-0.20	-0.22	0.17	1.00	
BA	0.51	-0.34	0.68	0.58	0.07	-0.27	1.00

GCODs having a correlation of greater than 0.50, or less than -0.50, is flagged as a highly correlated pair and indicated in bold in Table 7. Three of the GCIs of eq 2 are significantly correlated with ΔG . GC3 and GC4 are significantly correlated to one another and to ΔG . However, regression equations (QSAR models) constructed after removal of either one of these significantly correlated GCODs are statistically poor models ($r^2 \leq 0.30$). Moreover, deletion of any one of the GCIs in eq 2 leads to a poor model with r^2 in the range of 0.21–0.41. Thus, it appears that all six GCI(X)s of the best 4D-QSAR model provide requisite composite, and complementary, information about the variance in ΔG over the analogue training set.

The correlation between GCODs 3 and 4 involves two grid cells that are about 3 Å apart in space, yet both contribute to activity when occupied. The correlation between these two GCODs could possibly arise from an intramolecular "allosteric" effect where occupancy at one grid cell has a coupled interaction to yield an increase/decrease in the occupancy of the other grid cell.

Predictions of biological activities (ΔG in this study) using a 4D-QSAR model, but only employing the *predicted active conformation* (as defined in step 10 of the *Methods*) of each compound to estimate grid cell occupancy, permit entropic contributions of the ligand to activity (ΔG) to be estimated.⁴ The predicted ΔG for each active conformation of each compound was computed using eq 2. The difference between the predicted ΔG of each compound over its entire CEP and the predicted ΔG of its active conformation is plotted in Figure 3. The difference values in Figure 3 are mostly negative, indicating that the predicted ΔG values for the active conformations are higher than the corresponding predicted ΔG values obtained using the CEPs. The same is true if the observed ΔG values are used in place of the predicted ΔG values using the CEPs. This behavior reflects a decrease in the possible ΔG of the ligand binding to Gp β due to a loss in ligand conformational entropy upon binding. Each difference in ΔG in Figure 3 is, in fact, the estimated

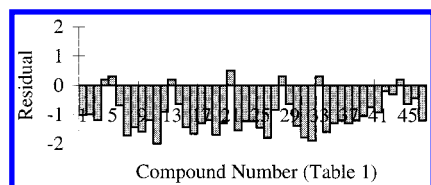


Figure 3. Residual plot of the ΔG calculated using eq 2 for the entire CEP minus the predicted active conformation for each compound in the training set.

loss in binding free energy due to a loss in ligand conformational entropy of the corresponding glucose analogue.

The composite set of conformations of each glucose analogue, completely sampled in nature, and approximately sampled by the CEP, correspond to *partial* occupancies of both the binding (ΔG) enhancing grid cells and grid cells that either decrease binding or do not contribute to ΔG . The active conformation of an analogue will likely occupy the grid cells which enhance binding, and not occupy the grid cells that decrease binding, or do not contribute to ΔG . Thus, compounds with relatively high, but not necessarily the highest, observed ΔG values, but also having large negative

residuals in Figure 3, should be good templates for new ligand design. Compounds with predicted high ΔG values in their predicted active conformations may better characterize specific (enthalpic) interactions between the ligand and its receptor than compounds with higher observed ΔG values, but lower predicted binding free energies for their postulated active conformations. Up to 2 kcal/mol of possible binding free energy is predicted to be lost due to the change in ligand entropy upon binding to Gpb, for the analogues of the training set; see Figure 3.

Graphical representations of the 4D-QSAR model given by eq 2 are shown in Figures 4 and 5. The 4D-QSAR model has been developed for a grid cell resolution of 1 Å, which is the diameter of the spheres used to portray the grid cells of eq 2 in Figures 4 and 5. From an inspection of eq 2 and Figure 4, the significance of occupying GC2, GC6, and GC5 can be inferred. Occupancy of GC2(np) and GC6(any) decreases ΔG owing to the negative regression coefficients for these GCODs in eq 2. The methyl group of compound 22 (Figure 4A) occupies GC6, and the methyl group of compound 23 occupies GC2. This results in both compounds

