Influence of Conformation on GRIND-Based Three-Dimensional Quantitative Structure−Activity Relationship (3D-QSAR)

Giulia Caron and Giuseppe Ermondi*

CASMedChem Laboratory, Dipartimento di Scienza e Tecnologia del Farmaco, Università di Torino, Via P. Giuria, 9, 10125 Torino, Italy

Received April 19, 2007

To address the dependence of GRIND-based 3D-QSAR on data set flexibility, we investigate a series of oxidosqualene cyclase (OSC) inhibitors. The results indicate that statistical models are determined independently of the data set but that despite identification of the same outliers and the acceptable test set prediction, not all models show good predictive correlation coefficient (q^2). Moreover, the best model was obtained using a data set of the lowest energy conformers generated by a conformational analysis.

Introduction

The grid-independent descriptors (GRIND[a]) were published in the year 2000 and implemented in the ALMOND software.[1] Since then, GRIND-based 3D-QSAR models have been successfully used to describe a number of biological topics.[2] The appeal of these descriptors is due to three main characteristics: GRIND are alignment-independent, chemically interpretable, and easy and quick to compute. Moreover, the whole procedure can be fully automated. One of the greatest criticisms of the standard GRIND methodology is that it takes no account of ligand flexibility, and thus, we selected a series of non-terpenoid oxidosqualene cyclase (OSC, interesting target[3,4] in the search for drugs that reduce plasma cholesterol levels in man) inhibitors reported by Lenhart et al.[3] and Dehmlow et al.[4] to quantitatively evaluate the influence of flexibility on GRIND-based 3D-QSAR models. The Hoffman-La Roche series is a difficult data set; it is very flexible, and many small changes can affect activities but not always consistently.

The first step of this study was to convert SMILES codes into 3D structures. This was first done by using CORINA[5] and Omega[6] to compare a 3D structure generation without conformational analysis (CORINA) with a method based on conformational analysis (Omega). Then, to evaluate the influence of nitrogen chirality on the final results, we also produced additional data sets of conformers obtained by manual modification of stereochemistry. A final data set was obtained by manually modifying ligand conformations.

Finally, the six data sets of conformers were submitted to the ALMOND package. Statistical and graphical results indicate that the fundamental information about OSC binding can be captured by using any data set but that a complete understanding of the pattern of interactions and the precise prediction of activity data require the use of the conformers with the lowest energy resulting from a conformational study. Since one of the main advantages of the GRIND-based 3D-QSAR approach lies in its speed and high level of automation, we suggest combining ALMOND with Omega, one of the most suitable tools for conformational analysis because it achieves an excellent balance between speed and performance.

* To whom correspondence should be addressed. Phone: +39 011 6707282. Fax +39 011 6707687. E-mail: giuseppe.emondi@unito.it.

[a] Abbreviations: OSC, oxidosqualene cyclase; GRIND, grid-independent descriptors; SMILES, simplified molecular input line entry system; PLS, partial least squares; LV, latent variables; VIP, variable importance in the projection.

Results and Discussion

The Data Set Investigated. All compounds (the chemical structures are reported in Supporting Information) belonging to the data set contain three distinct domains indicated as A, B, and C (Table 1). Of particular interest for this study is domain B, a spacer that connects domain A to domain C, since it is highly flexible with a number of single rotatable bonds ranging from 5 (for (2-methylcyclopropyl)methyloxy and but-2-enyloxy chains) to 8 (for the hexyloxy chain). Compounds **17**, **22**, **31**, and **42** were excluded from this study because they are structurally very different from the others and are not very active. The pIC_{50} ($-\log IC_{50}$) values of the OSC inhibitors (Table 1) range between 8.72 for **1**, the most potent, and 5.72 for **44**, the least potent. The error in pIC_{50} value for most OSC inhibitors is about ± 0.07 units.[3,4]

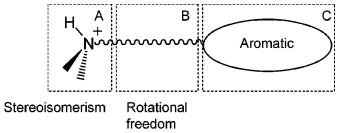
Data Set Preparation. Data set 1 (named DB01) was obtained by submitting the SMILES codes to CORINA[5] without any human contribution. Data set 4 (DB04) was prepared by submitting the SMILES codes to Omega[6] and then by selecting for any compound the conformer with the lowest energy.

