Comparative Molecular Field Analysis Using GRID Force-Field and GOLPE Variable Selection Methods in a Study of Inhibitors of Glycogen Phosphorylase *b*

Gabriele Cruciani*,[†] and Kimberly A. Watson[‡]

Department of Chemistry, University of Perugia, Via Elce di Sotto, 8, 06100 Perugia, Italy, and Laboratory of Molecular Biophysics, University of Oxford, South Parks Road, OX1 3QU, Oxford, England

Received March 1, 1994[®]

A primary goal in any drug design strategy is to predict the activity of new compounds. Comparative molecular field analysis (CoMFA) has been used in drug design and three-dimensional quantitative structure/activity relationship (3D-QSAR) methods. The CoMFA approach permits analysis of a large number of quantitative descriptors and uses chemometric methods such as partial least squares (PLS) to correlate changes in biological activity with changes in chemical structure. One of the characteristics of the 3D-QSAR method is the large number of variables which are generated in order to describe the nonbonded interaction energies between one or more probes and each drug molecule. Since it is difficult to know *a priori* which variables affect the biological activity of the compounds, much effort has been devoted to developing methods that optimize the selection of only those variables of importance. This work focuses on some of the aspects involved in the selection of such variables, applied to a series of glucose analogue inhibitors of glycogen phosphorylase *b*, using the program GRID to describe the molecular structures and using a method of generating optimal partial least squares estimations (program GOLPE) as the chemometric tool. This data set, consisting of over 30 compounds in which the three-dimensional ligand–enzyme bound structures are known, is well suited to study the effect of different data pretreatment procedures on the final model used for the prediction of new drug molecules. By relying on our knowledge of the real physical problem (i.e., using the combined crystallographic and kinetic results), it has been shown that suitable data pretreatment and variable selection have been found that does not result in a significant loss of relevant information. Moreover, by using an appropriate scaling procedure, GOLPE variable selection minimizes the risk of overfitting and overpredicting.

Introduction

Since the publication of the work of Cramer, comparative molecular field analysis (CoMFA)[1] has become widely used in drug design and QSAR methodologies. The approach provides acquisition of a large number of quantitative descriptors and uses PLS[2] methods to correlate changes in the observed biological activity with changes in the chemical structure for a series of potential drug molecules.

Although these methods have been of general use, there are a number of practical problems in their application to a dataset of interest. The results depend critically on the conformation and alignment criteria chosen for the drug molecules, on the chemical data used to describe the interactions (i.e., on the chosen probes), and on the validation method used in the chemometric tool. It is difficult to know *a priori* which of these considerations are more significant, since these problems are data set dependent and change their relative importance with changes in a series of drug molecules. Selection of the most informative variables, for example the interaction energies between a probe and a molecular structure, is a general problem which is always present in a QSAR study.

Traditional computational methods used in 3D-QSAR (CoMFA,[1] HINT,[3] COSMIC,[4] DOCK,[5] LUDI[6]) produce a large number of molecular descriptors (or variables such as force field parameters, interaction energies, or local minima distributions) which may or may not contribute to the final result. In some cases, it may be obvious which variables have a positive or negative effect on the prediction capability of the model. On the other hand, some variables may produce only subtle changes in the final model or have a secondary but not easily distinguishable effect. It has been shown with a large number of variables that keeping irrelevant variables in the model can in fact have detrimental effects on the predictive ability of the model.[7] Therefore, it would be useful to find a method which successfully selects only those variables which have the most significant effect on the biological activity.

Developments in statistics have provided new methods of measuring the validity of a model. These methods are based on simulating the predictive power of the model and work by creating a number of slight modifications to the original data set and estimating parameters from each of these modified data sets. The effects of the modifications to the data set may be assessed according to how well compounds within the dataset are predicted by the model and, further, by calculating the variability of the predictions for novel compounds using each of the resulting models.[8,9] Using these tools it is possible to compare the predictive power of different models or, within the same model, to estimate a set of unique parameters (variables and optimal number of components) necessary to maximise the predictive power of the model.[10]

* Author to whom correspondence should be addressed.
[†] University of Perugia.
[‡] University of Oxford.
[®] Abstract published in *Advance ACS Abstracts*, June 15, 1994.

0022-2623/94/1837-2589$04.50/0

More recently, an advanced variable selection procedure called GOLPE[11] has been used in 3D-QSAR studies. The procedure is based on variable selection which evaluates the effects of individual variables on the model predictivity. This can lead to the determination of precisely which variables are relevant to the problem under study. Consequently, only a few significant variables (as determined by the GOLPE procedure) are extracted from large amounts of more redundant information produced by 3D-QSAR methods and are used in subsequent analyses.

The primary aim of this work was to determine whether or not there exist a set of general rules for the pretreatment of the data that will result in a reliable method of variable selection for optimal predictivity. The present paper discusses some of the aspects involved in the selection of such variables using the GOLPE procedure. This procedure has been shown to be a powerful tool for evaluating and selecting the important variables which contribute positively to the predictivity.[11-16] However, as with all statistical methods the results from GOLPE depend on some method of pretreatment of the data. Unfortunately, some of these methods can lead to overfitting and chance correlations.

Different procedures were compared using a series of 36 compounds whose enzyme inhibitory potencies have been measured and whose X-ray structures bound to the same enzyme (glycogen phosphorylase = GP) have been determined.[17,18] Consequently, problems arising from the conformation and alignment of the ligands have been directly addressed. Furthermore, since the three-dimensional structures of the complexes are known, it is possible to assess more accurately the predicted variables and regions important for inhibition.

Methods

Data Set for Analysis and Validation. An appropriate series of molecules was required in order to compare the different ways in which to perform the GOLPE analysis. This series should be without superimposition problems and with the correct conformation for each molecule; with a good range of biological activity; with sufficent accuracy in the activity values; and, finally, with information about both the ligand–receptor interactions and the nature of these interactions.

Such a data set has been provided by studies on the design, synthesis, kinetics, and three-dimensional X-ray crystallographic results of a number of ligands complexed to glycogen phosphorylase.[17,18] It is known that α-D-glucose is a weak inhibitor of GP*b* (K_i = 1.7 mM) and acts as a physiological regulator of hepatic glycogen metabolism.[19] Glucose binds to phosphorylase at the catalytic site and results in a conformational change that stabilizes the inactive T state of the enzyme. It has been suggested that in the liver, glucose analogues with greater affinity for GP may result in a more effective regulatory agent. Hence, all the ligands studied were glucose analogues.

Over 50 compounds have been tested, 36 of which show the same mechanism of action (by binding in a similar fashion whereby the T state of the enzyme is stabilized and to the same site of GP*b* as glucose itself since there is the possibility of multiple binding sites) and cover a wide range in the inhibitor activity (Table 1). The binding of each of 36 compounds to GP*b* in the crystal has been studied to 2.3-Å resolution, and the ligand–phosphorylase complexes have been refined using crystallographic least-squares minimization to *R* values less than 0.20 using X-PLOR energy.[20] Some of the factors and the regions in the active site of GP*b* that are important for effective inhibition have already been established.[17,18] In particular, hydrogen bonding either directly to the protein or through water molecules contributes significantly to an increase in the binding energy and to a decrease in the inhibition constant (K_i). The experimental errors in the K_i values[17,18,21] determined for each ligand have an average error of ±14%.

This series of molecules eliminates doubts arising from the alignment criteria and from uncertainties about the conformations of the bound ligands. The X-ray structures of the 36 glucose analogue–phosphorylase complexes show that the glucose ring in each case is maintained in the same region of the phosphorylase active site due to a complementary hydrogen bonding network between the glucose fragment of the ligand and the active site residues of the enzyme. The substituted atoms at the α- and β-C1 positions of glucose occupy different regions of the enzyme active site, and it is these regions of the enzyme that have been targeted in order to further enhance the inhibition.

This data set superimposed by the enzyme itself is well suited to study the effect of different procedures on the final model in the step of variable selection.

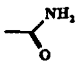
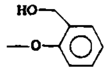
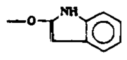
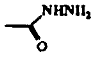
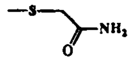
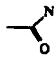
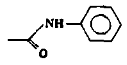
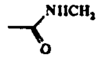
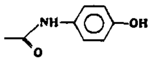
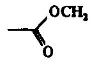
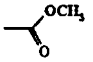
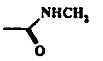
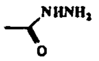
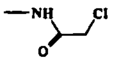
GRID Force Field. The program GRID[22-24] was used to calculate the interactions between a small chemical group (for example the phenol hydroxy probe) and each of the 36 compounds (the targets). GRID is a computational procedure for detecting energetically favorable binding sites on molecules of known three-dimensional structure. The energies are calculated as Lennard-Jones, electrostatic, and hydrogen bond interactions between the probe group and the target structures, using a position-dependent dielectric function in order to modulate the strong electrostatic interaction between charged centers, since solvent molecules were not explicitly included with the targets.