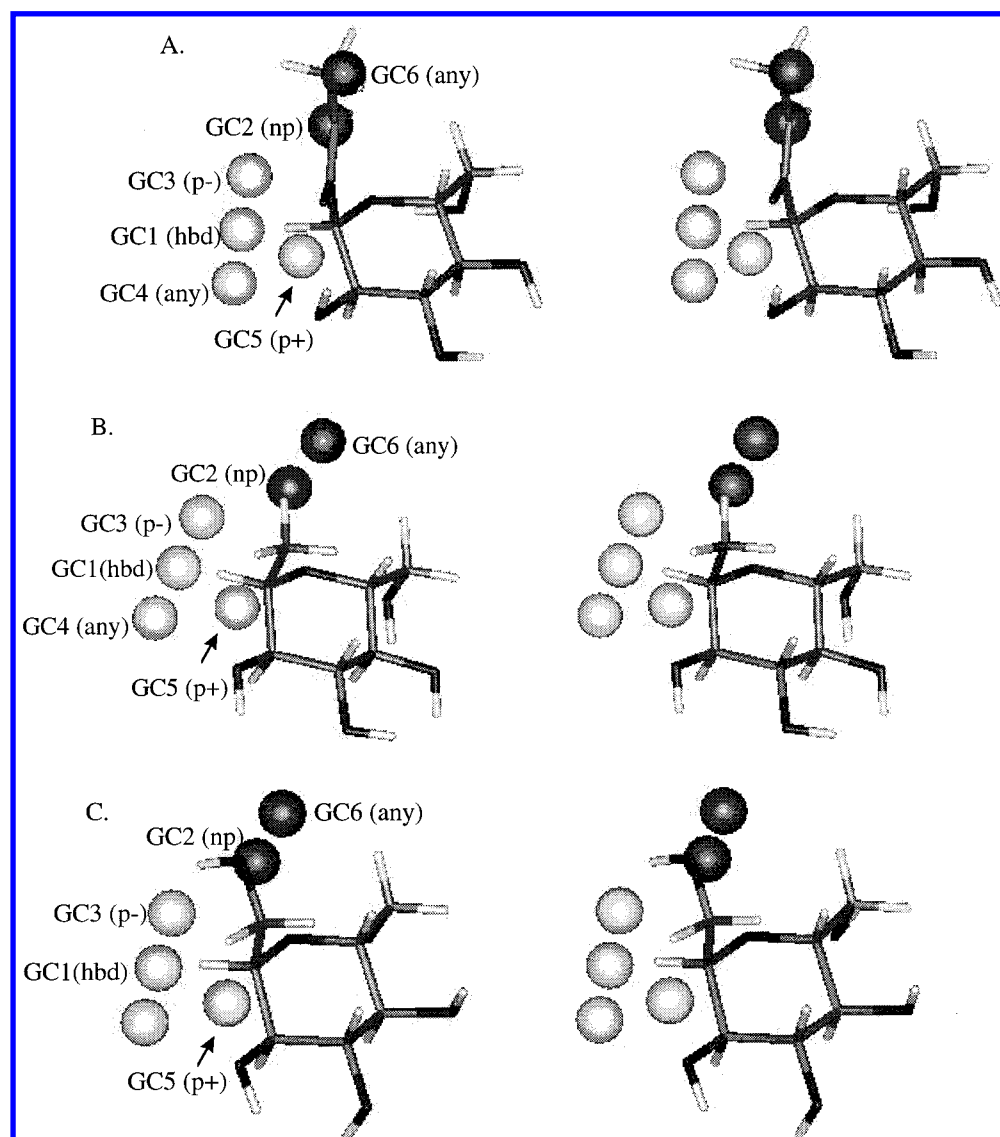


Figure 4. Stereoviews of certain analogues in their respective active conformations as predicted using eq 2. The significant grid cells of eq 2 are shown as spheres although the actual grid cells are cubes in space. ΔG enhancing grid cells are shown as lighter spheres, and grid cells which diminish ΔG are shown as dark spheres. (A) Compound 22. (B) Compound 23. (C) Compound 14. See Table 1.

