

Expedited Articles

Cross-Validated R^2 -Guided Region Selection for Comparative Molecular Field Analysis: A Simple Method To Achieve Consistent Results

Sung Jin Cho and Alexander Tropsha*

The Laboratory for Molecular Modeling, Division of Medicinal Chemistry and Natural Products, School of Pharmacy, University of North Carolina, Chapel Hill, North Carolina 27599

Received December 9, 1994[⊗]

Comparative Molecular Field Analysis (CoMFA) is one of the most powerful modern tools for quantitative structure–activity relationship studies. The CoMFA predictability is conventionally characterized by a cross-validated correlation coefficient R^2 (q^2). Our CoMFA investigation of 4 datasets, including 7 cephalotaxine esters, 20 5-HT_{1A} receptor ligands, 59 inhibitors of HIV protease, and 21 steroids reveals that the q^2 value is sensitive to the overall orientation of superimposed molecules on a computer terminal and can vary by as much as $0.5q^2$ units when the orientation is varied by systematic rotation. To optimize CoMFA, we have developed a new routine, cross-validated R^2 -guided region selection (q^2 -GRS). We first subdivide the rectangular lattice obtained initially with conventional CoMFA into 125 small boxes and perform 125 independent analyses using probe atoms placed within each box with the step size of 1.0 Å. We then select only those small boxes for which a q^2 is higher than a specified optimal cutoff value. Finally, we repeat CoMFA with the union of small boxes selected at the previous step. Four datasets described above were used to validate this new q^2 -GRS routine. In each case we have obtained an orientation-independent, high q^2 , exceeding the one obtained with the conventional CoMFA. This method shall be used routinely in the future CoMFA studies to guarantee the reproducibility of the reported q^2 values.

Introduction

Since its introduction in 1988, the comparative molecular field analysis (CoMFA)¹ approach has rapidly become one of the most powerful tools for three-dimensional quantitative structure–activity relationship (3-D QSAR) studies. Over the years, this approach has been applied to a wide variety of receptor and enzyme ligands (recently reviewed by Cramer et al.² and Thibaut³). Undoubtedly, the further development of this approach is of great importance and interest to many scientists working in the area of rational drug design.

CoMFA methodology is based on the assumption that since, in most cases, the drug–receptor interactions are noncovalent, the changes in the biological activities or binding affinities of sample compounds correlate with changes in the steric and electrostatic fields of these molecules. In a standard CoMFA procedure, all molecules under investigation are first structurally aligned, and the steric and electrostatic fields around them are sampled with probe atoms, usually sp³ carbon with +1 charge, on a rectangular grid that encompasses aligned molecules. The results of the field evaluation in every grid point for every molecule in the dataset are placed in the CoMFA QSAR table which therefore contains thousands of columns. The analysis of this table by the means of standard multiple regression is practically impossible; however, the application of special multivariate statistical analysis routines, such as partial least squares (PLS) analysis and cross-validation, ensures the

statistical significance of the final CoMFA equation. A cross-validated R^2 (q^2) which is obtained as a result of this analysis serves as a quantitative measure of the predictability of the final CoMFA model. The statistical meaning of the q^2 is different from that of the conventional r^2 : the q^2 value greater than 0.3 is considered significant.⁴

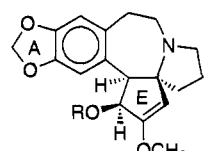
Experimenting with various datasets and CoMFA routines, we have discovered that in several cases we were unable to reproduce the q^2 values reported in the literature or obtained in our own laboratory. We found that the resulting q^2 values may vary greatly for the same molecular database with rigidly aligned molecules. Upon closer examination of research routines, we found that the only difference between the analyses was the orientation of rigidly aligned molecules on user terminals.

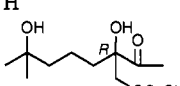
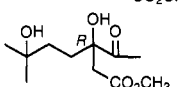
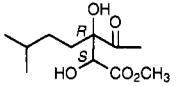
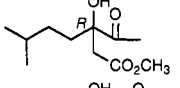
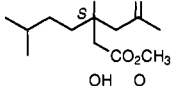
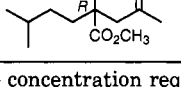
In this paper, we investigate this phenomenon systematically using 3 datasets of model compounds of different sizes: 7 cephalotaxine esters,⁵ 20 5-HT_{1A} receptor ligands,⁶ and 59 inhibitors of human immunodeficiency virus (HIV) protease.⁷ We find that in all cases the q^2 value is sensitive to the orientation of rigidly aligned molecules on the computer terminal and may vary with the orientation by as much as $0.5q^2$ units. We further propose a new CoMFA region optimization routine, which we call cross-validated R^2 -guided region selection (q^2 -GRS). We show that the application of this new routine to 3 datasets described above as well as to 21 steroids¹ leads to reproducible, high q^2 values that do not depend on the orientation of molecular aggregates on a computer terminal.

* To whom correspondence should be addressed.

[⊗] Abstract published in *Advance ACS Abstracts*, March 15, 1995.