GRID contains a table of parameters to describe each type of atom occurring in each of the ligand molecules. These parameters define the strength of the Lennard-Jones, hydrogen bond, and electrostatic interactions made by an atom and are used in order to evaluate the energy functions.

GRID probes are very specific.[25] They give precise spatial information, and this specificity and sensitivity are an advantage since the Probes may then be representative of the important chemical groups present in the active site provided that the statistical method used for the analysis can distinguish between different types of interactions.

In this work, a hydroxyl group bonded to an aromatic system was chosen as the primary chemical probe (the OH probe). 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

Table 1. Database of Glucose Analogue Inhibitors for Glycogen Phosphorylase *b*

Comp.	Subst. at C1 position		$\ln K_i$*exp*	$\ln K_i$*cal*	$\ln K_i$*pred*
	α	β			
1	—OH		0.53	0.67	0.85
2	—CH_3		3.97	2.63	2.33
3	—OCH_3		2.49	2.86	3.00
4		—OCH_3	3.21	3.02	2.50
5		—OCH_2CH_2OH	3.23	3.86	3.48
6	—CH_2N_3		3.11	2.90	2.40
7		—CH_2N_3	2.72	2.37	1.95
8	—OH	—CH_2N_3	2.00	2.56	2.60
9	—CH_2OH		0.41	1.38	1.86
10		—CH_2OH	3.09	2.60	2.24
11	—OH	—CH_2OH	2.76	2.37	2.08
12		—$CH_2OSO_2CH_3$	1.57	1.56	1.70
13	—OH	—$CH_2OSO_2CH_3$	1.31	1.20	1.32
14	—$CH_2NH_3^+$		3.54	2.61	2.07
15		—$CH_2NH_3^+$	2.82	3.09	2.47
16		—$CH_2CH_2NH_3^+$	1.50	1.67	1.66
17		—CH_2CN	2.20	2.48	2.38
18	—OH	—CH_2CN	2.03	2.19	2.10
19	-O-(1-6)-D-glucose		2.79	3.10	2.93
20			-0.99	-0.60	-0.07
21	5-thio-α-D-glucose		0.69	0.33	0.56
22		—SH	0.00	0.40	0.63
23			2.15	1.93	2.50
24			0.96	0.75	0.90
25			1.10	1.17	1.41
26			3.05	3.49	3.33
27			-0.82	-0.55	0.07
28			2.53	2.57	1.87
29			3.60	3.98	3.63
30			1.72	1.84	1.65
31			3.19	2.81	3.62
32	1,2-dideoxy-1,2-difluoro-α-D-glucopyranose		-1.61	-1.82	-0.98
33			0.92	-0.22	0.49
34			-1.61	-1.47	-1.14
35			-0.92	-1.07	-0.75
36			-3.22	-2.70	-0.72

similar oxygen probes and it makes strong hydrogen-bonding interactions which may account for the shape of the interaction regions with the target structures. This choice of probe was influenced by the presence of such residues in the GP*b* catalytic site.

Other probes selected for this study include the methyl group CH_3, the sp^2 carbonyl oxygen C=O, and the sodium cation Na^+ probe. The CH_3 probe has the electronic properties of an sp^3 carbon atom. The GRID parameters for this probe assume that it does not interact electrostatically with the target and it does not form hydrogen bonds. Thus the GRID calculation yields the energies of steric interaction between the target molecule and the probe. The carbonyl oxygen of the C=O probe can accept one or two hydrogen bonds. The choice of this probe (as with the OH probe) was influenced by the known residues in the active site of GP*b*. The Na^+ probe carries a +1.0 charge and has the potential for charge–charge interactions. This probe was selected due to the presence of glutamic and aspartic acid residues in the catalytic site thus providing the possibility of assessing charge–charge interactions within the protein.

The energy calculations were performed using both 1.0- and 2.0-Å spacings between the grid points in a rectangular box measuring 20 × 20 × 21 Å. The GRID origin and axes were chosen such that all the atoms of the target structures were included when they were maintained in the conformation when bound to the protein and an additional region which would include several active site residues of GP*b* (within 4 Å of the target structures) although the protein residues were not explicitly included in the calculations. A cut-off of +5 kcal/mol was used in order to make the data more symmetrically distributed about zero, since GRID energies can be very large and positive when the probe is close to the target.

With the chosen probe at the first grid point, (*xyz*), the overall nonbonded interaction energy E_{XYZ} between the probe and the target is calculated as

$$E_{XYZ} = E_{LJ} + E_{El} + E_{HB}$$

where E_{LJ} is the Lennard-Jones potential energy, E_{El} is the electrostatic, and E_{HB} is the hydrogen bonding energy.[24] The calculation is repeated with the probe at each successive grid point.

One GRID calculation produces 8400 interaction energies for each probe with each of the 36 compounds. Each set of calculated interaction energies contained in the resulting three-dimensional matrix from GRID can be rearranged as a one-dimensional vector of variables in two steps.[25] In the first step, the matrix planes are cut and positioned side by side producing a two-dimensional table with 20 rows and 20 × 21 columns. Then, in the second step, the 20 rows are juxtaposed such that a one-dimensional vector is produced. Thus, the interactions between each probe and each compound are described by this one-dimensional vector which is conventionally used as input to the program GOLPE and the computational analysis performed as described below.

The Partial Least Squares (PLS) Model. In the context of 3D-QSAR, the biological activity may be seen as a function of the physiochemical characteristics (such as electronic properties or energies of interaction within a given force field) of the compounds of interest. The need to convert such numerical data to useful information has led to the development of methodologies that rely on statistics and applied mathematics.

The PLS model is a two-block projection method that relates a matrix **X** (containing the chemical descriptors) to a matrix **Y** (containing the biological activities) with the aim of predicting the values in **Y** from the information contained in **X**.[26] The method provides an approximation of an **X** matrix in terms of the product of two smaller matrices **T** and **P′** as follows:

$$\mathbf{X} = 1\cdot\bar{x} + \mathbf{T}\cdot\mathbf{P}' + \mathbf{E} \qquad (1)$$

where **E** is a residual matrix, i.e., part of the data that is not explained by the model. The **Y**-block matrix is modeled in a similar manner to the **X**-block by,

$$\mathbf{Y} = 1\cdot\bar{y} + \mathbf{U}\cdot\mathbf{Q}' + \mathbf{E} \qquad (2)$$

The data matrices (**X** and **Y**) are projected down on the smaller matrices (**T** and **U**) with orthogonal columns. The projections can be calculated for any given number of variables, and, in fact, the projections become more stable for a given number of compounds the larger the number of relevant variables that are included. After the projection, the matrix **T** is used (instead of the original matrix **X**) to explain or predict **Y**, since **T** has fewer and orthogonal columns and therefore give rise to a numerically stable model. The relation between **U** and **T** (in eqs 1 and 2), the "inner relation", can be modeled by

$$\mathbf{U} = \mathbf{T}\cdot\mathbf{B} + \mathbf{H} \qquad (3)$$

where **B** is a diagonal matrix and **H** is a residual matrix.

Recalculated *y* values for each compound in the training set are obtained from the *x* data vector of each compound by insertion into the PLS model in the sequence,

$$\mathbf{X} \xrightarrow{\text{eq 1}} \mathbf{T} \xrightarrow{\text{eq 3}} \mathbf{U} \xrightarrow{\text{eq 2}} \mathbf{Y}$$

The prediction of *y* values for new compounds uses the same sequence but leaves out the *x* data vectors in the derivation of the model.

The matrices **T** and **P′** extract the essential information and any patterns contained in **X**. By plotting the columns in **T** (score plot), a picture of the dominant "object pattern" of **X** is obtained (i.e., a two-dimensional view of the compounds in the chemical property space). By analogy, plotting the rows of **P′** (loading plot) shows the complementary "variable pattern" which can give information about how the chemical properties should be modified in order to enhance the activity depending on the size and sign of the vectors *p*. The number of rows in **P′** and the columns in **T** are equal to the number of factors and can be determined by cross-validation to give the model optimal predictivity.[27]

Generating Optimal Linear PLS Estimation (GOLPE). GOLPE[11] is defined as an advanced variable selection procedure aimed at obtaining PLS regression models with the highest prediction ability which relies on the validation of a number of reduced models on variable combinations selected according to a factorial design strategy. It has been shown[7] that keeping irrelevant variables in the model gives rise to poorer predictions, since such variables represent only random variations. Therefore, in problems where the number of variables is large, it is necessary to use only relevant variables to improve the accuracy of the predictions. The power of a GOLPE procedure will depend on some method of pretreatment of the data, and this paper reports the results of five different data pretreatment methods.

