J. Mol. Model. **1998**, *4*, 276 – 283

Journal of Molecular Modeling

FULL PAPER

All-Orientation Search and All-Placement Search in Comparative Molecular Field Analysis

Renxiao Wang, Ying Gao, Liang Liu, and Luhua Lai

Institute of Physical Chemistry , Peking University, Beijing 100871, China. E-mail: arthur@ipc.pku.edu.cn

Received: 29 April 1998 / Accepted: 20 July 1998 / Published: 21 August 1998

Abstract Three sets of molecules have been used to study the conventional CoMFA procedure. For all the three test sets, the resulting q^2 values were observed to vary simply because of the change in the orientation or placement of the aligned molecules. The reason is believed to root in the imperfect sampling of the molecular field. We have introduced two new strategies, all-orientation search (AOS) and all-placement search (APS), to optimize the sampling process. By rotating and translating the molecular aggregate within the grid systematically, all the possible samplings of the molecular field are tested and subsequently the one with the highest q^2 value can be picked out. We have also demonstrated that the combined application of AOS/APS with GOLPE procedure can yield results better than the ones by using them respectively.

Keywords CoMFA, Field sampling, All-orientation search, All-placement search

Introduction

Since its advent in 1988 [1], the comparative molecular field analysis (CoMFA) has become one of the most powerful tools for three-dimensional quantitative structure-activity relationship (3D-QSAR) studies. Over these years, this approach has been widely applied to various receptors and ligands [2]. Utilization of this approach might assist pharmaceutical scientists in the design, selection, and development of potential therapeutic agents. The further enhancement of CoMFA is undoubtedly of great importance and interest.

CoMFA methodology is based on two basic assumptions: (1) at the molecular level, the interactions that occur between a receptor and its ligand which ultimately produce the biological effect are usually non-covalent in nature, and (2) a sampling of the steric and electrostatic field surrounding a set of ligands might provide the information necessary to understand their structure-activity relationships. In a standard CoMFA procedure, all molecules under investigation are first structurally aligned. Then, an evenly-spaced, rectangular grid is generated to enclose the molecular aggregate. A probe atom, e.g. sp^3 carbon with +1 charge, is placed on the grid and the steric and electrostatic interaction energies on each lattice point are calculated by using molecular mechanics. The results of the field sampling for every molecule in the dataset are input into a QSAR table for following analysis. Since this table usually has much more columns than rows, standard multiple regression is practically impossible. Instead, partial least squares (PLS) analysis is applied to deriving the final CoMFA model. A cross-vali-

Correspondence to: R. Wang

dated R^2 (q^2) usually serves as the quantitative measure of the predictivity. A CoMFA model with a q^2 value greater than 0.3 is usually considered to be significant [3].

We have noticed in the CoMFA studies of various datasets that the resulting q^2 value in a conventional CoMFA procedure may vary greatly for the same set of pre-aligned molecules. This phenomenon was first reported by Cho et al.[4] that "q^2 value is sensitive to the orientation of aligned molecules on the computer terminal and may vary with the orientation by as much as 0.5 q^2 units". They have developed a variable selection procedure, q^2-GRS, to achieve more consistent results in CoMFA studies. Kroemer and Hecht [5] also realize this problem and they have tried to obtain models of higher consistency by replacement of 6-12 steric potential by simple atom-based indicator variable. In this study, we demonstrate that not only the different orientations but also the different placements of the aligned molecules result in the variation of q^2 values. By rotating or translating the molecular aggregate systematically, we have developed two new strategies, all-orientation search and all-placement search, to find an orientation or placement that yields the highest q^2 value. Our study also shows that, by using this orientation/placement as the starting point, current variable selection procedures like GOLPE [6] could yield further optimized results.

