GRid-INdependent Descriptors (GRIND): A Novel Class of Alignment-Independent Three-Dimensional Molecular Descriptors

Manuel Pastor,[†] Gabriele Cruciani,[*,†] Iain McLay,[§] Stephen Pickett,[§] and Sergio Clementi[†]

Laboratory on Chemometrics, Department of Chemistry, University of Perugia, Via Elce di Sotto 10, 06123 Perugia, Italy, and CADD Department, Rhone-Poulenc Rorer, Dagenham, Essex RM10 7XS, U.K.

Traditional methods for performing 3D-QSAR rely upon an alignment step that is often time-consuming and can introduce user bias, the resultant model being dependent upon and sensitive to the alignment used. There are several methods which overcome this problem, but in general the necessary transformations prevent a simple interpretation of the resultant models in the original descriptor space (i.e. 3D molecular coordinates). Here we present a novel class of molecular descriptors which we have termed GRid-INdependent Descriptors (GRIND). They are derived in such a way as to be highly relevant for describing biological properties of compounds while being alignment-independent, chemically interpretable, and easy to compute. GRIND are obtained starting from a set of molecular interaction fields, computed by the program GRID or by other programs. The procedure for computing the descriptors involves a first step, in which the fields are simplified, and a second step, in which the results are encoded into alignment-independent variables using a particular type of autocorrelation transform. The molecular descriptors so obtained can be used to obtain graphical diagrams called "correlograms" and can be used in different chemometric analyses, such as principal component analysis or partial least-squares. An important feature of GRIND is that, with the use of appropriate software, the original descriptors (molecular interaction fields) can be regenerated from the autocorrelation transform and, thus, the results of the analysis represented graphically, together with the original molecular structures, in 3D plots. In this respect, the article introduces the program ALMOND, a software package developed in our group for the computation, analysis, and interpretation of GRIND. The use of the methodology is illustrated using some examples from the field of 3D-QSAR. Highly predictive and interpretable models are obtained showing the promising potential of the novel descriptors in drug design.

Introduction

Every science is based on mathematical representations of the objects studied. In the case of chemistry, many different mathematical representations are used to describe molecules, each one adequate for a certain purpose. For example, a simple chemical formula is enough to describe the stoichiometry of a molecule, while atom-based descriptors are more useful to compute certain physicochemical properties such as log *P*. A key concept in this respect is that no single mathematical descriptor is able to produce an exhaustive representation of a molecule and each one is used pragmatically according to need.

In the field of drug design, the identification of mathematical descriptions relevant to the pharmacological properties of compounds has been a challenge for decades. Early work of Hansch[1] demonstrated that it is indeed possible to obtain functions correlating mathematical descriptions with the biological properties of compounds. The simplest representations used in these studies (a few variables representing physicochemical properties) were intuitive and tried to represent what was known at that time about the ligand−receptor interaction. Since then, many other mathematical descriptions have been used in drug design, and more specifically in the field of quantitative structure−activity relationships (QSAR). Major milestones were the work of Goodford,[2] who introduced the concept of molecular interaction field (MIF) and the work of Cramer et al.,[3] who introduced the 3D chemical structure into the description of the compounds and hence developed the concept of 3D-QSAR.

With the development of combinatorial chemistry, drug design has changed dramatically. Pharmaceutical research demands a fast and accurate characterization of large series of compounds for a variety of purposes; sometimes it aims to evaluate the molecular diversity, sometimes the drug-likeness, etc. In other circumstances, the chemicals must be scored to prioritize their chemical synthesis or biological evaluation. Therefore the requirement of relevant mathematical descriptions is no longer exclusive of QSAR and should be adapted to meet the novel needs of modern drug research.

In this work we present a novel methodology for producing a mathematical description of molecules called GRid-INdependent Descriptors (GRIND). GRIND are highly relevant with respect to biological properties, are easy to obtain, even for large series of compounds, and are applicable in many different areas of drug design. An important characteristic of these novel descriptors is that they are insensitive to the position

* To whom correspondence should be addressed. Tel: +39 075 5855550. Fax: +39 075 45646. E-mail: gabri@chemiome.chm.unipg.it.
† University of Perugia.
§ Rhone-Poulenc Rorer.

and orientation of the molecular structures in the space. In the field of 3D-QSAR, the initial alignment of the compounds in the series is widely recognized as one of the most difficult and time-consuming steps. Therefore, since the GRIND need no alignment of compounds, 3D-QSAR analysis can be applied in a fraction of the time required using standard methodologies. However, the use of the novel descriptors is not limited to 3D-QSAR, and the very fact that they do not require a molecular superimposition allows them to be applied in a number of different fields such as in 3D searching, pharmacophore identification, and structure–metabolism relations.

The concept of alignment-independent molecular descriptors is not new, and a number of such descriptors have been reported in the past. Some of these approaches are based on autocorrelation functions, such as the approaches suggested by Broto,[4] Gasteiger et al.,[5,6] and Clementi et al.[7] Broto applied a classic autocorrelation transform on 2D and 3D structures to represent atomic properties, obtaining autocorrelation vectors that were used in QSAR.[8] Gasteiger described a more sophisticated approach, in which a spatial autocorrelation function was applied, on molecular surface properties. This approach combined with sensible application of neural networks yielded highly predictive models. Clementi's article represents early work made in our group leading to the results presented in this paper, but the mathematical transformation involved was complex and computationally demanding and the procedure worked well only for planar compounds. In all instances, the use of classic autocorrelation vectors had the disadvantage that the original information could not be reconstructed, thus making the interpretation of the results difficult. Other examples of alignment free molecular descriptors are the family of WHIM descriptors[9] and comparative molecular moment analysis.[10] Again, the main drawback of these methodologies is the difficulty to interpret both the descriptors and the results of the obtained models.

If alignment is such an important drawback, one might ask why alignment-free descriptors are not more popular. Surprisingly, in most cases the aforementioned descriptors are rarely used outside the research group that developed them. In our opinion, the main reason for this is related to the difficulties of understanding the descriptors and interpreting the results in terms that can aid in the design of novel compounds. The main objective of our study was to obtain descriptors that were both alignment-independent and easy to interpret by referring back to the molecules themselves. In this article we will describe in detail the procedure resulting from our efforts, as well as a few selected applications of the new descriptors in the field of 3D-QSAR in order to illustrate how to apply the method and interpret the results.

The first example deals with a subset of the series of glucose analogue inhibitors of glycogen phosphorylase (GP) described in previous methodological QSAR work made in our group.[11–13] This example is focused on demonstrating the alignment independence of the descriptors and will be used to show how to interpret a model obtained with them. In a second example, the classical series of steroids binding the corticosteroid-binding globulin (CBG)[3] is used to show an example of QSAR application on a well-known dataset. In a third example, the descriptors are applied in the study of a series of butyrophenones with serotoninergic affinities[14] obtaining good models with high predictive power. This last example illustrates the use of the method on semiflexible compounds and further exemplifies how to interpret models obtained with the novel descriptors.

Materials and Methods

MIF were obtained using the program GRID version 17.[15] GRIND were generated, analyzed, and interpreted using the program ALMOND version 2.0,[16] a software package developed in our group. This software is freely available for all academic and nonprofit institutions (see the Software Availability section at the end for further information).

Computations and graphical display were performed on a number of SGI O2 workstations (MIPS R5000 and MIPS R12000 processors).