The first step in the GOLPE procedure, as applied to a 3D-QSAR problem, is a normal linear PLS model using all the variables, followed by variable preselection according to a D-optimal design[28,29] in the loading space. With a large number of variables much of the information is redundant, and selection of variables in such a way that redundancy is reduced but collinearity is retained is a typical constrained mathematical problem.

D-optimal designs are appropriate for handling constrained problems, and the D-optimality criterion as implemented in GOLPE[11] enables variables to be selected such that most of the redundancy is reduced but sufficient collinearity among the remaining variables is maintained in order to satisfy the PLS algorithm. It should be noted that D-optimal designs usually work on objects in variable space;[30] however, in this case the objects are the original variables described by their loadings in LV space.

An important result of the first step in GOLPE is the determination of how many significant factors are present in the data. These factors, called latent variables (LV), are often directly interpretable as the number of independent chemical effects that influence the biological response. The LVs are linear combinations of all the original variables in which each variable participates according to its loading. In 3D-QSAR the loadings are usually rotated back into the original variable space of the molecules and take the form of coefficients of a linear polynomial for all the variables. In this study, the coefficients correspond to grid locations. These coefficients are displayed as contour maps which are representative of the important 3D regions characterized by the model.

The dimensionality of the LV is selected according to a cross-validation procedure[27] (whereby part of the data set is deleted from the modeling, a model is calculated, and predictions are made for the deleted data which are compared to the actual values, etc., until each data element has been kept out once) and verified graphically for homogeneities in the score plots. It is necessary to select the appropriate dimensionality since too few model dimensions will lead to a loss of information and too many will result in noise in the model. In either case, the prediction of the activities of new compounds will be less than optimal.

The second step in the GOLPE procedure is the building of a design matrix, which has the number of columns equal to the number of variables and the number of rows slightly greater than the number of variables and contains combinations of the variables according to a fractional factorial design (FFD) strategy. In the design matrix the combination corresponds to either the presence (+) or absence (−) of the original variables. The design matrix is then used to test the prediction ability of the reduced models, each involving a different combination of variables, including only the "plus" and excluding the "minus" variables. For each such combination, the prediction ability of the corresponding PLS model is evaluated by means of the standard deviation error of prediction (SDEP)[10] defined as follows:

$$\mathrm{SDEP} = [\sum_{i=1}^{N}(y_i - y_{i_{\mathrm{pred}}})^2/N]^{1/2} \qquad (4)$$

where y_i = experimental value (in this case, ln K_i), $y_{i_{\mathrm{pred}}}$ = predicted value, N = number of objects (in this case, number of compounds).

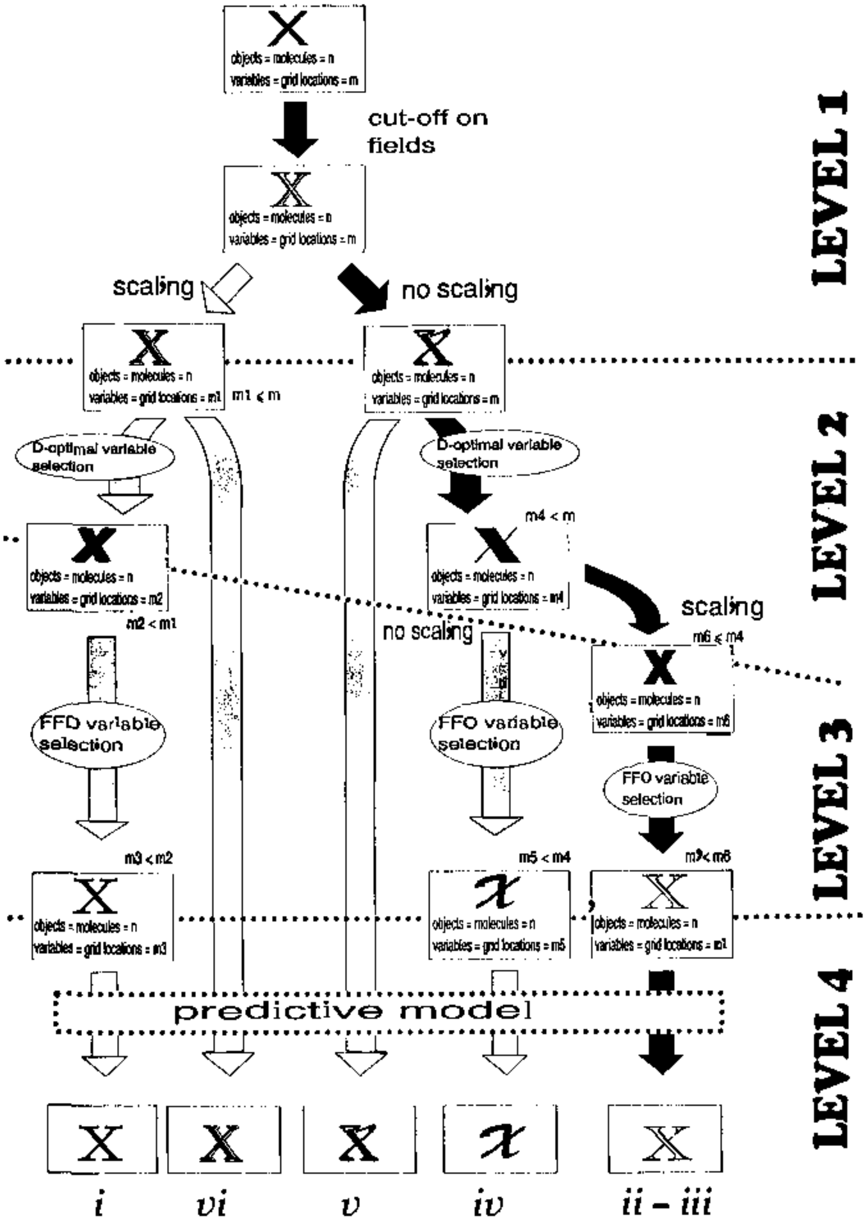
The value of SDEP is calculated using a combination of cross-validation and bootstrapping.[11] The data set is divided into several groups in a random way, and the computation is repeated several times, as in bootstrapping, but each group is excluded just once in each run, as in cross-validation. The higher the number of random ways of forming groups, the more stable the value of SDEP. The SDEP parameter is calculated for each excluded group based on the model derived from the remaining groups. The value of SDEP is never exactly reproducible, but it converges to an asymptotic value. The number of latent variables associated to the global minimum value of SDEP is selected as the model dimensionality that gives the best predictions.

The procedure based on FFD as outlined cannot work properly since there is a risk of selecting, as relevant, a variable which is not. In full factorial designs, the specific effect of a single variable can be evaluated unambiguously, however, at the expense of an increase in computation time. In order to estimate as precisely as possible the significance of a single variable effect on predictivity, a number of dummy variables can be introduced in the design matrix. These dummy variables are not actual numbers. The dummy variables are defined specific columns in the design matrix inserted among the true variables. Since these dummy variables are not true variables, they are not used in the variable combinations evaluating the predictivity of each row of the design matrix. However, the introduction of these dummy variables does allow comparison between the effect of a true variable and the average effect of the dummies.

The third step in the GOLPE procedure is the computation of the predictivity for each variable combination and the evaluation of the effect of each variable on the predictivity. Only the significant variables are selected for improving the predictivity of the model. The variables can be classified into four distinct categories: the dummy variables artifically introduced, variables with a definite positive effect on the predictivity, variables with a definite negative effect, and variables with uncertain effects. Variables exhibiting a positive effect on the model predictivity are those which show an effect statistically higher in absolute value than the reference value obtained by the dummies; consequently, these variables can be fixed within the variable combinations and always used in the estimation of the prediction in subsequent iterations. Conversely, those variables exhibiting a negative effect on the predictivity can always be excluded from the variable combinations. Thus, keeping fixed those variables with a positive effect and excluding those variables with a negative effect is an effective method of eliminating a number of variables and increasing the stability of the model. The iterative process continues until all the variables have been assigned and no variables remain to be fixed or excluded.

