

A Strategy for the Incorporation of Water Molecules Present in a Ligand Binding Site into a Three-Dimensional Quantitative Structure–Activity Relationship Analysis

Manuel Pastor,^{*,†} Gabriele Cruciani,[‡] and Kimberly A. Watson[§]

Department of Pharmacology, University of Alcala, 28871 Alcala de Henares, Spain, Department of Chemistry, University of Perugia, Via Elce di Sotto, 8, 06123 Perugia, Italy, Laboratory of Molecular Biophysics, The Rex Richards Building, South Parks Road, Oxford OX 3QU, U.K.

Received April 24, 1997[⊗]

Water present in a ligand binding site of a protein has been recognized to play a major role in ligand–protein interactions. To date, rational drug design techniques do not usually incorporate the effect of these water molecules into the design strategy. This work represents a new strategy for including water molecules into a three-dimensional quantitative structure–activity relationship analysis using a set of glucose analogue inhibitors of glycogen phosphorylase (GP). In this series, the structures of the ligand–enzyme complexes have been solved by X-ray crystallography, and the positions of the ligands and the water molecules at the ligand binding site are known. For the structure–activity analysis, some water molecules adjacent to the ligands were included into an assembly which encompasses both the inhibitor and the water involved in the ligand–enzyme interaction. The mobility of some water molecules at the ligand binding site of GP gives rise to differences in the ligand–water assembly which have been accounted for using a simulation study involving force-field energy calculations. The assembly of ligand plus water was used in a GRID/GOLPE analysis, and the models obtained compare favorably with equivalent models when water was excluded. Both models were analyzed in detail and compared with the crystallographic structures of the ligand–enzyme complexes in order to evaluate their ability to reproduce the experimental observations. The results demonstrate that incorporation of water molecules into the analysis improves the predictive ability of the models and makes them easier to interpret. The information obtained from interpretation of the models is in good agreement with the conclusions derived from the structural analysis of the complexes and offers valuable insights into new characteristics of the ligands which may be exploited for the design of more potent inhibitors.

Introduction

The objective of most drug design projects is to obtain compounds which bind tightly to a target bio-molecule. When the structure of this target is known it would be useful to incorporate all the information available in order to predict the binding strength of a potential ligand to its target. Unfortunately, molecular interactions are complex, and many different terms¹ should be taken into account in order to fully quantify the free energy of interaction that drives the binding process. Three-dimensional quantitative structure–activity relationship (3D-QSAR) techniques therefore provide an alternative approach. Originally, QSAR methods were intended to study the correlation between the activity and the structure in a series of congeneric compounds, but since the activity directly depends upon the affinity of the ligands for their target, they can be used to study ligand binding as well. With respect to structure-based drug design (SBDD) techniques, 3D-QSAR methods have the advantage of dealing only with the differences in affinity of a series of compounds. In this case, the energy of interaction of each compound is not relevant because some of the terms describing this energy (e.g. receptor desolvation, receptor entropy loss, etc.) take approximately the same value for every compound, and therefore cancel when only the differences are consid-

ered. For this reason, QSAR approaches are simpler and even very crude approaches such as the classical Hansch analysis gives valid results.² Typically, QSAR methods use all the information available, and more recent approaches^{3,4} incorporate the structures of the ligand–receptor complexes into the analysis.

QSAR techniques largely depend upon the assumption that every compound in the series interacts with the same target molecule and in the same way. Provided that the compounds are not too dissimilar, this is a reasonable assumption. However, receptors are not static entities, and they can change their conformation in order to accommodate and optimize ligand binding. Perhaps, an extreme example of receptor plasticity is the presence of water molecules at the ligand binding site. Water molecules are known to play a significant role in mediating ligand–protein binding.^{5–7} Since water molecules are not covalently bound to residues of a receptor, they are therefore susceptible to displacement from and/or movement within the receptor upon ligand binding. The effect of these water molecules present at the binding site in a QSAR analysis depends critically on the variability of their position. If the water molecules appear in exactly the same position for every compound in a series, they can be considered as a constant part of the receptor and do not require special methodological consideration. On the other hand, if they move upon ligand binding or are displaced in only some of the ligand–receptor complexes, then they

[†] University of Alcala.

[‡] University of Perugia.

[§] Laboratory of Molecular Biophysics.

[⊗] Abstract published in *Advance ACS Abstracts*, November 1, 1997.

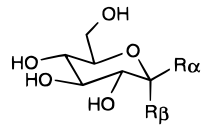
violate the basic assumption of the common receptor and therefore require special treatment in the QSAR methodology.

Recently, attention is being focused on water present at the ligand binding site,⁸⁻¹⁰ and there is a claim for novel drug design methodologies that would take into account the effect of such water molecules. The work reported here describes the GRID/GOLPE 3D-QSAR analysis of a series of 47 glucose analogue inhibitors of glycogen phosphorylase (GP) shown in Table 1, in which some water molecules present at the ligand binding site have been explicitly incorporated into the QSAR analysis. This series has the advantage that the structure of every compound in complex with GP has been solved to 2.4 Å resolution using X-ray crystallographic techniques.¹¹ From these studies, it appears that the ligand binding site is hydrated and contains several water molecules, some of which serve to bridge hydrogen bonds between the ligand and the enzyme. An initial inspection of the complexes revealed that most of the water molecules were present at approximately the same positions in every complex but others have been displaced by the ligands to different positions or completely removed. In such a situation, the 3D-QSAR analysis of the ligands alone will give an unrealistic picture of the receptor, because the position of the water changes from complex to complex. Therefore, instead of using only the ligands, the 3D-QSAR analysis was carried out using an assembly that included each ligand plus the water molecules directly interacting with the ligand and the receptor. The rationale behind this approach was that the assembly of ligand plus water would give a better representation of both the ligand-protein interactions and the water-mediated interactions than just the ligand on its own, and that this would yield better models.

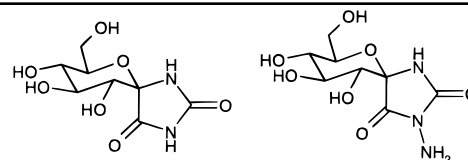
Objectives

In the present work, a 3D-QSAR analysis was performed using assemblies of ligands plus the surrounding water molecules which participate in the ligand binding. The objectives were 2-fold. First, the method was examined in order to determine whether there were advantages to the inclusion of water molecules in the QSAR model. Since the structures of the complexes are known, comparison between the areas that the QSAR model highlights as important for the activity and the crystallographic structures of the ligand-enzyme complexes should show the suitability of this new method. From this analysis it should be possible to assess whether this approach would be of general use and therefore used as an alternative to the classical 3D-QSAR analysis. Second, the QSAR models obtained were used to gain further insight into the ligand binding process. Although GP inhibitors are well-known and the structures of the complexes have been studied in detail,¹¹⁻¹⁶ the QSAR method approaches the problem from a different point of view and therefore has the potential to offer new information regarding ligand binding. Similarity in the information obtained from both methods (structural analysis of the complexes versus QSAR) should confirm the suitability of this new QSAR method, while differences may suggest further research in directions which the structural analysis alone may never have revealed. In addition, it is hoped that this new QSAR method will help to give a better

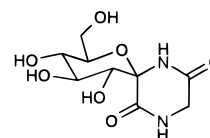
Table 1. Series of Inhibitors of Glycogen Phosphorylase b^{11-16}



no.	substituent at C1 position		pK_i (mM)
	R_α	R_β	
1	OH	H	2.77
2	C(=O)NH ₂	H	3.43
3	C(=O)NHNH ₂	H	2.52
4	COOCH ₃	H	1.62
5	CH ₂ NH ₃ ⁺	H	1.46
6	CH ₂ N ₃	H	1.65
7	CH ₂ OH	H	2.82
8	H	<i>O</i> -(1-6)-D-glucose	1.79
9	H	C(=O)NH ₂	3.36
10	H	C(=O)NHCH ₃	3.80
11	H	C(=O)NHCH ₂ CH ₂ OH	2.59
12	H	C(=O)NHPh	2.27
13	H	C(=O)NH-4-OHPh	2.36
14	H	C(=O)NHNH ₂	3.40
15	H	C(=O)NHNHCH ₃	2.75
16	H	C(=O)NHCH ₂ CF ₃	2.09
17	H	C(=O)NH-cyclopropyl	2.89
18	H	COOCH ₃	2.55
19	H	CH ₂ NH ₃ ⁺	1.78
20	H	CH ₂ CH ₂ NH ₃ ⁺	2.35
21	H	CH ₂ N ₃	1.82
22	H	CH ₂ CN	2.05
23	H	NHC(=O)NH ₂	3.85
24	H	NHC(=O)CH ₃	4.50
25	H	NHC(=O)CH ₂ CH ₃	4.41
26	H	NHC(=O)CH ₂ CH ₂ CH ₃	4.03
27	H	NHC(=O)CH ₂ Cl	4.35
28	H	NHC(=O)CH ₂ Br	4.36
29	H	NHC(=O)CH ₂ NH ₂	3.43
30	H	NHC(=O)Ph	4.09
31	H	NHC(=O)CH ₂ NHCOCH ₃	3.00
32	H	NHCOOCH ₂ Ph	3.46
33	H	CH ₂ OSO ₂ CH ₃	2.32
34	H	indoxyl- β -D-glucoside	2.59
35	C(=O)NH ₂	NHCOOCH ₃	4.80
36	OH	CH ₂ OH	1.80
37	OH	CH ₂ N ₃	2.13
38	OH	CH ₂ CN	2.12
39	OH	CH ₂ OSO ₂ CH ₃	2.43
40	H	SH	3.00
41	H	SCH ₂ C(=O)NH ₂	1.68
42	H	SCH ₂ C(=O)NHPh	2.44
43	H	SCH ₂ C(=O)NH-2,4-(F) ₂ Ph	1.72
44	5-thio- α -D-glucose		2.70



45 $pK_i = 5.52$ mM 46 $pK_i = 3.84$ mM



47 $pK_i = 4.22$ mM

understanding of the role of the water present at the active site, and will help assess which of these water molecules should be displaced and which should be preserved in order to obtain more active compounds.