The process of ligand–receptor interaction has often been represented with the help of the MIF.[2] When computed on biomolecules (e.g. protein active sites), the MIF identify regions where certain chemical groups can interact favorably, suggesting positions where a ligand should place similar chemical groups. MIF can also be computed starting from the ligands themselves. In this case, the regions showing favorable energy of interaction represent positions where groups of a potential receptor would interact favorably with the ligand. Using different probes, one can obtain for a certain ligand a set of such positions which defines a "virtual receptor site" (VRS). This abstract entity defines an ideal complementary site for a certain chemical compound and represents its potential ability to bind a biomolecule. If a compound is known to bind a certain receptor, some of the regions defined in its VRS should actually overlap groups of the real receptor site and, therefore, at least a subset of the VRS regions would be relevant for representing the binding properties of the ligand. For the last statement to be true the VRS must have been obtained from the bioactive conformation of the ligand and the probes used to compute it should represent chemical groups present in the binding site.

The molecular descriptors presented in this work are based on the concept of VRS. Basically, GRIND are a small set of variables representing the geometrical relationships between relevant regions of the VRS and as such are independent of the coordinate frame of the space where the MIF is computed. Using a metaphor, these variables represent the VRS in the same way that the measures a tailor obtains in order to make tailor-made clothes represent a person.

The procedure for obtaining GRIND involves three steps: (i) computing a set of MIF, (ii) filtering the MIF, to extract the most relevant regions that define the VRS, and (iii) encoding the VRS into the GRIND variables.

(i) Computing the MIF. MIF were computed using the program GRID.[15] To obtain relevant VRS, the probes used should represent potentially important groups of the binding site. For compounds interacting with proteins, it seems reasonable to use the DRY probe representing hydrophobic interactions, the O probe (carbonyl oxygen) to represent hydrogen bond acceptor groups, and the N1 probe (amide nitrogen) to represent hydrogen bond donor groups. In default mode a grid-spacing of 0.5 Å is used with the grid extending 5 Å beyond a molecule.

(ii) Filtering the MIF. Typically, for a drug molecule the MIF obtained with GRID contains between 10 000 and 100 000 nodes. However, this does not mean that there is the same order of independent pieces of information, and a close inspection allows one to identify a small number of "regions" within each field, representing interactions of a defined probe with different functional groups on the compounds.

For our purposes, we define the most interesting regions as those characterized by intense favorable (negative) energies of interaction. Unfortunately, a simple contouring procedure would not allow all such regions to be identified as a region
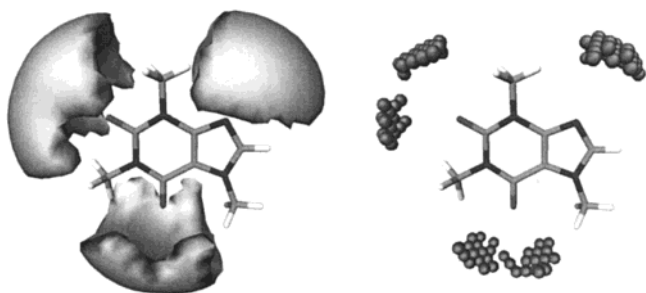
Figure 1. On the left-hand side: isocontour plot of a GRID MIF computed with an amide nitrogen (N1) probe around a caffeine molecule. The contour encloses points with negative values under −2.0 kcal/mol. On the right-hand side: the 100 nodes selected as a result of the MIF filtering described in the text.

Figure 2. Correlogram profile obtained for the caffeine molecule represented in Figure 1, using a DRY (hydrophobic) probe and a N1 (amide N) probe. From left to right the plot contains the auto-correlogram DRY−DRY, the auto-correlogram N1−N1, and the cross-correlogram DRY−N1.

Table 1. Physicochemical Meaning of the Six Correlograms in a Standard ALMOND Analysis

correlogram no.	probe 1	probe 2	interaction between nodes of type:
1	DRY	DRY	hydrophobic
2	O	O	hydrogen bond donor
3	N1	N1	hydrogen bond acceptor
4	DRY	O	hydrophobic and hydrogen bond donor
5	DRY	N1	hydrophobic and hydrogen bond acceptor
6	O	N1	hydrogen bond donor and hydrogen bond acceptor

representing a very intense interaction can mask other interactions induced by different parts of the ligand. Instead we have developed a procedure that extracts these "relevant regions" using an optimization algorithm, selecting from each MIF a fixed number of nodes optimizing a scoring function. This function includes two optimality criteria: the intensity of the field at a node and the mutual node−node distances between the chosen nodes. Therefore, the method extracts from each field a number of nodes (in the order of 100) that represent independent, favorable probe−ligand interaction regions. Figure 1 illustrates the procedure showing the nodes extracted from a MIF obtained using the probe N1 (amide nitrogen). The ensemble of these regions for all relevant probes defines the above-mentioned VRS.

Filtering employs a Fedorov-like optimization algorithm,[17] and the scoring function can be tuned to give different relative importance to each criterion. By default, the procedure provides a balanced solution, but sometimes, when the ligands contain charged groups or many polar substituents, the importance of the field values should be decreased, to extract all the recognizable independent regions. Another important parameter of the procedure is the number of extracted nodes. The method uses as a default a value of 100, which is enough in most cases, but for large compounds this number can be increased.

(iii) Encoding the VRS into GRIND. As mentioned above, GRIND encodes the geometrical relationships between the VRS regions in such a way that they are no longer dependent upon their positions in the 3D space. Basically, the encoding is an auto- and cross-correlation transform. The procedure works on the filtered nodes extracted by the previous step and computes the product of the interaction energy for each pair of nodes. The results of the products are handled according to the distance between the nodes. A discrete number of categories, each one representing a small rank of distances, are considered. In regular autocorrelation analysis, all computed terms are summed, and the result characterizes each category. In our approach, only the highest product is stored, while others are discarded. This important difference is responsible for the "reversibility" properties of GRIND. A sum cannot be reverted to all its terms but the nodes producing the maximum product can be stored in the computer memory and traced back when necessary. Accordingly, this method is called maximum auto- and cross-correlation (MACC) or, more specifically, MACC-2.[18]

The values obtained from the analysis can be represented directly in correlogram plots, where the products of the node−node energies are reported versus the distance separating the nodes. One single energy value is obtained for each of the categories considered and represents a small distance range. Figure 2 shows correlograms obtained for caffeine, a simple small molecule, with probes DRY and N1. Every peak in the correlogram indicates that the VRS contains two regions separated by a distance corresponding to the abscissa of the peak (short distances appear on the left-hand side and longer
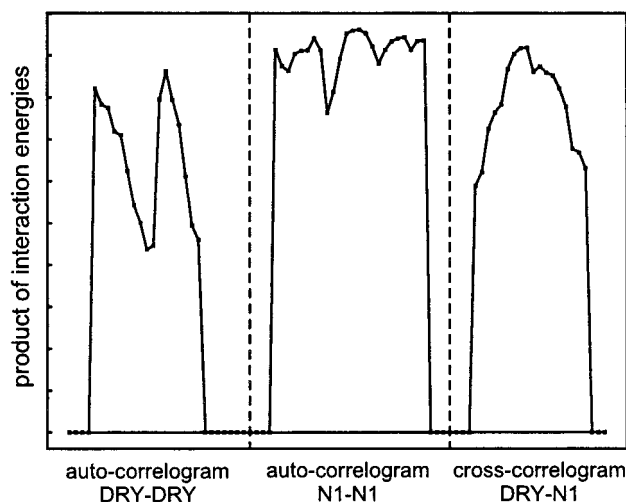
distances on the right-hand side). The height of the peak expresses the product of the intensity of the field on both nodes. The shape of the peak is also relevant: i.e., chemical groups producing intense interactions are represented, after filtering, by many contiguous nodes, and therefore the peaks are wider. Conversely, narrow peaks tend to represent weaker interactions.

The number of GRIND obtained depends on the extent of the GRID cage that defines the maximum node−node distance and on the size of the distance ranges considered. In general, the shorter the range the larger the number of discrete distance ranges considered, and therefore more variables are obtained.