Computational details and results

Datasets

We have tested three sets of compounds. The first set contained 21 steroids which Cramer et al. had used to develop CoMFA [1]. This set of compounds have already been modeled and are now supplemented as part of CoMFA tutorial in SYBYL. Therefore we extracted the pre-aligned structures of this set of molecules directly from SYBYL[7]. The second set contained 11 indole-based inhibitors of phospholipase A_2 [8] and the third set contained 31 growth hormone secretagogue mimics [9]. The latter two sets have been studied by conventional CoMFA in our lab before. Their 3D structures and the alignments were inherited into this study. The partial charges for all the molecules were calculated by Gasteiger-Huckel method. The bioassay data and molecular coordinates of all the three test sets are available in the supplementary material.

Conventional CoMFA

CoMFA was performed by using the QSAR module in SYBYL. All calculations were done on SGI O2/R10000 workstation. The CoMFA region was defined to extend beyond the van der Waals envelops of all molecules by 4.0 Å along the principle axes of the Cartesian coordinate system. The standard grid spacing of 2.0 Å was chosen unless as noted.

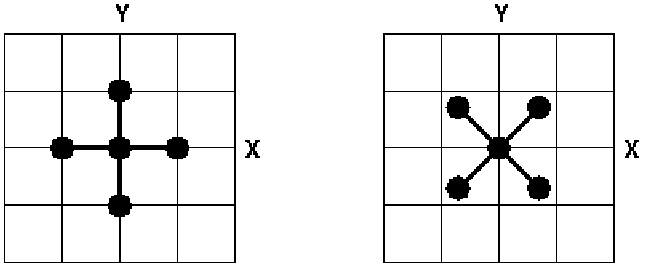


Figure 1 *Two different orientations*

The steric and electrostatic field energies were calculated using an sp^3 carbon probe atoms with +1 charge. Distance-dependent dielectric constant was adopted. Both steric and electrostatic fields were included in all CoMFA models and CoMFA standard scaling was applied. The steric and electrostatic energy cutoff were set to 30 kcal/mol and the electrostatics were dropped within the steric cutoff for each row. The standard deviation threshold for exclusion of columns from the PLS analysis was set to 2.0. The CoMFA QSAR equation was given by PLS analysis and leave-one-out cross-validation was performed to give the q^2 value.

Orientation dependence of q^2

Here, "orientation" means the direction to which the molecular aggregate is pointed on the grid. Figure 1 helps to illustrate this concept. We investigated the orientation dependence of q^2 values as follows. Starting from an arbitrary

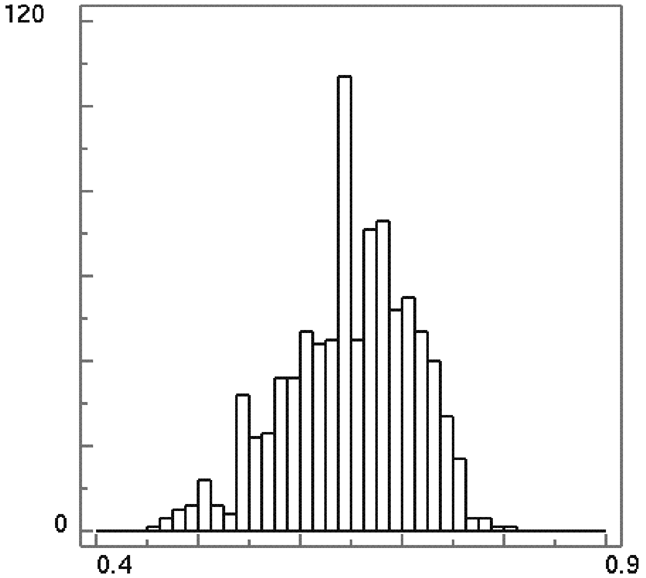


Figure 2a *Frequency distribution of q^2 values observed among all orientations for test set I (horizontal coordinate: q^2 value; vertical coordinate: population)*