Materials and Methods

Materials. The series of 47 glucose and thio-glucose derivatives reported in Table 1 were used for the QSAR analysis. The X-ray structure and biochemical data of each complex have been determined during the course of a long-term project aiming to obtain inhibitors of glycogen phosphorylase with potential therapeutic activity as antidiabetic drugs.^{11–16} Since the X-ray crystal structures of each ligand–GP complex have been solved, this series of complexes are particularly well suited for methodological studies, due to the lack of conformational and superimposition problems which constitute the main problem in most comparative molecular field analysis (CoMFA) and CoMFA-like analyses.¹⁷ Subsets of this series of ligand–GP complexes have been used previously in the development of novel 3D-QSAR methodologies.^{18,19}

The compounds in this series have in common a glucopyranose ring, with different substitution at the C1 position in the α and/or the β configurations. The size of the C1 substituents is very different and ranges from compounds as small as α -D-glucose **1** to others as large as gentiobiose **8**. Most C1 substituents are polar in nature, specifically designed to establish hydrogen bonds either directly to protein residues or through water molecules to the protein. Some hydrophobic moieties were designed to displace certain water molecules and to make favorable van der Waals interactions with the protein.

Biological Measurements. The compounds in this series were designed with the expectation of modulating the activity of GP in the same way that glucose, one of its natural regulators helps to control the activity of this enzyme. Glucose is an inhibitor that binds to the catalytic site in competition with substrate but also stabilizes the T state (inactive) form of the enzyme by making specific interactions with a loop of chain (280s loop) that blocks access to the catalytic site. Detailed discussion about the effect of glucose and the regulatory activity of GP in controlling hepatic glucose output can be found elsewhere.¹² For this series of compounds, the biological activity has been derived from kinetic studies measuring the changes in enzymatic activity that occurs upon incubation of GP with each of the different compounds at varying concentrations.¹³ From these studies, kinetic inhibition constants (K_i) have been calculated for each compound. The relationship between dissociation constants (K_D) obtained from calorimetric measurements and the inhibition constant for glucose has been discussed.^{13,20} Since the compounds in this series are primarily glucose derivatives, for comparative purposes, differences in K_i are likely to correlate with differences in binding energy.¹³ For the QSAR model, the negative base-10 logarithm of the K_i (pK_i) has been used as the dependent variable. The activities cover a wide range of more than five logarithmic units (from 1.46 to 5.52).

Crystallographic Analysis. The structures of each ligand–GP complex were obtained by X-ray diffraction experiments. The complexes were obtained in turn, by soaking approximately 100 mM of the ligand into a preformed rabbit muscle GPb crystal. The ligand position was determined by difference Fourier electron-density maps calculated with respect to the native GPb. In most cases, the analysis of the $F_o - F_c$ maps clearly showed the location and orientation of the ligand molecules ($>3\sigma$). The overall difference electron-density map was checked for any other changes, with particular attention to solvent molecules and protein residues within 20 Å of the active site center. The most significant changes were observed at the ligand binding site. The complexes were refined using X-PLOR energy and crystallographic least-squares minimization²¹ to R values under 0.19. Estimates of the precision of coordinates indicate errors of the order of 0.2–0.4 Å in positional coordinates. Further details of the crystallographic analyses have been reported elsewhere.^{11–16}

In this study, the structures of the ligand–enzyme complexes have been used at two different stages of the analysis. First, the ligands and water molecules were used, as found in the crystallographic structures, to construct the QSAR models. Second, knowledge of the structure of the ligand binding site was used to validate the interpretation derived from the analysis.

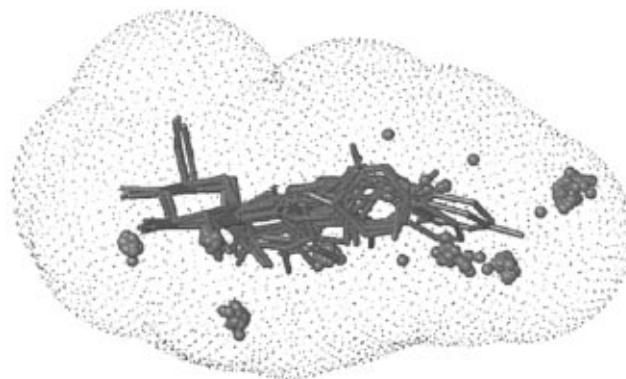


Figure 1. Superimposition of the ligands **1–47** plus the water molecules included in the analysis (small spheres). The surface delimited by the points represents the 4 Å distance criterion used to decide which water molecules to include in the QSAR analysis. The hydrogen atoms of the water molecules and of the ligands have been removed for clarity.

Water Molecules Present at the Ligand Binding Site.

In GP, the glucose analogues bind at the catalytic site, which is buried about 15 Å from the surface, and close to the cofactor, pyridoxal phosphate. In the T state form of the enzyme, access to the catalytic site is blocked by a loop of chain from residues 282 to 286 (termed the 280s loop). Upon activation, the shift to the active R state causes displacement of this loop and promotes conformational changes which replace the acidic side chain of Asp 283 from the catalytic site with the side chain of Arg 569, essential for the recognition of the phosphate moiety of α -D-glucose-1-phosphate (Glc-1-P). In the glucose–GP complex, the binding is dominated by hydrogen bonds from each of the peripheral hydroxyl groups to protein atoms, such that each hydroxyl is involved both as a donor and as an acceptor of hydrogen bonds.^{13,14} The bond between the α -1-OH and Asp 283 is mediated by Wat 872. Wat 897 also participates in the binding of the glucopyranose ring and mediates an hydrogen bond between 4-OH and the phosphate group of the pyridoxal phosphate cofactor. The ligand binding site is characterized by two pockets: a large channel adjacent to the 280s loop, which is accessible to β -C1-substituents (the β pocket), and a small water-filled channel accessible to α -C1-substituents (the α pocket). Both of these regions have been explored as targets for additional groups in order to obtain tighter binding inhibitors.¹⁴

At the ligand binding site, water either participates in the binding of the glucopyranose ring or acts to fill the two channels. Examination and comparison of each ligand and their surrounding water molecules show that in most cases each complex contains the same number of water molecules as observed in the glucose complex, and that these water molecules are in approximately the same positions. Differences appear when a ligand displaces one or more water molecules from the active site and only the binding of **45** gives rise to a significantly different water pattern.

To build the assemblies of ligand plus water it is necessary to define a set of criteria with which to assess whether to include a particular water molecule in the assembly. Ideally, the assemblies should include water molecules that participate directly in ligand binding and therefore can be easily identified by their proximity to each ligand. The ligands, as they appear in the crystal structures, were superimposed, and a volume extending 4 Å from their atomic centers was constructed (see Figure 1). The assemblies of ligand plus water included only the water molecules for which the position of its oxygen nuclei was enclosed by this 4 Å volume cutoff. In the present series, the volume cutoff criteria included a minimum of five and a maximum of eight water molecules. The water molecules included in the glucose–water assembly have been used as a standard for the nomenclature. The indexes assigned to these water molecules, as they appear in the glucose–GP complex deposited in the Brookhaven Protein Data Bank (entry PDB2GPB), are listed in Table 2. The coordinates and

Table 2. Water Molecules Included in the GRID/GOLPE Analysis for the Glucose-GP Complex

index ^a	residue ^b	x	y	z	B factor
847	HOH 25	38.005	24.258	31.448	22.54
872	HOH 279	30.543	25.687	28.630	25.56
887	HOH 421	31.460	23.749	33.403	29.31
890	HOH 453	37.827	27.399	33.321	44.09
892	HOH 473	41.041	26.410	36.863	52.64
897	HOH 522	29.372	21.633	26.075	13.71
987	HOH 422	35.641	27.214	37.089	20.44
990	HOH 458	36.163	24.588	23.957	29.75

^a Code used throughout this manuscript and in previous work.^{11–16} ^b Residue name and number as appears in the Brookhaven Protein Data Bank entry PDB2GPB.

isotropic *B* factors of these water molecules are also given in Table 2 for clarity.

These assemblies were then analyzed with particular attention to the presence or absence of the water molecules in each case. In most of the complexes, the presence or absence of each water molecule was easily justifiable, but some differences were observed:

1. Some water molecules in the complex of compound **45** (the most active) follow a slightly different pattern. Wat 890 is closer to the ligand, and there is a water molecule not found in other complexes near His 341, within hydrogen-bonding distance of Wat 847 (2.95 Å), O of Ala 383 (3.17 Å) and N2 of His 341 (2.85 Å).

2. Complex **24**-GP lacks Wat 872. From an inspection of the structure, and a comparison with other similar complexes, there seemed to be no apparent reason for the absence of this particular water molecule.

3. Wat 890 is absent in 19 of the 47 complexes. Inspection of these complexes shows a clear displacement by the ligand in only three of them (**32**, **42**, and **43**). The mobility of this water molecule is reflected in its *B* factor for the various ligand-enzyme complexes (from 36.9 to 69.1).

It should be noted that the result of a crystallographic analysis represents an observable average; therefore, the problem of "missing" water will not be an error in the structures but a consequence of the nature of the solution.²² It has been shown that buried waters generally are conserved either in crystal structures of homologous proteins or in different crystal structures of the same protein.²³ However, some differences among different crystal structures (even of the same protein) occasionally arise; these differences may be a result of partial occupancy and/or difficulty in assigning the water oxygens at medium resolution since inherent noise levels in the data prevent unambiguous assignment.^{22–24} In this present series of inhibitor-GP complexes, it is possible that both of these effects are responsible for the absence of Wat 890 in some complexes and for the absence of Wat 872 in complex **24**-GP.