Correlograms can be obtained analyzing node−node interaction when both nodes belong to the same MIF (auto-correlograms) or to different MIF (cross-correlograms). For example, when the analysis involves the three probes we recommend above (DRY, O, and N1), six correlograms are obtained: three auto-correlograms that represent node−node interactions in each MIF and three more corresponding to the interactions between the first and second, the first and third, and the second and third. The meaning of the six correlograms obtained in a standard GRIND analysis is summarized in Table 1. Usually, each compound is represented by the collection of all the computed correlograms, set side-by-side, as in Figure 2.

Analysis and Interpretation of the GRIND. The method can be applied either on one single compound or on a set of compounds. When more than one compound is studied, the number of variables extracted should be adjusted in such a way that even the longest node−node interactions found in the analysis can be represented. At the end, a matrix of descriptors is obtained, with rows representing compounds and with as many blocks of variables as correlograms extracted.
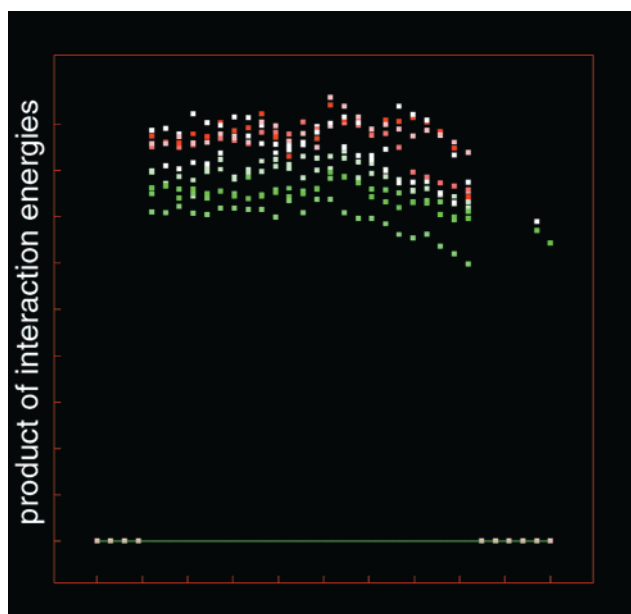
Figure 3. Set of 10 superimposed correlograms, corresponding to the series of glucose analogue inhibitors of glycogen phosphorylase. These are N1–N1 auto-correlograms, representing interactions between ligand hydrogen bond acceptor regions. Points are color-coded according to the biological activity of the corresponding compounds: active compounds are colored in red, intermediate in white, and inactive in green. A simple visual inspection shows that the strength of the interaction is correlated with the biological activity.

The GRIND can be used in a variety of chemometric analyses, using standard analysis methods such as PCA and PLS, but even a simple visual inspection of the correlogram of a series of compounds, color-coded according to their activity, chemical family, etc., can be very informative (Figure 3).

It is important to stress again that every point in the correlogram represents the product of two particular nodes for a certain compound. In this sense, a MACC transform differs from other autocorrelation analyses that accumulate the values of a large number of node–node interactions. In the case of MACC, when one is interested in the meaning of a certain variable, it is possible to trace back the transform and identify the pair of nodes involved. Indeed, with the help of appropriate computer software, it is possible to store the identity of these nodes for each value computed and identify them later.

Application of GRIND Analysis to the Example Series. In all the examples, GRIND were obtained using the program ALMOND.[16] This program starts from a set of 3D structures, interfaces the program GRID to run the calculation, reads the GRID output (MIF), and performs all the required computations, yielding a matrix of descriptors in which each row represents one of the structures. The whole process takes less than 30 s/compound in the worst case, on an O2 SGI workstation (MIPS R12000 270 MHz). ALMOND defaults were used in a first exploratory analysis and after visual inspection of the results (filtered nodes and histograms of the PLS pseudo-coefficients) some probes were removed and the parameters of the analysis were slightly modified. All MIF were computed using a grid spacing of 0.5 Å. The number of nodes was set between 100 and 150 and the relative weight of the fields between 35% and 75%. The width of the node–node distance range used to discretize the distances was always set to 0.8 grid unit.

Model Building, Validation, and Variable Selection on the Example Series. GRIND were analyzed within the program ALMOND. No scaling or pretreatment was applied. In all example series, internal validation was performed by cross-validation using five groups of approximately the same size in which the objects were assigned randomly. The whole

procedure was repeated 20 times. This cross-validation procedure provides a safer alternative to the more widely preferred leave-one-out (LOO) method and gives more conservative results: a smaller cross-validated squared correlation coefficient (q^2) and a higher standard deviation of error of predictions (SDEP).[19] Internal validation, as reported, was used in the models obtained in the example series for evaluating the optimal model dimensionality. However, we report also q^2 obtained by LOO (q^2_{LOO}), because this method does not depend on random selections and therefore is reproducible.

It is obvious that not all node–node interactions are relevant to describing the activity of the compounds. In this sense, GRIND models benefit from standard variable selection procedures, like the fractional factorial designs (FFD) variable selection procedure previously developed in our group.[20] The risk of obtaining an overfitted model by an improper use of variable selection is much less than in other 3D-QSAR methods, because the variables/objects ratio is much smaller. In the examples described below, FFD variable selection was applied as appropriate using two latent variables (LV), using random groups cross-validation, and keeping in the model uncertain variables. In all instances, the FFD variable selection was made using the ALMOND software. The FFD implementation in ALMOND is essentially the same as the FFD implemented in GOLPE.[20]
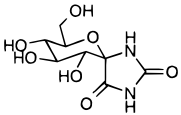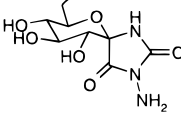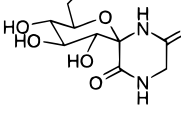
Results and Discussion

Glucose Analogue Inhibitors of the Glycogen Phosphorylase. In the past years our group has published several GRID/GOLPE models using a dataset of glucose analogue inhibitors of the glycogen phosphorylase.[11–13] This dataset is especially well-suited for methodological purposes because crystal structures of the receptor–inhibitor complexes are available for every compound in the series, and therefore it lacks the conformational and superimposition problems associated with most 3D-QSAR studies. The original dataset[13] contains 47 compounds. In this example, the small set of 10 compounds shown in Table 2 was chosen. The compounds were selected to include the main structural features of the series but avoiding explicitly very large compounds, which are known to bind poorly. The biological activity shown in Table 2 corresponds to kinetic inhibition constants (K_i). Further details relative to this series can be found in ref 13 and references therein.

To demonstrate the alignment independence of the novel descriptors, we built a series including three copies of each original structure; one corresponds with the original orientation extracted from the crystal complexes, while the other two were obtained by orienting randomly the original structure (random rotations and displacements in the three axes). This series of 30 molecular structures was analyzed in ALMOND, using DRY, O, and N1 probes, 150 seeds, and 70% field weight. The remaining parameters were set to default values. Six correlograms were obtained, containing 35 variables each, producing a matrix with 30 objects and 210 variables. PCA was directly applied to this matrix with no pretreatment or scaling. Figure 4a shows the scores plot resulting from the analysis. The plot shows clearly that structures representing the same compound in different orientations are closely clustered. This is particularly evident in compounds such as **1**, **5**, and **10**, while the three structures of compounds **8** and **9** show a slightly larger spread.

A perfect agreement between orientations would not be expected. Though GRIND are independent of the

Table 2. Series of 10 Glucose Analogue Inhibitors of Glycogen Phosphorylase

	substituent at C1 position		
no.	Rα	Rβ	pK_i (mM)
1	OH	H	2.77
2	C(=O)NH$_2$	H	3.43
3	H	C(=O)NH$_2$	3.36
4	H	COOCH$_3$	2.55
5	H	CH$_2$CN	2.05
6	H	NHC(=O)NH$_2$	3.85
7	C(=O)NH$_2$	NHCOOCH$_3$	4.80
8			5.52
9			3.84
10			4.22

position of the molecules, the results of the GRID program are not. The MIF obtained from GRID can be seen as a discrete sample of a continuous field, and the values of energy computed at the grid nodes can exhibit small differences when molecules adopt a different position inside the grid. This effect is minimized when using a smaller grid spacing, and in most cases, the default grid spacing (0.5 Å) is enough to produce a reasonable alignment independence.