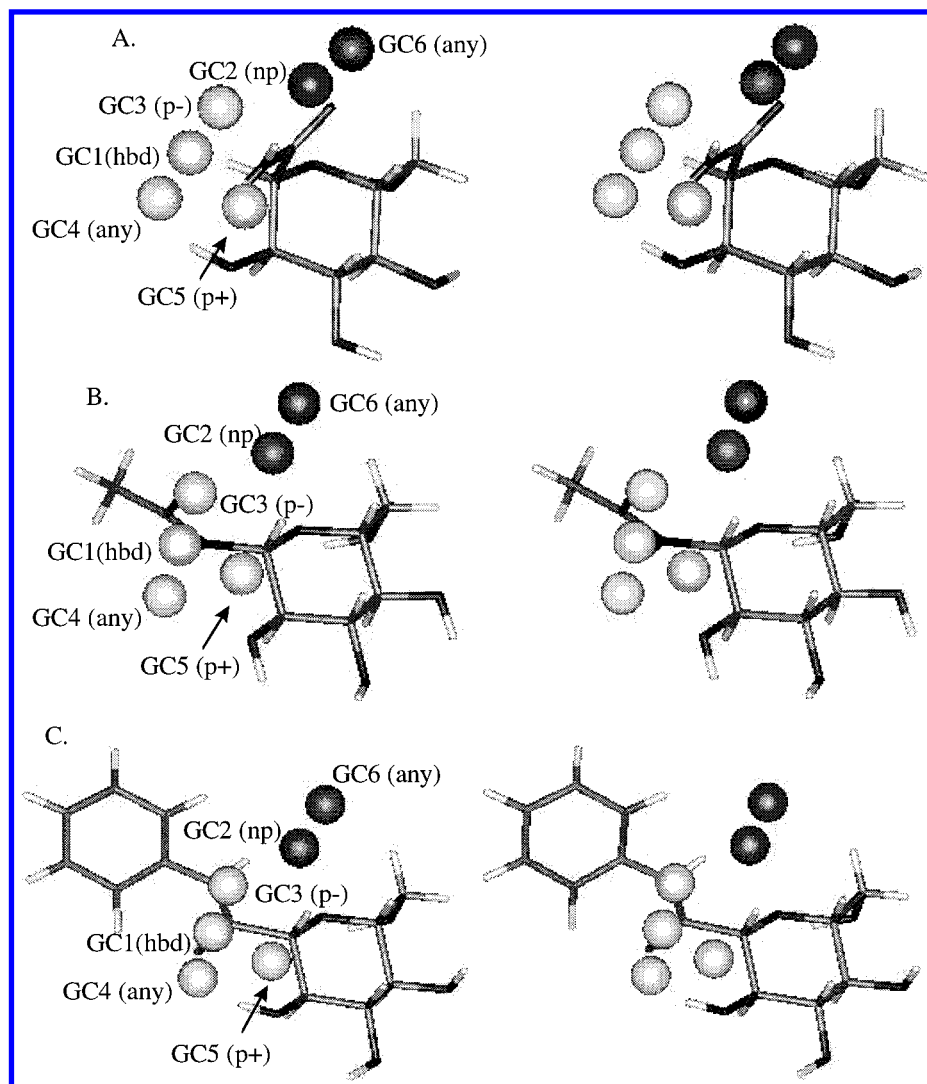


Figure 5. The same schematic plots as in Figure 4, but for (A) compound 10, (B) compound 1, and (C) compound 16. See Table 1.

22 ($\Delta G = 1.99$ kcal/mol) and 23 ($\Delta G = 1.77$ kcal/mol) being relatively weak binders. On the other hand, the hydroxyl group of the substituent of compound 14 (Figure 4C) (α -CH₂OH, β -H, $\Delta G = 3.92$ kcal/mol) occupies GC2, and is a better binder than both compounds 22 and 23 (Figure 4B). The α -amide substituent of compound 10 does not occupy the GC2 or GC6, and the H (NH₂) of the amide occupies GC5, which when occupied by a polar(+) atom enhances activity (Figure 5). Thus, compound 10 (Figure 5A) is found to have a relatively high binding free energy ($\Delta G = 4.76$ kcal/mol).

The significance of GCODs GC1 and GC3 can be demonstrated using Figure 5. The H of NH of compound 1 ($\Delta G = 6.23$ kcal/mol) occupies GC1 (Figure 5B). The occupancy of GC1 by a hydrogen bond donor (hbd) IPE is predicted to increase ΔG . The oxygen of the C=O of the β -substituent of compound 1 occupies GC3(p-) (Figure 5B). In contrast, for compound 16 (Figure 5C) a C-amide has GC1 occupied by the O of C=O, and GC3 is occupied by the H of NH. Thus, the lower ΔG value (3.14 kcal/mol) of compound 16 relative to compound 1 is "explained" by the 4D-QSAR model. This type of "incorrect" IPE grid cell occupancy of compound 16 is also observed for compound 8 (α -H, β -C(=O)NHCH₃, $\Delta G = 5.26$ kcal/mol). Also, the occupancy of GC4 by any atom leads to an increase in ΔG .

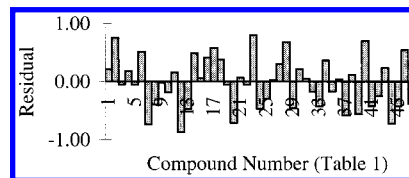


Figure 6. Observed ΔG minus the predicted ΔG using the CEP of each compound and eq 2.