GRID Calculations for Validation of Water Molecules at the Active Site. Differences in the presence or absence of individual water molecules for each complex structure contained in the QSAR model will have a negative effect on the quality of the model and will lead to inconsistencies in the final analysis. Since the purpose of this new strategy was to incorporate the water molecules into the QSAR model, then in those regions where the displacement of water molecules was not explained by the X-ray structural analysis, the program GRID^{25,26} was used as an alternative approach to assess the presence of these water molecules. The intention was not to carry out a solvent simulation analysis, but to assess whether the presence of a water molecule in its observed crystallographic position was energetically favorable using an objective criterion. Particular attention was focused on those regions in which a water molecule was present in the majority of complexes but inexplicably absent in only a few complexes.

GRID works by defining a 3D grid of points over an area of interest. At each node of the grid, the energy between the target molecule and a probe (E_{xyz}) is calculated as indicated in eq 1

$$E_{xyz} = \sum E_{EL} + \sum E_{HB} + \sum E_{LJ} \quad (1)$$

where E_{EL} is the appropriately modified electrostatic energy, E_{HB} is the hydrogen-bonding energy, and E_{LJ} is the Lennard-Jones potential energy between the constituent atoms of the probe and all the atoms of the target. The first GRID calculation focused on the region where the water molecule Wat 890 is located. Calculations were carried out in a box with dimensions $4 \times 4 \times 4$ Å centered on the positions where Wat 890 was found in the X-ray crystal structures of the ligand-GP complexes, using a grid spacing of 0.1 Å and a water (OH2) probe. As expected, the calculations show no unique minimum, but instead show a bulky area where water molecules can interact favorably with the different residues of the protein within this region. In particular, inside the GRID cage selected, water molecules can potentially make as many as eight different hydrogen bonds directly to the protein or through other water molecules (O Ala 383; N Phe 285; O Phe 285; O Asn 282; N His 341; Wat 847; Wat 892; and Wat 987). In the next step, the program MINIM, included in the GRID package,²⁵ was used to find favorable minima inside the same GRID box. A minimum is defined by program MINIM as a grid point which is completely surrounded by points with larger GRID energies (i.e. more positive values). For the different ligand-GP complexes, the number of minima, using an energy cutoff of -5.0 kcal/mol, ranged from 3 to 36. Complexes with compounds **32**, **42**, and **43**, in which the ligands clearly overlap the position occupied by Wat 890 in other complexes, had no minima under a $+5.0$ kcal/mol cutoff. There were no apparent differences in the results between the complexes containing Wat 890 and those for which this water molecule was not present.

To simulate the positions of the water molecules as observed in the crystal structures, the position derived from the weighted average (using the minus interaction energy as the weight) of the GRID minima was used. It is obvious that no perfect correlation should be expected between the observed and the calculated positions of the water molecules, therefore in order to estimate the uncertainty associated with the calculated water positions the isotropic *B* factors were used to compute the mean displacement \bar{U} using eq 2.⁸

$$B = 8\pi^2 \bar{U}^2 \quad (2)$$

The results of the calculations are represented in Figure 2. The simulation using the GRID minima gives positions for Wat 890 within 1.0 Å of the experimental positions, for most of the compounds, and an average distance of 0.7 Å. Most of the estimations are within the distance defined by the mean displacement \bar{U} . These results clearly show that in this particular series of inhibitors of GP the simulations can be considered acceptable estimations of the experimental positions. Figure 3, for example, shows the results of the GRID energy calculation obtained for the glucose-GP complex. The figure highlights a broad area enveloping the grid points where the method predicts a favorable interaction energy with a water molecule under -9.0 kcal/mol. The experimental position for Wat 890 in the crystal structure and the position calculated for the simulated Wat 890 were included. Neighboring residues, able to make polar interactions with this water molecule, were also included in the figure as a reference.

With respect to Wat 872 in the ligand-GP complex of *N*-acetylglucopyranosylamine **24**, a similar approach was followed. GRID was used to calculate favorable positions for a water probe in a $2 \times 2 \times 2$ Å box centered on the crystallographic position of Wat 872 as located in the complex of urea-*ido*-glucopyranose **23** with GP (the most similar compound to **24**). The procedure was carried out in parallel for both compounds, and Wat 872, as found in **23**-GP, was used as a reference in order to evaluate the accuracy of the simulation. As in the previous case for Wat 890, the results of the GRID calculation were analyzed using the MINIM option and the weighted averages were computed. The most favorable energies of interaction were comparable in both cases (-17.8 kcal/mol for **23** and -16.3 kcal/mol for **24**). With

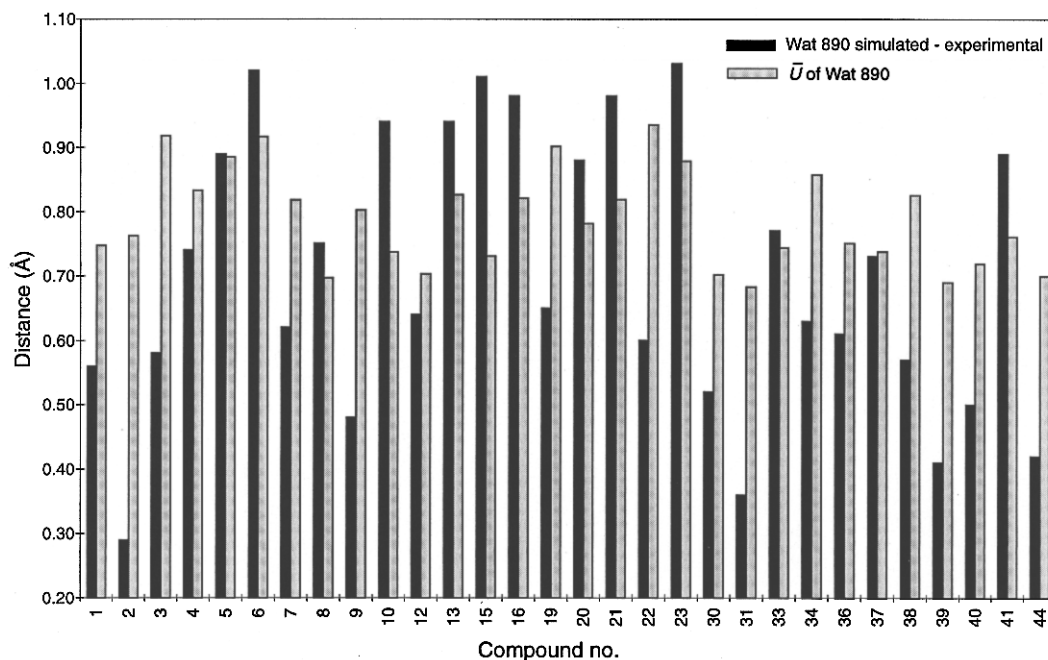


Figure 2. The black bars represent the distances (in Å) between the positions obtained in the simulation analysis (see text) for Wat 890 and the positions of Wat 890 as found in the crystal structure for each of the different compounds. The distances reported measure the separation in Å between the oxygen nuclei of both water molecules. The gray bars represent the mean displacement (\bar{U}) of Wat 890, calculated from the isotropic B factor of the individual molecule in each ligand–enzyme crystal structure, which represents an estimation of the mobility of Wat 890 in each complex.

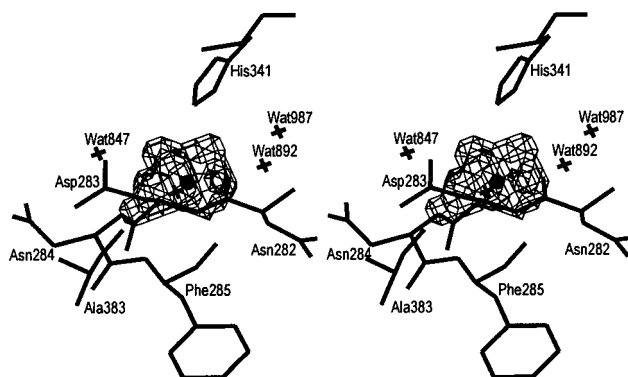


Figure 3. Stereoplot showing the results of the GRID calculations using a water probe for the glucose–GP complex. The grid region represents areas of the protein where the method estimates an energy of interaction with a water probe under -9.0 kcal/mol, corresponding to favorable regions for the interaction of a water molecule with the enzyme. The region encompasses both the experimental (+) and the predicted (●) positions of Wat 890. The figure also shows some residues and water molecules which are able to form potential hydrogen bonds or polar interactions with the water probe.

respect to the reference complex **23**, the calculated position of Wat 872 was only 0.37 Å away from the position found in the X-ray crystal structure.

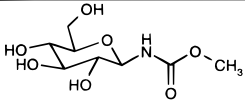
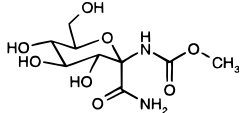
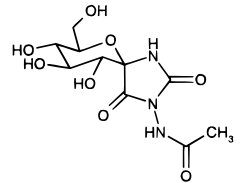
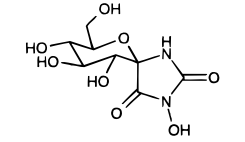
GRID/GOLPE Analysis. To investigate the effect of the incorporation of water into the QSAR analysis, three different data sets were studied. The first data set (DRY) included only the ligand molecules. The second data set (WAT1) included the ligands plus all the water molecules as found in the crystal structures of the ligand–enzyme complexes using the distance criteria described in the previous section (i.e. within 4 Å from the positions of any atom of each compound included in the series). In the third data set (WAT2) the GRID simulation studies were used to account for the differences observed in the positions of the water molecules which fell within the distance criterion. In particular, the positions predicted by the simulation for Wat 890 were used to add a water molecule in those complexes where this water was not observed and to

replace the crystallographic water Wat 890 with the simulated position in the remaining compounds. Also using this strategy, a simulated Wat 872 was added to the complex **24**–GP.