Table 1. Inhibition of Protein Synthesis by Cephalotaxine Esters⁵



alkaloid	R	IC ₅₀ (mM) ^a
cephalotaxine (1)	H	440
homoharringtonine (2)		3
harringtonine (3)		1
isoharringtonine (4)		7.5
deoxyharringtonine (5)		6.6
pseudodeoxyharringtonine (6S)		1100
pseudododeoxyharringtonine (6R)		1100

^a Results are expressed as the concentration required for 50% inhibition of protein synthesis in cell-free lysates.

Computational Details

SYBYL molecular modeling software⁸ was used for structure generation and CoMFA. All calculations were done on IBM RS6000 Model 340. For all steps of conventional CoMFA, the default SYBYL settings were used except as otherwise noted. For each CoMFA analysis, the minimum σ was set to 2.0 to expedite the calculation, and sp³ carbon with +1 charge was used as the probe atom.

Representative Datasets and Structure Alignment. For this report, we have selected 4 sets of rigidly aligned model compounds of different size: 7 cephalotaxine esters,⁵ 20 5-HT_{1A} receptor ligands analyzed by Taylor et al.,⁶ 59 inhibitors of HIV protease, analyzed by Waller et al.,⁷ and 21 steroids analyzed by Cramer et al.¹ The files with prealigned inhibitors of HIV protease and 5-HT_{1A} receptor ligands were kindly provided by Drs. Waller and Taylor, respectively. The file with prealigned steroid molecules and the corresponding region file were obtained with SYBYL⁸ package as part of CoMFA tutorial.

The alignment of seven cephalotaxine esters (Table 1) was generated as follows. The cephalotaxine rigid ring structure was generated from the fractional coordinates of cephalotaxine *p*-bromobenzoate.¹⁰ The computer models of all other molecules were generated by modifying this template; the geometry of each molecule was optimized with the standard Tripos force field.⁸ Systematic conformational search (10 deg increment, 0.1 maximum energy difference, and no electrostatic option) was performed on the side chain of each molecule to obtain the lowest energy conformers. The charges were then calculated using MNDO method and electrostatic potential (ESP) fit routine as implemented in MOPAC 6.0.¹¹ MOPAC calculations were done on the Convex C-220 minisupercomputer at the North Carolina Super Computing Center. Since 3 (Table 1) was the most active among seven compounds, it was used as a template. In order to align 2–7, the following common atoms were selected: the methylene carbon on the A ring, the methoxy oxygen on the E ring, the nitrogen, and the chiral carbon of each ester side chain (Table 1). For 1, which

Table 2. Number of Occurrences of Different q^2 as a Result of Systematic Rotation of Seven Rigidly Aligned Cephalotaxine Esters

q^2	number of occurrences					
	6 ^a	5 ^a	4 ^a	3 ^a	2 ^a	1 ^a
0.000–0.099	1	1	0	0	0	1
0.100–0.199	1	1	1	0	0	6
0.200–0.299	15	4	3	4	0	21
0.300–0.399	41	11	5	12	1	50
0.400–0.499	61	17	23	27	3	79
0.500–0.599	89	18	29	24	1	58
0.600–0.699	35	7	8	9	0	4
0.700–0.799	4	0	2	0	0	0
0.800–0.899	1	0	0	0	0	0

^a The optimal number of components.

does not have an ester side chain, the hydroxy oxygen on the E ring was used instead. The compounds were superimposed by rigid rms fit followed by the field fit optimization as implemented in SYBYL. Finally, the geometry of all compounds was reoptimized with the FIELDFIT option turned off to adjust internal geometry which might have been distorted due to field fit optimization.

Conventional CoMFA. Conventional CoMFA was performed with the QSAR option of SYBYL. The steric and electrostatic field energies were calculated using sp³ carbon probe atoms with +1 charge. The CoMFA grid spacing was 2.0 Å in all three dimensions within the defined region, which extended beyond the van der Waals envelopes of all molecules by at least 4.0 Å. The CoMFA QSAR equations were calculated with the PLS algorithm. The optimal number of components (ONC) in the final PLS model was determined by the q^2 value, obtained from the leave-one-out cross-validation technique. For small datasets, in order to maximize the q^2 value and minimize the standard error of prediction, the number of components was increased only when adding a component raised the q^2 value by 5% or more.¹² For HIV protease inhibitors, the number of components with the lowest standard error of prediction was selected as the ONC.

The Orientation Dependence of q^2 . The overall orientation of the molecular aggregate of rigidly aligned molecules was varied as follows. Starting from an arbitrary orientation, the whole aggregate was rotated by 30° at a time around *x*, *y*, and *z* axes using the SYBYL STATIC command. For each orientation, the conventional CoMFA was performed with 10 components, using 7 cross-validation groups for cephalotaxine esters, 20 cross-validation groups for 5-HT_{1A} receptor ligands, and 59 cross-validation groups for HIV protease inhibitors. The region files were generated automatically. After each CoMFA analysis, the q^2 value and the ONC were recorded.

q^2 -GRS Routine. The q^2 -GRS process is illustrated in Figure 1. A conventional CoMFA is performed initially using an automatically generated region file (step 1). The rectangular grid encompassing aligned molecules is then broken into 125 small boxes of equal size (step 2). This is achieved by using the Cartesian coordinates of the upper right and lower left corners of this initial lattice in order to compute the Cartesian coordinates of the upper right and the lower left corner of each of 125 small boxes. For each of these newly generated region files, a separate CoMFA is performed with the step size of 1.0 Å (step 3). The resulting q^2 from each analysis is compared to a specified cutoff, and only those regions with the q^2 greater than the cutoff are selected for further analysis (step 4). Finally, the selected regions are combined to generate a master region file (step 5), and the final CoMFA is performed (step 6).