The computational procedure (as outlined in steps 1−3 above) has been applied to the GP*b* database shown in Table 1. The first step of the GOLPE procedure was carried out on the full data set without scaling of the variables and with an energy cut-off on positive interactions ≥+5 kcal/mol (for this dataset). The D-optimal variable selection procedure was then applied in three runs, obtaining a reduction of variables from 8400 to 2000, from 2000 to 1000, and from 1000 to 500 for each run, respectively. In the second step, an FFD strategy was used in order to build the combination matrix using a 2:1 ratio of combinations/variables and a 4:1 ratio of true/dummy variables. And finally, in the third step the fixing/excluding procedure was used for the final vari-

Scheme 1. Flow-Chart Showing the Multiple Levels of Data Modification Generally Used in a GOLPE Procedure or Other 3D-QSAR Methods[a]



[a] A force field calculation produces an **X** descriptor matrix which may be modified at levels one, two, or three and finally at level four to obtain the predictive statistical model. The solid arrows indicate the most appropriate procedure found in this study for a GRID force field and GOLPE variable selection. Data pretreatment procedures i–vi correspond to those given in Table 2.

able selection starting from the set of 500 variables and using a 2024 column combination matrix. The final model selects only 165 out of the original 8400 variables.

Results

Effect of Different Data Pretreatment Methods on the Predictivity. Chemometric analyses give different results when different weights are given to the variables. When prior information is available about the relative importance of the variables, weights should be assigned proportional to this contribution. It is known that the results obtained from projection methods such as PLS depend on scaling of the data and that the initial variance of a variable partly determines its importance in the final model.[31] However, in general, no prior information regarding the contribution of each variable is known, and therefore a general method of weighting data is required.

Autoscaling is a standard weighting method which generates the values of the weights that are equal to the inverse of the standard deviations of the variables. After autoscaling, each variable has a unit variance. This unit variance gives all the variables the same initial importance. Consequently, autoscaling may assign significance to those variables which exhibit only small variations and therefore do not reflect real structural variations. To avoid this problem, the minimum-σ cut-off procedure is used. Technically, this procedure eliminates from the analysis those variables which have a variance less than the minimum-σ value.

In 3D-QSAR analyses the resulting unsymmetrical distribution of the interaction energies requires a simple

Table 2. Summary of the Data Set Pretreatment Used in This Work and of the Final Models Obtained by GOLPE Variable Selection

data pretreatment: field cut-off (kcal/mol), scaling, and minimum-σ cut-off (kcal/mol)	no. of active variables in the final model[a]	optimal dimension[b]	SDEC[c]	R^2	SDEP[d]	Q^2
(i) field cut-off = +5 on positive variables, autoscaling at level 1, minimum-σ cut-off = 0.5	101	4	0.44	0.93	0.74	0.81
(ii) field cut-off = +5 on positive variables, no scaling at level 1, autoscaling at level 2 of the variables selected by the D-optimal algorithm, minimum-σ cut-off = none	165	3	0.48	0.92	0.78	0.80
(iii) field cut-off = +30 on positive variables, no scaling at level 1, autoscaling at level 2 of the variables selected by the D-optimal algorithm, minimum-σ cut-off = none	169	2	0.73	0.82	0.96	0.67
(iv) field cut-off = +5 on positive variables, no scaling at level 1, no scaling at level 2, minimum-σ cut-off = none	134	4	0.50	0.91	0.97	0.68
(v) field cut-off = +5 on positive variables, no scaling, no variable selection, minimum-σ cut-off = 1.0	1050	2	0.68	0.84	1.62	0.05
(vi) field cut-off = +5 on positive variables, autoscaling, no variable selection, minimum-σ cut-off = 1.0	1050	3	0.72	0.82	1.64	0.04

[a] Number of active variables in the final model were obtained using the fractional factorial selection procedure with 25% dummy variables and a variable combination ratio equal to 2. [b] Optimal dimension is the number of LV for a model for which there is the minimum estimated prediction error. [c] SDEC is the standard deviation of error of calculation in the fitting procedure. [d] SDEP is the standard deviation of error of prediction computed as in ref 10.

method of centering the data. Using an appropriate cut-off on the field values, whereby only high interaction energy values are reduced, a well-distributed shape of the data may be obtained.

Scheme 1 shows an overall view of the different data pretreatment and variable selection procedures. It should be noted that the scaling procedure may be performed at level one on data pretreatment or, alternatively, at level two only after a reasonable amount of noise has been eliminated by D-optimal variable selection. This work shows that this is not a trivial point since the results obtained depend critically on the procedure selected.

In an effort to assess the effect of the different procedures on producing a model that accurately represents the data with optimum predictive power, each of the pretreatments shown in Scheme 1 were performed: (i) ordinary autoscaling performed using the entire data set at level one, (ii) no scaling at level one but autoscaling at level two on a subset of variables selected using a D-optimal algorithm, (iii) no scaling at level one but autoscaling at level two on a subset of variables selected using a D-optimal algorithm using an initial cut-off of +30 kcal/mol on the interaction energy in the GRID calculation (as opposed to +5 kcal/mol used in all the other calculations), (iv) no scaling procedure at level one or at level two, (v) autoscaling at level one with a minimum-σ cut-off = 1 kcal/mol without performing any variable selection, and finally, (vi) using no scaling or variable selection, but only a minimum-σ cut-off of 1.0 kcal/mol. The results of the data pretreatment procedures illustrated in Scheme 1 are presented in Table 2.

Each data pretreatment procedure produces a final model which can be evaluated by either SDEP (eq 4) or Q^2 defined as

$$Q^2 = 1 - \sum_{i=1}^{N}(y_i - Y_{i_{\text{pred}}})^2/\sum_{i=1}^{N}(y_i - y_{\text{im}})^2 \qquad (5)$$

where y_{im} = mean value.

These parameters are analogous to the standard deviation of error of calculations (SDEC) defined as

$$\text{SDEC} = [\sum_{i=1}^{N}(y_i - y_{i_{\text{calc}}})^2/N]^{1/2} \qquad (6)$$

where $y_{i_{\text{calc}}}$ = calculated value and R^2 defined as

$$R^2 = 1 - \sum_{i=1}^{N}(y_i - y_{i_{\text{calc}}})^2/\sum_{i=1}^{N}(y_i - y_{\text{im}})^2 \qquad (7)$$

Both R^2 and Q^2 parameters are dimensionless with values which lie between 0 (no correlation in the data) and 1 (maximum correlation in the data). On the other hand, SDEC and SDEP have units of the actual y values, and since they represent the errors in either fitting (SDEC) or prediction (SDEP) these should be minimized.

Examination of the results of procedures i–iv in Table 2 shows that the accuracy of the model in fitting (SDEC, R^2) is generally not influenced by different data pretreatments while in prediction (SDEP, Q^2) different results are obtained. However, the similarity of the values obtained by procedures i–ii and iii–iv in Table 2 illustrates the difficulty in selecting the most appropriate data pretreatment procedure only on the basis of R^2 and Q^2. Therefore, it is necessary to find an alternative method for evaluating the effect of the scaling procedures on the final model. Plotting the LV coefficients corresponding to each data pretreatment procedure allows a direct comparison of each procedure on the prediction of the important three-dimensional regions. The regions determined by each calculated model may then be compared to the actual experimental regions determined by the X-ray crystallographic binding studies. In this way, the effect of the different data pretreatment procedures on the predictive ability of the final model can be assessed.

Data Pretreatment Procedure i. A detailed examination of the important three-dimensional regions found by the different data pretreatment procedures shows that procedure i produces chance correlations which are reflected in the overestimation of regions where it is known from the three-dimensional structure that there are no possibilities of such interactions. Comparison of Figures 1 and 2 illustrate that interac-