The GRID/GOLPE analysis as performed in this and other work^{11,18,19} is essentially a 3D-QSAR, CoMFA-like methodology.²⁸ As a first step, the program GRID^{25–27} was used to compute the energy of interaction between a small molecule (the probe) and each compound of the series (the targets) within the nodes of a 3D grid. In this study, a phenolic hydroxyl group probe (OH) was used. This group is capable of donating and accepting one hydrogen bond. The electronic configuration of the OH probe is defined such that it interacts with the π -system of the aromatic ring, making the hydrogen-bonding pattern different from that of an aliphatic hydroxyl probe. The OH probe shows an intermediate polarizability value between those of other similar oxygen probes, and it makes strong hydrogen-bonding interactions, which may account for the shape of the interaction regions with the molecular structures. In a GRID analysis, molecular flexibility is allowed and controlled by a special directive in the program. However, in the version used for this analysis, the heavy atoms are considered in fixed positions, but thermal motion of the hydrogen-bonding hydrogen atoms and lone-pair electrons are taken into account.²⁷ The OH probe was chosen because the binding site of GP is known to be predominantly hydrophilic. Also, in previous studies^{11,18,19} the OH probe was shown to give the best results for this series and therefore serves as a good choice for evaluating this new procedure involving water molecules. The size of the GRID box used for the calculations was defined in such a way that it extended approximately 4 Å beyond each of the ligand molecules in each dimension. This resulted in a box with dimensions $22 \times 20 \times 18$ Å. GRID calculations were performed using a grid spacing of 1 Å, thus giving 7920 probe–target interactions for each compound, which were unfolded to produce a one-dimensional vector of variables. Each of these vectors (one for each compound in the series) were assembled into a matrix of 47 rows and 7920 columns, called the \mathbf{X} matrix. A cutoff of $+5$ kcal/mol was applied to the GRID data to produce a more symmetrical distribution of energy values.

This matrix was then imported into GOLPE 3.0 and pretreated by zeroing those values with absolute values smaller than 0.1 kcal/mol, and removing any variables with a

Table 3. Set of Inhibitors of Glycogen Phosphorylase Used for the External Prediction Analysis

no.	Compound	Activity (pK_i)			
		experim.	DRY	WAT1	WAT2
48		4.07	4.21	4.35	4.26
49		3.50	3.89	4.42	4.01
50		3.26	1.73	4.40	3.75
51		4.41	3.40	3.88	3.93
SDEP ^a		-	0.94	0.79	0.43

^aStandard Deviation of Error of Predictions.

standard deviation below 0.1. In addition, variables that take only two or three values and present a skewed distribution (for example, when one of the values is taken by only one or two molecules in the series) were also removed. After this pretreatment, the data sets still contained between 2000 and 2900 variables, and the ratio of variables to molecules was still large (42:1 and 62:1 respectively). At this point the smart region definition/GOLPE (SRD/GOLPE), a novel variable selection procedure,¹⁹ was used to carry out the variable selection on groups of variables chosen according to their positions in 3D space. The SRD algorithm was applied as described in ref 19 to the pretreated **X** matrix with the following parameters: 457 seeds selected in the PLS weights space, a critical distance cutoff of 1.0 Å, and a collapsing distance cutoff of 2.0 Å. The regions found were then used in a fractional factorial design (FFD) variable selection procedure,^{29–31} in successive iterations (typically two or three), until the ratio of variable to molecules decreased below 10:1.

Model Validation. The predictive ability of the models was first evaluated by cross-validation, using five groups of approximately the same size in which the objects were assigned randomly. The whole procedure was repeated 20 times. This cross-validation procedure provides a safer alternative to the more widely preferred leave one out (LOO) method and gives more conservative results: a smaller cross-validated squared correlation coefficient (q^2) and a higher standard deviation of error of predictions (SDEP).³¹

In addition to the internal validation, four new compounds, not available at the beginning of this work, were later used as an external validation or prediction set (Table 3). Compounds **48–51** underwent exactly the same treatment described for the training set **1–47**, and three data sets equivalent to DRY, WAT1, and WAT2 (named PDRY, PWAT1, and PWAT2) were generated for these compounds. PDRY does not contain water molecules, PWAT1 contains only those waters as found in the crystal structures of the complexes, and PWAT2 was altered (as in data set WAT2) for differences in the water positions for each of the ligand–GP complexes. For

prediction data set PWAT2, positions for the waters which were absent (**48**, Wat 887; **49**, Wat 987; **50**, Wat 847, Wat 892, and Wat 987; **51**, Wat 847 and Wat 887) were simulated following exactly the same procedure described for Wat 872 in complex **24**–GP. Similarly, a simulated position for Wat 890 was determined for every compound in this prediction data set, as described for data set WAT2.

Results

Comparison of the Models. For each of the three data sets described above (DRY, WAT1, WAT2), partial least squares (PLS) models were obtained both without variable selection and with SRD/GOLPE variable selection. The predictive ability of the models was evaluated by internal validation (cross-validation) and by external validation using the four compounds in the prediction data sets PDRY, PWAT1, and PWAT2. The results of the different PLS analyses are shown in Table 4 and Figure 4. The quality of the models in fitting is expressed as the squared correlation coefficient (r^2). The result of the internal validation is expressed as the cross-validated squared correlation coefficient (q^2), and as the standard deviation of error of predictions (SDEP). These quantities have been described in detail elsewhere.²⁹ The external SDEP expresses the ability of the models to predict the activity of the compounds in the external prediction sets (PDRY, PWAT1, and PWAT2).

The results show that application of SRD/GOLPE variable selection improves the quality indexes for all the data sets (Table 4). This is consistent with previous reports, which also show that these models are easier

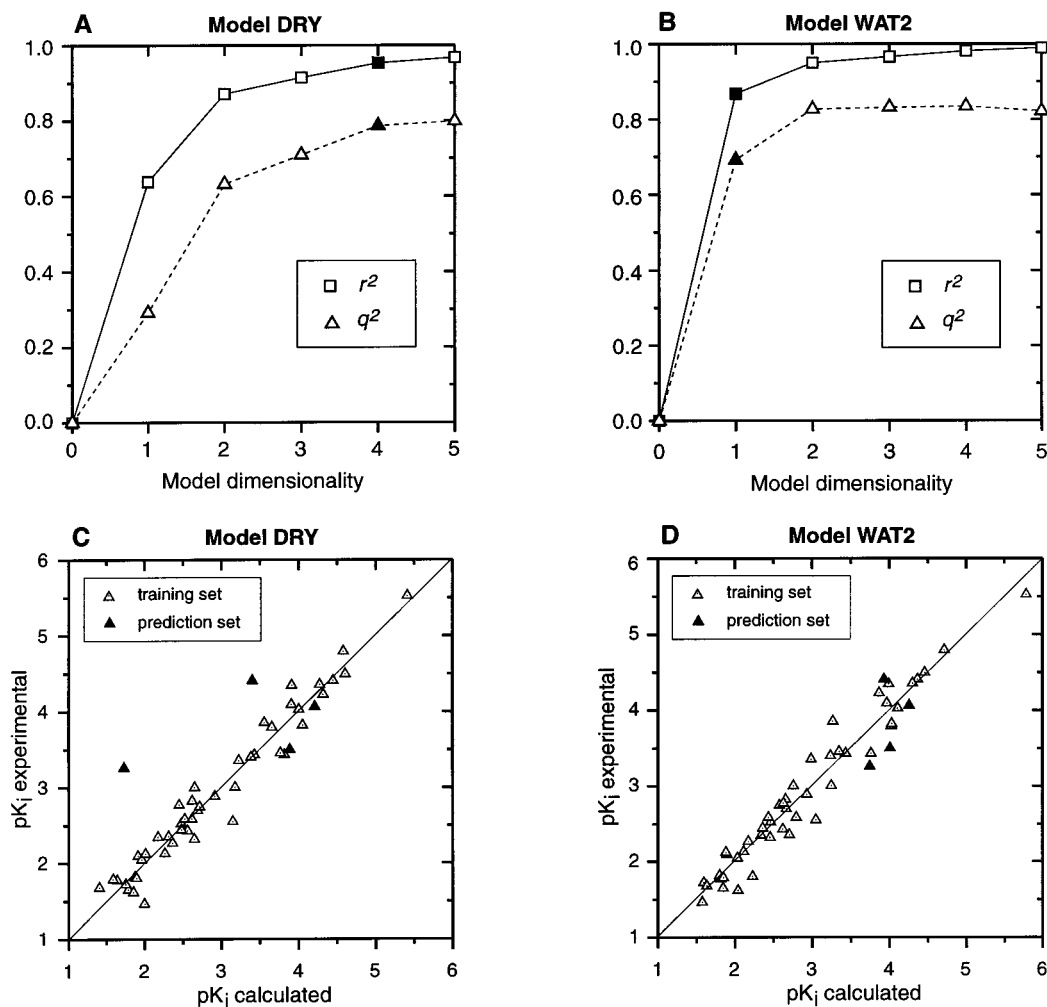


Figure 4. Squared correlation coefficient (r^2) and cross-validated square correlation coefficient (q^2) versus different dimensionalities for the model DRY (A) and for the model WAT2 (B). Scatter plot showing the calculated versus experimental activity values for the training set (open triangles) and for the prediction set (black triangles) using the model DRY (C) and the model WAT2 (D).