Results

The Orientation Dependence of q^2 . The number of occurrences of different ranges of the q^2 values and the ONC obtained for different orientations of various molecular aggregates are summarized in Tables 2–4. The frequency distribution of q^2 values observed for

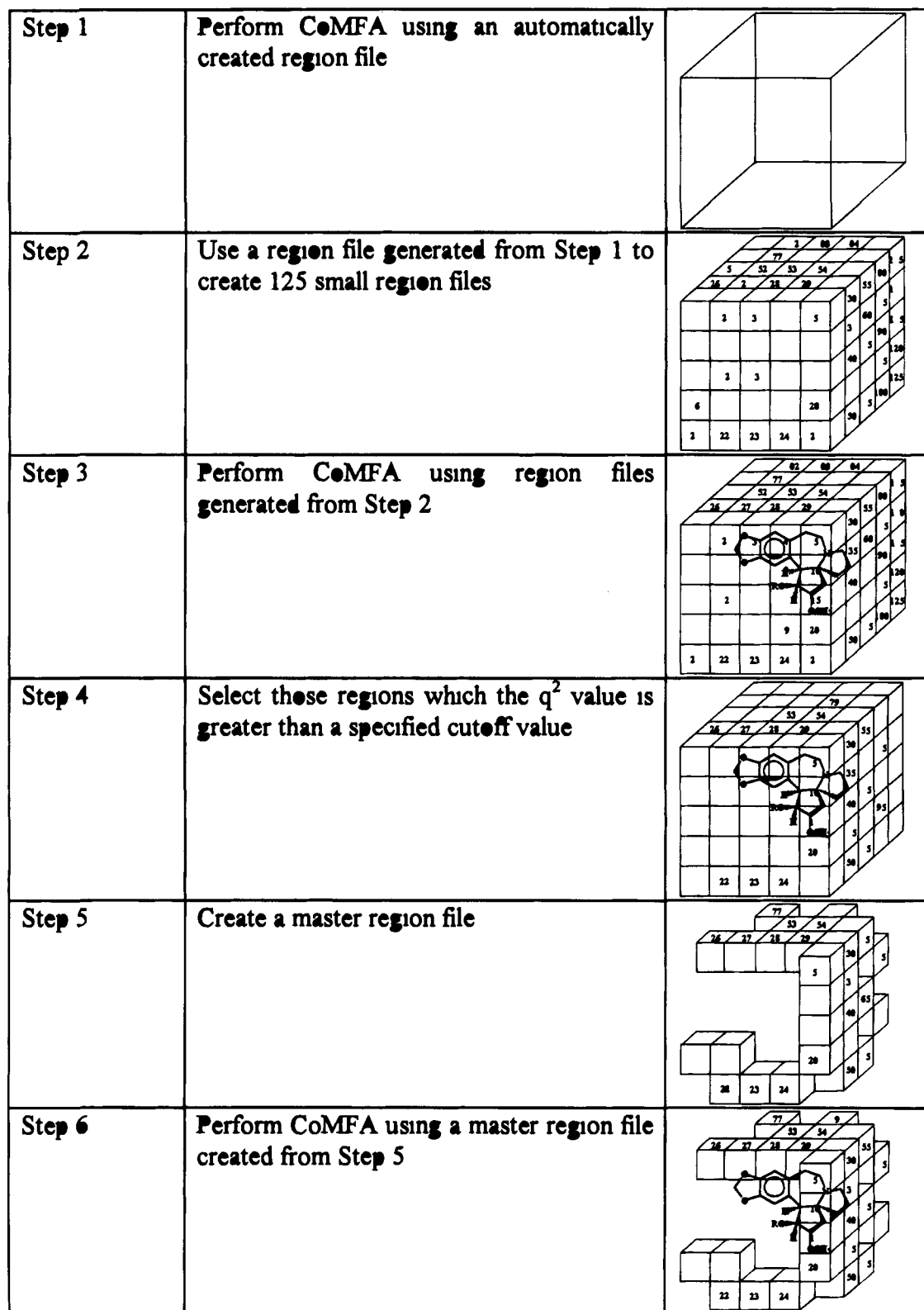


Figure 1. The cross-validated R^2 -guided region selection (q^2 -GRS) routine.

different datasets as a result of rotations are given in Figures 2–4. (Due to the large number of CoMFA runs, the number of components with the highest q^2 is selected as the ONC rather than employing the 5% increase rule described under Computational Details.) For cephalotaxine esters, the highest (0.819) and lowest (0.050) q^2 s were obtained with the ONC of 6 (Table 2 and Figure 2). For 5-HT_{1A} receptor ligands, the highest (0.607) and lowest (−0.015) q^2 s were obtained with the

ONC of 10 and 1, respectively (Table 3 and Figure 3). For HIV protease inhibitors, the range of q^2 value was much more narrow (Table 4 and Figure 4). The highest (0.802) and lowest (0.586) q^2 s were obtained with the ONC of 10. It is obvious from these results that a single orientation gives an arbitrary value of q^2 which most probably would fall into the region with the highest frequency of occurrences of the q^2 values. As discussed below, this happens with most of the datasets.