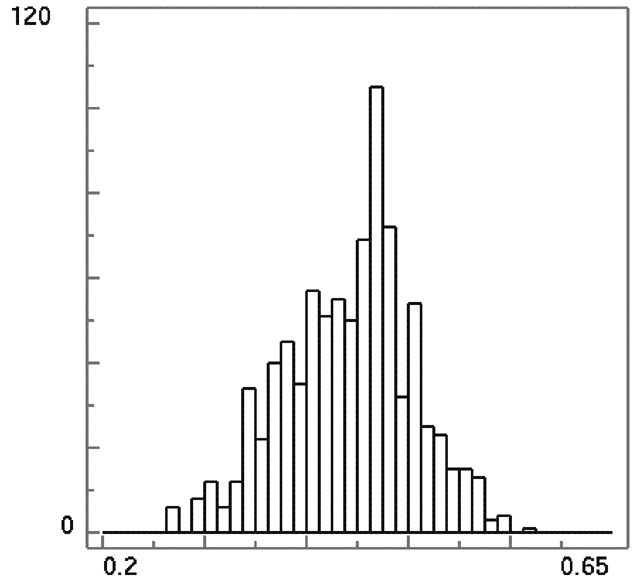


Figure 2b *Frequency distribution of q^2 values observed among all orientations for test set II*

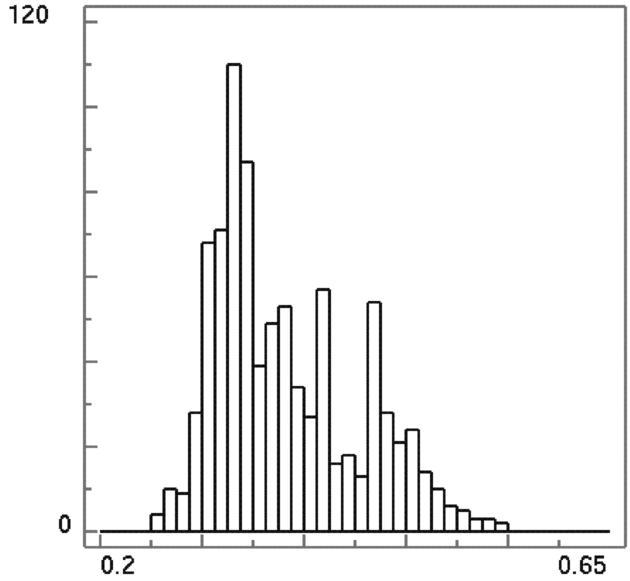


Figure 2c *Frequency distribution of q^2 values observed among all orientations for test III*

orientation, the whole aggregate was rotated around x, y, and z axes in an increment of 30° with the SYBYL STATIC ROTATE command. For each orientation, a conventional CoMFA was performed and the q^2 value was recorded. Thus, totally 12×12×6 = 864 orientations were explored for each test set. We call this strategy all-orientation search (AOS). A SYBYL SPL script was written to do AOS automatically. All the results were input into a spreadsheet and analyzed in SYBYL. The frequency distributions of q^2 values observed among all orientations for the three test sets are shown in Figure 2.

To study the influence of grid spacing on the variation of q^2 values, we performed AOS at several different grid spacings, i.e. 1.5 Å, 2.0 Å, 3.0 Å, and 4.0 Å. The results of such experiment for test set II are shown in Figure 3.