Table 4. Results of the PLS Modeling for Different Data Sets Using Different Variable Selection Procedures

set	var selection	no. of vars ^a	dimens ^b	r^2 ^c	q^2 ^d	SDEP ^e	ext SDEP ^f
DRY	none	2086	4	0.91	0.43	0.75	0.98
WAT1	none	2977	2	0.84	0.45	0.73	0.84
WAT2	none	2934	2	0.86	0.41	0.76	0.59
DRY	SRD/GOLPE	428	4	0.95	0.79	0.46	0.94
WAT1	SRD/GOLPE	416	2	0.94	0.82	0.42	0.79
WAT2	SRD/GOLPE	403	2	0.95	0.83	0.41	0.43

^a Number of variables used in the PLS model. ^b Dimensionality of the model. ^c Squared correlation coefficient. ^d Cross-validated squared correlation coefficient. ^e Standard deviation of error of predictions (for internal validation). ^f Standard deviation of error of predictions (for the external prediction set).

to interpret.¹⁹ Therefore, further analysis and evaluation of the method is focused on only those models obtained using SRD/GOLPE variable selection. Comparison of the models that include water molecules shows that WAT2 is better than WAT1. The differences are particularly remarkable with respect to the external validation and are a consequence of the differences in the number and position of the water molecules included in both prediction data sets (PWAT1 and PWAT2). These results support the use of the GRID simulation procedure for these water molecules and suggest that waters present in the majority of ligand binding (even

if not observed in every crystal structure) should be incorporated into the analysis.

A comparison of the best models obtained with and without water (WAT2 and DRY) shows some significant differences. First, the optimal dimensionality is different in both cases. As shown in Figure 4B, the amount of variance explained by model WAT2 increases rapidly. The first principal component (PC) explains nearly 90% of the variance, and the model reaches a plateau in its predictive ability after 2 PC and is not improved by the addition of further components. In contrast, for data set DRY, the PLS model requires as many as 4 PC to explain the same variance (Figure 4A). Thus, the model DRY needs 4 PC to obtain the same result that WAT2 obtains with only 2 PC. However, both models DRY and WAT2 have quite similar squared correlation coefficients, which indicates no difference in their ability to fit the data. The model WAT2 shows slightly better predictive ability than DRY in the cross-validation test, but again the difference is more remarkable in the results of the external validation. Such differences are mainly a consequence of the fact that model WAT2 gives an accurate prediction for compound **50** while model DRY fails to predict the activity of this compound (see Table 3 and Figure 4C). The poorer results of model DRY in prediction are not surprising since **50** is a unique compound, which places the *N*-acetamide sub-

stituent in a position not previously explored by any other compound in the series.¹⁹ On the contrary, the model WAT2 contains a description of the contribution to the activity of the substituents present in this region, because the *N*-acetamide moiety is not far from the position where Wat 887 is present in most of the compounds. Consequently, in the data set WAT2, the complex 50-GP is not so different from the rest of the complexes and the prediction given is significantly better (see Figure 4D). It should be noted that the external predictions given by model WAT2 have a standard error in the same range as in the internal predictions (external SDEP, 0.43; cross-validation SDEP, 0.41). This further supports the predictive ability of this model.

Interpretation of the Models. One of the most important features of a GRID/GOLPE procedure is the possibility of translating back the PLS coefficients assigned to each variable to the 3D positions they occupy in real space. These values, contoured and displayed graphically at particular significance levels, are essential in highlighting those areas in which the PLS model has found a high correlation between the ligand-probe interaction energy and the activity. The representations are called "grid plots of PLS coefficients" and represent the weighted pseudocoefficients obtained for a certain dimensionality of the PLS model.^{28,29} Such coefficients can be assigned to each grid interaction energy variable in the polynomial fashion: $Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_nX_n$. These coefficients represent the influence of a single variable on the response (in this case the biological activity), so it is appropriate to revert to coefficients when the final global model is studied. It should be noted that the effect of the coefficients on the response is a combination of their algebraic sign and of the algebraic sign of the grid interaction point (*X* variable). Accordingly, negative coefficient regions associated with negative interaction energies give rise to an increase in the biological activity, because the product b_nX_n is positive. And conversely, positive coefficient regions associated with negative grid interaction energies give rise to a decrease in the biological activity. For this reason, interpretation of the coefficient plots must consider the characteristics of the probe used to compute the energies of interaction.

In this study the phenolic OH probe, which is a polar group with the ability to participate in hydrogen bonds both as a donor and as an acceptor, was used. Consequently, areas containing negative coefficients correspond to areas in the receptor where energetically favorable (negative) hydrogen bonds or polar interactions produce an increase in the activity. These regions tend to coincide with the presence of polar groups in the receptor site where the ligands can bind with increased affinity. In contrast, an area with negative coefficients corresponds to a region where the presence of energetically favorable (negative) hydrogen bond or polar interactions lead to a decrease in the activity. Such regions can appear for various reasons. For example, they may represent a hydrophobic region where the presence of a polar group is unfavorable because the energy required to desolvate this group is not compensated for by any other interaction. Alternatively, these regions may highlight a ligand-receptor interaction that may only be present through a different

binding mode, which is less favorable for the activity. Unfortunately, the situation is often more complex, because it should be also considered that the energy of interaction takes positive values when the distance between the probe and the ligand is short. Therefore, coefficients that are situated very close to some of the ligands may mean that ligands which overlap these areas have an unusual activity. This activity may be high when coefficients are positive or low when the coefficients are negative.

In this particular study, it is possible to compare the results obtained from the PLS models to the crystal structure of the enzyme. The advantage of this comparison is 2-fold: methodological and practical. First, from a methodological point of view, the accuracy of the model interpretation can be evaluated in light of the structure of the receptor. In this regard, the interpretation gives information about the suitability of the method. Second, from a practical point of view, the results of the comparison can be contrasted to the previous observations obtained from the structure-based analysis of the complexes. Any agreement found between both methods could provide important confirmation of the observations, since they have been obtained from two completely different approaches. In addition, any new observations derived from the QSAR models could provide scope for future development of better inhibitors of GP.

The Model DRY. Figure 5 shows the grid plot of the PLS coefficients for the model DRY, using 4 PC (see Figure 4A). The magenta contours represent negative coefficients under -0.006 while the blue contours represent positive coefficients over 0.006 . The most important areas which show major effects on the activity have been labeled in the figure from A to F.

Region A, on the left-hand side of Figure 5, is in the innermost part of the active site and encloses a few of the residues involved in the binding of the glucosyl moiety, for example, the backbone nitrogen of Gly 675 and Wat 897. Superimposition of all the crystal structures of the series shows that the α and β substituted compounds exhibit slight differences in the position of the glucopyranose ring. In general, glucose, gentiobiose, and the compounds with an α substituent appear to be more deeply located in the active site and in a slightly higher position than the β -substituted compounds. The coefficients in region A do not reflect that one binding mode is more favorable for the activity than the other but rather they act to discriminate between the α and the β series of compounds. In particular, since the average activity of the β series is higher, the coefficients that identify the β compounds assign a higher activity score.

Region B has a two layer shape, with negative coefficients on the top and positive coefficients on the bottom. In most complexes, Wat 872 occupies an intermediate position between these two layers. The values of the field energy induced by the different ligands in these areas depends on: the position of C1, which changes slightly from one complex to another, and the size and conformation of the α C1 substituent, when it is present. Compounds with small α substituents, like the OH of glucose, induce positive field values in the region occupied by the upper layer of negative coefficients, thus leading to a decrease in the activity

