Incorporating Molecular Shape into the Alignment-free GRid-INdependent Descriptors

Fabien Fontaine, Manuel Pastor,* and Ferran Sanz

Research Unit on Biomedical Informatics (GRIB), IMIM, Universitat Pompeu Fabra, C/Dr. Aiguader, 80, E-08003 Barcelona, Spain

Received December 4, 2003

The recently introduced GRid-INdependent Descriptors (GRIND) were designed to provide a suitable description of a series of ligands for 3D-QSAR studies not requiring the spatial superimposition of their structures. Despite the proven usefulness of the method, it was recognized that the original GRIND failed to describe appropriately the shape of the ligand molecules, which in some cases plays a major role in ligand−receptor binding. For this reason, the original descriptors have been enhanced with the addition of a molecular shape description based on the local curvature of the molecular surface. The integration of this description into the GRIND allows the generation of 3D-QSAR models able to identify both favorable and unfavorable shape complementarity in a simple and alignment-independent way. The usefulness of the new GRIND-shape description in 3D-QSAR is illustrated using two structure−activity studies: one performed on a set of xanthine-like antagonists of the A_1 adenosine receptor; another performed on a series of *Plasmodium falciparum* plasmepsin II inhibitors.

Introduction

The success of quantitative structure−activity relationships (QSAR) has been linked to the development of appropriate molecular descriptors. The choice of simple but relevant descriptors can be seen as one of the keys for the success of the original Hansch method.[1] More recently, CoMFA[2] and other 3D-QSAR methods[3,4] have extended the original possibilities of the QSAR as a result of the use of much more sophisticated molecular descriptors based on 3D molecular interaction fields (MIF). However, most 3D-QSAR methods suffer from the drawback of requiring the superimposition of the 3D structures of the ligands according to a hypothesis of their binding mode. The 3D alignment of structures is often time-consuming and always subjective, thus increasing the probability of obtaining a low-quality model and limiting the applicability of 3D-QSAR methodologies. Indeed, the methodology for the 3D superimposition is still being improved.[5]

Recently, we developed a new class of molecular descriptors called GRid-INdependent Descriptors (GRIND)[6] that aim to overcome the 3D-QSAR drawbacks described above. The GRIND calculation starts by computing several MIF using the GRID program.[7] These MIF characterize the potential of interaction between a molecule of interest (e.g., a steroid) and particular chemical groups of the receptor, represented by chemical probes (e.g., water, amine nitrogen, etc.). The GRIND approach aims to extract the information enclosed in the MIF and to encode it into new types of variables whose values are independent of the spatial position of the molecule studied (Figure 1). This encoding performed in two steps: a filtering procedure based on the energy of the nodes and the distance between them is applied in order to extract relevant regions of favorable interaction. Then the product of the energy of interaction for each pair of nodes is computed and assigned to

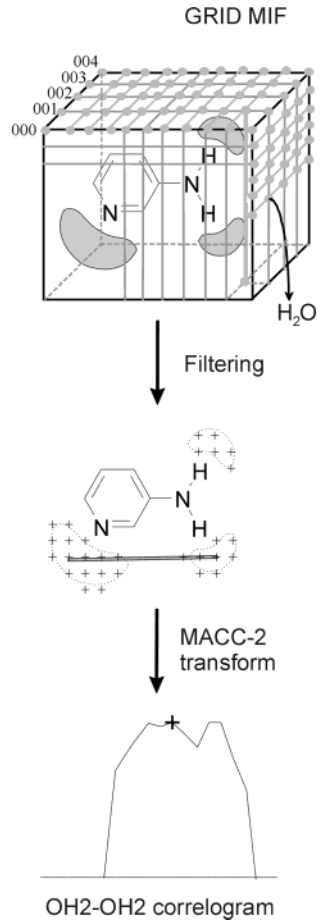


Figure 1. Computation of the GRIND. A molecular interaction field is computed with the GRID force field, and the most relevant interactions are filtered and encoded into MACC-2 correlograms.

* To whom correspondence should be addressed. Phone: +34 932 240 302. Fax: +34 932 240 875. E-mail: mpastor@imim.es.

10.1021/jm0311240 CCC: $27.50
Published on Web 04/24/2004

a distance bin according to the node separation. For each distance bin, only the highest product is kept, thus allowing its representation in the original 3D space as a line linking two specific MIF nodes. This possibility of representing the descriptors graphically makes the GRIND particularly well suited to studies requiring a structural interpretation of the models.

GRIND variables are grouped into blocks representing interactions between couples of nodes generated by the same probe (autocorrelograms) or combination of probes (cross-correlograms) (Figure 1). Such variables constitute a matrix of descriptors that can be analyzed using multivariate techniques, such as principal component analysis (PCA)[8] and partial least squares (PLS) regression analysis.[9]

A typical procedure involves the calculation of three fields, each one using a specific chemical probe highly relevant to a particular kind of interaction (e.g., hydrogen-bond acceptor, hydrogen-bond donor, and hydrophobic probe). Most of the interactions between a ligand and a protein binding site are covered by the MIF of these three probes, making the GRIND a powerful tool for the characterization of receptor-binding properties. Their use in the areas of 3D-QSAR,[6,10−12] data mining,[13] and molecular diversity[14] has been recently published.

However, the practical use of the GRIND has shown that the descriptions they provide can be incomplete in some cases. In its original formulation, the GRIND included no explicit description of the molecular shape. Instead, it was assumed that the MIF produced by hydrophobic, hydrogen-bond acceptor and donor probes would provide a comprehensive description of the different regions of every molecule, thus representing the molecular shape, albeit indirectly. Unfortunately, the fields generated by the DRY probe around some aliphatic hydrophobic regions are so weak that often the GRIND method fails to represent these regions. As a consequence, most aliphatic nonpolar areas remained completely "invisible", and the implicit molecular shape description presumed in the original method is left incomplete.

The shape of a ligand is crucial to its ability to bind to a receptor. On one hand, the binding strength of the ligands having appropriate shape complementarities with the binding pocket is enhanced by favorable van der Waals and hydrophobic interactions, some of which are difficult to quantify by classical 3D-QSAR methods.[15] On the other hand, an inappropriate shape complementarity might prevent some ligands from binding, purely for steric reasons. This negative effect of the shape is extremely important, since it is not additive, nor can it be compensated for by other effects. In many cases steric effects can explain the presence of outliers in the models.

Indeed, the importance of the shape description has been recognized by many authors, and a wide set of methodologies for describing the molecular shape in the context of drug design have been published.[16−23] The aim of the present study is to incorporate the shape description within the overall GRIND formalism. Ideally, shape should be represented ultimately in a correlogram-like form (Figure 2) where the autocorrelograms would describe the distance between certain regions defining the spatial extent of the molecule

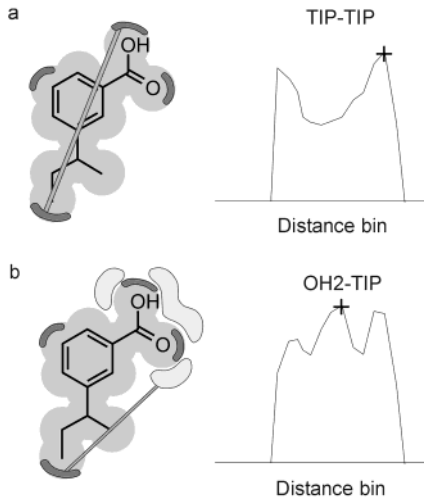


Figure 2. Aim of the shape descriptors: (a) autocorrelogram (TIP−TIP) encodes the geometrical relationships between the spatial extents of the molecule; (b) cross-correlogram (OH2−TIP) encodes the geometrical relationships between the spatial extents and some regions of favorable interaction with a GRID probe (e.g., water probe).