Table 3. Number of Occurrences of Different q^2 as a Result of Systematic Rotation of 20 Rigidly Aligned 5-HT_{1A} Receptor Ligands

q^2	number of occurrences									
	10 ^a	9 ^a	8 ^a	7 ^a	6 ^a	5 ^a	4 ^a	3 ^a	2 ^a	1 ^a
0.000-0.049	0	0	0	0	0	0	0	1	0	8
0.050-0.099	2	0	0	0	3	0	0	0	11	12
0.100-0.149	2	0	0	0	5	2	0	5	5	1
0.150-0.199	4	1	2	0	12	4	1	4	9	0
0.200-0.249	4	1	3	0	18	3	6	6	1	0
0.250-0.299	20	6	5	3	17	5	1	6	0	0
0.300-0.349	17	9	2	3	23	10	3	6	1	0
0.350-0.399	24	13	9	0	16	9	2	2	0	0
0.400-0.449	22	7	2	0	23	5	3	0	0	0
0.450-0.499	19	6	3	4	20	7	1	0	0	0
0.500-0.549	9	3	1	1	8	2	0	0	0	0
0.550-0.599	5	1	0	0	2	0	0	0	0	0
0.600-0.649	1	0	0	0	0	0	0	0	0	0

^a The optimal number of components.

Table 4. Number of Occurrences of Different q^2 as a Result of Systematic Rotation of 59 Rigidly Aligned Inhibitors of HIV Protease

q^2	number of occurrences					
	10 ^a	9 ^a	8 ^a	7 ^a	6 ^a	4 ^a
0.575-0.599	1	0	0	0	0	0
0.600-0.624	0	0	0	1	2	0
0.625-0.649	0	0	0	1	4	0
0.650-0.674	13	1	0	9	16	0
0.675-0.699	34	0	10	18	32	1
0.700-0.724	51	2	17	26	32	0
0.725-0.749	76	3	17	23	20	0
0.750-0.774	50	2	14	7	3	0
0.775-0.799	15	0	2	1	0	0
0.800-0.824	1	0	0	0	0	0

^a The optimal number of components.

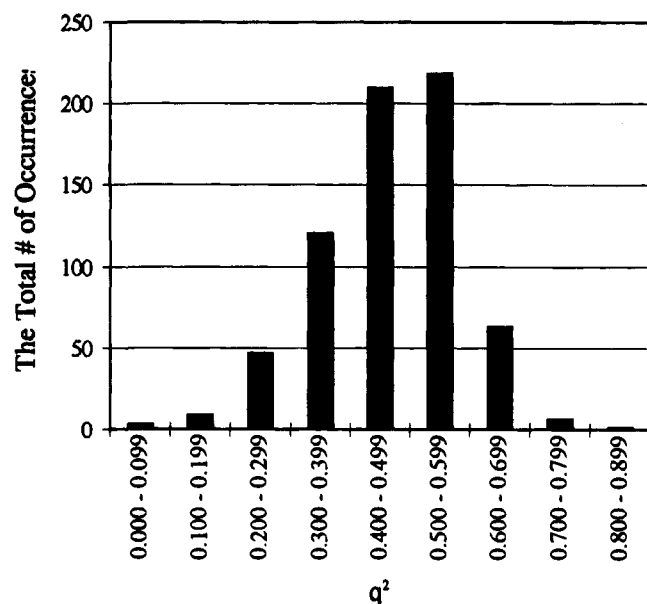


Figure 2. The q^2 values observed for different orientations of seven rigidly superimposed cephalotaxine esters. The molecular aggregate was systematically rotated by 30° at a time around x, y, and z axes.

It was suggested² that increasing the grid resolution may improve the CoMFA results. Tables 5-7 show q^2 s obtained as a result of CoMFA with the grid spacing of 1.0 vs 2.0 Å; for comparison, we have included the results obtained with the different number of components. Indeed, lowering the step size from 2.0 to 1.0 Å narrowed the distribution of q^2 s (cf. the differences