All compounds of the OSC inhibitors investigated bear a protonated nitrogen that is chiral (four different substituents) and thus can assume either *R* or *S* configuration. GRIND are sensitive to this kind of chirality because they produce different internal geometries.[1] Examination of DB01 and DB04 (obtained without any chiral specification) revealed that for some compounds Omega generates the *R* configuration whereas CORINA gives the *S* configuration. This difference is due to the fact that Omega and CORINA use their own implemented rules to assign nitrogen stereochemistry. To check for the relevance (in ALMOND models) of nitrogen chirality, we prepared data sets 2, 5, and 6 (DB02, DB05, and DB06) in which the nitrogen stereochemistry automatically assigned by the conformer generator was manipulated by the operator (see Experimental Section) so that the acidic hydrogen pointed in the same direction for all compounds. For DB02, CORINA was used, whereas DB05 and DB06 were prepared with Omega. The lowest energy conformers were included in DB05, whereas to obtain DB06, one low-energy conformer per molecule (within 3 kcal/mol of the lowest) was randomly chosen.

Data set 3 (DB03) was created to check whether expert manual data set optimization could in some way improve the quality of ALMOND results. Briefly (details are given in Experimental Section), in DB03 similar chemical features

10.1021/jm0704651 CCC: $37.00
Published on Web 08/31/2007

Table 1. Representation of the Main Structural Features Shared by the Compounds of the Series and the Experimental and Calculated pIC_{50} (OSC Inhibitor Activity)



		pIC_{50} (calcd)					
no.	pIC_{50} (exptl)	DB01	DB02	DB03	DB04	DB05	DB06
		Training Set					
1	8.72	8.11	8.18	8.19	8.39	8.47	8.10
2	8.54	8.40	7.89	8.51	8.07	8.15	8.45
4	8.46	8.28	8.34	8.64	8.36	8.28	8.39
6	8.39	8.75	8.78	8.47	8.47	8.73	8.54
7	8.34	8.23	8.33	8.15	7.99	8.14	7.78
8	8.27	8.26	8.16	8.09	8.13	8.06	8.16
10	8.21	8.07	8.37	7.93	8.47	8.48	8.24
11	8.20	8.28	8.40	8.18	8.55	8.34	8.46
12	8.19	7.76	7.75	8.18	8.47	7.85	8.48
13	8.17	8.32	8.12	7.86	8.07	8.46	8.02
14	8.11	7.69	7.63	7.89	7.77	7.95	7.94
15	8.10	8.43	8.42	8.53	7.94	8.15	8.39
16	8.06	7.80	7.72	8.02	8.17	8.02	7.88
18	7.94	7.34	7.60	7.78	7.76	7.77	8.40
19	7.91	8.15	7.88	8.22	8.04	8.05	8.12
20	7.87	7.30	7.26	7.84	7.63	7.57	7.56
23	7.72	7.39	7.57	7.30	7.91	7.81	7.96
24	7.71	7.90	8.00	7.85	7.88	7.71	8.18
25	7.66	7.29	7.07	7.48	7.59	7.65	7.73
26	7.65	7.88	7.88	7.45	7.59	7.71	7.85
28	7.54	8.21	7.87	7.72	7.49	7.57	7.50
29	7.50	7.72	7.99	8.05	7.69	7.83	7.39
30	7.45	7.74	7.97	7.62	7.64	7.47	7.47
33	7.32	7.06	7.28	7.24	7.27	7.37	7.22
34	7.15	7.24	7.24	6.86	6.97	7.05	7.66
35	7.01	7.31	7.21	6.93	7.19	7.20	7.24
37	6.62	6.60	6.64	6.51	6.77	6.67	6.50
39	6.36	6.29	6.49	6.81	6.65	6.31	6.46
40	6.30	6.66	6.87	6.81	6.69	6.57	6.62
41	6.21	6.60	6.36	6.70	6.47	6.41	6.54
43	5.73	6.43	6.16	5.68	5.53	5.66	6.21
44	5.72	5.64	5.68	5.63	5.53	5.65	5.53
		Test Set					
5	8.39	7.78	8.01	8.11	8.18	8.37	7.74
21	7.80	7.96	8.26	7.66	8.27	8.06	7.77
27	7.54	7.34	7.12	7.37	7.37	7.54	6.92
32	7.41	8.50	7.88	7.60	8.07	7.96	7.94
36	6.65	7.25	6.98	6.92	6.84	6.48	6.82

(aromatic rings, equal or similar substituents on corresponding atoms, and acidic hydrogens) have been oriented in the same direction.

