Smart Region Definition: A New Way To Improve the Predictive Ability and Interpretability of Three-Dimensional Quantitative Structure−Activity Relationships

Manuel Pastor, Gabriele Cruciani,* and Sergio Clementi

Laboratory of Chemometrics, Department of Chemistry, University of Perugia, Via Elce di Sotto 8, 06123 Perugia, Italy

Received November 21, 1996[⊗]

This report describes a new methodology aimed at grouping 3D-QSAR interaction energy descriptors into regions of neighbor variables bearing the same chemical and statistical information. These regions represent the structural variability of the series better than individual descriptor variables and can advantageously replace them in the chemometric analysis. The algorithm used to generate such regions is described, together with their application for improving the quality of GOLPE variable selection. The method is illustrated on a series of 47 glucose analogues, inhibitors of glycogen phosphorylase *b*, and is shown to improve both the predictive ability and the interpretability of the 3D-QSAR models obtained, comparing favorably with other methods previously described.

Introduction

In 3D-QSAR methodologies[1−4] the series of compounds are described by a vast number of grid-field variables. These variables are obtained by defining a three-dimensional grid around the compounds and calculating at each node the energy of interaction between the compound and a “probe”, which represents a chemical group. The idea behind these methodologies is that changes in the structure of the compounds will induce changes in the field variables, which somewhat represent the interaction of the compounds with the receptor and can be correlated with the activity by a partial least squares (PLS) model.

It is evident that even the smaller structural change in the compounds will not be reflected in a single variable but rather in a group of field variables that are spatially contiguous. These groups represent portions of the space surrounding the compounds which are affected in the same way by the structural variations in the series, and as a consequence, all variables inside the group bear the same information. However, when the 3D matrices of energies are unfolded into vectors to build the matrix of descriptors **X** the grid variables are considered individually and therefore the information contained in their positions in the 3D space is lost.

This report describes an original methodology called smart region definition (SRD)[5,6] aimed at extracting from a matrix of 3D-QSAR descriptors groups of neighbor variables in the 3D space (*regions*) bearing the same information. Such groups take into account the spatial continuity constraint (neighbor variables containing similar information) in order to produce more stable models, less prone to chance correlations and easier to interpret. Here we describe their use in GOLPE[7−9] variable selection where they replace the role of the individual variables.

There are many ways in which the *X* variables (grid nodes) can be grouped. The state-of-the-art in the field is represented by the methods reported by Cho and Tropsha[10] and Norinder,[11] who group the variables into square boxes of fixed size following only a geometrical criterion. On the contrary, the SRD procedure works by extracting a subset of highly informative *X* variables and then partitioning the space around the molecules amongs them. The regions formed following such a scheme have a shape and a size which depends upon the amount of information contained by the variables; areas rich in information contain many informative variables which compete for the space, thus producing many small regions, while areas poor in information will contain few large regions. Consequently, the regions produced by the SRD method tend to contain single, independent pieces of information. In this sense SRD represents a major improvement with respect to the present methods of grouping variables. The Cho and Tropsha[10] and Norinder[11] approaches do not guarantee that each box contains a single different piece of information; some boxes will contain little or no information, while others will contain many distinct pieces of information. Moreover, some pieces of information can be split into two or more contiguous boxes.

The regions generated by SRD have been used to improve variable selection in the GOLPE[7−9] procedure. The combined SRD/GOLPE method evaluates the effect of regions of variables, instead of individual variables, on the predictive ability of the PLS model. Finally, the regions (rather than the individual variables) not contributing to increasing the predictive power of the model will be removed from the model. The advantage of using regions in the procedure is 2-fold: first, the analysis takes into account the information about their 3D position, thus introducing a new continuity constraint which minimizes the risk of chance effects and leads to more predictive models. Second, the selected variables are grouped in the space and so are the results of the PLS analysis, thus greatly increasing its interpretability. In addition, as the number of regions is significantly smaller than the number of variables, the SRD/GOLPE method does not require any more variable preselection and the computations are performed in a fraction of the time required for the regular fractional factorial design (FFD) variable selection.[7−9]

On the other hand, it is doubtful that the boxes

[⊗] Abstract published in *Advance ACS Abstracts,* March 15, 1997.

generated by the Cho and Tropsha and Norinder methods can be successfully used in variable selection because, as mentioned above, they do not contain unique information. Norinder, although using a design criterion in a GOLPE-like fashion, reported only marginal improvements in the predictive ability. The Cho and Tropsha method can be criticized also because the effect of each box on the predictive power of the model is evaluated individually (one box at a time) without using any design criteria for selecting a representative number of box combinations.

The SRD Algorithm

With the aim of obtaining a grouping of variables with the aforementioned characteristics, we developed a computational algorithm which involves three major steps: (1) selecting the most informative variables (*seeds*) in an initial PLS model, (2) building Voronoi polyhedra around the seeds containing neighboring variables in 3D space, (3) merging of the polyhedra containing similar information into larger regions. It should be noticed that step 1 is performed in the chemometric space of the PLS weights on the whole **X** matrix, while steps 2 and 3 are performed in the real 3D space around the molecules and are repeated separately for each field or probe used to describe the compounds.

Step 1. Selecting Seeds. The idea is to select a set of variables (grid nodes) important for explaining the activity and as independent as possible of each other. In the subsequent steps of the algorithm they will be used to generate regions, hence their name, seeds. The number of seeds determines the number of regions, and therefore it is important to select a number of seeds large enough to single out every area involved in the interaction.

In order to select these grid nodes, the method has to evaluate the amount of information contained in the variables they represent. The algorithm starts from an initial PLS model and extracts a given number of variables following a D-optimal design criterion in the chemometric space of the PLS weights. Variables selected in such a way represent grid nodes that have a high importance for the model (with high absolute values of the PLS weights) and are, to a large extent, independent of each other.

Step 2. Building Voronoi Polyhedra. The seeds selected in the previous step are placed in the real space around the molecules, in the field to which they belong. Then, each *X* variable in the dataset is assigned to the nearest seed following a Euclidean distance criterion in the 3D space, thus producing a number of Voronoi polyhedra (VP). Variables which are farther than a certain cutoff from each seed are assigned to a special region (region 0) and removed from the analysis. The areas containing a large amount of important information will be populated by many seeds, and this will result in more and smaller VP, while areas which contain less information, if not removed from the analysis, will contain fewer and larger VP. The number of VP can be different in each field: fields containing more information will produce more VP, while the fields less important for the activity will produce fewer VP.

The region 0 contains variables which are far away from any seed. As the seeds are supposed to represent all the important information, variables belonging to region 0 are considered not important for the model and are therefore removed from further PLS analysis. It is interesting to plot the variables in this region in 3D space: they usually highlight areas far away from the compounds, where no interaction is possible, or positions where the compounds in the series exhibit no chemical variation (see Figure 1).