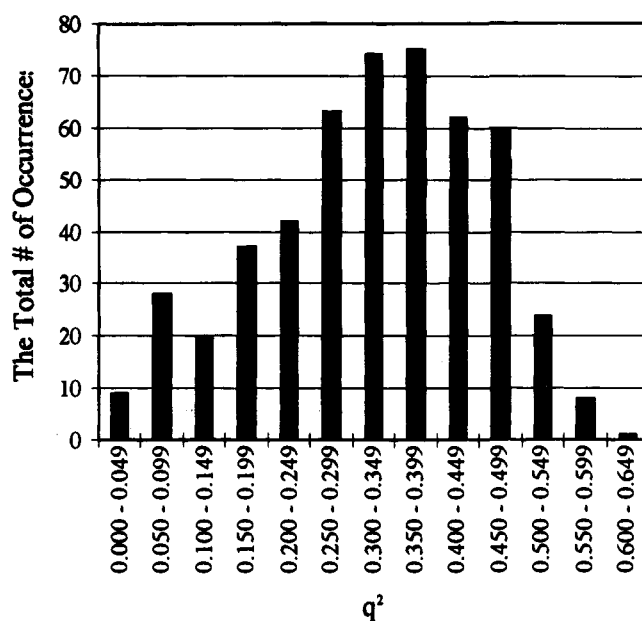


Figure 3. The q^2 values observed for different orientations of 20 rigidly superimposed 5-HT_{1A} receptor ligands. The molecular aggregate was systematically rotated by 30° at a time around x, y, and z axes.

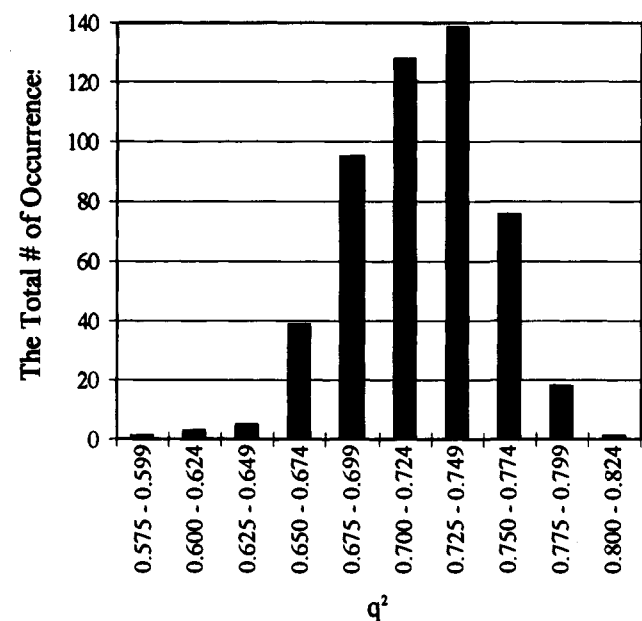


Figure 4. The q^2 values observed for different orientations of 59 rigidly superimposed HIV protease inhibitors. The molecular aggregate was systematically rotated by 30° at a time around x, y, and z axes.

between the lowest and the highest values of q^2 for 2.0 Å CoMFA runs vs 1.0 Å CoMFA runs in Tables 5-7). However, for each dataset (Tables 5-7), the highest q^2 obtained with 1.0 Å grid resolution is consistently lower than the highest q^2 obtained with the 2.0 Å step size.

The q^2 -GRS Routine. We have applied the q^2 -GRS routine to three different orientations of cephalotaxine esters, 5-HT_{1A} receptor ligands, and HIV protease inhibitors obtained in the course of the systematic rotation of molecular aggregates (see the previous section): random (i.e., some initial orientation), "best" (i.e., the one with the highest value of the q^2), and "worst" (i.e., the one with the lowest value of q^2). Initially, we have used the q^2 cutoff value of zero, and the results are summarized in Tables 5-7. Evidently,

Table 5. Comparison of Conventional CoMFA and the q^2 -GRS Method as Applied to Seven Cephalotaxine Esters (Numbers in bold represent the q^2 values for the ONC)

orientation ^a	CoMFA method	step size (Å)	q^2		
			1 ^b	2 ^b	3 ^b
random	conventional	2.0	0.594	0.594	0.595
worst	conventional	2.0	0.001	-0.055	0.001
best	conventional	2.0	0.562	0.537	0.536
random	conventional	1.0	0.464	0.441	0.464
worst	conventional	1.0	0.433	0.401	0.416
best	conventional	1.0	0.385	0.359	0.396
random	q^2 -GRS ^c	1.0	0.647	0.619	0.642
worst	q^2 -GRS ^c	1.0	0.587	0.568	0.577
best	q^2 -GRS ^c	1.0	0.651	0.619	0.641

^a The "best" and the "worst" orientations are defined as those that generate the highest and the lowest q^2 s, respectively. ^b The number of components. ^c q^2 cutoff = 0.

the application of this routine immediately leads to very consistent values of q^2 , regardless of the orientation of molecular aggregates. With the q^2 cutoff of zero, the resulting q^2 values were fairly close to the best q^2 values obtained with the 2.0 Å step size (cf. Tables 5–7). We further investigated the effect of various cutoff values on the resulting q^2 . The results for the series of 5-HT_{1A} receptor ligands and HIV protease inhibitors are summarized in Tables 8 and 9, respectively. For both

5-HT_{1A} receptor ligands and HIV protease inhibitors, the highest q^2 values (0.807 and 0.798 for 5-HT_{1A} receptor ligands and HIV protease inhibitors, respectively) were obtained with the q^2 cutoff of 0.4. The q^2 value obtained for 5-HT_{1A} receptor ligands was substantially higher (by more than $0.3q^2$ units) than the reported value of 0.481⁶ (Table 8). For HIV protease inhibitors, the q^2 obtained with the q^2 -GRS routine was 0.789, which is slightly higher than the reported value of 0.778⁷ (with six components in both cases). Finally, the q^2 -GRS method was applied to two steroid datasets with different biological activity measurements (Tables 10 and 11). A slight increase in the q^2 values was observed as compared to reported values¹ of 0.555 and 0.662 for the two datasets, respectively.