Finally, the six data sets were checked for their diversity in ALMOND descriptors by PCA analysis (data not shown). As expected, the six data sets explore different regions of the space determined by GRIND.

GRIND-Based 3D-QSAR Analysis: Statistical Results. GRIND were related to the inhibition potency values, expressed as pIC_{50}, by means of PLS analysis. The PLS analyses performed on the six data sets did not produce any model. By a careful analysis of the **T**−**U** score plots, the presence of three outliers (**3**, **9**, and **38**) was detected for all data sets (in-depth analysis of why these compounds are identified as outliers is beyond the scope of this study). The PLS analyses were then repeated, excluding the three outliers, and three-latent-variables (three-LV) models were obtained for all the data sets (data not shown). With the ability of ALMOND to find a statistical model, the data sets were then randomly split into a training set and a test set (**5**, **21**, **27**, **32**, and **36**) to perform analysis according to standard procedures.[7]

The PLS analysis was then repeated for the six training sets, and three-LV models were obtained (Table 2). The models show a r^2 value ranging from 0.82 to 0.94 and a q^2_{LOO} varying from 0.11 to 0.53. The test set prediction was also successful (Table 1): the difference in pIC_{50}(OSC) between experimental and calculated values was acceptable. Table 2 also shows the coefficients (i.e., the slope *a* and the *Y*-intercept *b*) of the linear relationships between experimental and calculated values (the corresponding graphs are in Supporting Information). Values close to 1 for the slope and close to 0 for the *Y*-intercept confirm the acceptable quality for all test set predictions.

Taken together, the statistical results indicate that 3D-QSAR models have been found for all the data sets investigated. With a deeper analysis, however, some relevant differences in the quality of the six models emerge. First, models based on CORINA data sets and obtained without manual alignment (DB01 and DB02) are of lower statistical quality than those based on Omega data sets (Table 2). Moreover, the introduction of nitrogen chiral specification (DB02) does not improve the quality of the model obtained using CORINA default settings (DB01). To obtain a significant statistical improvement in CORINA-based models (in particular in test set prediction), human expertise (DB03) must be introduced. However, this approach neglects the main advantage of ALMOND, i.e., it being fully automated.

The best models resulted from DB04 and DB05, both based on the lowest energy conformers generated by Omega. In particular, introduction of the chiral specification slightly improved the statistics of ALMOND models and enabled DB05 to be designated as the "best" data set. DB06, Omega-based but with low-energy conformers within 3 kcal/mol of the lowest, gives modest results, slightly better than CORINA-based DB01 model (Table 2).

Summing up, statistical results are poorly influenced by nitrogen chirality, whereas the most important factor for correctness of an ALMOND model is to respect some consistency in the generation of conformers. In particular, the use of the lowest energy conformer resulting from a conformational analysis may explain why Omega works better than CORINA.

GRIND-Based 3D-QSAR Analysis: Graphical Results. Each GRIND variable represents the presence and the intensity of a pair of nodes present at a certain distance, and thus, GRIND variables encode two pieces of information: the numerical value and the distance between two nodes. The numerical values are used in the PLS analysis (see above) to select the GRIND able to correlate with biological activity, whereas the distances are used to visualize the pair of nodes that have been used to assign a value to the GRIND variable.

Figure 1 shows the graphical results for the six ALMOND models. For each model (A−F) the filtered MIFs, which represent favorable probe−target interaction regions, are shown for the most active compound **1**. TIP, N1, O, and DRY nodes are shown, respectively, in green, blue, red and yellow. GRINDs with VIP values above 2.0 (this threshold being determined from a number of trials) and a positive PLS coefficient (see Experimental Section) are also shown by their distance ranges.