(Figure 2a) and the cross-correlograms would describe the distance between these regions and other regions representing relevant interactions of the compounds (Figure 2b). From a 3D-QSAR point of view, the variables from these correlograms can easily be used as descriptors and have a straightforward interpretation: variables having a positive contribution in structure−activity models indicate that the regions defining the corresponding spatial extents probably fit well into the receptor binding pocket, and conversely, variables with a negative contribution indicate steric hindrance for the same regions.

To be consistent with the GRIND formalism, the molecular shape description must start by selecting a reduced number of highly representative nodes. In the proposed method, these are extracted from the molecular surface and further selected according to a criterion based on the local curvature of the molecular surface. In the context of QSAR, where we intend to describe the differences in a series of structurally related compounds, a relevant description should be sensitive to the introduction of substituents, changes in ring size, elongation of chains, etc. In these situations, the local curvature of the molecular surface describes particularly well the structural changes, since most of these changes are characterized by producing "protrusions" in the surface. Consequently, the local surface curvature was the criterion for the selection of the most relevant surface nodes, which were then processed using the same method as for other GRID nodes.

The present article describes in detail the computational method used for generating a molecular shape description compatible with the GRIND formalism and shows the usefulness and relevance of the novel descriptors in two practical 3D-QSAR applications. In the first example, a 3D-QSAR model of adenosine receptor antagonists[24] is used to demonstrate the additional information provided by the shape field and how the good ligand−receptor shape complementarity of the high-affinity ligands can be detected. In the second example, a 3D-QSAR model of *Plasmodium falciparum*

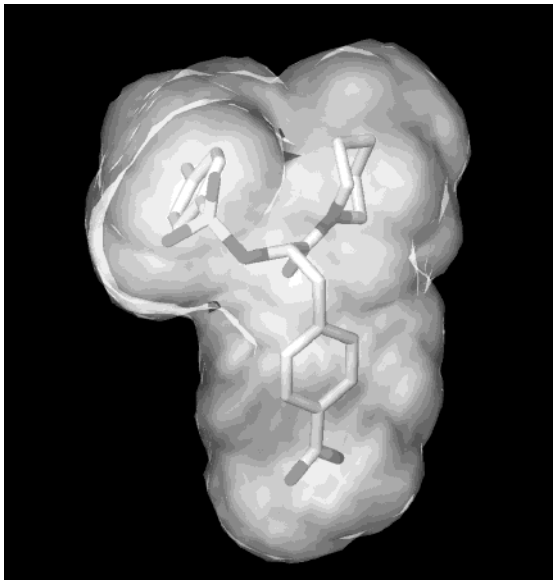


Figure 3. Positive 1 kcal/mol isosurface of 4-TAPAP computed with the O probe of the program GRID. The isosurface defines the shape of the inhibitor in its bound conformation.

plasmepsin II inhibitors issued from a statine combinatorial library,[25] the shape field gives us relevant structural insights about unfavorable shape features of some compounds of the series.

Materials and Methods

The development of a molecular shape descriptor suitable for integration into the GRIND involves two major steps: the development of a "shape field" and its incorporation within the overall GRIND methodology.

Shape Field. First of all, it must be clarified that the molecular shape description introduced herein is not actually a MIF, although the results of the shape analysis are expressed in a MIF-like format in order to facilitate its integration within the GRIND methodology, and for this reason, it was termed the "shape field". Also, to maintain consistency with the terminology commonly used for GRID fields (often called after the name of the probe used to generate it, e.g., DRY or N1), the molecular shape field was also referred to as TIP. The name of the probe makes reference to the fact that the regions getting more extreme values for this field are often located at the "tips" of the 3D molecular structures.

As mentioned in the Introduction, the rationale behind this shape field is the extraction of some regions in the surface of the molecule describing at best the spatial extent of the molecule. These regions are selected by considering the local curvature, with the idea that the most convex regions are the most descriptive of the spatial boundaries of the molecule and the most suitable for representing the structural diversity of a typical series used in QSAR. Therefore, the analysis involves three steps, which are detailed below: (i) an approximation to the molecular surface is obtained from a GRID MIF, (ii) then a set of nearest neighbors is selected for each surface node, and (iii) the surface curvature coefficient of each node is estimated.

(i) Molecular Surface. There are diverse methods for computing the surface of a molecule. Hard sphere models are used to compute several type of surfaces, e.g., the van der Waals outer surface, the solvent accessible surface[26] (the surface traced out by the center of a solvent probe rolling over the spherical atoms), and the solvent excluded surface[27−29] (the topological boundary of the union of all possible solvent molecules having no intersection with the atomic spheres). Isosurfaces calculated from a charge density distribution[30] or from a molecular interaction field are another way of defining the shape of a molecule. Since the GRIND method already

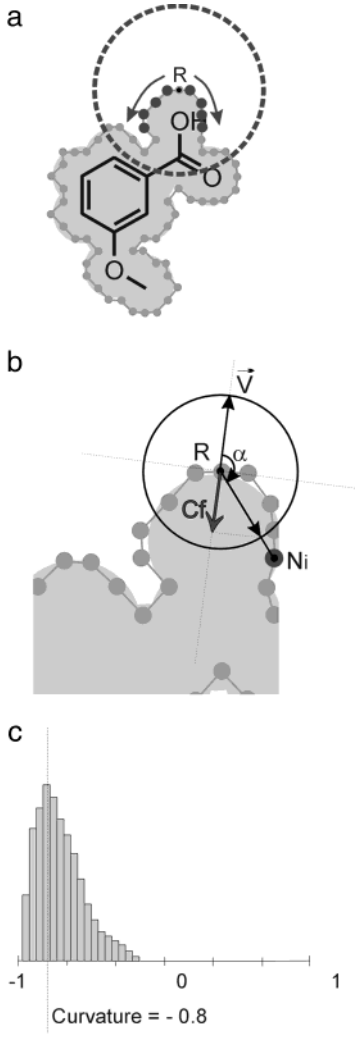


Figure 4. Surface curvature calculation: (a) recursive nearest-neighbor finding for each surface node *R*; (b) calculation of the partial curvature coefficient (Cf) for each nearest neighbor N_i; (c) curvature coefficient value of node *R* is equal to the median of the Cf distribution.

requires the computing of GRID MIF, the last strategy is convenient from a computational point of view. Any field generated with a GRID probe can be used for the computation of a molecular surface at a given energy level. To describe the shape of the molecule, the method uses the isosurface generated by the O probe at an energy threshold of 1 kcal/mol. An example of such a surface is shown for 4-TAPAP (Figure 3), a well-known thrombin inhibitor in its bound conformation. The surface is computed by identifying the grid nodes with a value at the boundary of the energy threshold of 1 kcal/mol. The nodes are then connected to define a molecular surface that is fitted to the grid. In this way, the shape field can be used and depicted in a similar way for any regular GRID field. Of course, the quality of the surface obtained depends on the grid resolution and the shape of the surface. GRID potential values change rather smoothly so that a grid step of 0.5 Å is sufficient for obtaining a reasonable approximation of the isosurface. The algorithm also computes the direction of the normal vector at each surface node. Each normal vector is smoothed by averaging it with the nearby normal vectors. Smoothing improves the accuracy of the normal vectors and, consequently, the accuracy of the curvature computation.

(ii) Nearest-Neighbors Selection. The method used for curvature calculation is summarized in Figure 4. It has been designed to fulfill three requirements: (a) curvature has to be computed from a discrete representation of the molecular surface; (b) the curvature can be estimated at different scales

according to the type of surface change looked for; (c) the method should be fast enough to be integrated into a normal GRIND calculation.