Discussion

The research presented in this paper was spurred by our incidental discovery that the values of the q^2 obtained as a result of conventional CoMFA are sensitive to the overall orientation of rigidly aligned molecules on the computer terminal. Puzzled by this observation, we have conducted systematic investigation of this phenomenon. We have performed systematic rotation of several molecular aggregates of different compounds in three-dimensional coordinate system,

Table 6. Comparison of Conventional CoMFA and the q^2 -GRS Method as Applied to 20 5-HT_{1A} Receptor Ligands (Numbers in bold represent the q^2 values for the ONC)

orientation ^a	CoMFA method	step size (Å)	q^2					
			1 ^b	2 ^b	3 ^b	4 ^b	5 ^b	6
random	conventional	2.0	0.127	0.274	0.358	0.418	0.460	0.477
worst	conventional	2.0	-0.015	-0.135	-0.383	-0.645	-0.674	-0.660
best	conventional	2.0	0.204	0.347	0.436	0.487	0.557	0.589
random	conventional	1.0	0.129	0.231	0.304	0.334	0.383	0.396
worst	conventional	1.0	0.119	0.215	0.266	0.284	0.329	0.345
best	conventional	1.0	0.137	0.219	0.275	0.292	0.341	0.357
random	q^2 -GRS ^c	1.0	0.209	0.405	0.512	0.567	0.603	0.612
worst	q^2 -GRS ^c	1.0	0.225	0.358	0.405	0.459	0.511	0.525
best	q^2 -GRS ^c	1.0	0.237	0.394	0.506	0.560	0.590	0.592

^a The "best" and the "worst" orientations are defined as those that generate the highest and the lowest q^2 s, respectively. ^b The number of components. ^c q^2 cutoff = 0.

Table 7. Comparison of Conventional CoMFA and the q^2 -GRS Method as Applied to 59 HIV Protease Inhibitors (Numbers in bold represent the q^2 values for the ONC)

orientation ^a	CoMFA method	step size (Å)	q^2						
			1 ^b	2 ^b	3 ^b	4 ^b	5 ^b	6 ^a	7 ^a
random	conventional	2.0	0.436	0.610	0.682	0.703	0.714	0.732	0.735
worst	conventional	2.0	0.335	0.473	0.562	0.574	0.562	0.571	0.572
best	conventional	2.0	0.415	0.574	0.701	0.747	0.767	0.785	0.793
random	conventional	1.0	0.379	0.526	0.653	0.696	0.713	0.732	0.739
worst	conventional	1.0	0.378	0.533	0.650	0.691	0.702	0.722	0.727
best	conventional	1.0	0.384	0.528	0.653	0.692	0.709	0.732	0.740
random	q^2 -GRS ^c	1.0	0.408	0.550	0.675	0.718	0.735	0.754	0.762
worst	q^2 -GRS ^c	1.0	0.389	0.538	0.668	0.711	0.727	0.747	0.755
best	q^2 -GRS ^c	1.0	0.388	0.539	0.664	0.705	0.716	0.738	0.742

^a The "best" and the "worst" orientations are defined as those that generate the highest and the lowest q^2 s, respectively, for conventional CoMFA. ^b The number of components. ^c q^2 cutoff = 0.

Table 8. Application of the q^2 -GRS Method to Random Orientation of 20 5-HT_{1A} Receptor Ligands Using Different q^2 Cutoff Values (Numbers in bold represent the q^2 values for the ONC)

q^2 cutoff	q^2					q^2 cutoff	α^2				
	1 ^a	2 ^a	3 ^a	4 ^a	5 ^a		1 ^a	2 ^a	3 ^a	4 ^a	5 ^a
0.0	0.209	0.405	0.512	0.567	0.603	0.4 ^b	0.374	0.615	0.749	0.807	0.808
0.2 ^b	0.261	0.517	0.646	0.701	0.723	0.5 ^b	0.203	0.712	0.744	0.787	0.801
0.3 ^b	0.300	0.582	0.709	0.753	0.762						

^a The number of components. ^b Minimum $\sigma = 0.0$.