Preliminary inspection of Figure 1 shows that the main favorable probe−target interaction regions (VIP values are reported in Supporting Information) determined by the six models are similar but not identical. All models indicate the presence of a blue region, separated by two green regions by different distances, as being significant. The distance between the blue and the green regions close to the nitro group is very similar for all models (7.6−10.4 Å). Conversely, the blue region

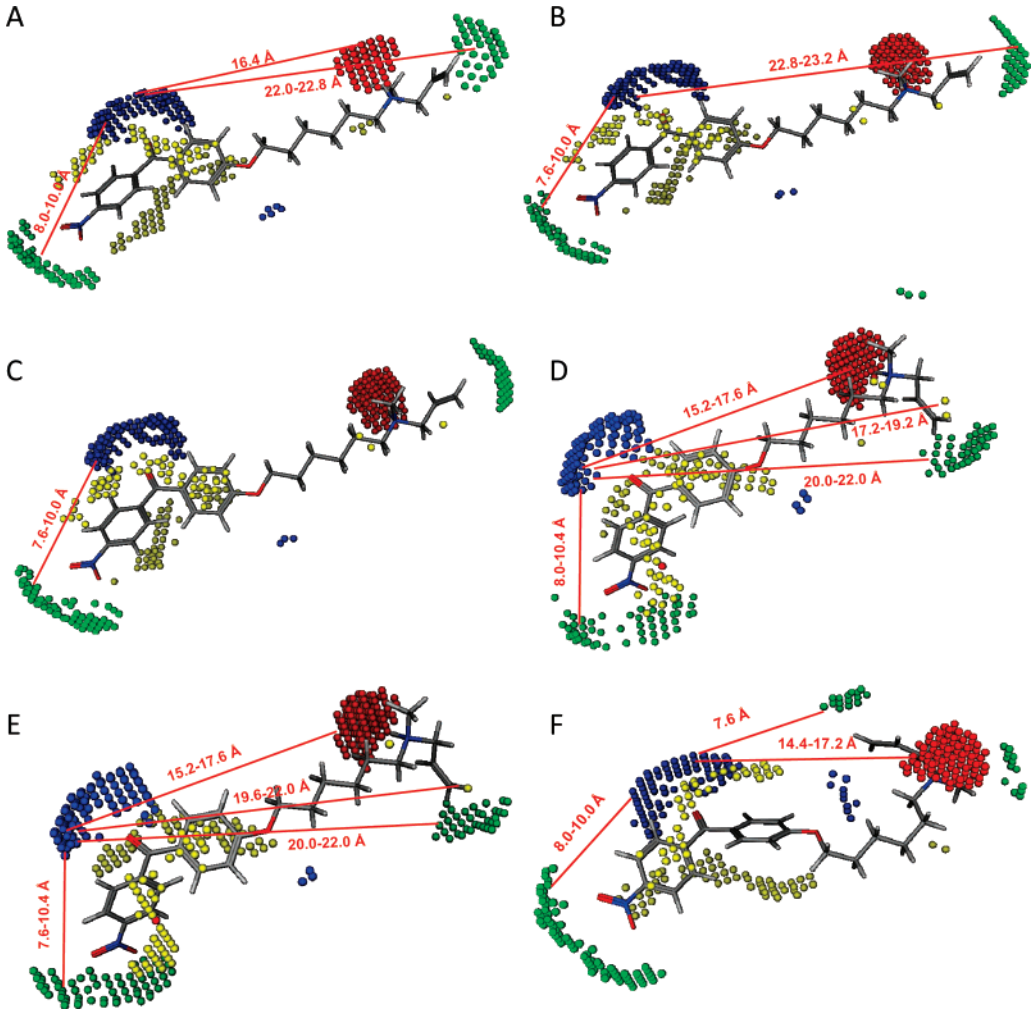


Figure 1. Filtered MIFs obtained with DRY (yellow), O (red), N1 (blue), and TIP (green) probes and the distance ranges corresponding to GRIND with VIP values greater than 2.0 of the most active compound, **1**: (A) DB01; (B) DB02; (C) DB03; (D) DB04; (E) DB05; (F) DB06.