Step 3. Collapsing of Polyhedra. The VP can be used directly as regions, but if neighboring regions do contain the same information they can be profitably combined together to produce larger regions. In order to check whether the regions do actually contain the same information, the algorithm computes the correlation of the information contained in the regions. Only those regions for which this information is strongly correlated are merged into a single common region. The operation is called "collapsing" and is performed as follows. Once the polyhedra are built, the information contained in each one of them is summarized by computing three new parameters. For each polyhedron *i* the following parameters are computed: P_i, average of the values for all the variables (grid node energies) included in the region; P_{+i}, average of the values for all the variables included in the region that take positive values; and P_{-i}, average of the values for all the variables included in the region that take negative values. Notice that P_i, P_{+i}, and P_{-i} are parameters that take a different value for each object molecule, thus defining three vectors throughout all the molecules. It can happen that for a molecule in a certain region one of the elements constituting the P vector never takes either a positive or a negative value. This situation occurs when in that region a molecule contains only positive (repulsive intaraction) variables. In this case P_i and P_{+i} are defined while P_{-i} does not exist and will be handled as a missing value in subsequent computations. Then the algorithm looks for the two nearest seeds (*i* and *j*) and makes pairwise comparisons of the missing value patterns for the plus and minus vectors (P_{+i} and P_{+j}, and P_{-i} and P_{-j}). When the patterns are different, collapsing is aborted. When the patterns are identical, the algorithm computes the correlation coefficient (Pearson's *r*) between the variables P_i, P_{+i}, and P_{+j}. The regions are merged into a new region only if $r(P_i,P_j) > 0.80$, $r(P_{+i},P_{+j}) > 0.50$, and $r(P_{-i},P_{-j}) > 0.50$.

The criterion for collapsing is conservative and prefers not to collapse two regions when they are slightly different. When two regions are merged, a new region is created which contains all the variables originally included in the old ones. In order to define its position in the 3D space, the algorithm calculates the coordinates of a "pseudoseed" as the weighted average of the original seeds coordinates. The procedure continues searching for the nearest seeds and merging them until the distance between them is higher than a given cutoff. If the cutoff distance is set to a high value, all of the regions are applicable for collapsing.

It often happens, when the collapsing cutoff distance is high, that regions far away from each other (even in opposite corners of the grid cage) are merged together. There is nothing wrong in this phenomenon, which reveals two areas that contain correlated information in the actual series (a change in structure in the first area is always accompanied by a similar change in

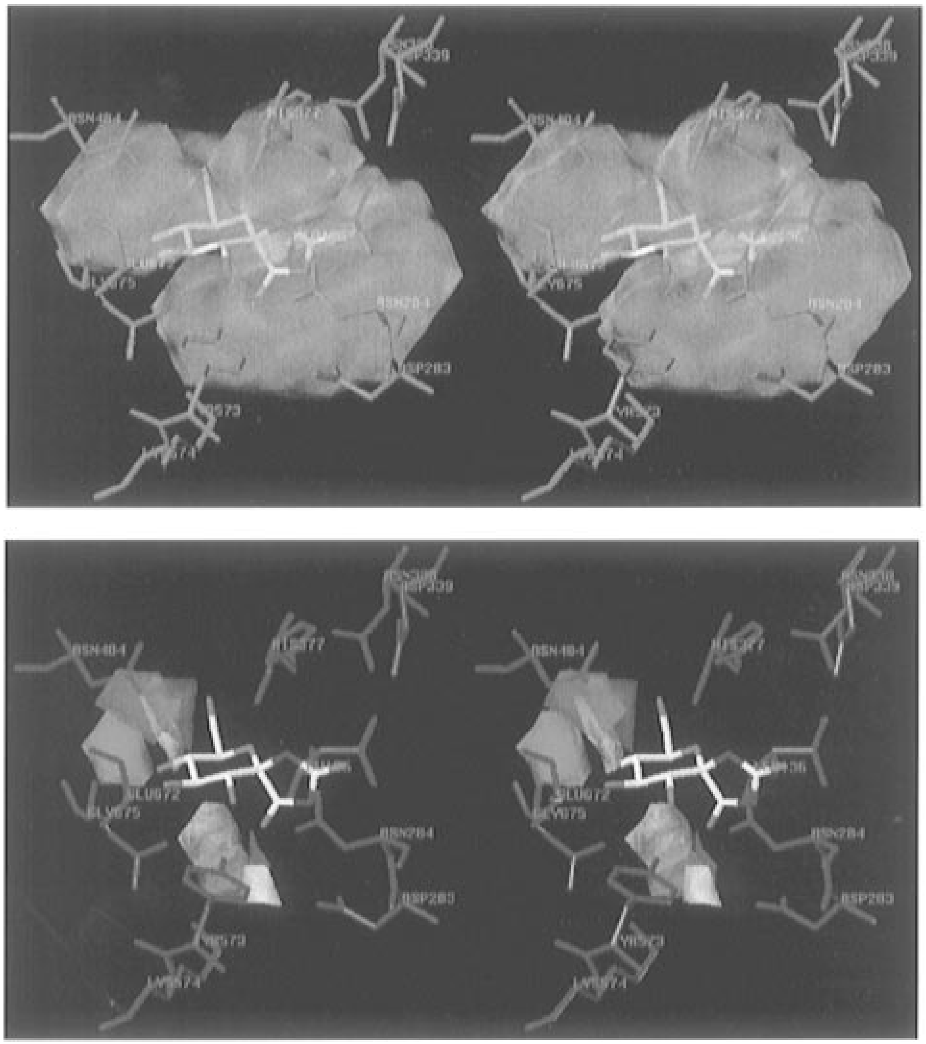


Figure 1. (a, Top) Stereoview of the regions generated by the SRD method. The green volume encloses all the active regions. The area outside of the green volume corresponds to region 0 and contains variables removed from further analysis. (b, Bottom) Stereoview of some regions generated by the SRD method (see text for explanation). In both pictures the structure of the most active compound (**45**) and some residues of the receptor were also included to aid in the interpretation.

structure in the second area). Indeed, it is extremely useful to unveil such internal correlations which otherwise might give rise to hidden misleading effects in the dataset.

Adjustment of the SRD Algorithm. The SRD algorithm contains a number of parameters which can be adjusted to obtain better results: (1) the number of seeds to be extracted, (2) the dimensionality of the chemometric space, (3) the critical distance cutoff for building the VP, and (4) the cutoff distance in the collapsing operation. The strategy that yields the best results for all the datasets we have tested so far is to extract a large number of seeds and a short critical distance cutoff, so many small VP are produced. All of them constitute an informative layer around the molecule, while variables far away from this layer are included in region 0 and then removed from the analysis. Optionally, some of the regions with useful information may be merged to simplify the picture, using high values for the collapsing cutoff distance. We have observed that there is an intrinsic limit in the number of regions: even when very different numbers of seeds are extracted at the beginning, they collapse in a nearly similar number of regions.