Table 9. Application of the *q*²-GRS Method to Random Orientation of 59 HIV Protease Inhibitors Using Different *q*² Cutoff Values (Numbers in bold represent the *q*² values for the ONC)

<i>q</i> ² cutoff	<i>q</i> ²								
	1 ^a	2 ^a	3 ^a	4 ^a	5 ^a	6 ^a	7 ^a	8 ^a	9 ^a
0.0	0.408	0.550	0.675	0.718	0.735	0.754	0.762	0.764	0.763
0.1	0.385	0.555	0.676	0.725	0.740	0.768	0.765	0.770	0.768
0.2	0.348	0.561	0.694	0.733	0.748	0.773	0.774	0.776	0.773
0.3	0.318	0.562	0.703	0.747	0.765	0.789	0.789	0.791	0.790
0.4	0.236	0.553	0.675	0.733	0.765	0.778	0.798	0.799	0.799
0.5	0.245	0.566	0.644	0.678	0.711	0.728	0.744	0.756	0.767

^a The number of components.

Table 10. Application of the *q*²-GRS Method to Random Orientation of 21 Steroids^a Using Different *q*² Cutoff Values (Numbers in bold represent the *q*² values for the ONC)

<i>q</i> ²	<i>q</i> ²				
	1 ^b	2 ^b	3 ^b	4 ^b	5 ^b
0.3	0.444	0.649	0.647	0.622	0.638
0.4	0.419	0.658	0.675	0.635	0.651
0.5	0.405	0.654	0.666	0.623	0.621

^a The biological activity is expressed as binding affinity to human testosterone-binding globulins. ^b The number of components.

Table 11. Application of the *q*²-GRS Method to Random Orientation of 21 Steroids^a Using Different *q*² Cutoff Values (Numbers in bold represent the *q*² values for the ONC)

<i>q</i> ² cutoff	<i>q</i> ²				
	1 ^b	2 ^b	3 ^b	4 ^b	5 ^b
0.3	0.624	0.756	0.727	0.694	0.662
0.4	0.697	0.777	0.726	0.691	0.678
0.5	0.661	0.790	0.787	0.782	0.733
0.6	0.602	0.691	0.662	0.632	0.537

^a Dependent variable: binding affinity to human corticosteroid-binding globulins. ^b The number of components.

generating a *q*² for each orientation of the aggregate. We found that the resulting *q*² values may differ by as much as 0.5*q*² units. We believe that the reason for this phenomenon lies in the core of the conventional CoMFA routine and may be explained as follows.

In the conventional CoMFA implementation, the steric and electrostatic fields, which theoretically form a continuum, are sampled on a fairly coarse grid. As a result, these fields are represented inadequately, and the results are not strictly reproducible.² Intuitively, decreasing the grid spacing may increase the adequacy of sampling as was suggested by Cramer et al.² Indeed, we report in this paper that decreasing the grid spacing from 2.0 to 1.0 Å minimizes the fluctuation in the observed *q*² values. Most probably, the reason for this phenomenon is that the decrease in grid spacing increases the number of probe atoms which in turn should raise the probability of placing the probe atoms in a region where the steric and electrostatic field changes can be best correlated with biological activity. However, as was noticed by Cramer et al.,² the increase in the number of probe atoms also increases the noise in PLS analysis and leads to a less statistically significant *q*².¹³ Furthermore, as mentioned above, decreasing the grid spacing from 2.0 to 1.0 Å decreased the highest *q*² value obtained for each dataset.

The grid orientation in CoMFA is fixed in the coordinate system of the computer; thus, every time when the orientation of the molecular aggregate is changed, the size of the grid may change but not its orientation. The orientation of the assembled molecules therefore

affects the placement of probe atoms which, in turn, influences the field sampling process. This leads to the variability of the *q*² values, mostly due to the reasons outlined above. We also noticed that the effect of variability of *q*² as a function of molecular aggregate orientation is more pronounced in the case of structurally diverse molecules, e.g., cephalotaxine esters and 5-HT_{1A} receptor ligands, than in the case of much less structurally diverse molecules, e.g., HIV protease inhibitors (cf. Tables 5 and 6 vs Table 7). This effect may be due to the fact that the pattern of probe atom placement with respect to the aligned molecules changes more dramatically when one changes the orientation of more structurally diverse molecules than it does when the dataset is comprised of structurally similar molecules.

An important feature of the conventional CoMFA routine is that it assumes equal sampling and *a priori* equal importance of all lattice points for PLS analysis whereas the final CoMFA result actually emphasizes the limited areas of three-dimensional space as important for biological activity. We have realized that the deficiencies of conventional CoMFA routine mentioned above may be effectively dealt with by eliminating from the analyses those areas of three-dimensional space where changes in steric and electronic fields do not correlate with changes in biological activity. We therefore devised the *q*²-GRS routine which eliminates those areas from the analysis based on the (low) values of the *q*² obtained for such regions individually. The major feature of this new routine is that it optimizes the region selection for the final PLS analysis. In this regard, it is intellectually analogous to the recently proposed GOLPE¹³ approach. The relative efficiency of both algorithms shall be compared using the same datasets. However, the *q*²-GRS method is substantially more straightforward, and it is implemented entirely within the SYBYL working environment. The latter feature makes the application of this routine transparent for SYBYL users.