The curvature is calculated for each surface node according to the position of its nearest neighbors. The curvature value depends on the distance cutoff used to determine whether a nearest neighbor should be considered in the calculation or not. This limit value is a crucial parameter because it affects the size of the region involved in the curvature calculation. For low-cutoff values, the computed curvature describes small surface irregularities and in particular those produced by the hydrogen atoms. Conversely, for high-cutoff values, the computed curvature provides a more global description of the molecular surface: e.g., the general shape of small molecules or a protein cavity. A Euclidian distance of 6 Å was empirically chosen as a suitable cutoff limit for the analysis of small molecules.

The selection of the nearest neighbors of a particular node R can be compared to expanding a net on the surface, starting from node R and following the node connections up to the cutoff limit L (Figure 4a). In other words, the selection procedure is equivalent to choosing all the neighbors N_i within a sphere S centered on the node R and having a radius L, but only if there is a path on the molecular surface that allows the neighbor N_i to be reached without going out of the sphere S.

(iii) Curvature Calculation. Once the nearest neighbors have been selected, a partial curvature coefficient is calculated for each neighbor N_i (Figure 4b). The partial curvature coefficient Cf is calculated as indicated in eq 1, which is the application of the scalar product in an orthonormal reference frame:

$$\mathrm{Cf} = \cos(\alpha) = \frac{xx' + yy' + zz'}{\sqrt{x^2 + y^2 + z^2}\sqrt{x'^2 + y'^2 + z'^2}} \qquad (1)$$

where x, y, and z are the components of the normal vector $\vec{v}$ of the surface node R. x', y', and z' are the components of the vector $\overrightarrow{RN_i}$. α is the angle between $\vec{v}$ and $\overrightarrow{RN_i}$. If the two vectors are perpendicular, the surface between the node R and the node N_i is considered planar and the partial curvature coefficient Cf is zero. If α is lower than $\pi/2$, the surface is considered concave and Cf is positive. Conversely, if α is greater than $\pi/2$, the surface is considered convex and Cf is negative.

For every surface node, the total curvature was defined as the median of the partial curvature coefficients calculated for all its neighbors (Figure 4c). The use of the median has the advantage of being less affected than the arithmetic mean by extreme values obtained as a consequence of imperfections in the isosurfaces. Curvatures can take values between −1.0 and +1.0 (e.g., the distribution of highly convex surface node tends toward −1.0).

In highly convex patches, the value of the local curvature changes very smoothly, which normally leads to uninformative, almost flat, shape correlograms. For this reason, the shape field is mathematically transformed using a cubic function so that the rate of change of curvature is sharp enough to produce a useful MACC-2 correlogram, similar to the ones obtained using standard GRID probes.

Convex regions are more important for the description proposed than concave regions because the former regions are more prone to interact with potential binding pockets and also because they mainly reflect the structural changes in the compounds such as the introduction of substituents and the elongation of side chains. Consequently, only the negative curvature coefficients are preserved for the subsequent GRIND analysis. For consistency with the GRIND methodology, the curvature value is scaled in the same way as the normal GRID MIF and the values of the curvature coefficient are approximately normalized between 0.0 and 1.0.

Incorporation of the Shape Field into the GRIND. In a standard GRIND calculation, a filtering procedure is applied to retain only a fixed number of grid nodes n; usually n is set between 100 and 200 nodes. The filtering process is based on an optimization function designed to extract the "relevant regions" of the MIF. Up to now, no optimization algorithm has been judged necessary for the filtering of the TIP field. The filtering is just a selection of the n most convex nodes obtained after they are sorted according to its curvature value. The filtered nodes are sufficiently representative of the expected pockets of interaction, as exemplified by the thrombin ligand 4-TAPAP (Figure 5).

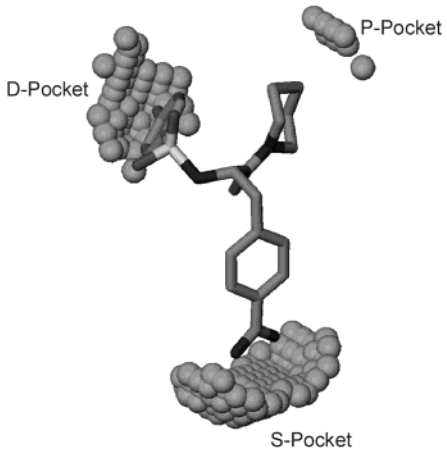


Figure 5. Filtered shape field of 4-TAPAP. Three shape patches identify the binding pockets of thrombin: the specific S pocket, the proximal P pocket, and the distal D pocket.

TIP correlograms integrated seamlessly into the GRIND methodology. The encoding of the filtered TIP fields is the same as the encoding of any filtered GRID MIF[6] (e.g., the MACC-2 transform): The products of the scaled curvature values of all the pairs of filtered nodes are calculated to generate a GRIND-like correlogram. Only the highest curvature product for each bin is stored to enable backtracking and interpretation of the information. In the same way as the original GRIND autocorrelograms, the variables included in the TIP−TIP autocorrelograms are associated with the distance between two TIP nodes in the original 3D space, and their values represent the curvature of both points. Cross-correlograms are obtained by multiplying the scaled curvature values with the scaled energy values, keeping only the highest products. Variables included in cross-correlograms are associated with the distances between a specific area of interaction (e.g., a hydrogen-bond acceptor site) and a convex part of the molecule.

Descriptors Computation. All the GRIND of this study were computed with the program Almond 3.2[31] on a Pentium IV 2.4 GHz running Linux Red Hat 7.3. Almond 3.2 has the shape field computation fully integrated as well as version 19 of the program GRID.[7] The 3D structures needed for the GRIND computations were automatically generated with the program CORINA 2.6[32] from the graph of the molecules in SDFiles or SMILES formats. The GRIND methodology is fairly robust to the small conformational inconsistencies sometimes produced by CORINA.[33] Since no major conformational difference has been observed in the 3D structures of the series studied in this work, the conformations obtained from CORINA were used directly for the calculation of the descriptors. Most of the Almond parameters were set to default values, e.g., the ALMD directive was equal to 1, the grid spacing was equal to 0.5 Å, the smoothing window of the correlograms was set to 0.8, and the size of the correlograms was automatically established by the program. The number of filtered nodes was adapted to each data set empirically from a representative compound of the data set studied. A total of 120 nodes and a weight of the field of 50% were required in order to cover all the regions of interaction of the A_1 receptor antagonists. For the plasmepsin II inhibitors, a good coverage of the regions of relevant interaction was obtained only after increasing the number of filtered nodes to 200 and reducing the weight of the field to 25% in the filtering algorithm. More

details on the Almond calculation parameters are given in the original GRIND article.[6] Regarding the TIP computation, the field of the O probe was used for the molecular surface computation and the distance cutoff used to estimate the curvature was set to 6 Å.

Chemometric Analysis. Almond 3.2 implements the chemometric methods used for the analysis of the GRIND generated in this article. PCA[8] is used to analyze the similarity of the compounds, while PLS regression analysis[9] is used to build models describing the relationships between the GRIND and experimental variables describing biological properties of the compounds. Both PCA and PLS are projection methods in which the original, highly dimensional descriptor space is projected onto a new low-dimension orthogonal space, where the axes are obtained as a linear combination of the original descriptors. In PCA, the new axes are called principal components (PC) and are chosen on the basis of their ability to retain at best the variance of the descriptor matrix (**X**). In PLS, the new axes are called latent variables (LV), and the criterion for their selection includes their ability not only to explain at best the variance of the descriptors but also to maximize their correlation with their biological properties. The PLS regression analysis can be used even when the number of descriptors is much higher than the number of compounds, and such descriptors are highly correlated, as it is normally the case for a GRIND model.