The same data matrix was analyzed by PLS. In this case, the three structures representing the same compound were assigned the same value of the biological activity. The PLS plot for the first LV (the *X*-scores vs *Y*-scores scatter plot, also called TU plot) is represented in Figure 4b, illustrating how the three different structures represent the same compound in the PLS model. Details about the PLS model obtained are not relevant, since the multiplicity of the objects bias the model validation.

The use of GRIND in PLS is better illustrated building a simplified matrix with a single structure representing each compound. This was done using the original crystal structures and the same ALMOND parameters indicated above. A preliminary inspection of the correlograms obtained showed that the auto-correlogram obtained with the N1 probe exhibits a clear correlation with the activity (Figure 3). We took advantage of this fact and decided to use only this probe, to obtain a model which is both simple and easy to interpret. The analysis was then repeated using only the N1 correlogram, obtaining a smaller matrix of 10 objects and 30 variables. The data were subjected to PLS analysis within the same ALMOND program without any scaling or pretreatment. FFD variables selection [20] was applied as described in Materials and Methods,
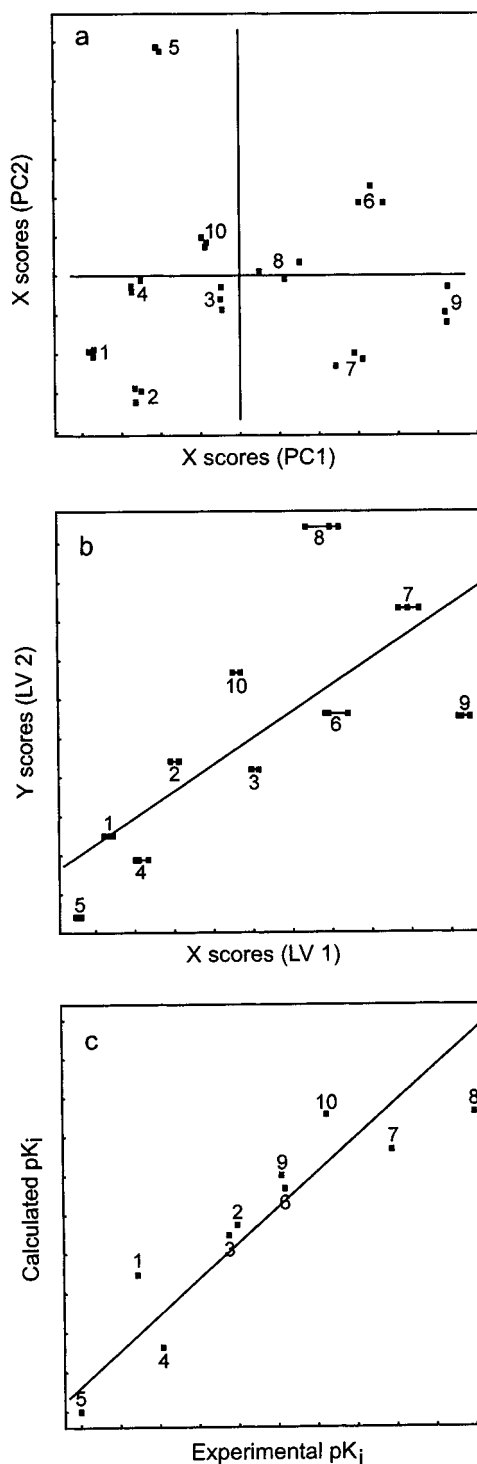

Figure 4. (a) PCA scores plot for a series of 30 molecular structures characterized by GRIND. The series contains three structures for each compound: one superimposed and two randomly oriented. In the plot, these are clustered together, showing the alignment invariance of the descriptors. (b) PLS plot (*X*-scores vs *Y*-scores) on the same series of 30 structures. Again, the representatives of the same structures appear closely clustered together. (c) Experimental vs calculated biological activity scatter plot for the GRIND model obtained on the series of 10 glucose analogue inhibitors of glycogen phosphorylase.

removing only four variables. At the end, one LV was found to be significant using cross-validation, and the final model, though simple, explained a significant amount of variance and survived model validation (r^2
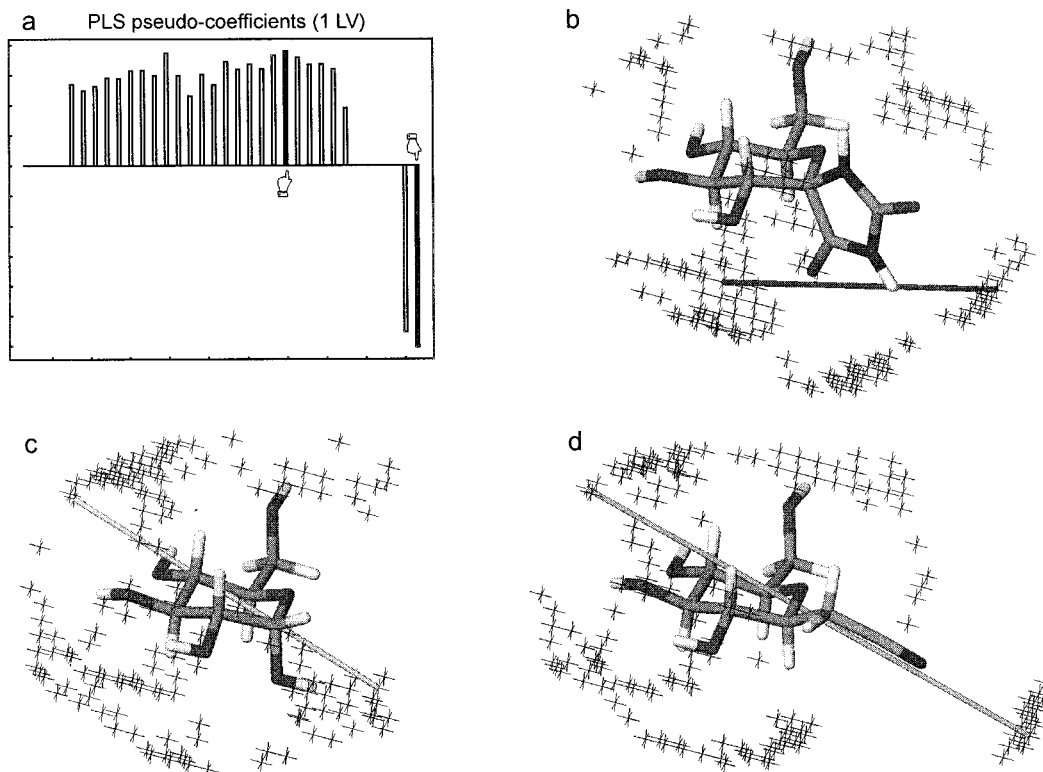
Figure 5. (a) PLS pseudo-coefficients histogram for the model represented in Figure 4c. The first hand points to the variable *A* represented in panels b and c. The second hand points to the variable *B* represented in panel d. (b) Interaction *A* present in compound **8**, the most active in the series. The interaction is intense, because the field represents interaction of the probe with carbonyl oxygen. (c) The same interaction *A* for glucose **1**, a weak inhibitor. The interaction is originated by hydroxyl groups and therefore is less intense. (d) Interaction *B*, typical compound **5**, the least active in the series. Note the negative PLS coefficients.

$= 0.83$, $q^2 = 0.60$, $q^2_{LOO} = 0.64$). A plot of the calculated vs experimental values is represented in Figure 4c. Better models can easily be obtained including more probes, but this would have made the model interpretation more complex.