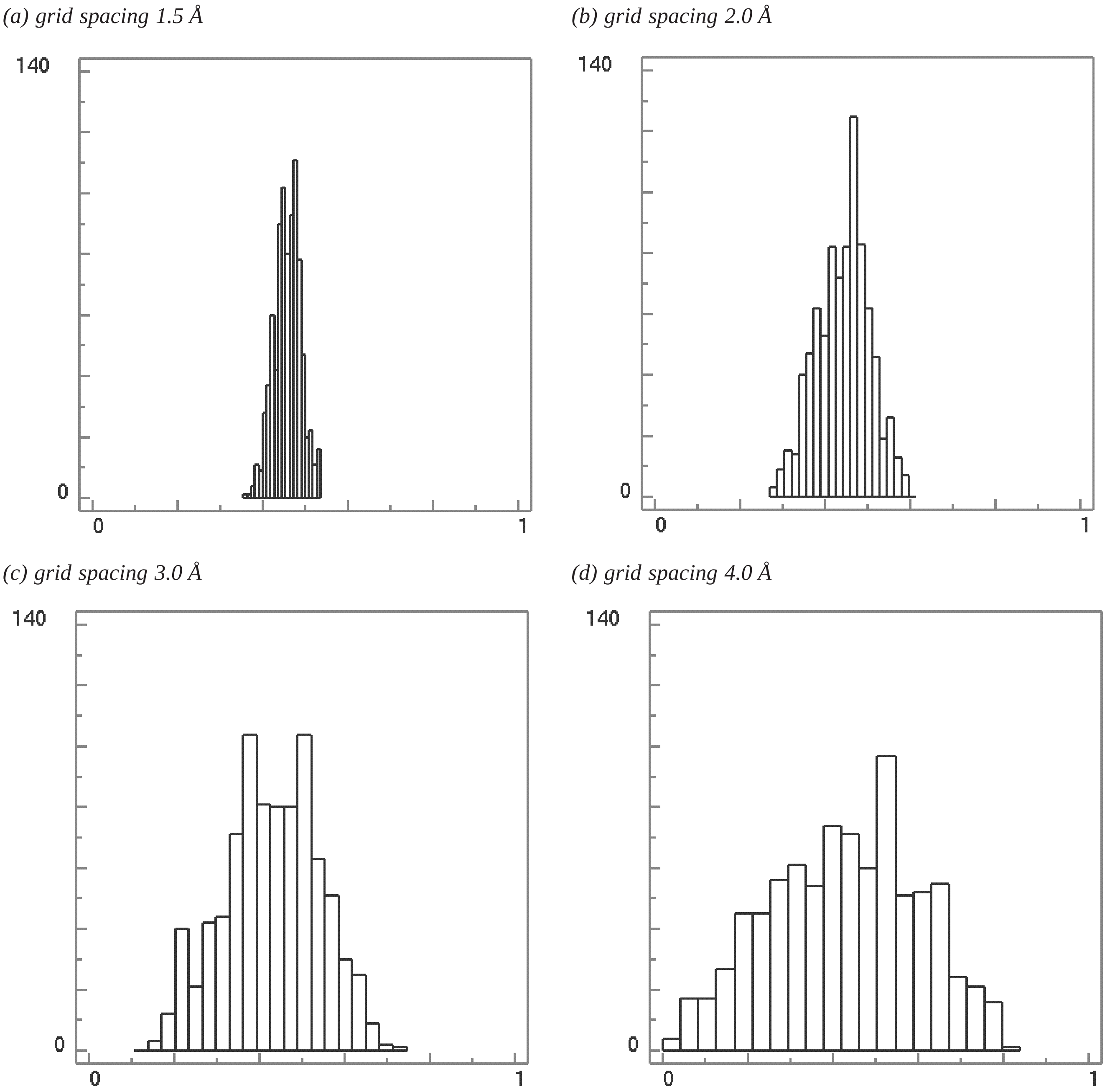
Placement dependence of q^2

Here, "placement" means the position at which the molecular aggregate is placed on the grid. Figure 4 helps to illustrate this concept. In a conventional CoMFA procedure, the CoMFA region extends beyond the van der Waals envelops of all molecules by a certain margin. The CoMFA grid is evenly spaced from one side to the other within the region in all three dimensions. If one changes the margins of the region, the whole grid will translate relatively upon the molecular aggregate. Translating the grid upon the molecular aggregate equals to translating the molecular aggregate within the grid. Therefore by this way we obtained different placements of the aggregate and we investigated the placement dependence of q^2 values as follows. Starting from the default region, we decreased the lower margin and increased the upper margin of the region simultaneously in an increment of 0.2 Å. A conventional CoMFA was performed by using the re-defined region and the q^2 value was recorded. This process ended when the overall change in the margin reached 2.0 Å since in this case the grid had overlapped on the original one. We performed the above process in all three dimensions. Therefore, 10×10×10 = 1000 different placements were explored for each test set. We call this strategy all-placement search (APS). A SYBYL script was written to perform APS automatically. The frequency distributions of q^2 values observed among all placements for the three test sets are shown in Figure 5. The variation of q^2 values observed in AOS and APS for all the three test sets are summarized in Table 1.