Results and Discussion

3D-QSAR of A_1 Adenosine Receptor Antagonists. This data set is an example of a series in which standard GRIND fail to explain correctly the activity of some compounds because of a lack of shape description. The series contains xanthine derivatives synthesized and tested by Strappaghetti et al.[24] Xanthines represent the largest family of adenosine receptors antagonists.[34] Adenosine receptors are membrane proteins that are activated by the nucleoside adenosine and belong to the G-protein-coupled receptors superfamily. Four subtypes of adenosine receptors have been cloned and studied so far: A_1, A_{2A}, A_{2B}, and A_3,[35] which are involved in different regulatory actions in the cardiovascular, renal, and central nervous systems.[34,36] In this example the authors were interested in finding antagonists of higher affinity and selectivity toward the A_1 subtype, since such antagonists are therapeutically interesting in the treatment of cognitive diseases,[37] renal failure,[38] or postinfarct treatment.[38]

Structures and affinities of the A_1 adenosine receptor antagonists are presented in Table 1. The compounds show three types of chemical modifications: the length of the aliphatic chain linking the pyridazinone ring and the xanthine scaffold (from 1 to 4), the substituent R in the 1-position of the xanthine (methyl or *n*-propyl), and the substituent R′ in the 6-position of the pyridazinone ring (chlorine, phenyl, or hydrogen). The 3D geometries for the compounds of the series were obtained automatically from their 2D structures using the program CORINA, as described in Materials and Methods. This method produces reasonable extended conformations for all the compounds, which is appropriate for the QSAR analysis of the congeneric series, as the present one.

Initially, classic GRIND descriptors were computed using the standard probes DRY (hydrophobic), O (hydrogen-bond acceptor), and N1 (hydrogen-bond donor). All the computational parameters were set to default values with the exception of the number of filtered nodes, which was fine-tuned to 120. In principle, these three probes should be enough to describe all the structural differences of the compounds of the series, but the results showed that this was not the case. Figure 6, depicting the filtered DRY field of compounds **8** and **18**, shows virtually no difference between the MIF of the two compounds. All the MIF selected nodes are concentrated around the heterocyclic rings, and the DRY probe produced no region around the *n*-propyl group, which therefore was not considered in the description. To confirm the ability of the GRIND to highlight the structural dissimilarities between the compounds in this series, a PCA was performed on the GRIND matrix. The two first principal components (PC) explain 62% of the

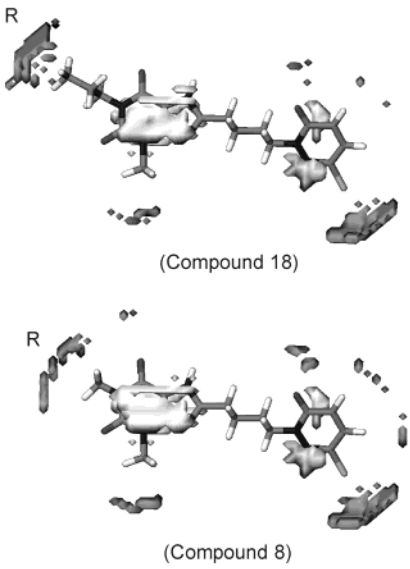
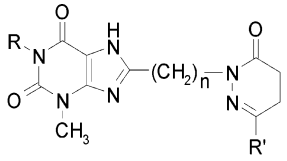


Figure 6. Comparison of DRY (light-gray) and TIP (dark-gray) filtered fields of 8-substituted xanthines **18** and **8**. The structures differ only in the length of the alkyl substituent at the 1-position of the theophylline nucleus (R substituent). Filtered DRY fields are situated at the same location for both structures, e.g., at each side of the conjugated rings. Filtered TIP fields are equivalent apart from the "R" patch, which is displaced according to the size of the R group.

Table 1. Structure and Affinity of 8-Substituted Xanthines as Antagonists of the A_1 Adenosine Receptor



compd	R	*n*	R′	$K_i(A_1)$ (μM)
5	CH_3	1	Cl	>100
6	CH_3	2	Cl	14.7
7	CH_3	3	Cl	8.8
8	CH_3	4	Cl	0.37
9	CH_3	1	Ph	53.8
10	CH_3	2	Ph	6.1
11	CH_3	3	Ph	10.8
12	CH_3	4	Ph	2.12
13	CH_3	2	H	12.8
14	CH_3	3	H	12.7
15	C_3H_7	1	Cl	9.2
16	C_3H_7	2	Cl	0.47
17	C_3H_7	3	Cl	1.2
18	C_3H_7	4	Cl	0.19
19	C_3H_7	1	Ph	2.27
20	C_3H_7	2	Ph	0.76
21	C_3H_7	3	Ph	0.81
22	C_3H_7	4	Ph	0.38

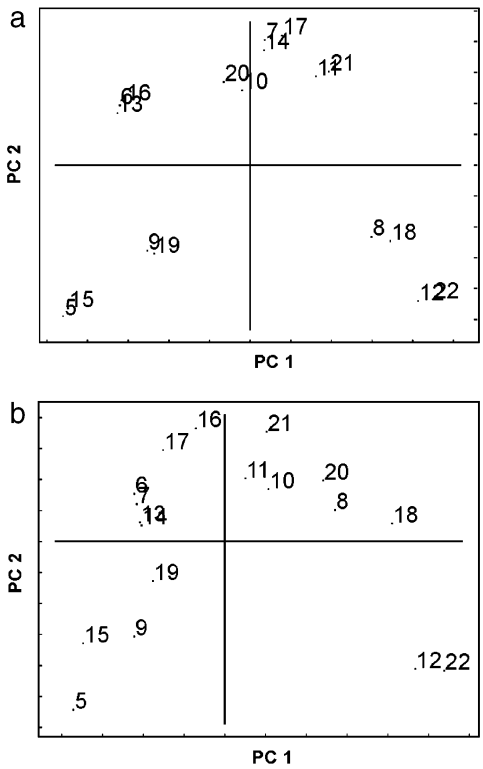


Figure 7. PCA score plots of the A_1 receptor antagonists: (a) DRY model; (b) TIP model.

variance of the original descriptors. The main factors revealed by the PC are the number of carbon atoms in the linker and the presence of a phenyl group at the R′ position (Figure 7a). As suspected, compounds with different R substituent (methyl or propyl) are tightly grouped in clusters of two or three members and compounds **8** and **18** shown in Figure 7 are closely clustered. Even worse, some compounds with very different activity are grouped together (e.g., **21** with **11** or **16** with **6**), which indicates the inability of the original descriptors to discriminate compounds with structural properties that affect their biological behavior.

The GRIND computation was repeated with the same parameters, but this time the DRY probe was replaced by the molecular shape field. The first two PC explain 47% of the variance of the original descriptors. The compounds were spread in the PCA score plot in a similar way (Figure 7b) according to the size of the linker and the presence or absence of a phenyl group in position R′, but interestingly, the clustering scheme has changed. Some low-activity compounds are now clustered (e.g., **6** and **7** or **13** and **14**) as well as some with high activity (e.g., **16** and **17**).

Regarding the PLS regressions, the use of the shape field improves dramatically the quality of the models: the best model obtained using the shape field requires 2 LV ($r^2 = 0.96$, $q^2_{(LOO)} = 0.85$). For the same number of LV, the model using the DRY probe has a rather poor predictive ability ($r^2 = 0.72$, $q^2_{(LOO)} = 0.11$) and in fact the value of q^2 only improves after the incorporation of 6 LV to the model ($r^2 = 0.98$, $q^2_{(LOO)} = 0.83$).

These results clearly show that the new shape field improves the description of the structural differences between the compounds in a way that leads to better structure−activity models. The main reason for such an improvement is that the new shape field is able to describe the different R substitution due to the changes that they produced in the convex patches (Figure 6), while the differences in the R substituent are not described at all by the DRY probe.