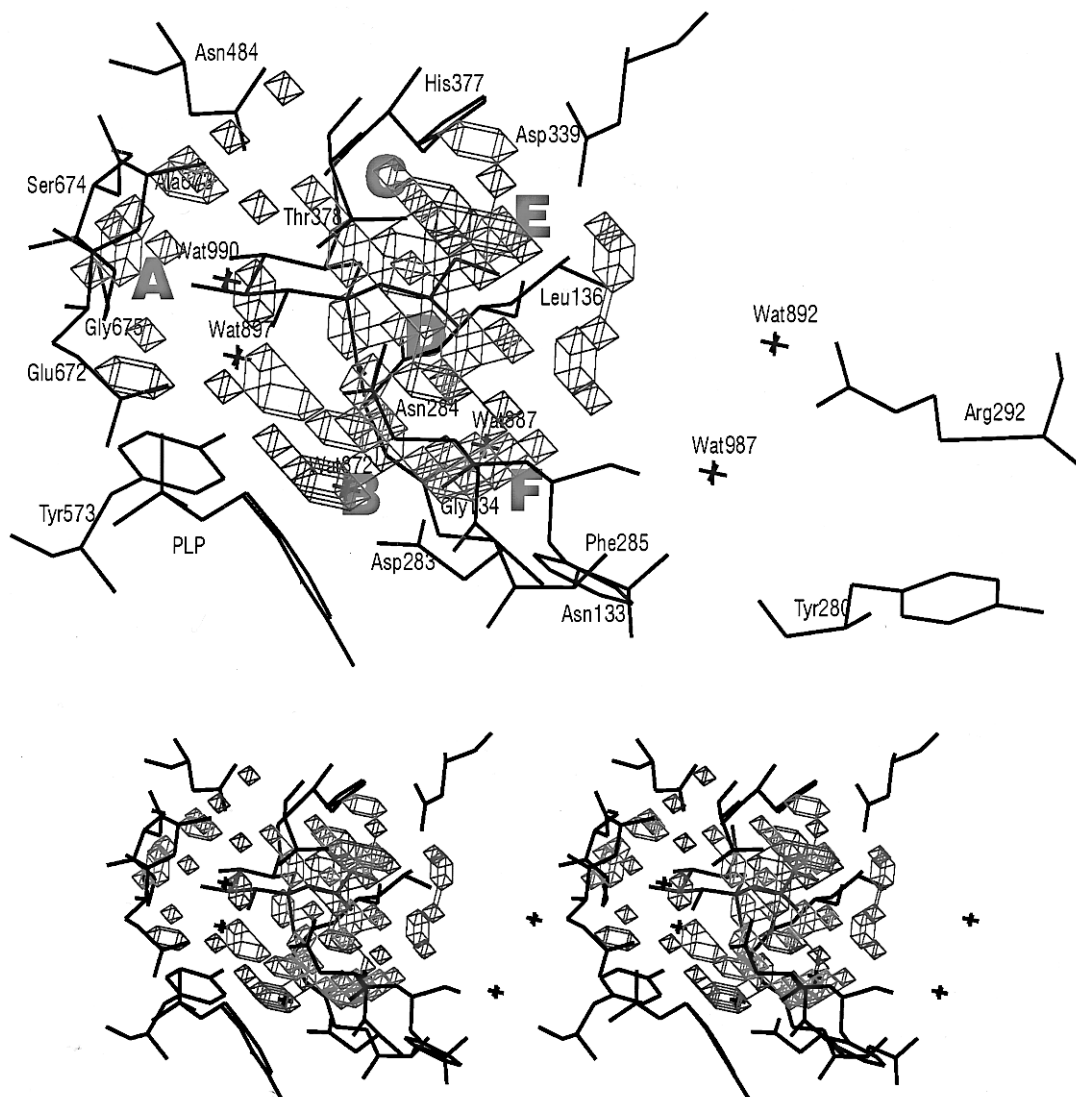


Figure 5. Grid plot of the PLS coefficients for the model DRY. The positive coefficients are represented in magenta and contoured at +0.006 and the negative coefficients are represented in cyan and contoured at -0.006. A selected subset of the protein residues and water molecules present in the active site, and the compound **35** (in red) are shown for reference. Regions A–F are described in detail in the text. A stereoview is provided for further clarity of the three-dimensional location of these regions with respect to residues in the active site of GP.

according to the PLS model. Conversely, compounds with larger α substituents such as **2** and **35** take positive interaction energies in the areas covered by the lower layer of positive coefficients, thus increasing the activity as calculated by the PLS model. With respect to the β compounds, those with the C1 atom occupying a lower position induce positive field values in the upper layer of negative coefficients and are represented by smaller activity values in the model. From an inspection of the structures of each ligand–GP complex, no unique interpretation can easily be obtained. For example, the size of the α substituent and the position of C1 appears to have an important effect in determining the position occupied by Wat 872, an important water molecule hydrogen bound to Asp 283, Gly 135, and Wat 891. Hence, these coefficients may represent the ability of the ligands to place Wat 872 in an optimum position. Alternatively, the coefficients may represent the changes in the position of the C1 atom produced by favorable binding in the upper region of the binding site (mainly to His 377) or by perturbation of the optimum hydrogen bond pattern of the glucosyl moiety.

Region C falls in the β pocket, overlapping the carbonyl oxygen of His 377 and extending toward the side chain of Thr 378 and His 377. This region contains only negative coefficients indicating that polar interactions here would increase the activity. In previous studies of ligand–receptor complexes the polar interactions between the ligand and the O of His 377 have been studied in detail^{11,13–16} and recognized as one of the most important polar interactions for increased inhibition. Indeed, one of the aims in the development of new β -substituted compounds has been to obtain compounds able to make a hydrogen bond with these residues.^{14,15}

Region D discreetly envelops the side chain nitrogen of Asn 284. The coefficients are negative, indicating that a polar interaction in this region would increase the activity of the compound. The ligand interactions with Asp 284 are also known to play an important role in the inhibition of GP, since interactions in this region stabilize the 280s loop and maintain the enzyme in its inhibited (T state) form.^{13,14}

Region E has a large area of positive coefficients which extends into the middle of the β pocket and

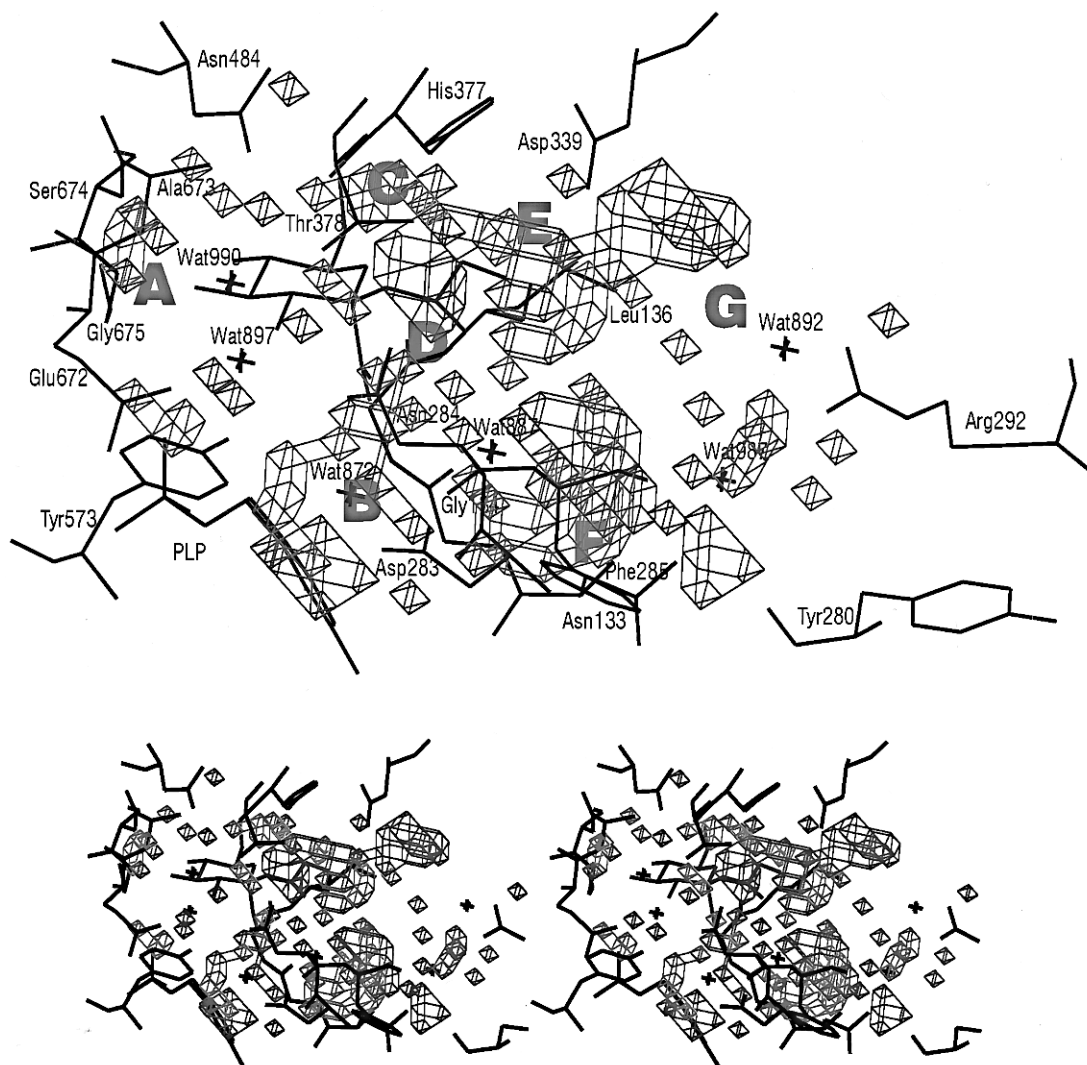


Figure 6. Grid plot of the PLS coefficients for the model WAT2. The positive coefficients are represented in magenta and contoured at +0.004 and the negative coefficients are represented in cyan and contoured at -0.004. A selected subset of the protein residues and water molecules present in the active site, and the compound **35** (in red) are shown for reference. Regions A–G are described in detail in the text. A stereoview is provided for further clarity of the three-dimensional location of these regions with respect to residues in the active site of GP.

overlaps the distal part of short β substituents (for example short C-amides such as **10** and **14** and N-amides such as **23**, **24**, **27**, **28**, and **29**). Small substituents in this region induce positive field values, thus producing an increase in the calculated activity according to the PLS model. Conversely, compounds with larger substituents (like **12** and **13**) fall into the areas with negative coefficients in the distal part of the pocket, and the PLS model predicts a decrease in the activity for these larger substituents. This result is consistent with previous observations based on the structural analysis of the ligand–enzyme complexes which indicated that medium size substituents gave the best results, and addition of bulky groups in this region often led to a diminished inhibitory effect.^{11,13} Moreover, this may suggest that the optimum size for a β substituent is just enough to displace Wat 847 from the complex, which as been pointed out as favorable for ligand binding due to entropic reasons.^{11,13}

Region F is on the bottom of the β pocket and contains mainly negative coefficients. This region corresponds to the area described by the nitrogen of the C-amide series and by the oxygen of the N-amide series of compounds. No compound in the series overlaps exactly

with this region. The energies of interaction are negative for those compounds having polar groups pointing in this direction and nearly zero for the remaining compounds of the series. This suggests that polar groups in this position are favorable for the activity, which is in agreement with preliminary models.¹¹ The structures of the ligand–enzyme complexes show that this region overlaps the positions occupied by Wat 887 in some of the compounds, which in turn is able to hydrogen bond to nearby polar groups of the protein. Previous analysis has suggested that interactions in this region may be responsible for the different inhibitory activities of the compounds in the C-amide and the N-amide series.¹¹

The Model WAT2. Figure 6 shows a grid plot of the PLS coefficients for model WAT2, using 2 PC (see Figure 4B). As a consequence of the incorporation of water molecules into this model, the area covered by the coefficients in this model is larger than in model DRY. The magenta contours represent negative coefficients under -0.004 while the blue contours represent positive coefficients over 0.004. The most important areas, which show major effects on the activity, have been labeled in the figure from A to G.