The PLS pseudo-coefficients histogram (Figure 5a) shows that nearly all N1—N1 interaction energies correlate positively with the activity, except for the ones corresponding to the longest distances, which show a negative interaction. As stated above, the variables represent pairs of nodes at different distances where the N1 probe has interacted favorably. Since the compounds are glucose derivatives, and hence rather polar, many such nodes were found, and indeed all compounds have a defined interaction for a pair of nodes at each distance considered. However, most active compounds in the series (**8**−**10**) are characterized by having a spiro ring with two carbonyl oxygens, while some of the less active compounds (**1** and **5**) do not have any such oxygen. Therefore, the product of the node−node interaction is higher when they represent pure carbonyl nodes (as the N1 probe interacts more strongly with carbonyl) than when they represent the product of hydroxyl and carbonyl nodes or between two hydroxyl nodes. These differences can be observed directly in the correlograms, coloring the points according to their activity. In Figure 3, red compounds (active) produce larger node−node energies than green compounds (inactive).

The peak at variable 20 in the PLS pseudo-coefficient histogram is particularly interesting: in the spiro compounds it represents node−node interactions between two clusters of carbonyl nodes (Figure 5b), while

in other compounds, like glucose **1**, they represent interactions between weaker hydroxyl nodes, situated in a lower position (Figure 5c). The regions present in the active compounds and shown linked in the figure are interesting targets for the introduction of polar, hydrogen bond acceptor groups.

The peak on the right-hand side of the histogram represents long-distance node−node interactions present only in compound **5**, which is the least active compound in the series. Therefore, the coefficients are negative, because the presence of such interactions correlates negatively with activity.

All in all, the information given by the model can be chemically interpreted in very simple terms as: (i) long substituents, like the cyano shown in Figure 5d, should be avoided and (ii) groups producing strong polar interactions in the regions linked in Figure 5b can produce stronger binding. In this particular series, due to the fact that the complexes are available, the interpretation hypothesis can be contrasted with the crystal structures of the receptor. Indeed, it can be seen that the lower of the two regions mentioned above is near the water 897 and the phosphate group of the pyridoxal phosphate cofactor, while the upper region represents regions where interaction with waters 847 and 890 is possible.

This interpretation, even if simplistic, captures most of the relevant information present in the series. To compare the model obtained with other 3D-QSAR methods we performed a GRID/GOLPE model on the same series. Exceptionally, since the structures of the compounds were extracted from the crystal structures,
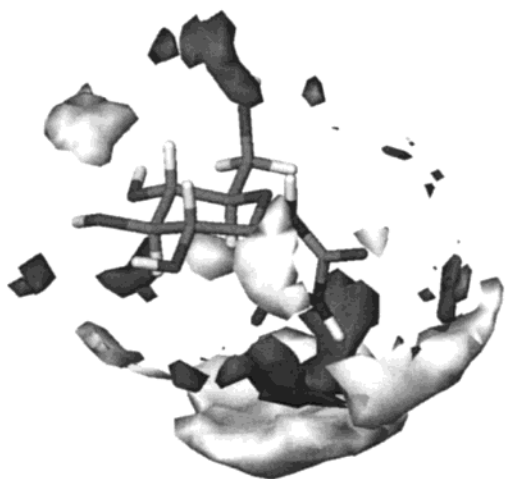
Figure 6. GRID/GOLPE model on the series of glucose analogue inhibitors of glycogen phosphorylase. Isocontour plot of the PLS pseudo-coefficients for 2 LV at levels +0.001 (light gray) and −0.001 (dark gray). Positive regions represent field interactions present in the most active compounds, much like the ones highlighted in Figure 5b. Conversely, negative regions represent field interactions present in less active compounds and correspond to the lower regions linked in Figure 5c.

no alignment step was required. The GRID analysis was as close as possible to the GRID analysis used in the ALMOND procedure (N1 probe, 0.5 Å grid spacing). The matrix contained 98 569 variables and 10 objects and was pretreated removing positive values (to mimic as much as possible the MIF used by ALMOND), zeroing those values with absolute values smaller than 0.01 kcal/mol and removing variables with standard deviation below 0.01. In addition, variables showing only two or three values and having a skewed distribution were also removed. After the pretreatment, the dataset still contained 17 763 variables, and SRD/FFD variables selection[13] was applied. At the end we obtained a 2-LV model with 847 variables ($r^2 = 0.94$, $q^2 = 0.64$, $q^2_{LOO} = 0.66$) on which we can try an interpretation and compare with the one made for the ALMOND model.

Figure 6 shows an isocontour plot of the PLS pseudo-coefficients (2-LV model). Light gray regions correspond to positions where a favorable interaction with N1 would enhance activity, while dark gray regions show regions where interaction with N1 would decrease activity. It is evident that interpretation is rather coincident with those of the ALMOND model. Examining variable 20 of the ALMOND model we can see how the regions linked in the active compound are nearly identical to the light gray (favorable) regions, while one of the regions present for the less active compound corresponds closely to the dark gray (unfavorable) region in the GRID/GOLPE model.

Steroid Binding to the Corticosteroid-Binding Globulin Receptor. This dataset was first used by Cramer et al.[3] and since then has been used by many authors as a kind of "benchmark" for measuring the quality of different 3D-QSAR methodologies. Indeed, the "CoMFA steroid" dataset has been the subject of a recent review[21] where the results obtained with about 15 QSAR methods on this dataset were compared. However, this review also showed that incorrect structures had sometimes been included. In this work, we have used the dataset compiled by Wagener et al.[5] (Table 3), which

Table 3. Series of 31 Steroids Binding to the Corticosteroid-Binding Globulin Receptor

no.	steroid	log K^a	pred log K^b
1	aldosterone	6.279	
2	androstanediol	5	
3	androstenediol	5	
4	androstenedione	5.763	
5	androsterone	5.613	
6	corticosterone	7.881	
7	cortisol	7.881	
8	cortisone	6.892	
9	dehydroepiandrosterone	5	
10	deoxycorticosterone	7.653	
11	deoxycortisol	7.881	
12	dihydrotestosterone	5.919	
13	estradiol	5	
14	estriol	5	
15	estrone	5	
16	etiocholanolone	5.255	
17	pregnenolone	5.255	
18	17-hydroxypregnenolone	5	
19	progesterone	7.380	
20	17-hydroxyprogesterone^c	7.740	
21	testosterone	6.724	
22	prednisolone	7.512	7.880
23	cortisol-21-acetate	7.553	7.729
24	4-pregnene-3,11,20-trione	6.779	6.873
25	epicorticosterone	7.200	7.557
26	19-nortestosterone	6.144	5.706
27	16α,17-dihydroxy-4-pregnene-3,20-dione	6.247	6.505
28	16α-methyl-4-pregnene-3,20-dione	7.120	7.050
29	19-norprogesterone	6.817	7.033
30	11β,17,21-trihydroxy-2α-methyl-4-pregnene-3,20-dione	7.688	7.606
31	11β,17,21-trihydroxy-2α-methyl-9α-fluoro- 4-pregnene-3,20-dione^c	5.797	

^{a} CBG affinity. ^{b} CBG affinity (log K) predicted by the GRIND model; see text. ^{c} Outlier, removed from analysis.

fixed some mistakes present in previous versions. In addition, we corrected a slight mistake in the activity of compounds **16** and **17** and used as biological activities the log K, instead of the pK, as has been suggested previously.[21] Table 3 does not include the 2D structures of the compounds, which can be found in the original references. The 3D structures used were downloaded from Gasteiger's internet homepage[22] and correspond with the structures generated by CORINA[23] from the 2D structures. We should emphasize that the 3D structures were used in the same orientation in which they appear in the original MDL SDFile and that they were not minimized or superimposed at all.