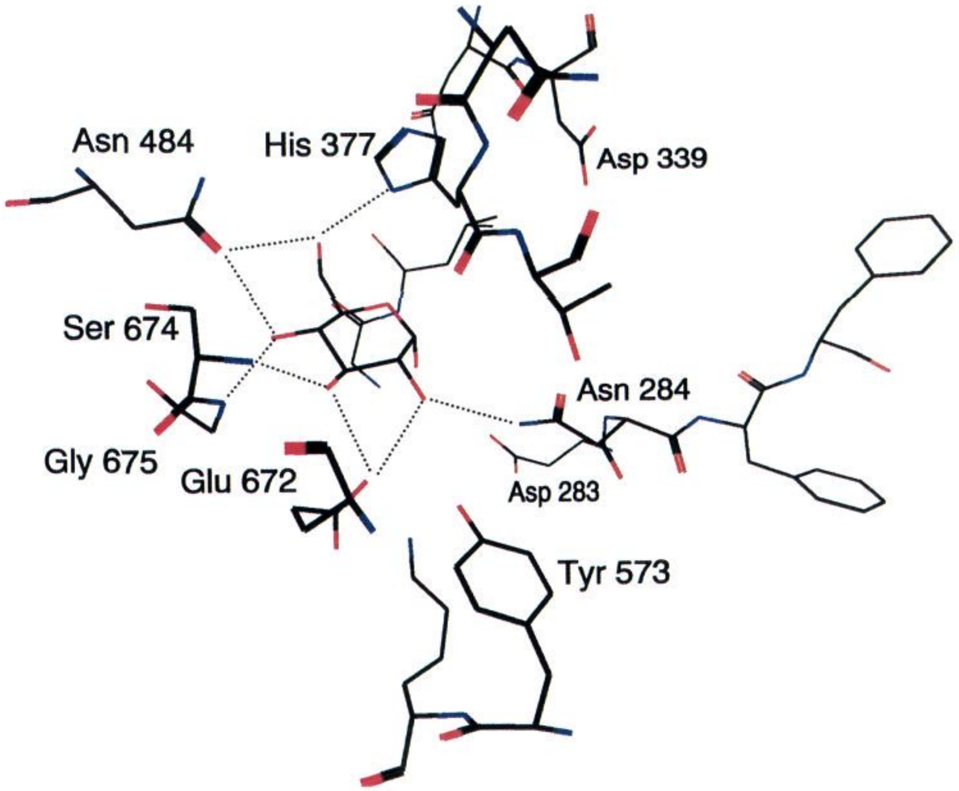


Figure 1. The active site of glycogen phosphorylase *b* showing the amino acid residues which interact with α-D-glucose (compound **1**) as a representative molecule. It has been shown[17,18] that maintaining these interactions (with the labeled residues) are important to the activity.

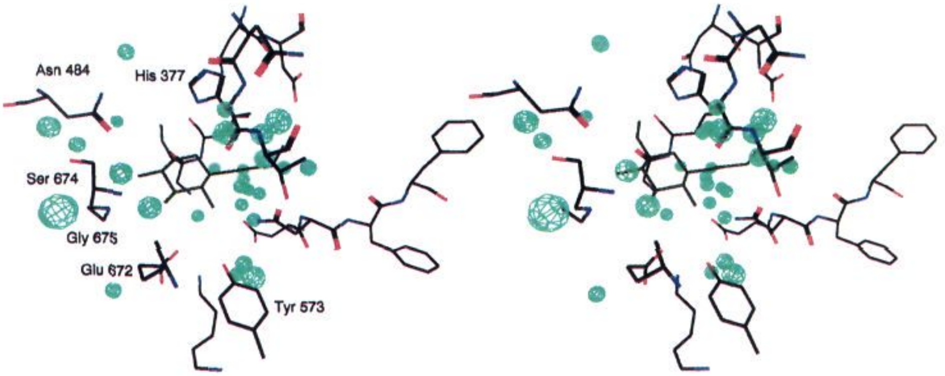


Figure 2. Contour map (in stereoview) of the PLS coefficient values (cut-off ≥ |0.4|) at the fourth PC for the model of the interactions between the OH probe and all 36 targets molecules using data pretreatment procedure i. The contour regions near the active-site residues are not well reproduced compared with Figure 3 produced by procedure ii. In particular, the regions near Tyr573, Glu672, Gly675, Ser674, Asn484, and His377 are not well modeled using this procedure and the new regions which appear are related to noise that results from using this pretreatment method. Only compound 36 is shown for clarity.

tions in the regions between the glucose analogue and GP*b* catalytic site residues Glu672, Gly675, Ser674, His377, Tyr573, and Asn484 are not well predicted by this method but are known from the binding studies to produce significant effects on the binding affinity. There are several regions between Asn284, Asp283, and Leu136 that are predicted but are known from the binding study to play no significant role. This is due to the fact that performing autoscaling at the beginning using all the data results in an overestimation of the small variables and the noise, consequently increasing the risk of chance correlations.

Data Pretreatment Procedure ii. All the projection methods, such as PLS, require autoscaling to give a good estimate of the latent variables and loading coefficients. However, data pretreatment i has shown the risk associated with autoscaling of all the data from the beginning. Thus, Scheme 1 shows that scaling may be performed after a preliminary variable selection. An efficient way for the purpose of this selection is to choose variables in the loadings space according to a D-optimal design. The D-optimality criterion selects variables in such a way that most of the redundant data is discarded, while enough useful information is retained.

Procedure ii illustrates the effect of performing autoscaling at level two (Scheme 1) after a reasonable amount of noise has been eliminated by a D-optimal variable selection. Comparison of Figures 1 and 3 shows

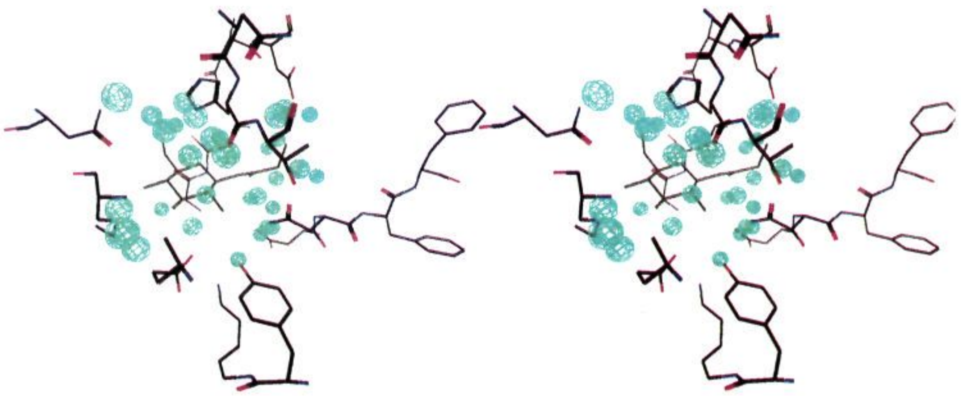


Figure 3. Contour map (in stereo view) of the PLS coefficient values (cut-off ≥ |0.4|) at the third PC for the model of the interactions between the OH probe and all 36 target molecules using data pretreatment procedure ii. The illustrated regions represent the predicted significant positions that may be related to a change in the inhibition constant. These contours are obtained without using any knowledge of the actual three-dimensional structure of the ligand−GP*b* complexes. These positions show good correlation to those active-site residues which interact with the ligands as shown in Figure 1. Only compound 36 is shown for clarity.

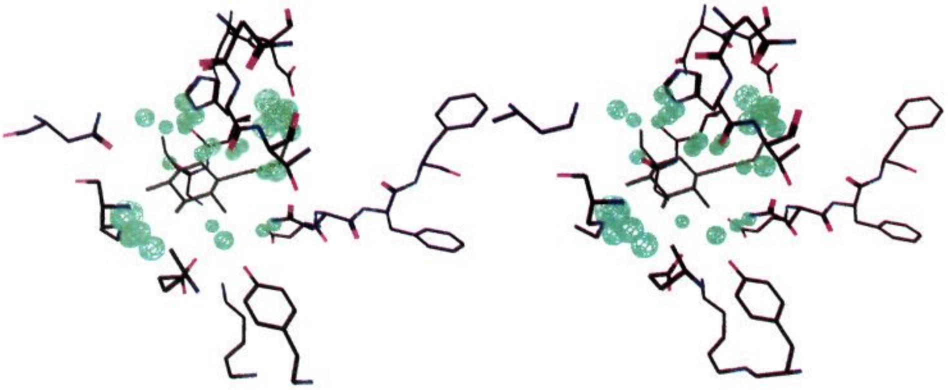


Figure 4. Contour map (in stereo view) of the PLS coefficient values (cut-off ≥ |0.4|) at the second PC for the model of the interactions between the OH probe and all 36 targets molecules using data pretreatment procedure iii. Shows reasonably good agreement with experimental regions identified in Figures 1 and 3 but results in a poor statistical model for the data ($Q^2 = 0.67$, Table 2).

that there is good agreement between the predicted regions and the experimental regions (as identified by the X-ray crystallographic binding studies) for the data pretreatment procedure ii. In particular, the interactions between residues His377, Glu672, Gly675, Ser674, Tyr573, and Asn484 are well reproduced and identified. This is a more effective procedure for improving the prediction capability of the model. Comparison of Figures 2 and 3 illustrates the importance of using procedure ii in transforming the original data and reproducing the information known from the three-dimensional X-ray structure, since there are clearly regions of importance that are not reproduced by procedure i, and subsequently these are absent from Figure 2. Once the best scaling procedure has been determined, then different models of the data (for example, using various probes or force fields to calculate the interaction energies) may be evaluated on the basis of the R^2 and Q^2 values.