Thus, the choice of β -substituents is guided by specific types of groups composing the substituent. Additional chemical groups beyond the amide group (relative to the glucose ring) on a β -substituent do not appear to significantly influence binding as can be seen for compounds 2, 3, and 4 (Table 1).

Only the postulated active conformations are shown and used in Figures 4 and 5 in making the 4D-QSAR spatial pharmacophore structure-activity analyses. The entire CEP of each compound is used to actually make activity (ΔG) predictions. The difference between the observed and the predicted ΔG values, using eq 2 and the CEP GCOD values, is plotted in Figure 6. In comparison to the predicted ΔG values using only the active conformations, see Figure 3 for reference, the predicted values using the CEPs are not significantly different from the respective observed ΔG measures. There are no outliers. These results suggest that changes in the ligand conformational entropy upon binding

Table 8. Key Ligand–Receptor Interactions between Ligand Groups Occupying the Grid Cells of the 4D-QSAR Model and the Active Site Residues

significant grid cell	ligand atom (RMS (Å), compd no.)	potential interactions with active site residues (RMS in (Å))	av distance between interacting groups (Å)
GC1	NH (0.22, 1)	main chain O of His 377 (0.30)	2.9
GC2	NH (0.21, 16) α -CH ₃ (0.38, 23) O (α -CH ₂ OH) (0.23, 14)	main chain O of His 377 (0.25)	3.6
GC3	O (C=O) (0.40, 10) O(C=O) (0.32, 1) O(C=O) (0.24, 16)	main chain N of Leu 136 (0.32) main chain N of Leu 136 (0.31) ND2 of Asn 284 (0.25)	2.9 3.0 3.0
GC5	H (NH ₂) (0.36, 10)	ND2 of Asn 284 (0.37)	3.5
GC6	CH ₃ (0.42, 22)	OD1 of Asp 283 (0.43) close to Gly 135 (0.78)	3.3 1.3–2.5

meaningfully influence the binding free energy for this class of analogues even though the “core” glucose ring is relatively rigid.

A comparison of the *receptor-independent* (RI) 4D-QSAR model, eq 2, constructed from the 4D-QSAR analysis, employing only the data in Table 1, to the geometry of the glucose analogue inhibitor–Gpb complex permits an exploration of the accuracy and information content of the RI 4D-QSAR model. The GCODs of the RI 4D-QSAR model were aligned to the geometry of Gpb by using the bound conformation of glucose (compound 15 of Table 1) as a reference. The crystal structure of the glucose–enzyme complex reveals that the α -pocket is a relatively small, water-filled pocket, and the β -pocket is lined by both polar and nonpolar groups. The potential interactions of the active site residues with ligand groups occupying the grid cells of the 4D-QSAR model are listed in Table 8. The average distances between these interacting ligand–receptor groups, and the root-mean-square (RMS) deviations of the interacting group, for the conformations explored during the MDS carried out as part of the FEFF 3D-QSAR analysis,⁷ are also given in Table 8. The ligands whose interactions with Gpb are graphically portrayed, and described below, are examples from the data set in Table 1 and have been selected to help in understanding the significance of the grid cells of the 4D-QSAR model relative to the binding model built from crystal structure information.

GC2, GC6, and GC5 are present in the α -pocket of the Gpb binding site. The significance of GC2(np) can be understood from analysis (Figures 7–10). According to the RI 4D-QSAR model, eq 2, occupancy of GC2 by any nonpolar group is detrimental to ΔG . The α -CH₃ group of compound 23 (Figure 7) occupies GC2, and the analogue is found to have a low ΔG of 1.77 kcal/mol. In contrast, the oxygen of the α -CH₂OH group of compound 14 (Figure 8), which is a moderate binding ligand, ΔG = 3.92 kcal/mol, also occupies GC2, and is within hydrogen-bonding distance from the main chain N (NH) of Leu 136. Similarly, the carbonyl oxygen of compound 10, which is shown bound to Gpb in Figure 9, is close to GC2 and is within hydrogen-bonding distance to the main chain N of Leu 136. Compound 10 binds relatively well to Gpb, ΔG = 4.76 kcal/mol.