In order to show graphically the results of the SRD algorithm, we have carried out the described procedure using the set of glucose analogues which will be described in detail below. As this run was produced only for demonstrative purposes, only 40 seeds were used. Some of the regions defined in the analysis are shown in Figure 1, where the structure of one of the glucose analogues and some important residues of the receptor are also represented. The green volume in Figure 1a encloses all the regions used in the analysis. The excluded area, which extends from this volume to the boundaries of the box, is the so-called region 0 and contains nonimportant variables. When this volume is compared with the structure of the receptor (not included in the analysis), it can be seen that SRD has removed from the analysis the variables placed farther away from the residues which are those containing no information. Also, variables falling into that part of the structure which shows no variation within the series were included in region 0 and removed from the analysis.

In Figure 1b a few selected regions have been represented in different colors. It is worth noting how areas rich in information are represented by many small VP while areas less important are represented by fewer and larger VP. For instance, the small white and purple regions at the bottom of Figure 1b represent one of the areas bearing more information, the location where the

2-hydroxyl and 1α-glucosyl substituents interact with Asp 283. On the contrary, the large orange region represents an area where the glucosyl ring interacts weakly with the protein. The structure of the inhibitors exhibits much less variation in this area, and consequently the region contains much less information. The claim that such regions represent potential interactions is supported by the finding that some of them clearly overlap water molecules present in the crystal structure of the complexes. All of these results show how the SRD regions contain single, independent pieces of information and how the results are in agreement with the structure of the receptor, not included in any step of the analysis.

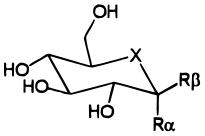
Methods

The method presented has been implemented in the GOLPE program[12] and successfully applied in our group to a few different datasets. Further details of these investigations will be reported in due course. In this paper, the method has been applied in a GRID/GOLPE, CoMFA-like study onto a new set of recently synthesized glucose analogs which are inhibitors of the glycogen phosphorylase *b* (GP*b*) enzyme[13−17] (Table 1). The purpose of this study is mainly comparative because we wanted to assess the suitability of the method in real datasets and to compare the different models obtained either in the absence of any variable selection or by using other well-established variable selection methodologies.

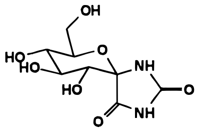
This set is especially suitable for 3D-QSAR methodological research, because high-resolution crystallographic structures of the enzyme−ligand complexes are available for every compound in the series.[17] Therefore, both the conformation and the superposition of the compounds have been experimentally determined, and it is possible to investigate the effect of other different parameters on the quality of the models. Figure 2 shows the compounds included in the series superimposed in the conformation and the position in which they appear in the crystallographic structures.

It is important to remark that these conformations are not the minimum energy ones but the X-ray conformations in the refined ligand−enzyme complexes. In this sense, they give a much better representation of the bioactive conformations than the outcome of whatever energy minimization procedure which considers the ligands on their own. Besides, the ligands are in energetically reasonable conformations because the enzyme complexes were refined using X-PLOR.[18−20] The ligands appear superimposed in approximately the same position for every complex since the crystallographic procedure used to determine their structure is essentially the same. However, the crystallographic analysis superimposes the ligand−enzyme complexes and not only the ligands. As a consequence the ligands are not actually superimposed but rather placed in equivalent positions with respect to their interaction within the receptor. This fact can be observed in Figure 2, where some compounds with large C1 substituents have been placed in positions where the glucosyl group is slightly shifted with respect to the rest of the compounds, reflecting the change in the position of the glucosyl moiety required to accommodate the bulky C1 substituent. More details about the crystallographic analysis are given in ref 17 and references quoted therein. The biological activities (K_i) have been determined by kinetic studies and reflect the ability of the compounds to inhibit the enzymatic activity (phosphorylase) of rabbit glycogen phosphorylase enzyme. For the regression analysis the negative decimal logarithm of the experimental K_i was used (pK_i).

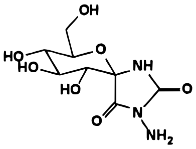
Table 1. Series of Glucose Analog Inhibitors of Glycogen Phosphorylase *b*[13−17]



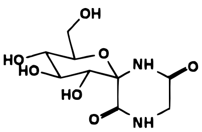
		substituent at C1 position		
no.	X	Rα	Rβ	K_i (mM)
1	O	OH	H	1.7
2	O	$C(=O)NH_2$	H	0.37
3	O	$C(=O)NHNH_2$	H	3.0
4	O	$COOCH_3$	H	24.2
5	O	$CH_2NH_3^+$	H	34.5
6	O	CH_2N_3	H	22.4
7	O	CH_2OH	H	1.5
8	O	H	*O*-(1−6)-D-glucose	16.3
9	O	H	$C(=O)NH_2$	0.44
10	O	H	$C(=O)NHCH_3$	0.16
11	O	H	$C(=O)NHCH_2CH_2OH$	2.6
12	O	H	$C(=O)NHPh$	5.4
13	O	H	$C(=O)NH$-4-OHPh	4.4
14	O	H	$C(=O)NHNH_2$	0.4
15	O	H	$C(=O)NHNHCH_3$	1.8
16	O	H	$C(=O)NHCH_2CF_3$	8.1
17	O	H	$C(=O)NH$-cyclopropyl	1.3
18	O	H	$COOCH_3$	2.8
19	O	H	$CH_2NH_3^+$	16.8
20	O	H	$CH_2CH_2NH_3^+$	4.5
21	O	H	CH_2N_3	15.2
22	O	H	CH_2CN	9.0
23	O	H	$NHC(=O)NH_2$	0.14
24	O	H	$NHC(=O)CH_3$	0.032
25	O	H	$NHC(=O)CH_2CH_3$	0.039
26	O	H	$NHC(=O)CH_2CH_2CH_3$	0.094
27	O	H	$NHC(=O)CH_2Cl$	0.045
28	O	H	$NHC(=O)CH_2Br$	0.044
29	O	H	$NHC(=O)CH_2NH_2$	0.37
30	O	H	$NHC(=O)Ph$	0.081
31	O	H	$NHC(=O)CH_2NHCOCH_3$	0.99
32	O	H	$NHCOOCH_2Ph$	0.35
33	O	H	$CH_2OSO_2CH_3$	4.8
34	O	H	1*H*-indol-2-yloxy	2.6
35	O	$C(=O)NH_2$	$NHCOOCH_3$	0.016
36	O	OH	CH_2OH	15.8
37	O	OH	CH_2N_3	7.4
38	O	OH	CH_2CN	7.6
39	O	OH	$CH_2OSO_2CH_3$	3.7
40	O	H	SH	1.0
41	O	H	$SCH_2C(=O)NH_2$	21.1
42	O	H	$SCH_2C(=O)NHPh$	3.6
43	O	H	$SCH_2C(=O)NH$-2,4-F_2Ph	18.9
44	S	OH	H	2.0