Data Pretreatment Procedure iii. The results obtained by procedure iii using a high cut-off of positive energy variables ($E_{\max} = 30$ kcal/mol) in the GRID calculation followed by data pretreatment procedure ii show regions in relatively good agreement with the experimental results (Figure 4), although the numerical results are by comparison much poorer ($Q^2 = 0.67$, Table 2). This demonstrates the importance of a good distribution of the energy variables. In fact, a high positive cut-off leads to instability in the coefficients of the model, while a positive energy cut-off similar to the maximum negative energy value (abut −7 kcal/mol in this example) results in more symmetrically distributed variables and increases the stability of the model. This result is in agreement to that found by Klebe et al.[32] using the CoMFA force field.

Data Pretreatment Procedure iv. Performing GOLPE on the data set without autoscaling (procedure iv in Table 2) shows that no all the experimental regions are identified. Comparison of Figures 1 and 5 shows good agreement between the experimental regions and the important variables. However, comparison of Figures 3 and 5 shows that not all the experimental regions are identified using procedure iv. Specifically, the regions close to residues Asn484, His377, and Tyr573 are not identified when no autoscaling of the data is performed. The numerical results confirm a loss of predictivity in the case of procedure iv ($Q^2 = 0.68$, Table 2).

When autoscaling is not performed, the initial importance of the variables is proportional to the relative magnitude of the interaction energy. It may be the case that there are some strong interactions between a given probe and a target; however, the corresponding three-dimensional regions may not necessarily relate to the activity. Therefore, without scaling the model is forced to take into account this effect. And since the model is influenced by those regions with large energy values,

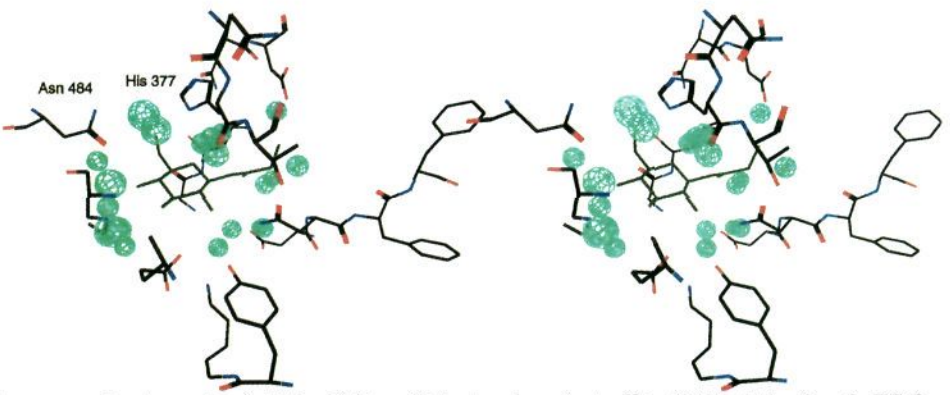


Figure 5. Contour map (in stereoview) of the PLS coefficient values (cut-off ≥ |0.4|) at the fourth PC for the model of the interactions between the OH probe and all 36 targets molecules using data pretreatment procedure iv. Not all the experimental regions that relate to the activity are identified as in procedure ii shown in Figure 3. In particular, the 3D regions near His377 and Asn484 are not predicted by this model.

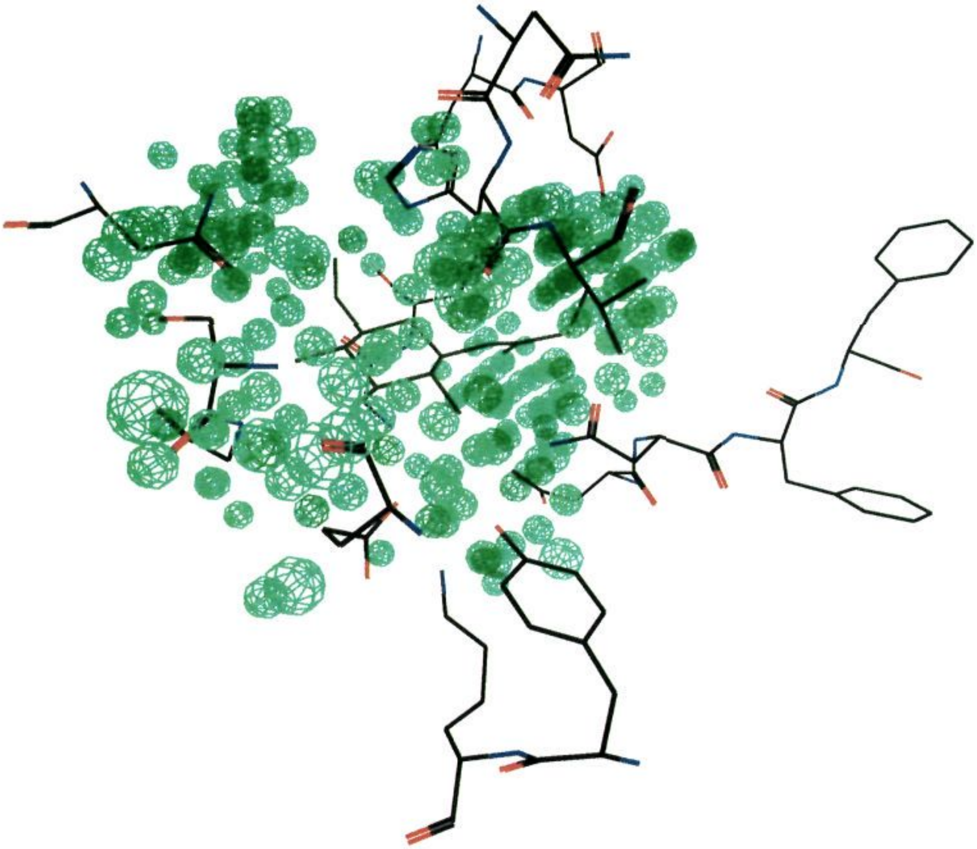


Figure 6. Contour map of the PLS coefficient values at the third PC for the model of the interactions between the OH probe and all 36 target molecules using the minimum-σ approach v. Only high-value coefficients (cut-off > |0.4|) are shown for clarity. The large amount of information retained by this method make it impossible to distinguish between significant regions and noise.

regions with small energy values may not be picked up by the model even though they may be significant.

Data Pretreatment Procedures v and vi. Finally, the models obtained by elimination of small (1.0 kcal/mol) standard deviation variables without any variable selection (procedures v and vi in Table 2) are shown to contain not only the important experimental regions but also an elevated number of regions which do not fit the information known from the crystallographic data. Numerical and graphical analyses show that both models behave in a similar way (Table 2, Figure 6). The numerous regions found by procedures v and vi (illustrated in Figure 6 for pretreatment v) render it impossible to interpret the significance of these predictions. Clearly, the elimination of variables with a small standard deviation does not eliminate enough noise, and autoscaling is ineffective when applied under this condition.

Choice of the Most Appropriate Pretreatment. It is evident that a single numerical comparison based on either the Q^2 or SDEP parameters between models obtained from different pretreatments of the same data set, as in Table 2, is not sufficient to select the best model. However, led by knowledge of the real physical problem which allows comparison of the calculated to the actual experimental regions, it is possible to suggest

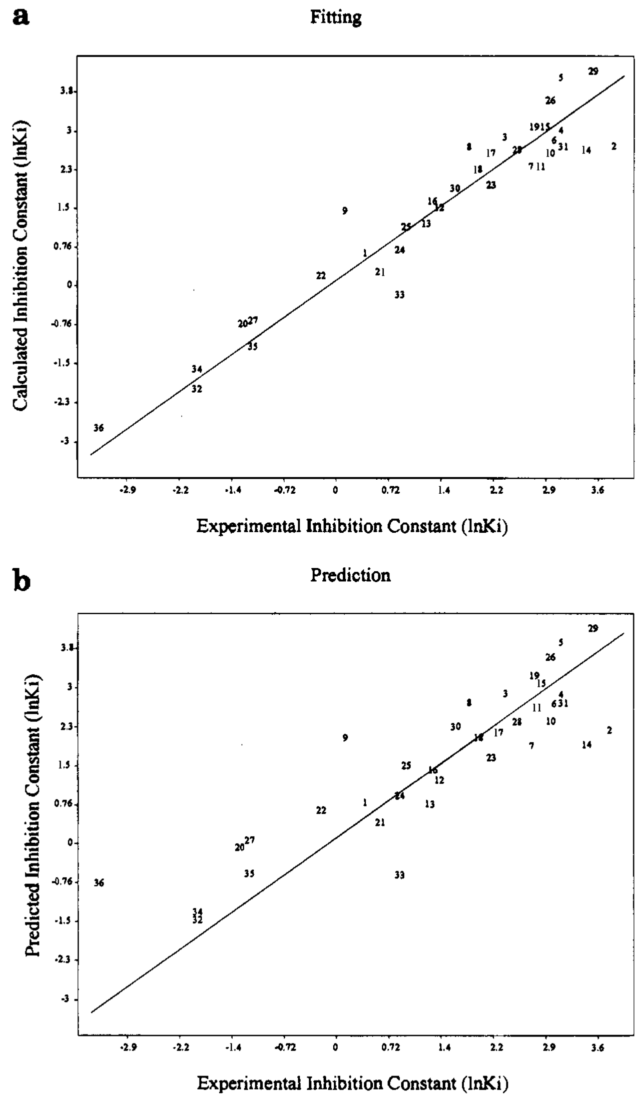


Figure 7. Experimental inhibition constant versus (a) calculated and (b) predicted values from the model calculated using procedure ii in Table 2. There is good agreement between the experimental and calculated values both in the fitting (Figure 7a) and in the prediction (Figure 7b) capability of the model. Figure 7b indicates that compound **36** is not well predicted. This is expected since this compound is chemically different from all the other compounds in the dataset.