Compound 10 also has the amide H (NH₂) occupying GC5 (Figure 9), which has a binding enhancing effect for such polar positive atoms. Compound 10 has a ΔG about 1 kcal/mol higher than that of compound 14 where the occupancy of GC5 by the hydroxyl proton is lower than that of the amide H of compound 14. The amide N (NH) of compound 10 is within hydrogen-bonding distance from OD1 of Asp

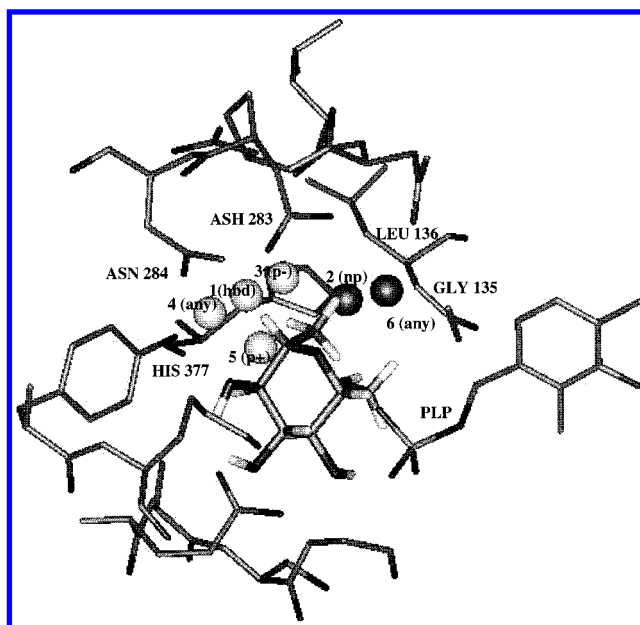


Figure 7. A representative low-energy structure of the receptor–inhibitor complex for compound 23 and the grid cells of eq 2, shown as spheres as in Figures 4 and 5, superimposed on the complex.

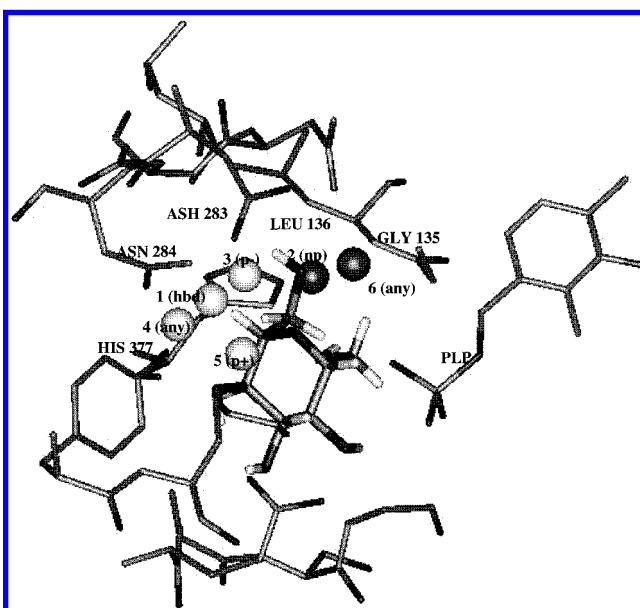


Figure 8. The same structure—4D-QSAR representation—as in Figure 7, but for compound 14 as the inhibitor.

283 (Figure 9). However, an inspection of the crystal structures of the Gpb–ligand complexes suggests that the amide N of compound 10 is hydrogen bonded through a water molecule bridge to the N (NH) of Gly 135.⁹ Thus, the

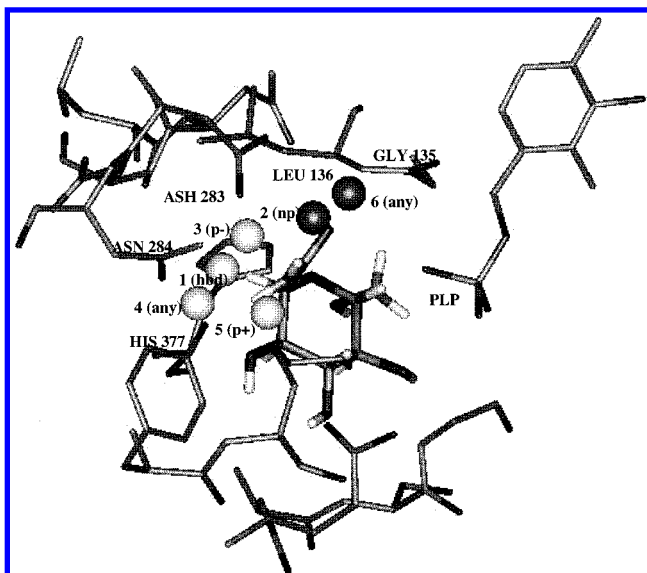


Figure 9. The same structure—4D-QSAR representation—as in Figure 7, but for compound 10 as the inhibitor.