The successful development and application of the *q*²-GRS method to several datasets illustrates several important aspects of the present and future applications of CoMFA in drug design. Our discovery that the results of conventional CoMFA are sensitive to the overall orientation of molecular aggregates on computer terminal shows that, for a given alignment, the single *q*² value obtained from standard CoMFA will most likely fall within the region of the highest frequency of *q*² (cf. Figures 2–4). For instance, the reported *q*² values for 5HT_{1A} receptor ligands and HIV protease inhibitors were 0.481⁶ and 0.778,⁷ respectively. In both cases, this values are within the highest frequency regions of the *q*² orientational distribution (cf. Figures 3 and 4,

respectively). On the other hand, the low q^2 value obtained from conventional CoMFA (which in many cases will not be reported in the literature) may not necessarily be a result of a poor alignment but may be caused merely by the poor orientation of molecular aggregate on user terminals. As we show in this paper, the reorientation of the aggregate may significantly improve the results. For instance, Taylor et al.⁶ have reported the q^2 value of 0.481 which, as we show (Table 6), is not the best value possible for their alignment.

Another important aspect of our work is that reporting the single value of q^2 and associated CoMFA fields as a result of standard CoMFA method appears inadequate. In general, scientists who use standard CoMFA routines should present the range of possible q^2 values (similar to our Figures 2–4) instead of one number. Furthermore, the presentation of associated CoMFA fields becomes ambiguous because the shape of CoMFA fields varies with the q^2 .

In summary, the new q^2 -GRS routine developed in this work generates an orientation-independent, high q^2 , generally exceeding the one obtained with the conventional CoMFA. We therefore conclude that this novel routine, which eliminates the major deficiency of conventional CoMFA method, shall be applied both to the future analyses and previously reported CoMFA studies in order to guarantee the reproducibility of CoMFA results.

Acknowledgment. We first noticed the phenomenon of nonreproducibility of q^2 values during the laboratory sessions of the introductory molecular modeling class taught by the senior author of this paper at the University of North Carolina. We therefore acknowledge the contribution of the Molecular Modeling Class of 1993 to the initiation of this research. The authors appreciate the software grant from Tripos Associates and the computer time provided by the North Carolina Supercomputing Center. The authors are very grateful to D. Patterson and R. D. Cramer, III (Tripos), for their interest to this work and helpful critique, and Dr. S. Wyrick for helpful comments. We also would like to acknowledge Drs. E. W. Taylor and C. L. Waller for

providing us with their datasets. This investigation was supported in part by the University of North Carolina Research Council Research Grant.

Supplementary Material Available: The SPL script that automatically performs the q^2 -GRS routine can be obtained from the authors (jin@gibbs.oit.unc.edu) upon request.

References

- (1) 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.
- (2) Cramer, R. D., III; DePriest, S. A.; Patterson, D. E.; Hecht, P. The Developing Practice of Comparative Molecular Field Analysis. In *3D QSAR in Drug Design: Theory, Methods, and Applications*; Kubinyi, H., Ed.; ESCOM: Leiden, 1993; pp 443–485.
- (3) Thibaut, U. Applications of CoMFA and Related 3D QSAR Approaches. In *3D QSAR in Drug Design: Theory, Methods, and Applications*; Kubinyi, H., Ed.; ESCOM: Leiden, 1993; pp 661–696.
- (4) Agarwal, A.; Pearson, P. P.; Taylor, E. W.; Li, H. B.; Dahlgren, T.; Herslof, M.; Yang, Y.; Lambert, G.; Nelson, D. L.; Regan, J. W.; Martin, A. R. Three-Dimensional Quantitative Structure-Activity Relationships of 5-HT Receptor Binding Data for Tetrahydropyridinylindole Derivatives: A Comparison of the Hansch and CoMFA Methods. *J. Med. Chem.* **1993**, *36*, 4006–4014.
- (5) Huang, M. T. Harringtonine, an Inhibitor of Initiation of Protein Biosynthesis. *Mol. Pharmacol.* **1975**, *11*, 511–519.
- (6) Taylor, E. W.; Agarwal, A. 3-D QSAR for Intrinsic Activity of 5-HT_{1A} Receptor Ligands by the Method of Comparative Molecular Field Analysis. *J. Comput. Chem.* **1993**, *14*, 237–245.
- (7) Waller, C. L.; Oprea, T. I.; Giolitti, A.; Marshall, G. R. Three-Dimensional QSAR of Human Immunodeficiency Virus (I) Protease Inhibitors. 1. A CoMFA Study Employing Experimentally-Determined Alignment Rules. *J. Med. Chem.* **1993**, *36*, 4152–4160.
- (8) The program SYBYL 6.0 is available from Tripos Associates, 1699 South Hanley Road, St. Louis, MO 63144.
- (9) Arora, S. K.; Bates, R. B.; Grady, R. A.; Germain, G.; Declercq, J. P.; Powell, R. G. *J. Org. Chem.* **1976**, *41*, 551–554.
- (10) Arora, S. K.; Bates, R. B.; Grady, R. A.; Powell, R. G. *J. Org. Chem.* **1974**, *39*, 1269–1271.
- (11) Mopac 6.0 available from Quantum Chemistry Program Exchange.
- (12) Personal Communication with Dr. David E. Patterson at Tripos Associates Inc.
- (13) 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. Rel.* **1993**, *12*, 9–20.

JM9408272