Table 3. Comparison between the Average Experimental Error in the Inhibition Constant Expressed as the Natural Logarithm of K_i in mM and SDEC and SDEP Estimated Errors from the Final Model with an OH Probe Refined by GOLPE

experimental error expressed in terms of ln K_i	SDEC	SDEP
0.68	0.48	0.78

that the pretreatment procedure ii in Table 2 (no scaling from the beginning but before the FFD variable selection) is most suitable for GOLPE variable selection with a GRID force field description of the present system.

With the exception of compounds **2** and **36**, the errors in fitting and in prediction for the final model (Figure 7, Table 3) are in good agreement with the average experimental error in the K_i values. This indicates that GOLPE variable selection with suitable data pretreatment minimizes the risks of overfitting and overpredicting.

Table 4. Influence of Probes and Grid Spacing on the Prediction Capability of GOLPE Models within the Same Data Set

probes	grid spacing (Å)	optimal dimension	SDEC	R^2	SDEP	Q^2
OH	2	3	0.88	0.73	1.21	0.49
OH	1	3	0.48	0.92	0.78	0.80
CH_3	1	3	0.54	0.90	1.10	0.62
C═O	1	3	0.42	0.94	0.78	0.80
Na^+	1	3	0.65	0.85	0.99	0.65

It was expected that compound **36** would be poorly predicted by the best model since it is spatially different from all the other compounds in the data set. This may illustrate the importance of selecting as many structurally different compounds (as well as several of each type of compound) in order to obtain an accurate description of all the compounds in the data set. In addition, compound **2**, the poorest inhibitor, was not well predicted by the chosen model. The crystallographic result showed that compound **2** bound very weakly to the enzyme. The difficulty in accurately fitting the ligand to the active site of the enzyme may therefore be reflected in the inability of the model to accurately predict this compound.

Effect of Different Descriptions on the Predicitivity. With the correct scaling procedure, the parameters Q^2 and SDEP, obtained by GOLPE, can be used to measure the degree of information given back from different descriptions of the same data set. Table 4 shows SDEP and Q^2 parameters for the glucose analogue inhibitor data set described using different probes (hydroxy OH, methyl CH_3, carbonyl oxygen C═O, and cationic Na^+) and different grid spacings (1-Å or 2-Å spacing).

Important information may be lost when either the GRID spacing is too large or the GRID probes are inadequately described. Examination of the values of R^2 and Q^2 in Table 4 show that if the grid spacing is increased from 1 to 2 Å, both the fitting and the predicting capability drop dramatically. The chemical interpretation of this finding is that the 2-Å spacing is too large for sensitive and highly directional interactions such as those found in multiple hydrogen bonds to be adequately defined, and therefore, these interactions are poorly described using the larger grid size or less specific GRID probes. In fact, the 2-Å spacing makes an immense difference to the strength of the hydrogen-bonding interactions, and it is exatly these interactions that confer specificity to the ligands. This work shows that the GOLPE variable selection procedure is so powerful that a 1-Å spacing using good GRID probes is sufficient for eliminating noisy variables while retaining only relevant information.

Higher values of R^2 and Q^2 (Table 4) for the OH and C═O probes illustrate that these probes are more favorable for making predictions in this system than either the Na^+ or CH_3 probes. This again demonstrates the importance of considering hydrogen-bond interactions and implies that there are more interactions between hydroxyl groups on both the active-site residues of the protein and the ligands. In this particular case with GP*b* the results indicate that the active site contains many more polar regions than nonpolar or charged regions. Using the knowledge of the three-dimensional structure of this enzyme, it may be concluded that the GOLPE predictions are consistent with what is experimentally observed for the ligand–enzyme bound structures. Thus, it seems apparent that the appropriate GOLPE procedure accurately predicts the type and location of known regions of importance (for biological activity) using only the information contained in the ligand molecules. The significance of such a result has direct application toward the possible construction of probable active-site residues in an unknown protein.

The effect of selecting different dielectric constants for the GRID calculations was also investigated. In the GRID force field the dielectric constant influences the interaction energies and the shape of the field. These results indicate that the dielectric constant may also influence the predictivity of the models. For this data set, the best Q^2 prediction value was obtained using a dielectric constant between 10 and 20. Generally, a dielectric constant of 80 is chosen to mimic the effect of bulk water; however, it is known that the active site of GP*b* is only partially hydrated in the presence of the drug molecules. Thus, the value obtained for the dielectric constant is consistent with that expected for a buried active site within a protein, as is the case in the GP*b* structure.

Conclusions

This work has shown that the selection of variables for making predictions is affected by different data pretreatments. Detailed examination of the calculated and experimental three-dimensional regions has shown the importance of selecting an appropriate scaling procedure. Comparison of the different data pretreatments illustrates the difficulty in relying on only the values of R^2 and Q^2 to select the best strategy.

In this case, led by the knowledge of the three-dimensional crystallographic structure, it was possible to determine unequivocally the best computational method for the pretreatment of the data that accurately reproduced the experimental result. This was obtained using the D-optimal preselection with no autoscaling at level 1 followed by the FFD selection with autoscaling performed at level 2 (Scheme 1).

It was clear that, in a 3D-QSAR context, autoscaling can give too much importance to those variables which in fact have only a small influence and therefore do not reflect real structural variations. However, it was evident that once most of the noise has been eliminated from the data, autoscaling improves the performance of the PLS models and GOLPE variable selection. It was also apparent that using the minimum-σ cut-off is not sufficient in eliminating all the noise from the data, and therefore autoscaling after such a procedure should be used with caution.

It has been shown that the power of a GOLPE procedure is not only in the capacity to improve the predictivity of a model but also in the ability to retain only relevant information from large amounts of redundant data. Both the fitting and the prediction errors in the final model have been shown to be similar to the average experimental error in the observed inhibition constants. This demonstrates that GOLPE variable selection minimizes the risks of overfitting and overpredicting when preceded by an appropriate data pretreatment.

By choosing the same data pretreatment procedure for the selection of the variables, the prediction ability of different models (produced by using different dielectric constants, probes, and grid spacing) may be compared and the most accurate model for a given study may be chosen. Only at this stage may the values of R^2 and Q^2 be used to assess the quality of the individual models. Naturally, this cannot be generalized by these results since the best model will depend on the target structures, the probes, the force field, and the data set under investigation.

From the X-ray crystallographic binding studies it has been possible to use a structure-based ligand design strategy to determine which interactions have a significant contribution to the biological activity. This work shows that using the program combination GRID/GOLPE (GRID to describe the molecular structures and GOLPE performed with the appropriate data pretreatment and variable selection) may provide an objective method of obtaining similar information. However, since the latter method gives quantitative results, the possibility arises of observing features which may have been previously overlooked in the crystallographic analyses.

The information gained from this analysis offers the possibility of predicting new molecules in a rational manner using the best model possible for the GP*b* data set. These results show that the model has good predictive ability and has reliably chosen regions of biological significance consistent with both the X-ray crystallographic and kinetic results.

Experimental Section

GRID/GOLPE Analysis. All programs and computations were carried out on an R4000 CRIMSOM SGI computer. A UNIX version of GOLPE (Version 1.0)[11] takes 10 h to refine a model with 8400 total initial variables, single user. The program GRID (Version 9.0)[22-24] produces the three-dimensional matrix for all of the 36 compounds with the OH probe in 1 min, single user.