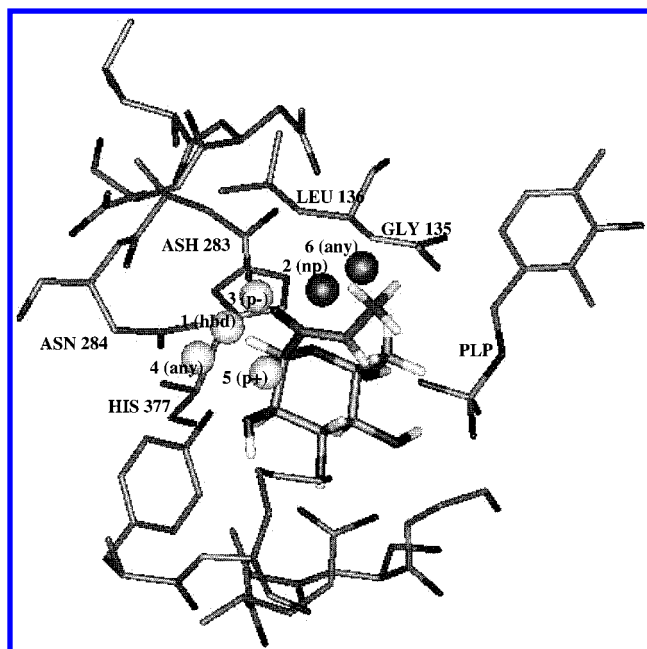


Figure 10. The same structure—4D-QSAR representation—as in Figure 7, but for compound 22 as the inhibitor.

presence of an NH_2 in the region of GC5 is significant, but the specific interaction is not identified in the 4D-QSAR model since explicit water molecules and receptor geometry are not used in RI 4D-QSAR analysis.

Grid cell GC6 is identified as a region in ligand-receptor binding space that is detrimental to ΔG when occupied by any IPE atom type. GC6 is located very close to Gly 135. The CH_3 group of the α -substituent of compound 22, for example, occupies GC6 (Figure 10).

The GCODs that correspond to the β -pocket include GC1-(hbd), GC3(p-), and GC4(all). Each of these GCODs enhances binding according to the 4D-QSAR model quantitatively expressed by eq 2. The β -substituent of compound 1 has its NH occupying GC1, and the oxygen of the carbonyl group is in the region of space near GC3. The N of the amide of compound 1 is 2.9 Å from the main chain O ($\text{C}=\text{O}$) of His 377, and the carbonyl oxygen of the substituent is 3.0

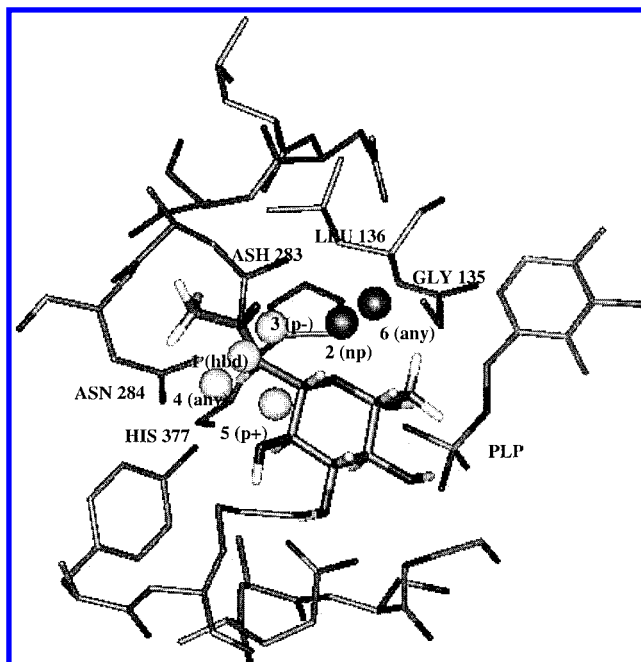


Figure 11. The same structure—4D-QSAR representation—as in Figure 7, but for compound 1 as the inhibitor.