45 K_i = 0.003 mM



46 K_i = 0.146 mM



47 K_i = 0.0597 mM

The energy calculations were performed by the GRID program,[21] using the phenolic hydroxyl group probe (OH). This group is capable of donating and accepting one hydrogen bond. The electronic configuration of the OH probe is defined so 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. For the GRID analysis, the heavy atoms are considered in fixed positions, but the thermal motion of the hydrogen-bonding hydrogen atoms and lone-pair electrons is taken into account. When the target molecule donates a hydrogen bond, then the bond direction is determined by the hydrogen position as computed from the heavy atom structure of the molecules. When the OH probe donates the bond, it is assumed that the probe can orient itself to form the most effective hydrogen-bond interaction with the acceptor atom of the target molecule.[22] The OH probe was chosen because the binding site is predominantly hydrophilic. In particular, the α-pocket, in the absence of inhibitor, is a water-

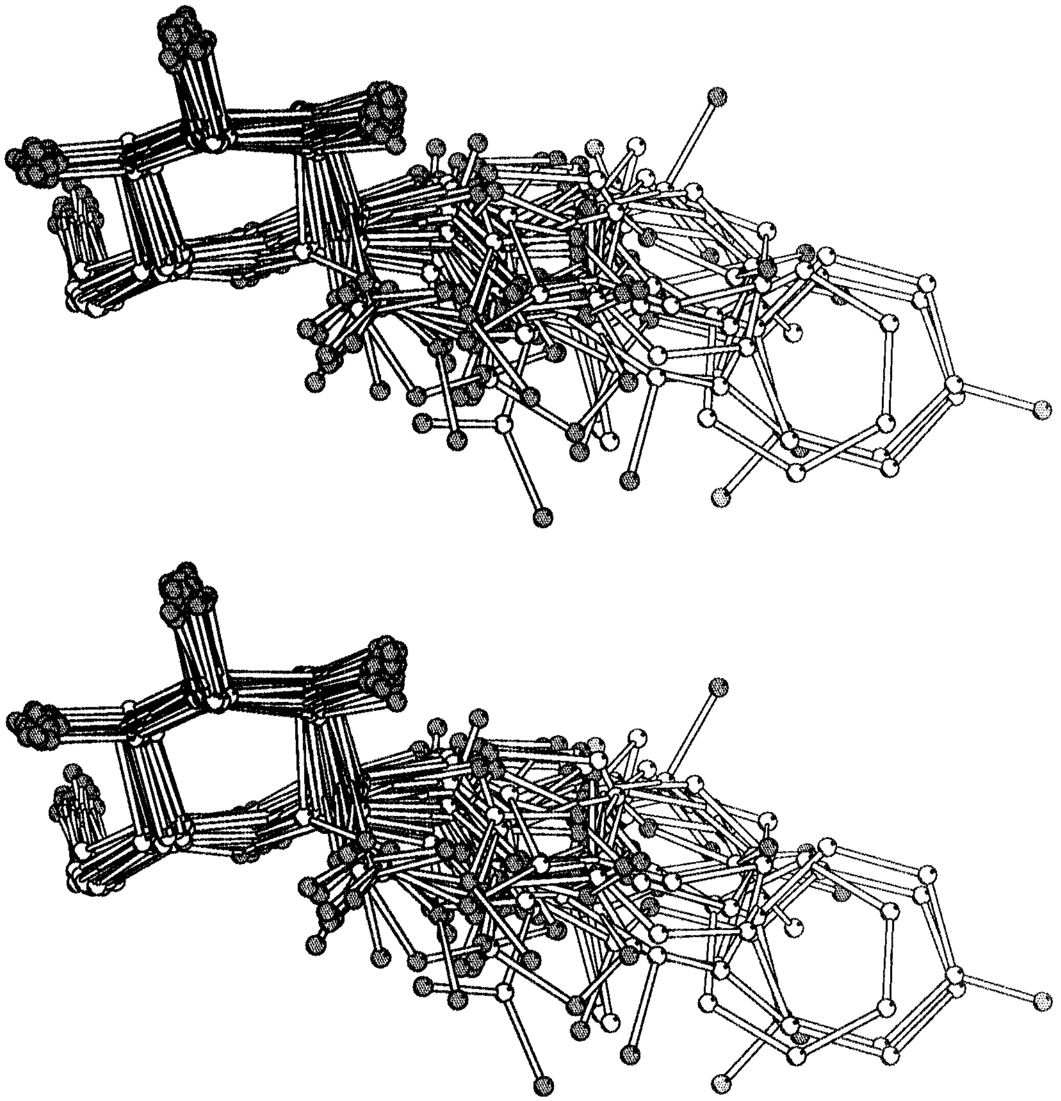


Figure 2. Stereoview of the series of 47 glucose analogues analyzed, superimposed in the conformation, and the position where they were found in the crystallographic analysis. White balls represent carbon atoms and gray balls heteroatoms.

filled channel.[14] Besides, in previous studies[17,23] the OH probe was chosen among many others because it gave better results.

The size of the box was defined in such a way that it extends about 4 Å from the structure of the inhibitors, resulting in a box of 22 × 20 × 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. A cutoff of +5 kcal/mol was applied to produce a more symmetrical distribution of the **X** matrix.

This matrix was then imported into GOLPE 3.0 and further pretreated by zeroing those values with absolute values smaller than 0.1 kcal/mol and removing any variables with standard deviation below 0.1. In addition, variables which take only two or three values and present a skewed distribution (one of these values is taken by only one or two molecules) were also removed.

The SRD algorithm was applied on this matrix as described above, with the following parameters: 457 seeds selected on the PLS weights space, critical distance cutoff of 1.0 Å, and collapsing distance cutoff of 2.0 Å. The regions found were used at a later stage in a FFD variable selection procedure.[7−9]

In order to compare them with SRD, Norinder's and Cho and Trophsa's methodologies were applied to the same matrix. Both methods were implemented in the GOLPE 3.0 software, where the same cross-validation technique and PLS algorithms used in the derivation of the SRD/GOLPE models were employed. Following Norinder's domain mode variable selection (DMVS)[11] approach, the original grid cage was divided into 125 small boxes of approximately the same size, using five divisions per axis. These boxes were included in a FFD variable selection procedure as described in the original reference.[11] In Cho and Trophsa's cross-validated r^2-guided region selection (q^2-GRS)[10] approach, the original grid cage was also divided into the same 125 small boxes, and for each box an independent PLS analysis was carried out. The model dimensionality chosen was that showing the best predictive ability. Only those boxes with a q^2 value higher than a predefined cutoff were used in the derivation of the final model. Four different cutoffs were evaluated (0.0, 0.1, 0.2, and 0.3), and that yielding the highest q^2 was chosen (0.1). The GOLPE implementation differs from the original procedure[10] mainly in two respects: single GRID field values were used, instead of the steric and electronic fields proposed by Cho and Tropsha, and random groups cross-validation rather than leave-one-out (LOO) was used (*vide infra*).