Display of the GOLPE contour maps and structures were performed using the molecular graphics routine in GOLPE implemented on a Silicon Graphics CRIMSON terminal. Details of the individual complexes are published elsewhere.[17,18]

X-ray Crystallographic and Kinetic Binding Studies. Crystals of T state GP*b* were grown as detailed previously.[33] The inhibition constants were established from kinetic binding experiments in which the enzyme was assayed in the direction of glycogen synthesis and activity determined from measuring the inorganic phosphate released.[17,18] The crystallographic binding studies were carried out by soaking T state GP*b* crystals for 1 h in a buffered solution (10 mM *N*,*N*-bis(2-hydroxyethyl)-2-aminoethanesulfonic acid (BES), 0.1 mM ethylenediaminetetraacetic acid (EDTA) pH 6.7) containing approximately 100 mM of the ligand under study.

Data (to 2.3-Å resolution) were collected on a Nicolet IPC multiwire area detector using a Rigaku RU-200 rotating anode X-ray source and processed (Program XENGEN[34]) to give data sets that typically were 80% complete with merging *R* values of 7% as detailed elsewhere.[17,18]

The data were scaled to native to produce a set of structure factors, which were used to generate difference Fourier electron density maps using the CCP4 package of crystallographic programs. Difference maps were calculated and examined for the presence of the ligand using the molecular graphics program FRODO[35-36] implemented on an Evans and Sutherland PS390 graphics terminal.

Final refinement of the phosphorylase–ligand complexes was performed on a Convex C210 using X-PLOR energy[20] and crystallographic least-squares minimization. Individual atomic *B*-factor refinement was performed in the final cycles resulting in final *R* values less than 0.20. Potential hydrogen bonds were assigned if the distance between two electronegative atoms was less than 3.3 Å and if the angle formed between these two atoms and the preceding atom was greater than 90°.

Acknowledgment. This work has been supported by the MRC (K.A.W.), the Oxford Centre for Molecular Sciences (K.A.W.), and by the Italian Government MURST (G.C.). We thank Professors L. N. Johnson and S. Clementi for helpful comments on the manuscript. Dr. G. W. J. Fleet and co-workers for the synthesis of the glucose analogues and Dr. N. G. Oikonomakos and co-workers for kinetic studies are gratefully acknowledged.

References

(1) Cramer, R. D. III; Patterson, D. E.; Bunce, J. D. Comparative Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.

(2) Wold, S.; Ruhe, A.; Wold, H.; Dunn, W. J. III. The collinearity problem in linear and non linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM J. Sci. Stat. Comput.* **1984**, *5*, 735–743.

(3) Kellogg, G. E.; Abraham, R. J. *J. Mol. Graph.* **1992**, *10*, 212–217.

(4) Morley, S. D.; Abraham, R. J.; Haworth, I. S.; Jackson, D. E.; Saunders, M. R.; Vinter, J. G. COSMIC (90): An Improved Molecular Mechanics Treatment of Hydrocarbons and Conjugated Systems. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 475–504.

(5) Shoichet, B.; Bodian, D. L.; Kunitz, I. D. Molecular Docking Using Shape Descriptors. *J. Comput. Chem.* **1992**, *13*, 380–397.

(6) Bohm, H. J. The Computer Program LUDI: A New Method for the *novo* Design of Enzyme Inhibitors. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 61–78.

(7) Baroni, M.; Clementi, S.; Cruciani, G.; Costantino, G.; Riganelli, D. Predictive Ability of Regression Models. Part II: Selection of the Best Predictive PLS Model. *J. Chemom.* **1992**, *6*, 347–356.

(8) Wold, S. Validation of QSAR's. *Quant. Struct.-Act. Relat.* **1991**, *10*, 191–193.

(9) Cramer, R. D. III; Bunce, J. D.; Patterson, D. E.; Frank, I. Crossvalidation, bootstrapping, and partial least squares compared with multiple regression in conventional QSAR studies. *Quant. Struct.-Act. Relat.* **1988**, *7*, 18–25.

(10) Cruciani, G.; Baroni, M.; Clementi, S.; Costantino, G.; Riganelli, D.; Skagerberg, B. Predictive Ability of Regression Models. Part I: Standard Deviation of Prediction Errors (SDEP). *J. Chemom.* **1992**, *6*, 335–346.

(11) 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. *Quant. Struct.-Act. Relat.* **1993**, *12*, 9–20.

(12) Allen, M. S.; La Loggia, A. J.; Dorn, L. J.; Martin, M. J.; Costantino, G.; Hagen, T. J.; Koeheler, K. K.; Skolnick, P.; Cook, J. M. Predictive Binding of *β*-Carboline Inverse Agonists and Antagonists via the CoMFA/GOLPE Approach. *J. Med. Chem.* **1992**, *35*, 4001–4010.

(13) Good, A. C.; So, S. S.; Richards, W. G. Structure Activity Relationships from Molecular Similarity Matrices. *J. Med. Chem.* **1993**, *36*, 433–438.

(14) Cruciani, G.; Clementi, S.; Baroni, M. Variable Selection in PLS Analysis. In *3D QSAR in Drug Design*; Kubinyi, H., Eds.; ESCOM: Leiden, 1993; pp 551–564.

(15) Good, A. C.; Peterson, S. J.; Richards, W. G. QSAR's from Similarity Matrices. Technique Validation and Application in the Comparison of Different Similarity Evaluation Methods. *J. Med. Chem.* **1993**, *36*, 2929–2937.

(16) Cocchi, M.; Menziani, M. C.; De Benedetti, P. G. Use of advanced chemometric tools and comparison of different 3D descriptors in QSAR analysis of prazosin analogs alpha-1-adrenergic antagonists. In *Trends in QSAR and Molecular Modelling 92*; Wermuth, C. G., Eds.; ESCOM: Leiden, 1993; pp 527–529.

(17) Martin, J. L.; Veluraja, K.; Ross, K.; Johnson, L. N.; Fleet, G. W. J.; Ramsden, N. G.; Bruce, I.; Ochard, 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.

(18) 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.; Papageorgiu, A. C. The Design of Inhibitors of Glycogen Phosphorylase: A Study of α and *β* C-Glucosides and 1-Thio-*β*-D-Glucose Compounds. *Biochemistry*, in press.

(19) 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.

(20) Brunger, A. T. A Memory-Efficient Fast Fourier Transformation Algorithm for Crystallographic Refinement on Supercomputers, *Acta Crystallogr.* **1989**, *A45*, 42–50.

(21) Oikonomakos, N. G. Unpublished results.

(22) Goodford, P. J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28*, 849–857.

(23) Boobbyer, D. N. A.; Goodford, P. J.; McWhinnie, P. M.; Wade, R. C. New Hydrogen-Bond Potentials for Use in Determining Energetically Favorable Binding Sites on Molecules of Known Structure. *J. Med. Chem.* **1989**, *32*, 1083–1094.

(24) Wade, R.; Clerk, K. J.; Goodford, P. J. Further development of hydrogen bond function for use in determining energetically favourable binding sites on molecules of know structure. Ligand probe groups with the ability to form two hydrogen bonds. *J. Med. Chem.* **1993**, *36*, 140–147.

(25) Cruciani, G.; Goodford, P. J. A Search for Specificity in DNA-Drug Interactions. *J. Mol. Graph.*, in press.

(26) Hoskulsson, A. PLS Regression Methods. *J. Chemom.* **1988**, *2*, 211–218.

(27) Wold, S. Cross-Validation Estimation of the Number of Components in Factor and Principal Components Models. *Technometrics* **1978**, *20*, 397–405.

(28) Mitchell, T. J. An Algorithm for the Construction of "D-Optimal" Experimental Designs. *Technometrics* **1974**, *16*, 203–210.

(29) Steinberg, D. M.; Hunter, W. G. Experimental Design Review and Comment. *Technometrics* **1984**, *26*, 71–76.

(30) Baroni, M.; Clementi, S.; Cruciani, G.; Kettaneh, N.; Wold, S. D-Optimal Design in QSAR. *Quant. Struct.-Act. Relat.* **1993**, *12*, 225–231.

(31) Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52.

(32) Klebe, G.; Abraham, U. On the prediction of binding properties of drug molecules by comparative molecular field analysis. *J. Med. Chem.* **1993**, *36*, 70–80.

(33) Oikonomakos, N. G.; Melpidou, A. E.; Johnson, L. N. Crystallization of Pig Skeletal Phosphorylase *b*. *Biochim. Biophys. Acta* **1985**, *832*, 248–256.

(34) Howard, A. J. *A Guide to Macromolecular X-Ray Data Reduction for the Nicolet Area Detector*: *The Xengen System, version 1.3*: Protein Engineering Dept., Genex Corp.: Gaithersburg, MD, 1988.

(35) Jones, T. A. A Graphics Model Building and Refinement System for Macromolecules. *J. Appl. Crystallogr.* **1978**, *11*, 268–272.

(36) Jones, T. A. Interactive Computer Graphics: FRODO. *Methods Enzymol.* **1985**, *115*, 157–171.