Region A on the left-hand side of Figure 6 is equivalent to Region A of model DRY. As in the previous model, the coefficients in region A envelop a few residues in the receptor participating in the binding of the glucosyl moiety, and serve to discriminate between the α - and β -substituted compounds since each series bind the glucopyranose ring in a slightly different way. However, this model better discriminates between the residues which play a major role in the interaction. The negative coefficients overlap the amide nitrogens of Gly 675 and Ala 673 and the phosphate group of the cofactor pyridoxal phosphate (PLP), which is bound to the glucosyl moiety of the ligand through Wat 897.

Region B is again equivalent to region B as described by the model DRY but, as a result of the incorporation of water it loses the characteristic two layer shape and looks more like two hemispheres that surround Wat 872 (which now is considered part of the ligand). The upper hemisphere contains negative coefficients and is oriented toward His 571 and Tyr 573. The lower hemisphere containing positive coefficients is oriented to the cofactor PLP and encloses the backbone of Gly 135. Both regions appear responsible for assigning higher inhibitory activities to the ligands that place Wat 872 in a precise position. More specifically, when Wat 872 is too high, the energies of interaction are negative in the lower areas (with positive coefficients) and positive in the upper areas (with negative coefficients) so the model predicts a decrease in the activity. Conversely, when Wat 872 is in a lower position the situation is reversed, i.e. positive field values in the positive coefficient areas and negative field values in the negative coefficient areas; therefore the model predicts an increase in the activity. As mentioned previously for model DRY, the coefficients in this region probably represent the optimum position for Wat 872, which is related to the position of the C1 atom. In other words, the positions of Wat 872 and the C1 atom appear to be correlated and effect the activity.

Region C is at the top of the β pocket overlapping the carbonyl oxygen of His 377 and close to the side chain polar groups of Thr 378 and His 377. The negative coefficients in this region represent favorable hydrogen bonds and polar interactions, which lead to an increase in the inhibitory effect. This region is similar to region C in the model DRY, but again in this model it is more discreetly defined, particularly with respect to potential interactions with Thr 378. The importance of polar interactions in this region of the β pocket for enhanced activity has been discussed previously for the model DRY.

Region D is equivalent to region D in the model DRY. The negative coefficients envelop the side chain polar groups of Asn 284, indicating that polar binding to this residue enhances the inhibition. In model WAT2, this region is smaller and concentrated specifically over the polar group of this residue.

Region E is located in the β pocket, halfway between the glucopyranose ring and Wat 847. The positive coefficients in the center of the pocket assign an increased activity to β -substituted compounds that place a substituent in this area, since for these compounds the field values are positive. These coefficients identify the presence of a β substituent in the ligand, which is in general associated with greater activity. In addition,

ligands with substituents in this position are able to make van der Waals contacts with Thr 378. In the region near Leu 136, there is a small area of negative coefficients. These coefficients identify a hydrophobic patch where the presence of a polar interaction will not improve the activity. In earlier work,¹¹ the 10-fold increase in the inhibition observed for *N*-glycyl **29** compared to *N*-ethylacetyl **25** was rationalized as a consequence of the energy required to desolvate the amine functionality of **25** and bury it in an hydrophobic environment. For the α -substituted compounds, the presence of Wat 847 induces negative field values in this region, and the model predicts a smaller inhibitory effect.

Region F surrounds Wat 887 and encloses some of the polar groups to which it is bound, for example, the backbone nitrogens of Gly 134 and Asn 133. The coefficients are negative, which means that polar interactions with these residues result in an increase in the activity. To date there have not been compounds designed specifically to interact with residues in this region.

Region G is in the distal part of the β pocket. Two of the largest areas with positive coefficients fully envelop two water molecules which have a distinct position in compound **45**, the most active in the series. In this complex, **45**-GP, the water overlapping this region induces positive field values, which according to the model, explain its high inhibitory effect. The presence of different water positions for complex **45**-GP has been previously described,¹⁵ and it may be that this is one of the factors responsible of the high activity of this compound. However the absence of more compounds in the series with a similar water pattern prevents any general statements regarding the exact role of this water disposition in the inhibition of GP. Apart from these areas, this region contains some patches of coefficients around Wat 987 that, like region B, seem to be defining the optimum position of this water. The coefficients are close to the side chains of Arg 292, Asn 133, and the main chain carbonyl O of Tyr 280. The coefficients in the most distal part of the region appear to identify the largest ligands in the series (**42** and **43**) and are assigned a corresponding decrease in activity.

Discussion

In most QSAR drug design approaches, the water present at the binding site is completely ignored. Recently, the incorporation of water molecules into the structures of the receptor has been suggested.¹⁰ However, for QSAR studies where a common receptor is assumed, it is more appropriate to include the water molecules as an integral part of the various ligands. It may be argued that water is not a constituent part of the ligand and therefore the incorporation of water molecules to the ligand moiety would be a misrepresentation. However, in QSAR studies comparisons are made between different compounds; therefore, the study should characterize as many differences as possible between each ligand structure and correlate these differences with their associated activity. Any potential source of variation upon binding of the compounds that is not considered in the QSAR analysis may lead to a potential source of error in the final activity estimations. Since water molecules are less tightly bound (than

ligands) at the receptor site and are more susceptible to movement or displacement upon ligand binding, their incorporation into the QSAR study can help to remove such sources of error in the estimations. It should be stressed that the water molecules incorporated into the analysis should not be seen as a permanent constituent of the ligand, but instead as a constituent part of the complex formed upon binding of each ligand that gives rise to a particular biological activity.

In this study, the presence and the position of the water molecules were determined experimentally using X-ray crystallography. However, it is recognized that this technique might fail to give a consistent and complete picture of the hydration for different complexes of the same protein, particularly for medium- to low-resolution structures. Consequently, with particular attention to the water structure of each complex, the program GRID was used to calculate (on average) energetically favorable positions for the water molecules in order to build up a consistent picture of the ligand–water–enzyme structure. The positions obtained for the water molecules from the GRID calculations are not intended as the “true” positions of the water molecules, nor is it appropriate to describe the presence of water in such terms generally. However, it has been shown from this work to be a valid procedure that in this particular series improves the quality of the QSAR models.

Inspection of the water molecules present in each of the complexes shows mobile water molecules in positions where GRID indicates the possibility of several minimums. These are areas where a water molecule can establish a large number of hydrogen bonds. These findings are consistent with previous reports.⁸ Poornima et al. used the isotropic *B* factors of the water molecules in different crystallographic complexes to account for the mobility of the water. They showed that when the different positions accessible to such a mobile water were examined, it appeared that a large number of the water molecules present at the binding site could establish as many as six hydrogen bonds, according to standard measurements of hydrogen bond angles and distances.⁸ However, since water is not able to participate simultaneously in more than four hydrogen bonds, these findings suggest the presence of some close energetic minimum, such as those found in the present analysis of Wat 890 (Figure 3). It would seem that the presence of such an attractive environment for the water may have some implications with respect to the energy of solvation in these regions. For example, it is conceivable that this situation is entropically favorable, since in such an environment water can move between the different minima and is not anchored to a unique, fixed position.

The results of this study using inhibitors of GP clearly indicate that incorporation of water molecules into the QSAR analysis improves the quality of the model in many different ways. With respect to the predictive ability, the models obtained with water compare favorably with the classic “dry” model, particularly in the predictions given for the compounds in the prediction sets. These results are not surprising, since incorporation of water allows the model to predict the activity of more dissimilar compounds, like **50**, which displace water molecules that are present in other complexes.

One of the most striking observations arising from the comparison of the models is in the differences in their optimum dimensionality. This reflects the different ability of both data sets to express in a generally applicable way, the characteristics of the ligands relevant for binding. With respect to data set DRY, the ligands are interacting with the receptor through different water patterns. Consequently, the PLS method needs many successive PC to obtain a model which will be valid for different representations of the receptor that arise from compounds of different size. Accordingly, the grid plots of the PLS weights in real space (not included here) show that each PC represents the complexes of different size with different hydration patterns. These observations confirm that incorporation of the water molecules, as in data set WAT2, gives a more consistent picture of the ligand interaction. In terms of the interpretability of the models, data set WAT2 shows a much more detailed picture of the receptor. The differences are more obvious in the region of the β pocket where differences in the hydration pattern seem more important. Model DRY, on the other hand, does not present a unique picture of the receptor but instead an average of how different kinds of ligands see the receptor.

This work demonstrates the advantages of incorporating water into the training set used for a QSAR analysis; however in most QSAR studies, availability of the structure of the complexes is more the exception than the rule. The explicit incorporation of water molecules can be applied to those studies where some experimental evidence (X-ray diffraction, NMR, etc.) shows that the ligand binding is hydrated and that this hydration layer is perturbed upon ligand binding. If the structure of at least one complex is available, it should be possible to apply solvent simulation techniques^{22,32,33} in order to obtain approximate positions for the water molecules in the remaining complexes.

QSAR methodologies, like the one described here, can be used as a tool for the structural analysis of several complexes. There are occasions where a strictly structure-based approach fails to explain the differences in activity, and this is where a QSAR analysis offers advantages since (1) the comparison is performed over a whole series of compounds simultaneously, while such a structural analysis is more difficult and comparisons are often based on pairwise differences between structurally related compounds; (2) the QSAR models are objective and therefore unaffected by subjective interpretations; and (3) the QSAR models are quantitative and therefore give information about the relative importance of different factors (for example GRID interaction energies) that are correlated to the activity.