Results and Discussion

For the comparison, PLS models were derived without variable selection, with regular GOLPE variable selection (here, two D-optimal preselections plus a FFD variable selection), and with SRD/GOLPE region selection (a single FFD selection performed on regions). The predictive ability of the models was evaluated by cross-validation, using five groups of approximately the same size to which the objects were assigned randomly. The

Table 2. Results of the PLS Modeling with Different Variable Selection Procedures

vars sel	var[a]	dimens[b]	r^2 [c]	q^2 [d]	SDEP[e]
none	2087	4	0.92	0.50	0.69
GOLPE	379	4	0.93	0.73	0.51
SRD/GOLPE	457	4	0.95	0.79	0.45
DMVS	1542	4	0.94	0.60	0.63
q^2-GRS	626	4	0.91	0.65	0.58

[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.

whole procedure was repeated 20 times. This cross-validation procedure provides a safer alternative to the more widely preferred LOO and gives more conservative results, smaller q^2 and higher standard deviation of error of preditions (SDEP). Table 2 shows the results of the variable selection and the PLS analyses. Figure 3 contains plots of experimental versus calculated values for the five models, and Figures 4 and 5 illustrate the grid plot of the PLS coefficients, for models with four principal components (PC's).

Comparisons between the PLS models clearly show that both the GOLPE and the SRD/GOLPE variable selection procedures improve the quality of the models in terms of fitting and predictive ability. Table 2 shows large q^2 increases (23% and 25% increases) and slight r^2 increases as a result of the application of these variable selection methods. The improvement on the model quality is also noticeable in the experimental vs calculated plots represented in Figure 3. On the other hand, when comparing the classical GOLPE approach and the new SRD/GOLPE, it turns out that the new technique yields slightly better models, with higher q^2 (about 5% higher) and r^2 values. Even if the increase of these values is not large, in order to make a fair comparison it should be taken into account that the new method does not require the D-optimal preselection (two runs were required in the classical approach) and is performed in a single run. In fact, the purpose of the D-optimal preselection was to reduce the variables down to a number that can be handled by the FFD variable selection procedure. With the new SRD/GOLPE methodology this variable preselection is not required any more, and the method uses FFD variable selection from the very beginning. In this way the new procedure is safer since no variables are removed without assessing the impact of their removal on the predictive power of the model. It is noticeable that, as it is shown in the results, this improvement on the computational aspects of the methodology not only produces no decrement on the quality of the models obtained but, on the contrary, gives rise to an increase of their fitting and predictive power. With respect to the DMVS and q^2-GRS, both methods produce a slight increment in q^2 (10% and 15% increases), smaller than the increment produced by the use of either GOLPE or SRD/GOLPE. The r^2 obtained after DMVS is slightly better while, remarkably, it decreases after the use of q^2-GRS.

The predictive ability of the models has been further tested using the set of four new compounds reported in Table 3. The procedure described above has been used to obtain the structures of the complexes between these compounds and the GP*b* enzyme. Their predicted activities, according to each model, together with their experimental activities are reported in Table 3. The

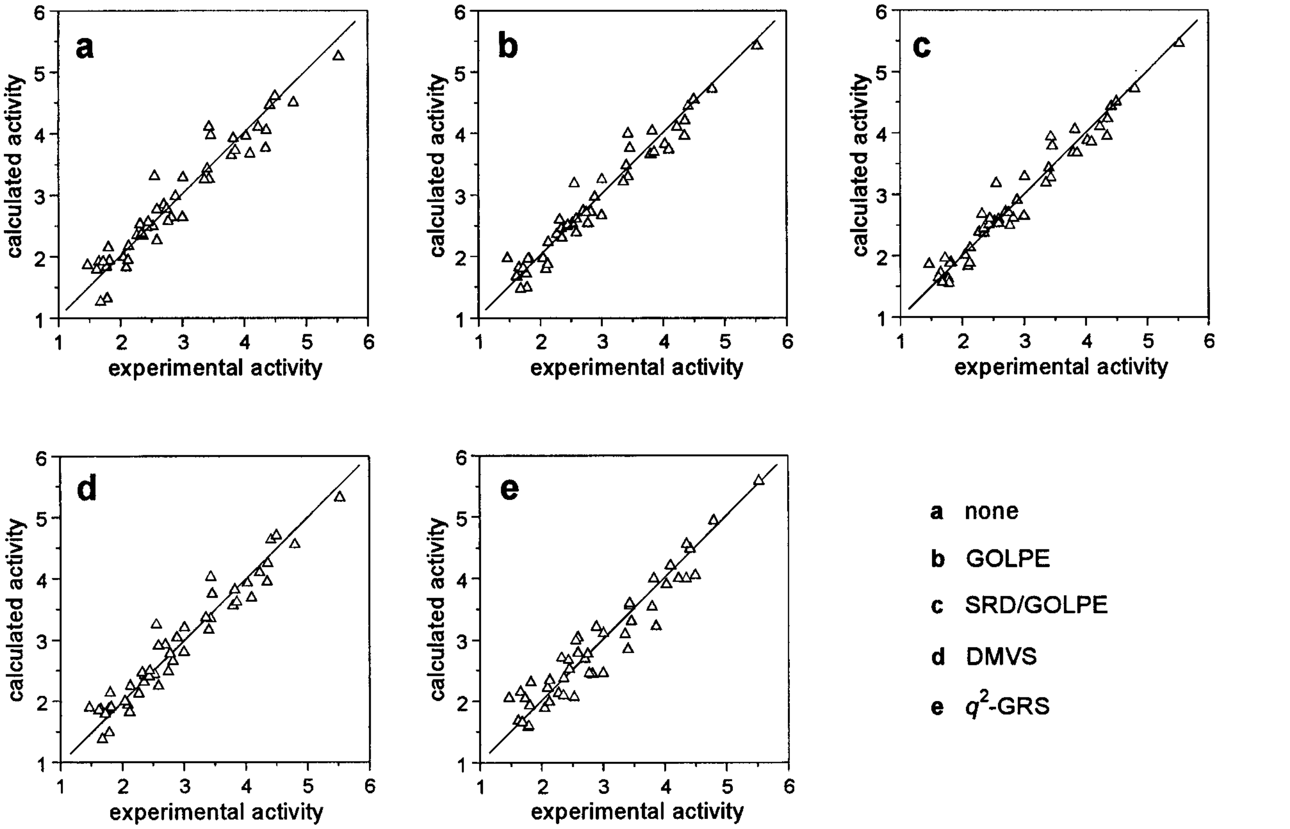


Figure 3. Scatter plot of the experimental (horizontal) versus calculated (vertical) activity values (pK_i) for the five models being compared: (a) model with no variable selection, (b) model with classical GOLPE variable selection, (c) model with SRD/GOLPE variable selection, (d) model with Norinder's DMVS, and (e) model with Cho and Tropsha's q^2-GRS.