Combined application of AOS/APS with GOLPE

GOLPE [6] is a variable selection procedure aiming at obtaining PLS regression models with the highest prediction ability. Key steps in the procedure include a preliminary variable selection by means of D-optimal design and an iterative evaluation of the effects of individual variables on the predictivity of the model. GOLPE has been widely applied to CoMFA studies [10] and in general it can yield model with higher prediction ability than conventional CoMFA study. Recently, this procedure was supplemented by a new methodology SRD [11]. SRD builds contiguous grid-field variables that contain single pieces of chemical and statistical information into groups and thus yields models which are easier to interpret.

However, GOLPE does not account for the orientation/ placement of the molecular aggregate either. In this study,



(a) grid spacing 1.5 Å

(b) grid spacing 2.0 Å

(c) grid spacing 3.0 Å

(d) grid spacing 4.0 Å

Figure 3 *Result of all-orientation search for test set II at grid spacing (a) 1.5 Å, (b) 2.0 Å, (c) 3.0 Å, and (d) 4.0 Å (horizontal coordinate: q^2 value; vertical coordinate: population)*

we have also investigated the influence of orientation/placement on the GOLPE results. For each test set, we picked out the "best" (with the highest q^2), the "worst" (with the lowest q^2), and a random orientation/displacement based on the results of AOS and APS. Then we processed these orientation/placements by using software package GOLPE 4.0. For each GOLPE procedure, the region file and the CoMFA grid were imported from SYBYL/CoMFA. The default GOLPE settings were used for all steps, i.e. data pretreatment, D-optimal preselection, and FFD variable selection. The q^2 value of leave-one-out cross-validation was recorded. For each orientation/placement, both of the classical GOLPE and GOLPE/SRD were applied. The results of AOS-GOLPE are summarized in Table 2 and the results of APS-GOLPE are summarized in Table 3.

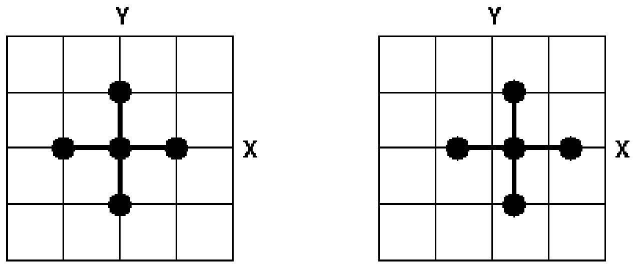


Figure 4 *Two different placements*

Discussion

As the results have shown, the q^2 values given by conventional CoMFA procedure for different orientation/placements of the molecular aggregate do vary. For all the three test sets, roughly bell-shaped frequency distributions of q^2 values are observed both in AOS and APS (see Figure 2 and Figure 5). For a given set of molecules, the q^2 value may vary as much as 0.4 units (see Table 1). Therefore, it is obvious that a conventional CoMFA which is usually performed using an arbitrary orientation/placement gives a somewhat arbitrary q^2 value. This value would probably fall into the region with the highest frequency of occurrences (the peak in the distribution). And, it is possible that the low q^2 value obtained from conventional CoMFA which often frustrates the researcher may be caused simply by the poor orientation/placement of the molecular aggregate.