ALMOND was first applied using standard parameters. Six correlograms were obtained containing 45 variables each. The first PLS models showed that the probe DRY contributed almost nothing to the model, and two outliers were identified (**20** and **31**). Compound **31** was, indeed, recognized as an outlier in previously published models,[5] while the substituents at position 17 in compound **20** are somewhat particular (present only in compounds **18**, **27**, and **20**). Some tests were made using different probes and parameters, and compound **31** consistently behaved as an outlier, while in order to represent properly compound **20** an additional probe modeling steric interactions was needed. It is good chemometrical practice not to include variables representing only a single compound, since the models obtained are artificially good, and accordingly, we decided to remove both compounds from the dataset. Analyses were repeated using only probes O and N1,
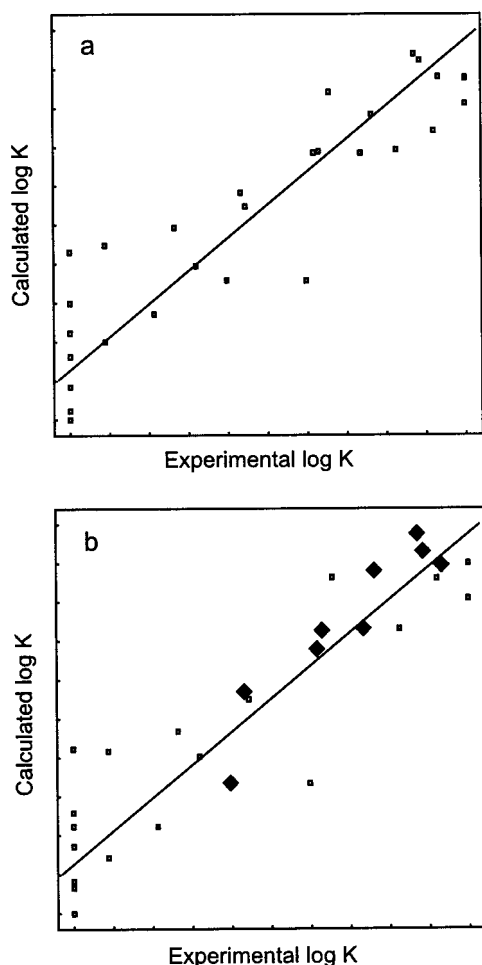
Figure 7. (a) Experimental vs calculated biological activity scatter plot for a 2-LV model made using GRIND on the series of steroids. (b) Same as in panel a, but using only 20 compounds as the training set and the other 9 as an external prediction set. Filled diamonds represent the compounds in the external set and the values of activity predicted.

Table 4. Series of 25 Butyrophenones with Serotoninergic Activity

compd	structure[a]	substituent[b]	pK_i(5-HT$_{2A}$)
10a	VIII	a	7.58
10e	VIII	e	7.34
10f	VIII	f	6.05
23a	IX	a	8.15
23b	IX	b	8.76
23e	IX	e	7.37
24a	X	a	8.84
24b	X	b	8.56
24e	X	e	6.54
24f	X	f	6.84
29a	XII	a	7.95
29b	XII	b	8.17
35a	XI	a	7.24
35b	XI	b	8.33
35e	XI	e	6.97
35f	XI	f	5.82
Ia	I	a	8.80
Ie	I	e	7.75
If	I	f	6.29
IIa	II	a	8.60
IIe	II	e	6.91
IIIa	III	a	8.11
IIIe	III	e	7.14
IIIf	III	f	7.39
IVa	IV	a	7.88
haloperidol[c]			7.70

[a] See Chart 1. [b] See Chart 2. [c] See Chart 3.

18) are rather high. This explains why the second model gets much better SDEP in the external prediction set than in the internal one, since the 9-compound prediction set contains no inactive compound.

From these results, we can see that the method compares reasonably well with other 3D-QSAR models obtained for the same dataset. For example, in a recent review of models obtained for this dataset against CBG binding[21] Coats reported 14 methods with q^2_{LOO} ranging between 0.23 and 0.93 (average q^2_{LOO} = 0.71), which compares well with the q^2_{LOO} of 0.76 obtained using GRIND. It should be stressed that the GRIND model was obtained without the need to perform an alignment step and using descriptors which can be related back to the input structures as shown in the previous section.

Butyrophenones with Serotoninergic (5-HT$_{2A}$) Affinities. In the two previous examples we have shown the use of our methodology for two well-studied series: the first was used for demonstrative purposes, the second for comparison with other methods. In this example we would like to illustrate the use of GRIND on a novel series of compounds, extracted from the literature.[14] Raviña et al. present a series of 25 butyrophenones with serotoninergic (5-HT$_{2A}$, 5-HT$_{2C}$) affinities (Table 4). The compounds were defined as "conformationally constrained" by the authors and indeed are more rigid than the average serotonin-binding compounds, but even so they contain between three and eight rotatable bonds, introducing remarkable conformational freedom. With respect to the biological activities, only the binding affinity for 5-HT$_{2A}$ is available for the whole series. The pK_i values were obtained from in vitro assays and reflect competitive binding, measured as displacement of [^{3}H]ketanserin in rat cerebral cortex.

The structures were built from the published 2D structures. All compounds were considered to be protonated, with a formal charge +1 placed on one nitrogen of the piperazine or piperidine ring. They were modeled
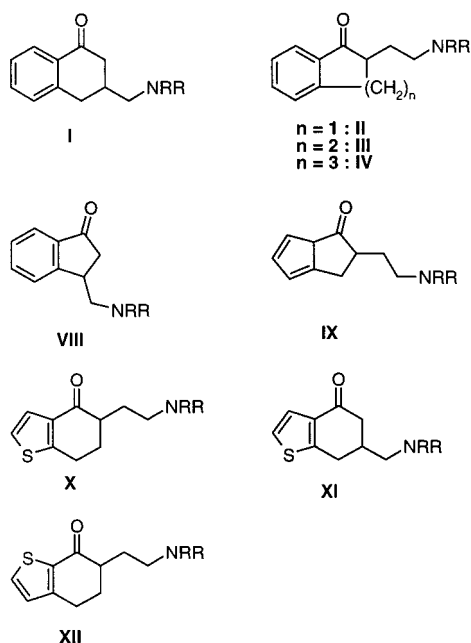
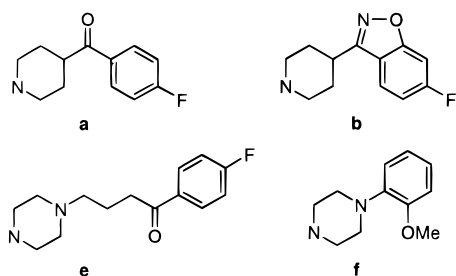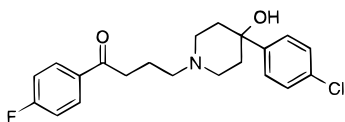100 seeds, and 75% weight of field values. The analysis produced three correlograms of 45 variables each, giving a matrix of 135 variables and 29 objects. FFD variable selection kept 69 variables. A final model with an optimal dimensionality of 2 LV was obtained (69 variables, r^2 = 0.83, q^2 = 0.75, q^2_{LOO} = 0.76). Figure 7a shows a scatter plot of the calculated vs experimental biological activities for this dataset.