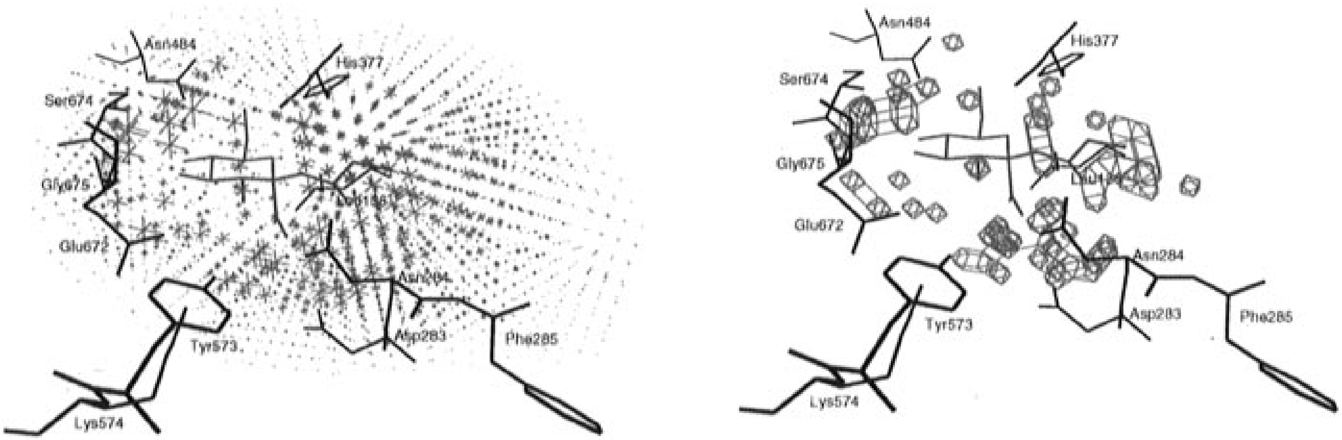
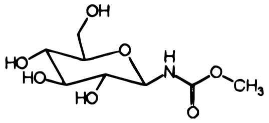
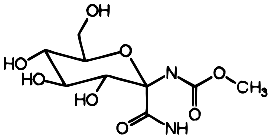
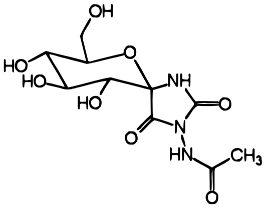
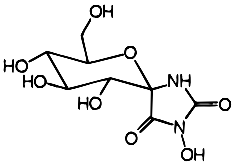


Figure 4. Grid plots of the PLS coefficients for the model with no variables selection, using four PC's. On the plot at the left-hand side (a) the coefficients are represented by crosses whose size is proportional to their absolute value. On the plot at the right-hand side (b) these coefficients have been contoured at appropriate levels (+0.006 and −0.006 for positive and negative values, respectively) in order to show only the most important coefficients. Positive coefficients are represented in yellow, and negative coefficients are represented in cyan in both pictures. The structure of compound **35** (in red) and some of the residues of the active site (in blue) have been included in the picture for reference.

Table 3. Set of New Glucose Analogues Used for External Predictions

		activity (pK_i)					
no.	compound	exp	none	GOLPE	SRD/GOLPE	DMVS	q^2-GRS
48		4.07	4.04	4.09	4.10	4.25	3.78
49		3.50	3.35	3.32	3.20	2.96	2.63
50		3.26	1.40	1.57	1.82	1.11	1.19
51		4.41	3.38	3.31	3.24	3.06	2.34
SDEP[a]			1.07	1.01	0.93	1.30	1.53

[a] Standard deviation of error of predictions.

SDEP values computed for each model provide an index which expresses how good those predictions are.

With respect to the external predictions the original, GOLPE and SRD/GOLPE models give rather similar SDEP values (Table 3), but again the best result is obtained when the SRD/GOLPE method is applied. A closer inspection of the data reveals that the three models give quite good predicted values for three of the four compounds (**48**, **49**, and **51**), but they fail completely to predict the activity of compound **50**. This is not surprising since **50** is quite a peculiar compound which places a substituent in a position not explored by any compound in the series. This singularity is clearly highlighted in Figure 6, which represents the structure of **50** superimposed on the rest of the compounds. The models cannot know the contribution to the activity of the *N*-acetamide substituent linked to the N2, because no substituent in such a position has been tested before, and therefore the error in the prediction is greater than 1.5 logarithmic units (30-fold in terms of activity). Without this compound, the total SDEP for the SRD/GOLPE model is 0.69, not very distinct from the internal SDEP of 0.45.

The DMVS and q^2-GRS methods did not improve the results of the external SDEP at all and produced much worse predictions than the original model and other variable selection methods. Indeed the DMVS method was already reported to produce either small or no prediction improvements for different datasets.[11]

The most important advantage of the SRD/GOLPE methodology is the improved interpretability of the models produced by this novel technique of variable