The fact that the classic GRIND descriptors produce a PLS model of reasonably good quality after incorporating 6 LV requires further clarification, especially because this fact illustrates an indirect representation of the shape often observed in classic GRIND models. The presence of a propyl substituent is, in fact, described by the classic GRIND but not by the descriptors associated with the DRY probe. Compounds with this substituent have slightly more favorable interaction energies in the nodes representing the interaction of the carbonyl groups with the N1 probe, probably due to van der Waals interactions between the propyl substituent and the probe. This small change in the energy potential has a minor effect on the values of a few GRIND variables that is only recognized by the PLS model when enough LV have been incorporated. However, the PLS model using the DRY probes is rather complex (6 LV) and the aforementioned indirect effect can be understood only after a careful and time-consuming analysis. It is clear that the shape field both simplifies the model and clarifies its interpretation.

With respect to the model interpretation, the most important variable in the new model is a distance in the molecular shape autocorrelogram (TIP−TIP 38 Å) between two shape patches situated at the two extremities of the compounds: the propyl substituent at one side and the pyridazinone ring at the other. Since the PLS weight of this variable is positive, it can be expected that the propyl group in N1 has a favorable effect on the compound activity. In fact, this observation is not new and many xanthine antagonists of the A_1 adenosine receptor have a propyl substituent at this position (e.g., DPCPX, KW 3902, midaxifylline). Interestingly, such a TIP−TIP variable suggests a concerted effect between the R group and the linker size, which are the two principal structural features that determine the value of this variable. From an examination of the structure−activity table (Table 1), it becomes clear that if the R group is a propyl, the linker can be short (e.g., compound **16**, $K_i = 0.47\ \mu$M), while if the R group is a methyl, the linker must be long enough to preserve a submicromolar affinity (e.g., compound **8**, $K_i = 0.37\ \mu$M) and probably to let the pyridazinone ring reach a site of favorable interactions. Such an effect is generally difficult to detect because it is not additive but complementary. Complementary effects are particularly interesting because they offer several options for ligand design. For example, in this particular series, if a compound with a small R group and a large linker does not have the desired pharmacological profile, it is still possible to try an alternative compound, e.g., a compound with a large R group and a short linker. The fact that the shape descriptors are able to detect such effects further supports the use of the GRIND-shape methodology.

3D-QSAR of *Plasmodium falciparum* Plasmepsin II Inhibitors. In the previous data set, the shape field was used to detect favorable interactions between the ligands and its receptor. Conversely, the main

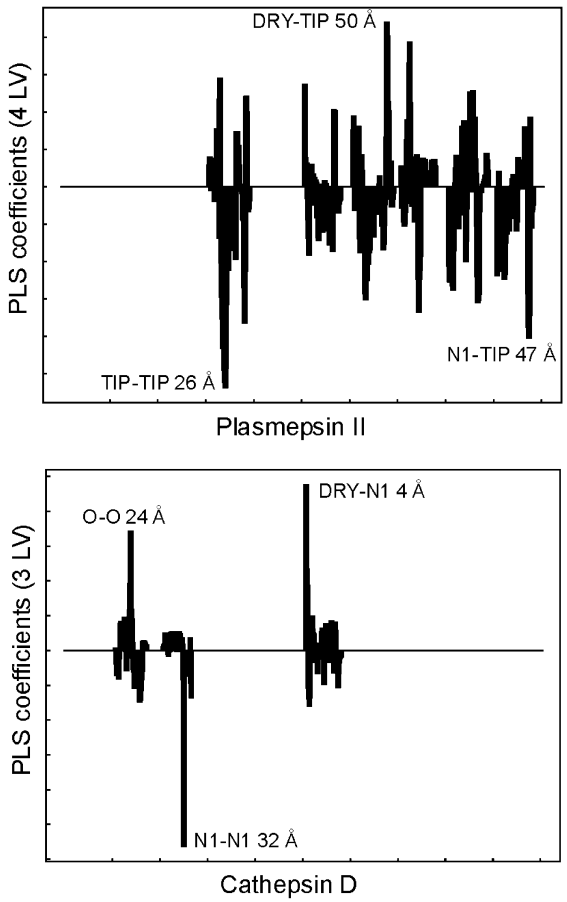


Figure 8. PLS coefficient profile of the plasmepsin II and cathepsin D models.

objective of this data set is to show how unfavorable interactions can be detected.

Plasmodium falciparum is the most lethal agent of malaria, an infectious disease that causes millions of deaths every year.[39] The rise of strains of *Plasmodium falciparum* resistant to conventional antimalarials such as chloroquine has reopened the need for new antimalarial drugs. *Plasmodium falciparum* aspartyl proteases of the plasmepsin family have emerged as promising targets for the development of new drugs.[40] These proteases are required by the parasite for the metabolism of hemoglobin, its principal source of nutrients.[40] In 1998, Carroll et al. published the first combinatorial library designed toward the inhibition of plasmepsin II,[25] which is one of the four plasmepsins localized in the food vacuole of the parasite.[41] The inhibitors of the plasmepsin II protease contained a statine core (Table 2), a common structural scaffold of aspartyl protease inhibitors. Carroll's library contained a total of 13 020 members that were screened for plasmepsin II and cathepsin D inhibition. Among the hit compounds identified from the screening, a few representatives (16 compounds) were selected for resynthesis and quantitative inhibition measurement (Table 2). The study resulted in the identification of compounds inhibiting the growth of the parasite in cell culture and more than 10-fold selectivity with respect to cathepsin D, a lysosomal aspartyl protease that is the most homologous human enzyme to plasmepsin II[25] (35% overall homology). Cathepsin D is present in large amounts in most cell types, and knockout mice lacking functional cathepsin D have only a 3-week life span.[40] Therefore, it is highly desirable to design inhibitors showing high affinity for plasmepsin II and no affinity for cathepsin D.

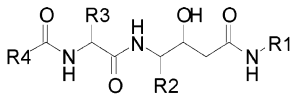
In this example, we intend to produce PLS models for the series represented in Table 2 for both plasmepsin II and cathepsin D activities and to use them for the prediction of the activities of all the hits from the screening. The shape probe is tested on both models, and its relevance is assessed. As in the previous example, the 3D structures of the compounds were obtained by automatic conversion from their 2D structures, which produces extended conformations of reasonably low energy. It should be emphasized that these structures, although suitable for 3D QSAR analyses, cannot be claimed to represent accurately their bioactive conformations. This fact limits the potential structural interpretation of the results in terms of the direct mapping of distances highlighted by the model to distances between receptor residues, but this is a generic limitation of 3D QSAR that cannot be overcome without additional structural information from external sources.

A first 3D-QSAR model for plasmepsin II inhibitors was attempted using only the standard GRIND probes DRY (hydrophobic), O (hydrogen-bond acceptor), and N1 (hydrogen-bond donor). Most of the parameters required for the calculation of the descriptors were set to their default values. Only the filtering parameters were adapted to the compounds of the series: after a few tests, 200 filtered nodes and a field weight of 25% produced a satisfying coverage of the relevant regions of interaction. No relevant structure−activity relationship was obtained after PLS regression ($q^2_{(LOO)} < 0$, models go to 5 LV).

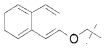
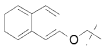
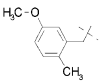
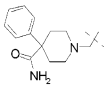
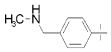
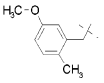
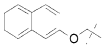
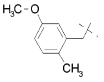
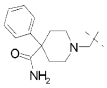
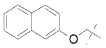
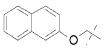
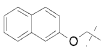
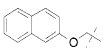
A second PLS model was attempted, this time adding the shape field to the initial set of standard probes (the DRY probe was also retained) and with the same parameters for the descriptor calculations. The predictive ability of the model improved dramatically, and the $q^2_{(LOO)}$ jumped to 0.35, with 4 LV. Auto- and cross-correlograms showing little contribution to the model after the visual inspection of the PLS coefficient profiles were removed (e.g., DRY−DRY, O−O, N1−N1, and DRY−O), and a new PLS regression was computed, leading to an acceptable QSAR model with 4 LV, an r^2 of 0.95 and a $q^2_{(LOO)}$ of 0.53. The PLS coefficient profile of a 4 LV model is shown in Figure 8. Many variables are important for the regression model, as indicated by the number of peaks present along the coefficient profile. The interpretation of all of them is out of the scope of this study, which will be limited to the explanation of the structural features pointed out by the most important variables involving the shape field (e.g., TIP−TIP 26 Å, N1−TIP 47 Å, and DRY−TIP 50 Å, shown in Figure 9).