To further test the predictive ability of the model, the dataset was then divided into a training set including the first 20 compounds and a prediction set including the last 9 compounds. The PLS model obtained on the training set was quite similar to the one obtained on the complete dataset (63 variables, r^2 = 0.82, q^2 = 0.64, q^2_{LOO} = 0.64). The standard deviation error of the predictions (SDEP), obtained by internal validation, was of 0.67. Surprisingly, the SDEP obtained for the external prediction set is much better (SDEP = 0.26), indicating that the model behaves better on making external predictions. Figure 7b shows a scatter plot of the calculated vs experimental activities, including the activities predicted for the external dataset. From the plots in Figure 7 it can be seen that both models produce reasonable fitting for most of the compounds, but the residuals of the inactive compounds (**2**, **3**, **9**, **13**−**15**, and

Chart 1

I

II
n = 1 : II
n = 2 : III
n = 3 : IV

VIII

IX

X

XI

XII

Chart 2

a

b

e

f

Chart 3

Figure 8. (a) Experimental vs calculated biological activity scatter plot for a 3-LV model made using GRIND on the series of butyrophenones with serotoninergic (5-HT$_{2A}$) affinities. The value of activity predicted for haloperidol was also included in the plot.

in extended conformation, and dihedrals were adjusted by hand in order to obtain consistent conformations through the series. Particular attention was paid to orient consistently the heteroatoms in both rings. The compounds were roughly superimposed using the piperazine or piperidine ring only to help graphic interpretation.

The modeled structures were then analyzed using ALMOND using default parameters except for the number of nodes (120) and the field weight (35%). Six correlograms of 54 variables were obtained, thus producing a matrix of 324 variables and 25 objects. The PLS analysis of the original matrix produced good models, but in order to simplify the interpretation, some of the correlograms were removed for subsequent analysis: the nodes extracted from the O MIF represent mainly the interaction with the charged nitrogen and the O–O correlogram contributes very little to the model. Moreover, cross-correlograms DRY–O and DRY–N1 exhibit only a small contribution to the model and make the interpretation unnecessarily complex. After the removal of these correlograms, the matrix contained 162 variables. FFD was applied removing 62 more
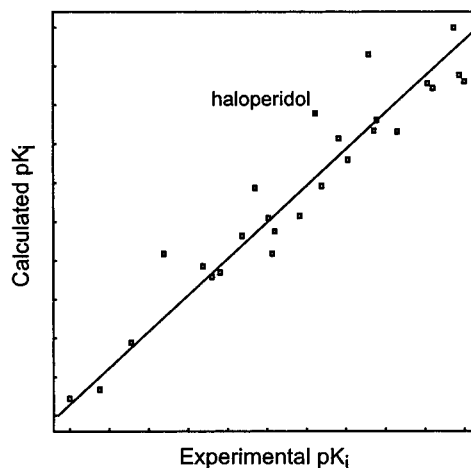
variables. The PLS analysis of this matrix (100 variables and 25 objects) produced a model of 3 LV ($r^2 = 0.93$, $q^2 = 0.81$, $q^2_{LOO} = 0.83$). Remarkably, the SDEP obtained for this series was 0.35, quite similar to the average experimental error reported for the pK_i (0.37). The model was then used to predict the activity of the antipsychotic haloperidol, whose 5-HT$_{2A}$ binding affinity is also reported in the article. The model predicted an activity of 7.92, very similar to the experimentally measured value of 7.7. Figure 8 shows a scatter plot of the calculated vs experimental activities, including also the activity prediction for haloperidol.

The model obtained can be interpreted with the help of the interactive ALMOND plots. Figure 9 shows the histogram of the PLS pseudo-coefficient for the 3-LV model obtained, together with some plots identifying node–node interactions important for the interpretation. In the histogram, we can easily identify variables belonging to three different correlograms: the first block corresponds with the DRY auto-correlogram, the second corresponds with the N1 auto-correlogram, and the third corresponds with the O–N1 cross-correlogram.

The variables of the DRY auto-correlogram represent the optimal distance that should separate the aromatic rings placed on both sides of the compounds. It should be observed that compounds in which the NRR substituents are linked by methylene bridges (**I**, **VIII**, and **XI**) are usually less active than structures linked by ethylene bridges. On the other hand, very long compounds, like those with **e** type substituents, are less active than equivalent compounds bearing different NRR substituents. According to the model, the most active compounds are characterized by hydrophobic patches separated by the distance shown in the positive peak (16.6 Å); see Figure 9b for reference.

The N1 auto-correlogram shows three peaks. A first small negative peak shows small, negligible effects. The second peak corresponds with interactions between the polar groups in the scaffold and in the NRR substituents (Figure 9c). These are stronger for NRR substituents bearing a carbonyl oxygen (**a** and **e** substituents) or isoxazole (**b**) than the methoxy group (**f**). According to the direction of the coefficients, the first set of substit-
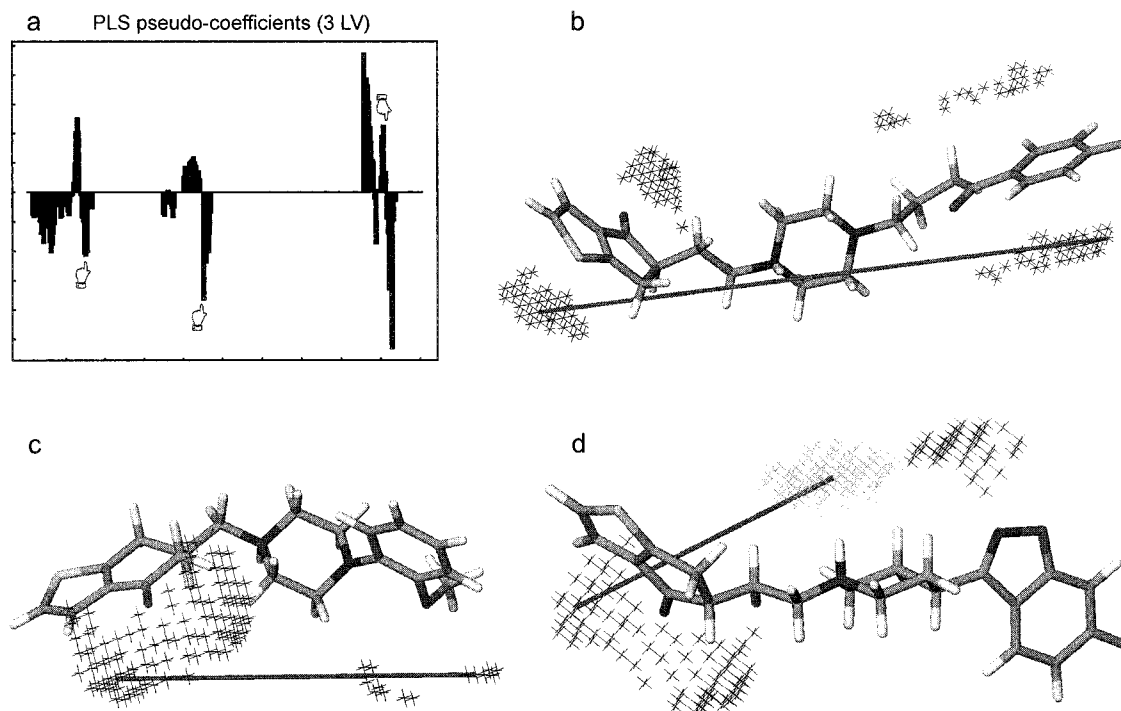
Figure 9. (a) PLS pseudo-coefficients histogram for the model of butyrophenones with serotoninergic affinities represented in Figure 8. The first hand points to the variable *A* represented in panel b, the second hand points to the variable *B* represented in panel c, and the third hand to the variable *C* represented in panel d. (b) Interaction *A*, showing an unfavorable interaction between hydrophobic regions present in long compounds such as the one represented (**23e**). (c) Interaction *B*, between hydrogen bond acceptor regions. The distance shown characterizes the presence of a methoxy group (NRR of type **f**) which is typical of less active compounds. The plot shows compound **35f**, the least active in the series. (d) Interaction *C*, representing interaction between hydrogen bond donors (mainly around the protonated nitrogen) and hydrogen bond acceptor groups. Interactions at the distance shown are present in active compounds such as the one shown (**23b**).

uents is preferable for the activity. Indeed, most compounds including **f** type NRR substituents are significantly less active. The last peak represents long-distance interactions, which are absent in compounds of **b** type. Since some of the most active compounds have **b** type substituents, the presence of interactions at this distance correlates negatively with the activity. All in all we can conclude that the N1 correlogram shows an optimal distance separating the regions favorable for hydrogen bond acceptor interactions and, also, that substituents able to produce interaction stronger than methoxy are preferable.