Table 2. ALMOND Statistical Results for the Six Data Sets ($n = 35$)

	CORINA			Omega		
	DB01 no chiral, automatic	DB02 chiral, automatic	DB03 chiral, manual	DB04 no chiral, automatic	DB05 chiral, automatic	DB06 chiral, automatic
NLV	3	3	3	3	3	3
q^2_{LOO}	0.20	0.11	0.40	0.50	0.53	0.28
r^2	0.82	0.82	0.89	0.92	0.94	0.88
a^a	1.00 (±0.09)	1.00 (±0.09)	1.00 (±0.09)	1.00 (±0.05)	1.00 (±0.05)	1.00 (±0.07)
b^a	0.00 (±0.66)	0.00 (±0.66)	0.01 (±0.66)	−0.01 (±0.41)	0.01 (±0.35)	−0.01 (±0.52)

[a] Coefficients of the relationship between experimental and calculated values according to the equation: $pIC_{50}^{exptl} = a\ pIC_{50}^{calcd} + b$. 95% confidence limits are given in parentheses.

and the second green region are separated by a distance that varies with the model, ranging from 22.0 to 23.2 Å for CORINA-based (DB01 and DB02) and from 7.6 Å (DB06) to 22.0 Å (DB04 and DB05) for Omega-based models. This difference reflects the fact that CORINA conformations are more elongated than Omega conformations. The absence of this distance in the model obtained for DB03 is due to the fact that the corresponding VIP values are slightly below 2.0 (see Supporting Information) and thus excluded from the representation. Parts A and D−F of Figure 1 show the presence of a red region at a distance of about 16 Å from the blue one. This distance is absent in the models whose statistics are of lower quality.

Taken together, graphical results indicate that Omega-based models are more detailed than CORINA-based models. Moreover Omega-based models are slightly different from the other as a consequence of the variation in the elongation of conformers.

Conclusions

This study represents the first investigation of the influence of 3D starting conformation in obtaining GRIND-based 3D-QSAR models. The results show that, as expected, ALMOND models are dependent on the starting conformation of the compounds investigated. However, this dependence is of differing relevance depending on the information required from the study. If what is desired is fast screening for good drug candidates, any reasonable 3D structure can be used as an input. If the predictive aspect of the model is more important than the interpretative aspect, predictions resulting from the lowest energy conformers obtained with a conformational analysis are

more convincing than single conformer generation approaches. Finally, if the computational strategy has been designed to combine predictive and interpretative aspects, a series of different starting conformations should be used, and the different models obtained should be carefully compared to extract all possible information encoded in the models. However, it must be pointed out that the comparative analysis of more than one model remains a field to be explored by academics rather than industrial researchers.

As a whole, this study indicates that this computational approach depends on the 3D conformation of the compounds investigated, but this dependence can easily be taken into account without too much computational and human effort and without any loss of automation.

Experimental Section

Data Sets Preparation. The SMILES codes were prepared through a semiautomatic procedure. The 2D structures were drawn in ChemDraw (version 10.0, CambridgeSoft Corporation), and then the ChemDraw conversion tool was used to generate SMILES codes that were then submitted to two widely used 3D structure generators, CORINA[5] and Omega.[6] CORINA generates by default one low-energy conformation for each SMILES code by combining monocentric fragments with standard bond lengths and angles and by using appropriate dihedral angles. Omega uses fragment templates along σ bonds to assemble initial models, after which it begins a torsion search with an assessment of freely rotatable bonds. Finally, Omega furnishes a list of conformations ranked by their MMFF energy.

DB01 and DB04 were obtained by submitting the SMILES codes without any chiral specification to CORINA and Omega, respectively. Since Omega gives a list of conformers, for any compound the conformer with the lowest-energy was selected (DB04) to be submitted to ALMOND runs.

Two additional sets of SMILES were then prepared by applying the chiral specification @ (neighbors listed anticlockwise) and @@ (neighbors listed clockwise) on the positively charged nitrogen. The two sets of SMILES were submitted to CORINA, and two data sets of compounds were thus obtained. One of the two 3D structures of the most active compound was randomly selected and used as a template to select, for all other compounds, the conformation for which the hydrogen atom linked to the positive nitrogen was in the same direction as the template. The data set obtained was DB02. DB05 and DB06 were similarly prepared using Omega. The lowest energy conformers were included in DB05, whereas to obtain DB06, one low-energy conformer per molecule (within 3 kcal/mol of the lowest) was randomly chosen.