The variable with the most negative weight on the model is TIP−TIP 26 Å. It has a particularly high value for the inactive compound PS 731167. Backtracking of the variable in 3D space highlights two shape patches of PS 731167, one at the tip of the butyl R1 substituent and the other at the tip of the glutamine R3 substituent (Figure 9a). The glutamine side chain protrudes much more from the scaffold than the side chain of isoleucine, which is the amino acid at the R3 position of the most active plasmepsin II inhibitors. This is an indication of a potential steric hindrance at the R3 position.

Table 2. Structure and Affinities of the Statine Inhibitors



P3 P2 P1 P1' P2'

Compound	R1	R2	R3	R4	Plas II K_i (nM)	Cat D K_i (nM)
PS 273800	C_4H_9	CH_2Ph	$CH(Me)C_2H_5$		50	320
PS 707194	C_4H_9	$CH_2CH(Me)_2$	$CH(Me)C_2H_5$		90	50
PS 189863	C_4H_9	$CH_2CH(Me)_2$	$CH(Me)C_2H_5$		110	500
PS 777621	C_4H_9	CH_2Ph	$CH(Me)C_2H_5$		180	560
PS 444035	C_4H_9	$CH_2CH(Me)_2$	$CH(Me)C_2H_5$		220	1200
PS 662477	C_4H_9	$CH_2CH(Me)_2$	$CH(Me)C_2H_5$		220	2400
PS 429694	C_4H_9	$CH_2CH(Me)_2$	$CH(Me)C_2H_5$		410	5500
PS 222036	C_3H_6Ph	$CH_2CH(Me)_2$	$CH(Me)C_2H_5$		440	1300
PS 725074	C_4H_9	$CH_2CH(Me)_2$	$CH(Me)_2$		500	1600
PS 361691	C_4H_9	$CH_2CH(Me)_2$	Ph		560	200
PS 154636	C_4H_9	$CH_2CH(Me)_2$	Ph		590	2100
PS 749213	C_3H_6Ph	$CH_2CH(Me)_2$	$CH(Me)C_2H_5$		600	3900
PS 679304	C_4H_9	CH_2Ph	$CH_2CH(Me)_2$		5600	920
PS 699506	C_4H_9	CH_2Ph	Ph		5800	110
PS 290351	C_4H_9	CH_2Ph	CH_2Ph		12000	1300
PS 731167	C_4H_9	$CH_2CH(Me)_2$	$C_2H_4CONH_2$		>30000	200

The second variable with the most negative weight is N1−TIP 47 Å. In the inactive compounds, such as PS 731167, this variable is associated with a particular distance between the extremity of the R4 group and the carbonyl moiety at residue P1 (Figure 9b). The variable indicates that the R4 substituent 2-naphthoxy probably suffers from some unfavorable steric hindrance with the binding site. However, this negative feature seems to be counterbalanced by favorable hydrophobic interactions as exemplified by the variable DRY−TIP 50 Å of the plasmepsin II inhibitor PS 189863 (Figure 9c).

In the same way as in the plasmepsin II model, a

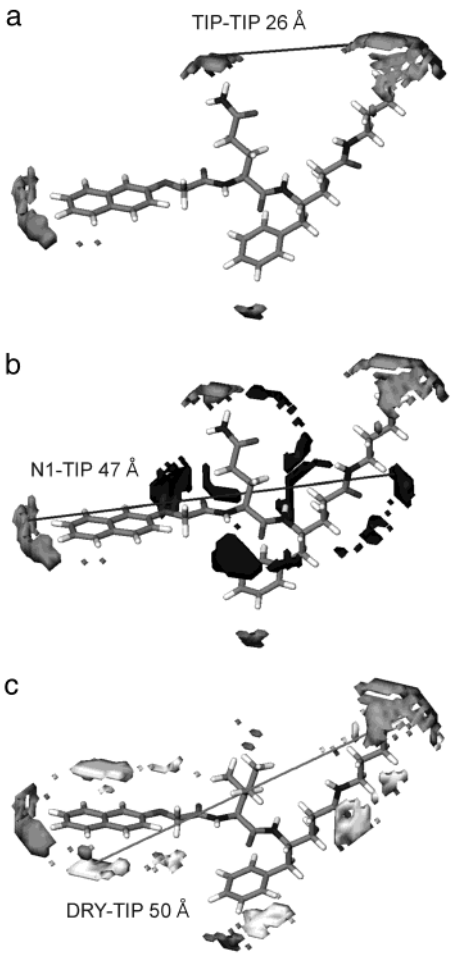


Figure 9. Backtracking of the important GRIND-shape variables of the plasmepsin II model: (light-gray) DRY field; (dark-gray) TIP field; (black) N1 field; (a) TIP−TIP 26 Å; (b) N1−TIP 47 Å; (c) DRY−TIP 50 Å.

model for the cathepsin D activities was generated with the three standard GRIND probes (DRY, O, N1) computed with the same set of parameters. The model had better statistical parameter values than the initial plasmepsin II model (1 LV, $r^2 = 0.55$, $q^2_{(LOO)} = 0.38$) and no relevant improvement was obtained after the incorporation of the shape field (2 LV, $r^2 = 0.7$, $q^2_{(LOO)} = 0.38$), which indicates that the differences of cathepsin D inhibition of the compounds of this series are mainly due to structural features other than shape complementarity. Interestingly, the inactive compounds for plasmepsin II PS731167 (inhibition greater than 30 μM) is one of the most active inhibitors of cathepsin D (K_i = 200 nM) in the data set. According to the plasmepsin II model, there must be some differences of shape in the binding site of the two proteins so that cathepsin D can accommodate the glutamine side chain while plasmepsin II cannot.

The X-ray structures of plasmepsin II (PDB entry 1M43) and cathepsin D (PDB entry 1LYB) have been published in complex with pepstatine A, a naturally occurring aspartyl protease inhibitor.[42,43] Since pepstatine A has the same residues at P2, P1, and P1′ as the plasmepsin inhibitor PS 725074, we have compared the binding sites of the two enzymes at S2, which is the subsite of the R3 substituent. Interestingly, the tip of the β-hairpin flap that covers the binding site is situated just over the P2 valine residue of pepstatine A (Figure