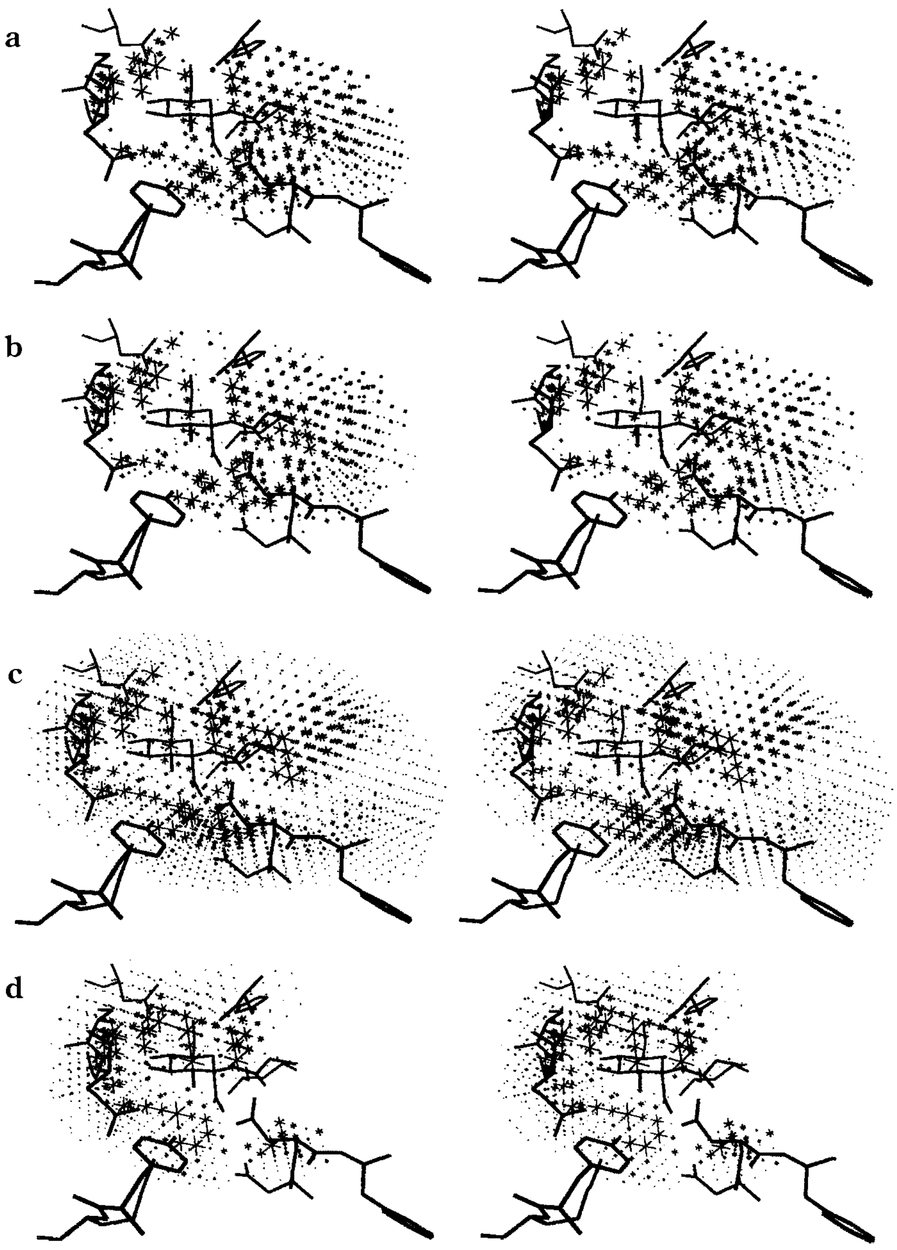


Figure 5. Stereoview of the PLS coefficients grid plot for models of four components. The coefficients are represented by crosses whose size represent their absolute value. The structure of compound **35** and some of the residues of the active site have been included in the picture for reference. The figure represents models obtained after different variable selection procedures: (a) classical GOLPE, (b) SRD/GOLPE, (c) Norinder's DMVS, and (d) Cho and Tropsha's q^2-GRS.

selection. The 3D-QSAR models are intepreted with the help of plots that show the values of the PLS weighted pseudocoefficients in the spatial position that they represent around the ligands. For each grid point, the values of such coefficients express how correlated the probe–ligand interaction energies are with respect to changes in the biological activity. Figure 4 represents such a plot for the PLS model obtained with no variables selection. The structure of one of the compounds (**35**) and some of the most important residues in the receptor site have been included in the plot in order to help the interpretation.

In most cases the information given by such plots is used to propose a hypothesis about the binding site and to design new compounds which best mimic the interactions that correlate positively with the activity while avoiding the ones that decrease the activity. In this particular study, in which the structures of the ligand–receptor complexes were already available, the comparison between the areas highlighted by the model and the residues actually present in the receptor have been used to "validate" the models. It may be considered that the models are better when the areas highlighted by this model are near or overlap real receptor residues where ligand–receptor interactions are stronger.

In Figure 4a the values of the coefficients have been represented by crosses of sizes which are directly proportional to their absolute values. It can be seen that the model with no variable selection contains so many small coefficients which make this method of representation not useful at all. In order to simplify the plot it is necessary to contour the areas containing the highest (and lowest) coefficients, as shown in Figure 4b. Unfortunately, this operation hides a considerable amount of information and the contributions to the model produced by a large number of grid variables are completely neglected. In particular, the interaction of the ligands with residues of the α-pocket, which in

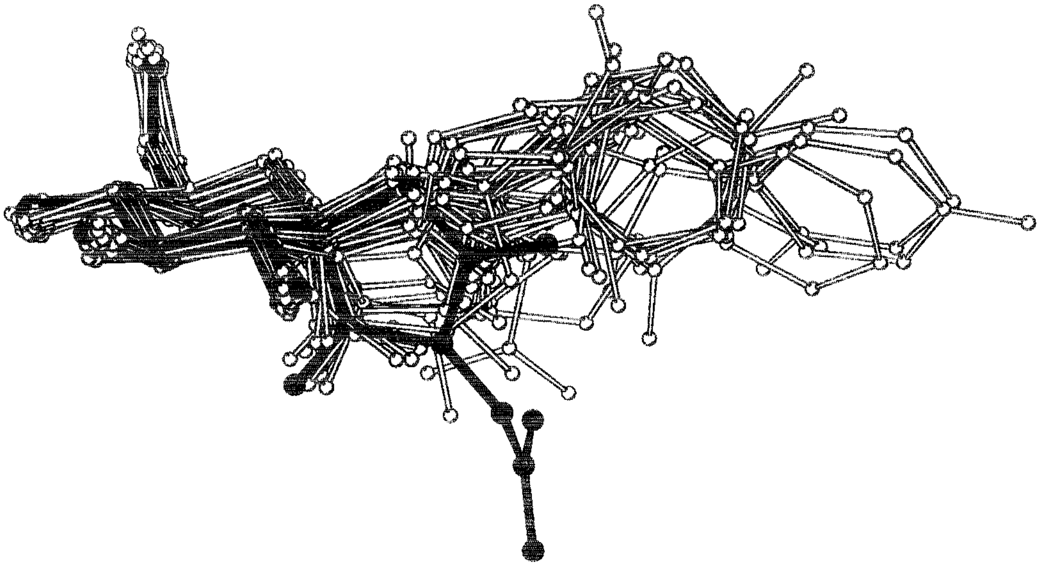


Figure 6. The figure represents the structure of one of the compounds used for the external predictions (**50**), in black, superimposed on the rest of the compounds, in gray. Notice how the *N*-acetamido substituent linked to the N2 (at the bottom) occupies an area not explored by any other compound.

Figure 4a are represented by a dense cloud of crosses at the right-hand side, are not represented at all in Figure 4b.

After variable selection, apart from other improvements, models became simpler and they can often be interpreted without contouring the plots. In Figure 5 we have represented stereoplots of the PLS coefficients for the models obtained after the application of different variable selection procedures: GOLPE (5a), SRD/GOLPE (5b), Norinder's DMVS (5c), and Cho and Tropsha's q^2-GRS (5d). The purpose of those plots is to compare the different way in which the different methods show the results of the PLS model in the space and not to actually interpret the model. Therefore, the signs of the coefficients have been omitted, and no contouring application has been applied.

The model with GOLPE variable selection, is, by far, easier to interpret than the model with no variable selection represented in Figure 4, and no contouring operation is strictly required. In Figure 5a the effect of the different substituents is represented by single leader variables which take large values and condense the information of many contiguous variables, which in Figure 4a appear as a cloud of small crosses. For instance, in Figure 5a it is easy to recognize the interaction of the ligands with the residue Asn284, which is known to play an important role in the interaction[14,17] and is represented here by the two large crosses overlapping the amide group of the residue. However, this simplification is too drastic and gives an unrealistic picture of the actual interactions, because the structural changes in the series are never reflected by the change of a single variable but by a few contiguous variables.