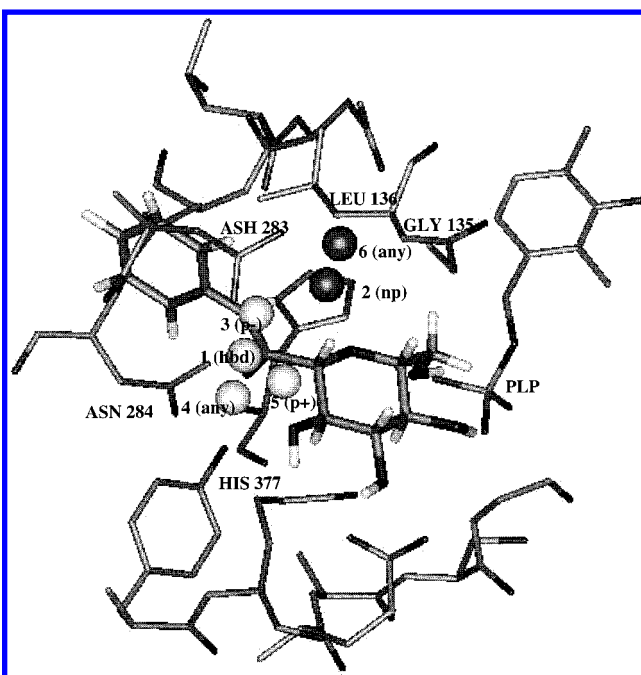


Figure 12. The same structure—4D-QSAR representation—as in Figure 7, but for compound 16 as the inhibitor.

Å from the ND2 of Asn 284 for the low-energy conformational state shown in Figure 11. On the other hand, for compound 16, whose ΔG is more than 3 kcal/mol less than that of compound 1, the β -substituent is a C-amide, the NH (amide) corresponds to GC3, and the O (carbonyl) corresponds to GC1 (Figure 12). The difference in binding free energy difference between compounds 1 and 16 can be partly attributed to the "wrong" IPE types of compound 16 occupying GC1 and GC3. The ligand O (carbonyl) is about 3.5 Å from the ND2 of Asn 284 and the amide N is about 3.6 Å from the main chain O of His 377 for the low-energy conformational state of the inhibitor-enzyme complex for compound 16 shown in Figure 12.

DISCUSSION

A comparison and demonstration of consistency of the RI 4D-QSAR model to the crystal structure complex of Gpb with bound inhibitors from the training set given in Table 1 can be considered a validation process of the 4D-QSAR model. That is the purpose for the comparison reported in this study. However, a 4D-QSAR model could also be accepted as valid, and then used as an alignment rule for defining the ligand–receptor binding mode if (a) no complex geometry is available and (b) there is reason to believe that the available geometry of the receptor is close to the ligand-bound receptor geometry. Defining ligand–receptor binding alignments for structure-based design studies is an additional and, potentially, very important use of 4D-QSAR models.

Interpretation and understanding from graphical representations of 4D-QSAR models can be difficult and confusing. This situation arises because the graphical representation corresponds to a single static conformation of a molecule while the 4D-QSAR model is developed for the ensemble of accessible conformations of the ligand. Work is in progress to devise composite graphical representations of the accessible ensemble of conformations of a molecule superimposed on a 4D-QSAR model.

The ability to estimate how the change in ligand conformational entropy upon receptor binding influences the binding free energy (or biological activity) is, perhaps, unique to 4D-QSAR analysis. Unfortunately, it is difficult to evaluate the accuracy of predicted changes in ΔG (biological activity) due to ligand entropy because the corresponding experimental data are virtually nonexistent in the open literature. However, the predictions reported in plots such as that of Figure 3, and that of Figure 10 in ref 4, may provide an impetus to determine entropic behavior in ligand–receptor binding experiments. These plots suggest, for a training set of compounds, the compound which best binds may not have the best enthalpy of binding. Other compounds may bind better with respect to enthalpy, but owing to changes in entropy upon binding, may have lower net free energies of binding. Still, the compound with the best enthalpy of binding should be the best structural template for new analogue/compound design.

Finally, 4D-QSAR models may be the best form, in terms of the type of descriptors employed, to use in the virtual high throughput screening, VHTS, of virtual libraries. This is the subject of the following paper in this issue.

ACKNOWLEDGMENT

This work was supported, in part, from resources of the Laboratory of Molecular Modeling and Design at the University of Illinois at Chicago. We appreciate the help of, and useful discussions with, Shen Wang and José Duca, of our laboratory, over the course of this study.