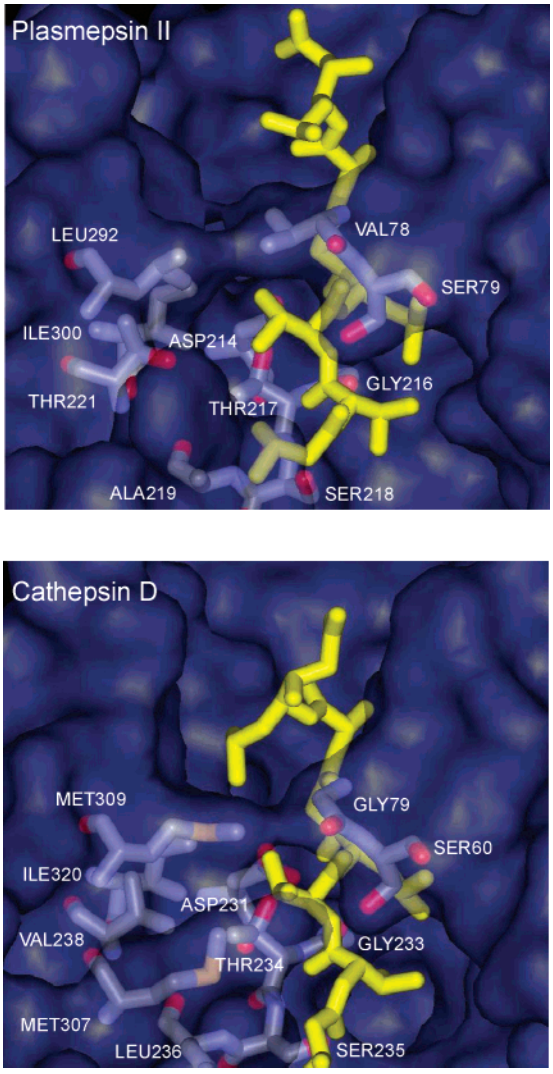


Figure 10. Binding site of plasmepsin II and cathepsin D complexed with pepstatine A (yellow). Only the residues situated at less than 8 Å from the Cβ of P2 valine are drawn. The binding site surface is the solvent excluded surface calculated with a probe radius of 1.4 Å.

10). The flap tip contains a valine residue (VAL78) in plasmepsin II instead of a glycine residue (GLY79) in cathepsin D. The space left by the glycine residue of cathepsin D is partially covered by a compensating mutation on the other side of the binding site: the leucine residue (LEU292) of plasmepsin II becomes a methionine residue (MET309) in cathepsin D. Although the P2 valine residue of pepstatine A is buried in both proteins (Figure 10), GLY79 confers greater flexibility to the β-hairpin flap than VAL78, and the side chain of MET309 has one more degree of freedom than the side chain of LEU292. Thus, it is highly probable that cathepsin D is better able to accommodate bulky amino acids of statine inhibitors at S2 than plasmepsin II. These observations are in agreement with the plasmepsin II PLS model and justify the absence of improvement after the inclusion of the shape field in the cathepsin D model.

The initial cathepsin D model was refined by keeping only the correlograms with a real contribution to the model (e.g., O−O, N1−N1, and DRY−N1), which led to a final model with 3 LV, an r^2 of 0.92, and a $q^2_{(LOO)}$ of 0.58. The PLS coefficients profile for a model with 3 LV

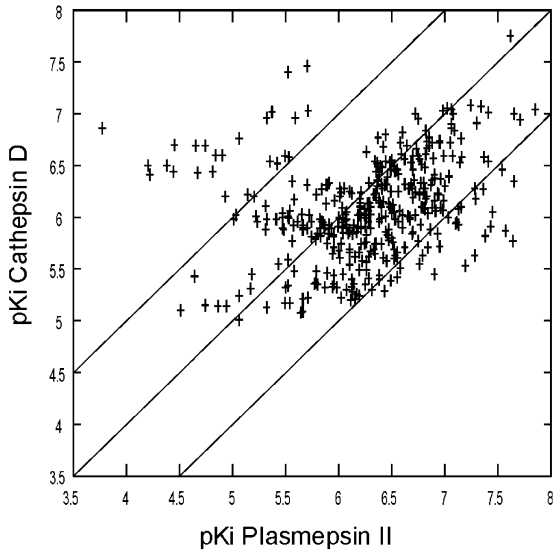


Figure 11. Predictions of the plasmepsin II and cathepsin D activities of the "hit-rich" statine library subset. Compounds between the two external lines are predicted to be less than 10-fold selective.

is shown in Figure 8. The model is much simpler than the plasmepsin model, with only three important peaks along the profile: two positives at O−O 24 Å and DRY−N1 4 Å and one negative at N1−N1 32 Å. All the variables highlight structural features of the R4 substituents. N1−N1 32 Å indicates where the presence of a hydrogen-bond acceptor atom is unfavorable for activity, for example, in compound PS 429694. The two other variables identify positive features of the 2-naphthyl moiety at R4 as in the most active compound PS 707194: O−O 24 Å indicates that the amide nitrogen at P2 must be fully accessible to the O probe, and DRY−N1 4 Å refers to favorable hydrophobic interactions next to the hydrogen-bond donor site of the P3 carbonyl group.

The GRIND computation has the advantage of being automatic and fairly fast and therefore is also of potential use for in silico screening. To give an idea of calculation speeds, the descriptors for the 13 020 compounds of the statine library[25] were computed in less than 36 h on a Pentium IV running at 2.4 GHz. One of the potential uses for the models obtained in the present work could be the prediction of the binding affinity of the hits obtained from the screening,[25] with the aim of selecting a subset of the most promising candidates for resynthesis and biological evaluation. Unfortunately, the original full list of hits from the screening is not available, but it is possible to build a hit-rich list of compounds using the hit frequencies published by Carroll et al.[25] simply by removing from the whole list the compounds that have an R group with a zero hit frequency for both plasmepsin II and cathepsin D. This produced a list of 329 compounds. The plasmepsin II and cathepsin D model were used to predict the pK_i for all the compounds of this subset, which are represented in the scatter plot shown in Figure 11.

Compounds falling near the diagonal region in this plot show little selectivity (less than 10-fold) for any receptor and are of no interest. Compounds on the left upper corner have more than 10-fold selectivity toward cathepsin D over plasmepsin II, while those on the right bottom corner have more than 10-fold selectivity toward plasmepsin II over cathepsin D and obviously are the most interesting compounds whose synthesis should be prioritized. The selectivity plot shows the importance of considering the predicted activities of plasmepsin II and cathepsin D inhibitors at the same time, since most of the compounds with a predicted submicromolar affinity for plasmepsin II have a similar order of predicted affinity for cathepsin D. The selectivity prediction considerably reduces the number of compounds of interest for further experiments.

Conclusions

We consider that the "molecular shape field" described in this article is an interesting enhancement of the original GRIND descriptors that overcomes one of the main drawbacks of the method. This field is not intended to provide an exhaustive description of the molecular shape but simply to give information about some highly relevant shape characteristics, like the overall size and spatial extents of the molecule. The fact that the GRIND method uses this molecular shape field to build cross-correlograms further enriches this information by describing the geometrical relationships between these spatial extents and other physicochemical features of the compounds. As required, the shape description is simple and fast to compute and can be seamlessly integrated within the GRIND methodology. Moreover, the resulting variables are simple to interpret.

The usefulness of the molecular shape field has been illustrated in the QSAR models presented in this study. In both of them, the consideration of the shape dramatically improved the model. The first one concerns antagonists of the A_1 receptor and illustrates how an improved receptor ligand shape matching can be detected. The second one describes inhibitors of the plasmepsin II aspartyl protease of *Plasmodium falciparum* and shows how unfavorable steric interactions can be identified. In both cases, the new shape field seems to be particularly useful in improving the description of aliphatic moieties that are not well described by the GRID hydrophobic field.

Acknowledgment. This work has been supported by a predoctoral grant from Multivariate Infometric Analysis (MIA) S.r.l. and additional funding from the "Fundació La Marató TV3". The authors also thank Hugo Gutiérrez-de-Terán, Ismael Zamora, and David Cosgrove for their useful advice and discussions.

References

(1) Hansch, C. A quantitative approach to biochemical structure−activity relationships. *Acc. Chem. Res.* **1969**, *2*, 232−239.
(2) Cramer, R. D.; 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) 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.
(5) Cramer, R. D. Topomer CoMFA: a design methodology for rapid lead optimization. *J. Med. Chem.* **2003**, *46*, 374−388.