The interpretability of the model is improved by the use of SRD; in Figure 5b the effect of the different substituents is now represented by small clusters of variables corresponding to the groups defined by the SRD algorithm. This solution represents a compromise between the requirement to simplify the plots and undesirable oversimplifications. It can be seen that inside these groups the most important variables have large coefficients, represented by large crosses in the plot, but smaller contributions from contiguous variables are also preserved. Examples of the improved interpretability of this model can be observed by comparison of parts a and b of Figure 5. For instance, the interactions between the ligands and the residues Ser674 and Gly675 (on the left-hand side of Figure 5a–d) are important because these residues participate in the network of hydrogen bonds which places the glycosidic ring in the active site. When parts a and b of Figure 5 are compared, it is clear that the effect of the interaction of the ligands with these residues is represented in the GOLPE model (Figure 5a) by a few variables which highlight only the major effect. Conversely in the SRD/GOLPE model (Figure 5b) the same regions are represented by a small group of variables. It should be noted how these regions represent, besides the major effects, some neighboring variables which have modest coefficients but which contribute to giving a more detailed and quantitative picture of the interactions present.

The model obtained after DMVS (Figure 5c) is not too different from the original PLS model, and therefore nearly no improvement on interpretability can be seen. The model obtained with q^2-GRS selection, on the contrary, has been pruned too much and, surprisingly, most of the important variables within the α-pocket (on the right-hand side of the Figure 5d) have been removed. Within the boxes retained, the aspect of the plot is quite similar to the original PLS model, and no improvement is apparent.

Conclusions

A new methodology aimed at building groups of contiguous grid-field variables that contain single pieces of chemical and statistical information has been described. Among the many possible uses of this methodology in 3D-QSAR, it has been successfully combined with the GOLPE variable selection procedure. The resulting SRD/GOLPE procedure improves the predictive ability and the fitting of the PLS models obtained, when compared with the models obtained with no variable selection. This method is also advantageous with respect to the classical GOLPE procedure, because variable preselection is no longer required and because the quality of the model obtained is slightly better. Moreover, the models produced by this novel methodology are easier to interpret than the models obtained with no variable selection and give a more realistic picture of the receptor than the models obtained by a classical GOLPE variable selection. The SRD/GOLPE methodology compares favorably with other methods of variable selection which use grouping of variables (Cho and Tropsha and Norinder methods) in terms of fitting, predictive ability, external predictive ability, and interpretability.

The methodology has been tested on a series of 47 glucose analogues which are inhibitors of glycogen phosphorylase *b*. The results, in terms of internal and external predictive power and interpretability, fully support the conclusions. The new SRD methodology has been implemented in version 3.0 of the GOLPE chemometric analysis program.[12]

Acknowledgment. We thank our colleagues L. N. Johnson, K. A. Watson, G. W. J. Fleet, and N. G. Oikonomakos for sending data regarding the compounds **35**, **46**, and **48**−**51** in advance of publication and the EC for providing financial support (project BIO2-CT943025) and a grant to M.P. The Computing Center of the Universidad de Alcala (Spain) provided a generous allowance of SG Power Challenge computer time. We thank M. Baroni and F. Gago for useful discussions, and the Italian funding agencies of MURST and CNR are thanked for financial support to S.C. and G.C. The authors also want to thank Dr. P. J. Kraulis for his software Molscript v1.4,[24] used for the generation of Figures 2 and 6 in this article.

References

(1) *3D QSAR in Drug Design, Theory Methods and Applications*; Kubinyi, H., Ed.; ESCOM: Leiden, 1993.

(2) 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.

(3) Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict Their Biological Activity. *J. Med. Chem.* **1994**, *37*, 4130−4146.

(4) Floersheim, P.; Nozulak, J.; Weber, H. P. Experience with comparative molecular field analysis. In *Trends in QSAR and Molecular Modelling 92*; Wermuth, C. G., Ed.; ESCOM: Leiden, 1993; pp 227−232.

(5) Clementi, S.; Cruciani, G.; Riganelli, D.; Valigi, R. GOLPE: Merits and Drawbacks in 3D-QSAR In *QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications*; Sanz, F., Giraldo, J., Manaut, F., Eds.; Prous Science Publishers: Barcelona, 1996; pp 408−414.

(6) Cruciani, G. From variables selection to region selection in 3D-QSAR studies. 4th Scandinavian Symposium on Chemometrics, Lund, Sweden, 1995.

(7) 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.

(8) Clementi, S.; Cruciani, G.; Pastor, M.; Riganelli, D.; Valigi, R. The Golpe Philosophy in 3D-QSAR Studies. To be submitted to *Quant. Struct.-Act. Relat.*

(9) 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.

(10) Cho, S. J.; Tropsha, A. Cross-Validated R^2-Guided Region Selection for Comparative Molecular Field Analysis: A Simple Method to Achieve Consistent Results. *J. Med. Chem.* **1995**, *38*, 1060−1066.

(11) Norinder, U. Single and Domain Mode Variable Selection in 3D QSAR Applications. *J. Chemometr.* **1996**, *10*, 95−105.

(12) GOLPE version 3.0, Multivariate Infometric Analysis, Perugia, Italy, 1996.

(13) Martin, J. L.; Johnson, L. N.; Withers, S. G. Comparison of the Binding of Glucose and Glucose-1-Phosphate derivatives to T State Glycogen Phosphoryase *b*. *Biochemistry* **1990**, *29*, 10745−10757.

(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.; Papageorgioui, 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 Spirodiketopiperazine at the Anomeric Position of Glucopyranose. *Tetrahedron Lett.* **1995**, *36*, 8281−8294.

(17) 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.

(18) Brünger, A. T. Crystallographic refinement by simulated annealing. Application to a 2.8Å resolution structure of aspartate aminotransferase. *J. Mol. Biol.* **1988**, *203*, 803−816.

(19) Brünger, A. T. A memory-efficient fast Fourier transformation algorithm for crystallographic refinement on supercomputers. *Acta Crystallogr.* **1989**, *A45*, 42−50.

(20) Brünger, A. T.; Karplus, M.; Petsko, G. A. Crystallographic refinement by simulated annealing: application to crambin. *Acta Crystallogr.* **1989**, *A45*, 50−61.

(21) GRID v. 14, Molecular Discovery Ltd., West Way House, Elms Parade, Oxford, 1996.

(22) 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.

(23) Cruciani, G.; Watson K. A. Comparative Molecular Field Analysis Using GRID Force-Field and GOLPE Variable Selection Methods in a Study of Inhibitors of Glycogen Phosphorylase *b*. *J. Med. Chem.* **1994**, *37*, 2589−2601.

(24) Kraulis, P. J. “MOLSCRIPT”: a program to produce both detailled and schematic plots of protein structures. *J. Appl. Crystallogr.* **1991**, *24*, 946−950.

JM9608016