The GRID/GOLPE method as described in this work has some limitations. One of the most important limitations is the inability of the method to account for entropic effects. For example, in this series of inhibitors of GP, the most active compound is a rigid spirohydantoin **45**. The binding of this compound to GP does not induce any loss of conformational freedom of the substituent at the C1 position and therefore produces no unfavorable entropic gain in energy. In fact, the gain in entropy has been recognized and has been used to rationalize the differences in activity between a subset of these compounds¹³ and between the rigid and semi-

rigid compounds **45** and **47**.¹⁶ Attempts to introduce variables which would account for entropic effects into the QSAR models so far have been without success, primarily because the series contains too few rigid compounds (**45**, **46**, and **47**) with which to evaluate their statistical significance. Clearly, entropic effects are also important with regard to water. It is generally accepted that the displacement of a water molecule present at a ligand binding site leads to an entropic gain in energy and therefore leads to a favorable interaction. Indeed, at different stages of this project, the displacement of water molecules has been attempted in order to obtain more active compounds.^{11,13,14} In the QSAR models, this effect is not accounted for directly, but within a series when a water molecule has been advantageously displaced, the QSAR model indicates that either its interaction, or its presence, is unfavorable for the activity. Nevertheless, it should be noted that the entropic gain obtained with the displacement of water depends upon the nature of the binding and for loosely bound water, like Wat 890 in this series, the change in free energy of the system would be modest.^{1,34}

Conclusions

This work has shown that, for this series of inhibitors of glycogen phosphorylase, incorporation of water molecules into the QSAR analysis has been valuable in providing an alternative understanding of the role of the water molecules associated with ligand binding. The QSAR models obtained, using the assemblies of ligand plus waters, are shown to have better predictive ability and are easier to interpret than the models obtained using only the ligands molecules. The water-inclusive QSAR models have also provided information which is consistent with that obtained from the X-ray structural analysis of the complexes. Moreover, this new QSAR model has highlighted a potentially important region at the catalytic site of GP which, to date, has not been previously explored. Thus, this new QSAR technique can be seen as a complement to the structure-based approach, which can indicate the most important areas of interaction for biological activity in a quantitative way.

This new QSAR procedure, as described in this work, requires knowledge of the position of individual water molecules present at the ligand binding site and is sensitive to differences in the water structure associated with different ligands. Thus, currently its application is limited to a series for which the crystal structures of the ligand-receptor complexes are available. However, utilization of solvent simulation techniques may extend the range of application to a series in which the structure of only one ligand-receptor complex is known. Work is in progress to evaluate the applicability of a similar approach to a series where less structural data is available.

Experimental Section

Computations were carried out in a R5000 INDY SGI workstation, using programs GRID V.14 for the energy calculations and GOLPE 3.0.5 for the chemometrical analysis. InsightII 95 was used for visualization.

Acknowledgment. This work has been supported by the European Community research project BIO2-CT943025. We acknowledge valuable contributions to

this work from our colleagues L. N. Johnson, G. W. J. Fleet, and N. G. Oikonomakos. We are also grateful to M. Kontou for valuable comments and discussion. The Italian funding agencies of MURST and CNR are thanked for financial support to G.C. The British Diabetic Association is acknowledged for financial support to K.A.W.

References

- (1) Ajay; Murcko, M. A. Computational Methods to Predict Binding Free Energy in Ligand-Receptor Complexes. *J. Med. Chem.* **1995**, *38*, 4953-4967.
- (2) Kubinyi, H. *QSAR: Hansch Analysis and Related Approaches*; Mannhold, R., Krosgaard-Larsen, P., Timmerman, H., Eds.; VCH: Weinheim, 1993.
- (3) Ortiz, A. R.; Pisabarro, M. T.; Gago, F.; Wade, R. C. Prediction of Drug Binding Affinities by Comparative Binding Energy Analysis. *J. Med. Chem.* **1995**, *38*, 2681-2691.
- (4) Head, D. H.; Smythe, M. L.; Oprea, T. I.; Waller, C. L.; Green, S. M.; Marshall, G. R. VALIDATE: A New Method for the Receptor-Based Prediction of Binding Affinities of Novel Ligands. *J. Am. Chem. Soc.* **1996**, *118*, 3959-3969.
- (5) Finney, J. L. The organization and function of water in protein crystals. *Philos. Trans. R. Soc. London Ser. B.* **1977**, *278*, 3-32.
- (6) Quioco, F. A.; Wilson, D. K.; Vyas, N. K. Substrate specificity and affinity of a protein modulated by bound water molecules. *Nature* **1989**, *340*, 404-407.
- (7) Meiering, E. M.; Wagner, G. Detection of long-lived bound water molecules in complexes of human dihydrofolate reductase with methotrexate and NADPH. *J. Mol. Biol.* **1995**, *247*, 294-308.
- (8) Poornima, C. S.; Dean, P. M. Hydration in drug design. 1. Multiple hydrogen-bonding features of water molecules in mediating protein-ligand interactions. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 500-512.
- (9) Poornima, C. S.; Dean, P. M. Hydration in drug design. 2. Influence of local site surface shape on water binding. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 513-520.
- (10) Poornima, C. S.; Dean, P. M. Hydration in drug design. 3. Conserved water molecules at the ligand-binding sites of homologous proteins. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 521-531.
- (11) Watson, K. A.; Mitchell, E. P.; Johnson, L. N.; Cruciani, G.; Son, J. C.; Bichard, C. J. F.; Fleet, G. W. J.; Oikonomakos, N. G.; Kontou, M.; Zographos, S. E. 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.
- (12) 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 b. *Biochemistry* **1990**, *29*, 10745-10757.
- (13) Martin, J. L.; Veluraja, K.; Ross, K.; Johnson, L. N.; Fleet, G. W. J.; Ramsden, N. G.; Bruce, L.; 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 for Diabetes. *Biochemistry* **1991**, *30*, 10101-10116.
- (14) 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.
- (15) Bichard, C. J. F.; Mitchell, E. P.; Wormald, M. R.; Watson, K. A.; Johnson, L. N.; Zographos, S. E.; Koutra, D. D.; Oikonomakos, N. G.; Fleet, G. W. J. Potent Inhibition of Glycogen Phosphorylase by a Spirohydantoin of Glucopyranose: First Pyranose Analogues of Hydantocidin. *Tetrahedron Lett.* **1995**, *36*, 2145-2148.
- (16) Krülle, T. M.; Watson, K. A.; Gregoriou, M.; Johnson, L. N.; Crook, S.; Watkin, D. J.; Griffiths, R. C.; Nash, R. J.; Tsitsanou, K. E.; Zographos, S. E.; Oikonomakos, N. G.; Fleet, G. W. J. Specific Inhibition of Glycogen Phosphorylase by a Spirodike-topiperazine at the Anomeric Position of Glucopyranose. *Tetrahedron Lett.* **1995**, *36*, 8291-8294.
- (17) *3D QSAR in Drug Design. Theory Methods and Applications*; Kubinyi, H., Ed.; ESCOM: Leiden, 1993.
- (18) Cruciani, G.; Watson, K. A. Comparative Molecular Field Analysis Using GRID Force-Field and GOLPE Variable Selection Method in Study of Inhibitors of Glycogen Phosphorylase b. *J. Med. Chem.* **1994**, *37*, 2589-2601.
- (19) Pastor, M.; Cruciani, G.; Clementi, S. Smart Region Definition (SRD): a New Way to Improve the Predictive Ability and Interpretability of 3D-QSAR Models. *J. Med. Chem.* **1997**, *40*, 1455-1464.

- (20) Steiner, R. F.; Greer, L.; Bhat, R.; Oton, J. Structural changes induced in glycogen phosphorylase by the binding of glucose and caffeine. *Biochem. Biophys. Acta* **1980**, *611*, 269–279.
- (21) Brünger, A. T. A memory-efficient fast Fourier transformation algorithm for crystallographic refinement on supercomputers. *Acta Crystallogr.* **1989**, *A45*, 42–50.
- (22) Smith, P. E.; Pettit, B. M. Modeling Solvent in Biomolecular Systems. *J. Phys. Chem.* **1994**, *98*, 9700–9711.
- (23) Finer-Moore, J. S.; Kossiakoff, A. A.; Hurley, J. H.; Earnest, T.; Stroud, R. M. Solvent Structure in Crystals of Trypsin Determined by X-Ray and Neutron Diffraction. *Proteins* **1992**, *12*, 203–222.
- (24) Screenivasan, U.; Axelsen, P. H. Buried Water in Homologous Serine Proteases. *Biochemistry* **1992**, *31*, 12785–12791.
- (25) GRID V.14, Molecular Discovery Ltd., West Way House, Elms Parade, Oxford, 1996.
- (26) Goodford, P. J. A Computational Procedure for Determining Energetically Favourable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28*, 849–857.
- (27) Boobbyer, D. N. A.; Goodford, P. J.; McWhinnie, P. M. New Hydrogen-Bond Potentials for Use in Determining Energetically Favorable Binding Sites of Molecules of Known Structure. *J. Med. Chem.* **1989**, *32*, 1083–1094.
- (28) 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.
- (29) Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems. *Quantum Struct.-Act. Relat.* **1993**, *12*, 9–20.
- (30) Clementi, S.; Cruciani, G.; Pastor, M.; Riganelli, D.; Valigi, R. (unpublished results).
- (31) Cruciani, G.; Clementi, S.; Baroni, M. Variable Selection in PLS Analysis. In *3D QSAR in Drug Design, Theory Methods and Applications*; Kubinyi, H., Ed.; ESCOM: Leiden, 1993; pp 551–564.
- (32) Wade, R. C.; Goodford, P. J. Further Development of Hydrogen Bond Functions for Use in Determining Energetically Favorable Binding Sites on Molecules of Known Structure. 2. Ligand Probe Groups with the Ability to Form More Than Two Hydrogen Bonds. *J. Med. Chem.* **1993**, *36*, 148–156.
- (33) Wade, R. C.; Bohr, H.; Wolynes, P. G. Prediction of Water Binding Sites on Proteins by Neural Networks. *J. Am. Chem. Soc.* **1992**, *114*, 8284–8285.
- (34) Dunitz, J. D. The Entropic Cost of Bound Water in Crystals and Biomolecules. *Science* **1994**, *264*, 670–670.

JM970273D