The last correlogram represents interaction between nodes extracted from the O field (basically, the region around the positively charged N) and nodes extracted from the N1 field (the same regions mentioned above). Shorter distances express interactions between O regions and N1 regions produced by the NRR substituents, while longer distances represent interactions with polar groups of the scaffold. Therefore, the first peak expresses the same information as the first positive peak of the N1 auto-correlogram: weaker O−N1 interaction produced by methoxy substituents in compounds of type **f** produces less active compounds. Much more interesting are the next two positive and negative peaks. They seem to express the ideal distance between O and N1 regions. The interactions are optimal for the distance expressed by the positive peak (9 Å), while for longer distances these interactions are associated with less active compounds. This is the reason scaffolds of type **IX** seem to produce optimal interaction while those of

type **XI** are the worst. Figure 9d shows such interactions for compound **23b**, one of the most active in the series.

Conclusions

In this work we have presented a novel type of molecular descriptor, GRid-INdependent Descriptors (GRIND), based upon the 3D molecular interaction fields around a ligand and the program ALMOND for the generation and visualization of the descriptors and the resulting models.

The GRIND present a number of interesting features which overcome some of the limitations of other descriptors used in 3D-QSAR. Their most important properties being: (i) alignment-independent, (ii) simple and fast to compute, (iii) highly relevant for describing biological properties, and (iv) ability to interpret models in the original descriptor space (molecular interaction fields of the compounds). Models obtained with GRIND are of comparable statistical quality with respect to models obtained with other 3D-QSAR methodologies, from the point of view of the fitting and predictive power of the models, and are readily interpretable. Moreover, from a chemometric point of view, they are quite robust and less prone to overfitting and other undesirable problems than models obtained with other methods. All these characteristics make GRIND a promising methodology. In the field of drug design, GRIND provides a fast and simple way to obtain structure−activity methods, not involving a time-consuming and difficult preliminary step related to compound superimposition.

The main limitation of the GRIND is the sensitivity to the conformation of the structures. This drawback is

not limited to GRIND, and in fact, they appear to be more robust to conformational changes than other methodologies. In this article, the descriptors were applied to series of rigid and semirigid compounds with some success. However, obtaining accurate QSAR models on highly flexible compounds is an ongoing area of active research in our group.

This article has introduced the GRIND for the first time, and only a few of the potential applications were presented. Preliminary results in our group demonstrate that GRIND are also able to characterize binding sites in biomolecules, in particular in selectivity studies. Other possible applications are also under study.

Software Availability. ALMOND is a commercial software distributed by Multivariate Infometric Analysis S.r.l. MIA offers free unsupported licenses to academic and nonprofit institutions. Licensing information is available at http://www.miasrl.com.

References

(1) Hansch, C. A quantitative approach to biochemical structure–activity relationships. *Acc. Chem. Res.* **1969**, *2*, 232–239.
(2) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849–857.
(3) 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.
(4) Broto, P.; Moreau, G.; Vandycke, C. Molecular structures: perception, autocorrelation descriptor and sar studies. Autocorrelation descriptor. *Eur. J. Med. Chem.* **1984**, *19*, 66–70.
(5) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of Molecular Surface Properties for Modeling Corticosteroid Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769–7775.
(6) Anzali, S.; Barnickel, G.; Krug, M.; Sadowski, J.; Wagener, M.; Gasteiger, J.; Polanski, J. The comparison of geometric and electronic properties of molecular surfaces by neural networks: application to the analysis of corticosteroid-binding globulin activity of steroids. *J. Comput.-Aided. Mol. Des.* **1996**, *10*, 521–534.
(7) Clementi, S.; Cruciani, G.; Riganelli, D.; Valigi, R.; Costantino, G.; Baroni, M.; Wold, S. Autocorrelation as a tool for a congruent description of molecules in 3D-QSAR studies. *Pharm. Pharmacol. Lett.* **1993**, *3*, 5–8.
(8) Broto, P.; Moreau, G.; Vandycke, C. Molecular structures: perception, autocorrelation descriptor and sar studies. Use of the autocorrelation descriptor in the QSAR study of two non-narcotic analgesic series. *Eur. J. Med. Chem.* **1984**, *19*, 79–84.
(9) Todeschini, R.; Gramatica, P. New 3D Molecular Descriptors: The WHIM Theory and QSAR Applications. In *3D QSAR in Drug Design*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; KLUWER/ESCOM: Dordrecht, 1998; Vol. 2, pp 355–380.
(10) Silverman, B. D.; Platt, D. E. Comparative molecular moment analysis (CoMMA): 3D-QSAR without molecular superposition. *J. Med. Chem.* **1996**, *39*, 2129–2140.
(11) 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.
(12) Pastor, M.; Cruciani, G.; Watson, K. A. A strategy for the incorporation of water molecules present in a ligand binding site into a three-dimensional quantitative structure–activity relationship analysis. *J. Med. Chem.* **1997**, *40*, 4089–4102.
(13) Pastor, M.; Cruciani, G.; Clementi, S. Smart region definition: a new way to improve the predictive ability and interpretability of three-dimensional quantitative structure–activity relationships. *J. Med. Chem.* **1997**, *40*, 1455–1464.
(14) Raviña, E.; Negreira, J.; Cid, J.; Masaguer, C. F.; Rosa, E.; Rivas, M. E.; Fontenla, J. A.; Loza, M. I.; Tristán, H.; Cadavid, M. I.; Sanz, F.; Lozoya, E.; Carotti, A.; Carrieri, A. Conformationally constrained butyrophenones with mixed dopaminergic (D_2) and serotoninergic (5-HT_{2A}, 5-HT_{2C}) affinities: synthesis, pharmacology, 3D-QSAR, and molecular modeling of (aminoalkyl)benzo- and -thienocycloalkanones as putative atypical antipsychotics. *J. Med. Chem.* **1999**, *42*, 2774–2797.
(15) GRID v.17; Molecular Discovery Ltd., West Way House, Elms Parade, Oxford, 1999.
(16) ALMOND v. 2.0; Multivariate Infometric Analysis, S.r.l., Viale dei Castagni, 16, Perugia, 2000.
(17) Fedorov, V. V. *Theory of Optimal Experiments*; Academic Press: New York, 1972.
(18) Clementi, M.; Clementi, S.; Clementi, S.; Cruciani, G.; Pastor, M. Chemometric Detection of Binding Sites of 7TM Receptors. In *Molecular Modelling and Prediction of Bioreactivity*; Gundertofte, K., Jorgensen, F. S., Eds.; Kluwer Academic/Plenum Publishers: New York, 2000; pp 207–212.
(19) Cruciani, G.; Clementi, S.; Baroni, M. Variables Selection in PLS Analysis. In *3D QSAR in Drug Design, Theory, Methods and Applications*; Kubinyi, H., Ed.; ESCOM: Leiden, 1993; pp 551–564.
(20) 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.
(21) Coats, E. A. The CoMFA Steroids as a Benchmark Dataset for Development of 3D QSAR Methods. In *3D QSAR in Drug Design*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; KLUWER/ESCOM: Dordrecht, 1998; Vol. 3, pp 199–213.
(22) Dataset of 31 steroids binding to the corticosteroid-binding globulin (CBG) receptor. http://www2.ccc.uni-erlangen.de/services/steroids/index.html.
(23) CORINA Molecular Networks, GmbH Computerchemie Langemarckplatz 1, Erlangen, Germany, 1997.

JM000941M