REFERENCES AND NOTES

- (1) Acharya, K. R.; Stuart, D. I.; Varvil, K. M.; Johnson, L. M. *Glycogen Phosphorylase b, Description of the Protein Structure*; World Scientific Press: Singapore, 1991.
- (2) Bollen, M.; Hue, L.; Stalmans, W. Effects of glucose on phosphorylase and glycogen synthase in hepatocytes from diabetic rats. *Biochem. J.* **1983**, *210*, 783–787.
- (3) Kasvinsky, P. J.; Schechosky, S.; Fletterick, R. J. Synergistic regulation of phosphorylase a by glucose and caffeine. *J. Biol. Chem.* **1978**, *253*, 9102–9106.
- (4) 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*, 10509–10524.
- (5) Albuquerque, M. G.; Hopfinger, A. J.; Barreiro, E. J.; deAlencastro, R. B. Four-dimensional quantitative structure–activity relationship analysis of a series of interphenylene 7-oxabicycloheptane oxazole thromboxane A₂ receptor antagonists. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 925–938.
- (6) Martin, J. L.; Johnson, L. N.; Withers, S. G. Comparison of the binding of glucose and glucose-1-phosphate derivatives to T-state glycogen phosphorylase. *Biochemistry* **1990**, *29*, 10745–10757.
- (7) Tokarski, J. S.; Hopfinger, A. J. Prediction of ligand–receptor binding thermodynamics by free energy force field (FEFF) 3D-QSAR analysis: Applications, to a set of peptidomimetic renin inhibitors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 792–811.
- (8) Martin, J. L.; Veluraja, K.; Ross, K.; Johnson, L. N.; Fleet, G. W. J.; Ramsden, N. G.; Bruce, I.; Orchard, M. G.; Oikonomakos, N. G.; Papageorgiou, A. C.; Leonidas, D. D.; Tsitoura, H. S. Glucose analogue inhibitors of glycogen phosphorylase: The design of potential drugs of diabetes. *Biochemistry* **1991**, *30*, 10101–10116.
- (9) Watson, K. A.; Mitchell, E. P.; Johnson, L. N.; Son, J. C.; Bichard, C. J. F.; Orchard, M. G.; Fleet, G. W. J.; Oikonomakos, N. G.; Leonidas, D. D.; Kontou, M.; Papageorgiou, A. Design of inhibitors of glycogen phosphorylase: A study of α - and β -C-glucosides and 1-thio- β -D-glucose compounds. *Biochemistry* **1994**, *33*, 5745–5758.
- (10) Watson, K. A.; Mitchell, E. P.; Johnson, L. N. Glucose analogue inhibitors of glycogen phosphorylase: from crystallographic analysis to drug prediction using GRID force-field and GOLPE variable selection. *Acta Crystallogr.* **1995**, *D51*, 458–472.
- (11) Pastor, M.; Cruciani, G.; Clementi, S. Smart Region Definition: A new way to improve the predictive ability and interpretability of three-dimensional quantitative structure–activity relationships. *J. Med. Chem.* **1997**, *40*, 1455–1464.
- (12) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
- (13) Pearlstein, R. A. CHEMLAB—II Users Guide, Version 11.1, Molecular Simulations Inc., 16 New England Executive Park, Burlington, MA 01803, 1991.
- (14) Doherty, D. C. MOLSIM User's Guide, The Chem21 Group, Inc., 1780 Wilson Dr., Lake Forest, IL 60045, 1997.
- (15) Allinger, N. L. Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V₁ and V₂ torsional terms. *J. Am. Chem. Soc.* **1977**, *99*, 8127–8134.
- (16) Hopfinger, A. J.; Pearlstein, R. A. Molecular mechanics force-field parametrization procedures. *J. Comput. Chem.* **1984**, *5*, 486–492.
- (17) Glen, W. G.; Dunn, W. J., III; Scott, D. R. Principal components analysis and partial least squares. *Tetrahedron Comput. Methods*, **1989**, *2*, 349–354.
- (18) 4D-QSAR User's Manual, Version 1.00, The Chem21 Group, Inc., 1780 Wilson Dr., Lake Forest, IL 60045 1997.
- (19) Rogers, D. G/SPLINES: A hybrid of Friedman's multivariate adaptive regression splines (MARS) algorithm with Holland's genetic algorithm. *The Proceedings of the Fourth International Conference on Genetic Algorithms*; San Diego, 1991; pp 38–46.
- (20) Rogers, D.; Hopfinger, A. J. Applications of genetic function approximation to quantitative structure–activity relationships and quantitative structure–property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854–866.
- (21) Rogers, D. WOLF Reference Manual, Version 5.5, Molecular Simulation Inc., 1994.
- (22) Friedman, J. Multivariate adaptive regression splines. Technical Report No. 102; Laboratory for Computational Statistics, Department of Statistics, Stanford University, Stanford, CA, November 1988 (revised August 1990).
- (23) Hansch, C.; Fujita, T. ρ - σ - π Analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616–1626.
- (24) (a) Rogers, D. Personal communication, 1997. (b) Applied in ref 5.

CI9900332