Finally, DB03 (CORINA-based) was the refinement of DB02 by aligning the structural features shared by the compounds in the series with the flex_alignment MOE tool[8] (modified version downloaded from http://svl.chemcomp.com/).

ALMOND. The six data sets were used as an input for ALMOND (version 3.2, software available from Molecular Discovery Ltd., 215 Marsh Road, HA5 5NE, Pinner, Middlesex, U.K., http://www.moldiscovery.com). In this study, standard ALMOND probes (DRY, O, N1, and TIP) and parameters were used. A scheme illustrating the ALMOND procedure is given in Supporting Information.

Chemometric Analysis. Chemometric analysis was carried out using the statistical tools included in ALMOND together with the SIMCA-P statistical package (SIMCA-P, version 10.0.2.0; Umetrics, 2003).

The GRIND calculated by ALMOND were related to OSC inhibition potency by means of partial least-squares (PLS) analysis. The GRIND (used without scaling) were the *X* variables, whereas the corresponding inhibitor potencies, after adequate transformation (discussed above), were the *Y* variables. The optimum number of PLS components (latent variables, LV) was chosen by monitoring changes in the model's predictivity index (q^2_{LOO}, leave one out) evaluated by applying the cross-validation procedure available in ALMOND. Outliers were determined by careful inspection of the plot of the *X* and *Y* scores (generally known as the **T** and **U** matrices, respectively) of the first latent variable. VIP plots generated by SIMCA were used to guide data interpretation.[9] To interpret the statistical model correctly, the variable importance in the projection (VIP) plot must be combined with the PLS coefficient table. The VIP plot displays the VIP values as a column plot sorted in descending order, with confidence intervals derived from jackknifing.[9] For each grid-independent descriptor, the VIP values reflect the importance of terms in the model with respect to *Y* and to *X* but they do not consider the sign of the coefficients. Conversely, the PLS coefficients represent the contribution of each single descriptor to the model only with respect to *Y*. Coefficients with positive values increase the inhibition potency of the compound and vice versa. The highest VIP values usually correspond to the highest peak values in the coefficient plot.

Acknowledgment. G.E. and G.C. are indebted to the University of Turin for financial support. The authors also thank the reviewers for their stimulating comments.

Supporting Information Available: Chemical structures of compounds, protonated nitrogen chirality, VIP tables for the six models, plot of the observed pIC_{50}(Exp) vs calculated pIC_{50}(ALMOND), PLS pseudocoefficients obtained by ALMOND for all data sets, schematic ALMOND calculation procedure, and the 10 correlograms obtained in this study. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) 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.* **2004**, *43*, 3233−3243.
(2) Cruciani, C. *Molecular Interaction Fields. Applications in Drug Discovery and ADME Prediction*; Wiley-VCH: Zurich, Switzerland, 2006.
(3) Lenhart, A.; Reinert, D. J.; Aebi, J. D.; Dehmlow, H.; Morand, O. H.; Schulz, G. E. Binding structures and potencies of oxidosqualene cyclase inhibitors with the homologous squalene-hopene cyclase. *J. Med. Chem.* **2003**, *46*, 2083−2092.
(4) Dehmlow, H.; Aebi, J. D.; Jolidon, S.; Ji, Y. H.; der Mark, E. M.; Himber, J.; Morand, O. H. Synthesis and structure−activity studies of novel orally active non-terpenoic 2,3-oxidosqualene cyclase inhibitors. *J. Med. Chem.* **2003**, *46*, 3354−3370.
(5) *CORINA*; http://www.molecular-networks.com/online_demos/corina_demo.html.
(6) *OMEGA*, version 2.1.0; OpenEye Science Software (3600 Cerrillos Road, Suite 1107): Santa Fe, NM, 2006.
(7) Hawkins, D. M.; Basak, S. C.; Mills, D. Assessing model fit by cross-validation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579−586.
(8) *MOE*, version 2006.08; Chemical Computing Group Inc.: Montreal, Quebec, Canada, 2006.
(9) *User's Guide to SIMCA-P, SIMCA-P+*; Umetrics: Kinnelon, NJ, 2003.

JM0704651