(6) Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. GRid-INdependent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors. *J. Med. Chem.* **2000**, *43*, 3233−3243.
(7) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849−857.
(8) Carey, R. N.; Wold, S.; Westgard, J. O. Principal component analysis: an alternative to "referee" methods in method comparison studies. *Anal. Chem.* **1975**, *47*, 1824−1829.
(9) Hoskuldsson, A. PLS regression methods. *J Chemom.* **1988**, *2*, 211−228.
(10) Benedetti, P.; Mannhold, R.; Cruciani, G.; Pastor, M. GBR compounds and mepyramines as cocaine abuse therapeutics: chemometric studies on selectivity using grid independent descriptors (GRIND). *J. Med. Chem.* **2002**, *45*, 1577−1584.
(11) Afzelius, L.; Masimirembwa, C. M.; Karlen, A.; Andersson, T. B.; Zamora, I. Discriminant and quantitative PLS analysis of competitive CYP2C9 inhibitors versus non-inhibitors using alignment independent GRIND descriptors. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 443−458.
(12) Brea, J.; Masaguer, C. F.; Villazon, M.; Cadavid, M. I.; Ravina, E.; et al. Conformationally constrained butyrophenones as new pharmacological tools to study 5-HT 2A and 5-HT 2C receptor behaviours. *Eur. J. Med. Chem.* **2003**, *38*, 433−440.
(13) Cruciani, G.; Pastor, M.; Mannhold, R. Suitability of molecular descriptors for database mining. A comparative analysis. *J. Med. Chem.* **2002**, *45*, 2685−2694.
(14) Fontaine, F.; Pastor, M.; Gutierrez-De-Teran, H.; Lozano, J. J.; Sanz, F. Use of alignment-free molecular descriptors in diversity analysis and optimal sampling of molecular libraries. *Mol. Diversity* **2003**, *6*, 135−147.
(15) Ajay; Murcko, M. A. Computational methods to predict binding free energy in ligand−receptor complexes. *J. Med. Chem.* **1995**, *38*, 4953−4967.
(16) Hall, L. H.; Kier, L. B. The molecular connectivity Chi indexes and Kappa shape indexes in structure−property modeling. *Reviews in Computational Chemistry*; VCH Publishers: New York, 1991; pp 367−422.
(17) Lemmen, C.; Lengauer, T. Computational methods for the structural alignment of molecules. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 215−232.
(18) Good, A. C.; Kuntz, I. D. Investigating the extension of pairwise distance pharmacophore measures to triplet-based descriptors. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 373−379.
(19) Nilakantan, R.; Bauman, N.; Venkataraghavan, R. New method for rapid characterization of molecular shapes: applications in drug design. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 79−85.
(20) Kotani, T.; Higashiura, K. Rapid evaluation of molecular shape similarity index using pairwise calculation of the nearest atomic distances. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 58−63.
(21) Cosgrove, D. A.; Bayada, D. M.; Johnson, A. P. A novel method of aligning molecules by local surface shape similarity. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 573−591.
(22) Kearsley, S. An alternative method for the alignment of molecular structures: Maximizing electrostatic and steric overlap. *Tetrahedron Comput. Methodol.* **1990**, *3*, 615−633.
(23) Waller, C. L.; Oprea, T. I.; Chae, K.; Park, H.-K.; Korach, K. S.; et al. Ligand-based identification of environmental estrogens. *Chem. Res. Toxicol.* **1996**, *9*, 1240−1248.
(24) Strappaghetti, G.; Corsano, S.; Barbaro, R.; Giannaccini, G.; Betti, L. Structure−activity relationships in a series of 8-substituted xanthines as A1-adenosine receptor antagonists. *Bioorg. Med. Chem.* **2001**, *9*, 575−583.
(25) Carroll, C. D.; Patel, H.; Johnson, T. O.; Guo, T.; Orlowski, M.; et al. Identification of potent inhibitors of *Plasmodium falciparum* plasmepsin II from an encoded statine combinatorial library. *Bioorg. Med. Chem. Lett.* **1998**, *8*, 2315−2320.
(26) Lee, B.; Richards, F. M. The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.* **1971**, *55*, 379−400.
(27) Greer, J.; Bush, B. L. Macromolecular shape and surface maps by solvent exclusion. *Proc. Natl. Acad. Sci. U.S.A.* **1978**, *75*, 303−307.
(28) Richards, F. M. Areas, volumes, packing and protein structure. *Annu. Rev. Biophys. Bioeng.* **1977**, *6*, 151−176.
(29) Connolly, M. L. Analytical molecular surface calculation. *J. Appl. Crystallogr.* **1983**, *16*, 548−558.
(30) Walker, P. D.; Arteca, G. A.; Mezey, P. G. A complete shape characterization for molecular charge densities represented by Gaussian-type functions. *J. Comput. Chem.* **1991**, *12*, 220−230.
(31) Cruciani, G.; Fontaine, F.; Pastor, M. *Almond*, 3.2.0; Molecular Discovery Ltd.: Perugia, Italy.
(32) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic generation of 3D atomic coordinates for organic molecules. *Tetrahedron Comput. Methodol.* **1990**, *3*, 537−547.
(33) Fontaine, F.; Pastor, M.; Sanz, F. Potential usefulness of the GRIND descriptors for obtaining 3D-QSAR models without supervision. Presented at the XIIth National Congress of the Spanish Society of Medicinal Chemistry, Sevilla, Spain, 2001; Poster.
(34) Müller, C.; Stein, B. Adenosine receptor antagonists: Structures and potential therapeutic applications. *Curr. Pharm. Des.* **1996**, *2*, 501−530.
(35) Fredholm, B. B.; Ijzerman, A. P.; Jacobson, K. A.; Klotz, K. N.; Linden, J. International Union of Pharmacology. XXV. Nomenclature and classification of adenosine receptors. *Pharmacol. Rev.* **2001**, *53*, 527−552.
(36) Poulsen, S. A.; Quinn, R. J. Adenosine receptors: new opportunities for future drugs. *Bioorg. Med. Chem.* **1998**, *6*, 619−641.
(37) Haas, H. L.; Selbach, O. Functions of neuronal adenosine receptors. *Naunyn-Schmiedeberg's Arch Pharmacol.* **2000**, *362*, 375−381.
(38) Suzuki, F.; Shimada, J.; Mizumoto, H.; Karasawa, A.; Kubo, K.; et al. Adenosine A1 antagonists. 2. Structure−activity relationships on diuretic activities and protective effects against acute renal failure. *J. Med. Chem.* **1992**, *35*, 3066−3075.
(39) Nezami, A.; Luque, I.; Kimura, T.; Kiso, Y.; Freire, E. Identification and characterization of allophenylnorstatine-based inhibitors of plasmepsin II, an antimalarial target. *Biochemistry* **2002**, *41*, 2273−2280.
(40) Coombs, G. H.; Goldberg, D. E.; Klemba, M.; Berry, C.; Kay, J.; et al. Aspartic proteases of *Plasmodium falciparum* and other parasitic protozoa as drug targets. *Trends Parasitol.* **2001**, *17*, 532−537.
(41) Nezami, A.; Kimura, T.; Hidaka, K.; Kiso, A.; Liu, J.; et al. High-affinity inhibition of a family of Plasmodium falciparum proteases by a designed adaptive inhibitor. *Biochemistry* **2003**, *42*, 8459−8464.
(42) Silva, A. M.; Lee, A. Y.; Gulnik, S. V.; Maier, P.; Collins, J.; et al. Structure and inhibition of plasmepsin II, a hemoglobin-degrading enzyme from *Plasmodium falciparum*. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 10034−10039.
(43) Baldwin, E. T.; Bhat, T. N.; Gulnik, S.; Hosur, M. V.; Sowder, R. C., 2nd; et al. Crystal structures of native and inhibited forms of human cathepsin D: implications for lysosomal targeting and drug design. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 6796−6800.

JM0311240