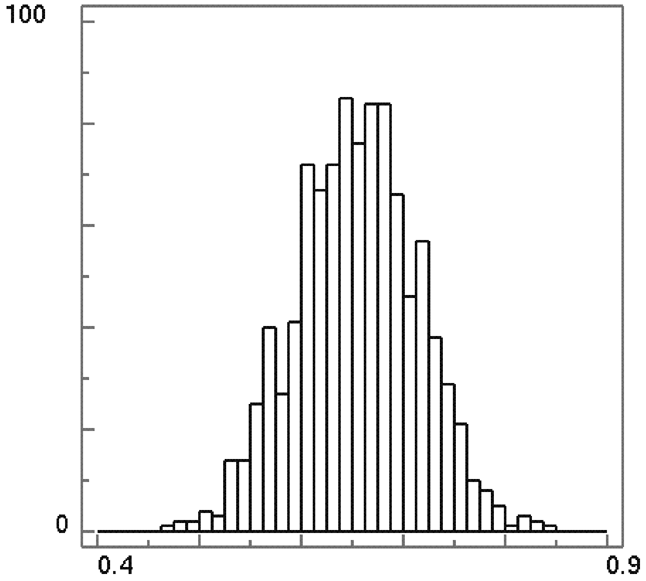


Figure 5a *Frequency distribution of q^2 values observed among all placements for test set I (horizontal coordinate: q^2 value; vertical coordinate: population)*

The reason of the variation of q^2 values roots in the field sampling routine adopted by conventional CoMFA. In such a routine, it is inevitable to use discrete grid to represent the continuous molecular field. And, the steric and electrostatic field on each lattice point are calculated with distance-sensitive functions, such as Lennard-Jones 6-12 potential. Thus

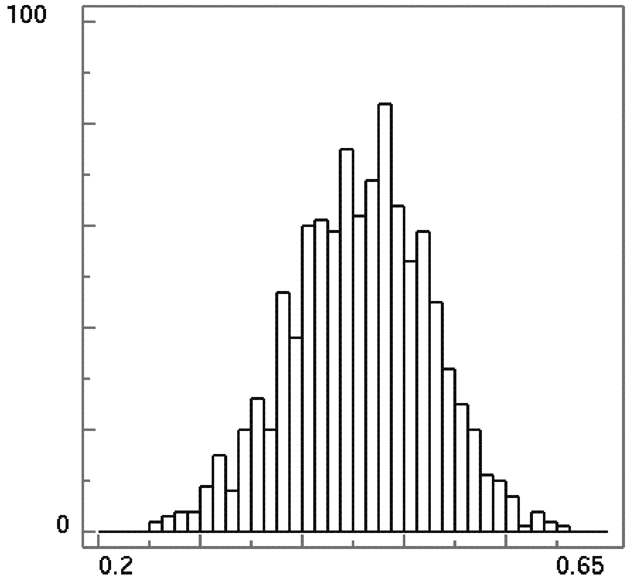


Figure 5b *Frequency distribution of q^2 values observed among all placements for test set II (horizontal coordinate: q^2 value; vertical coordinate: population)*

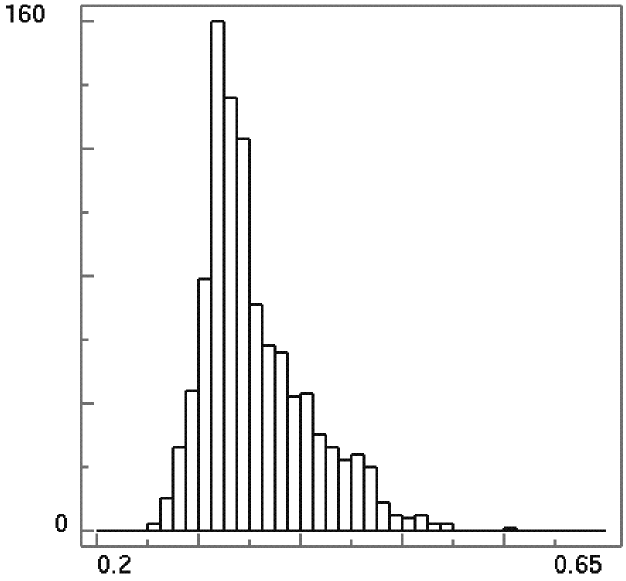


Figure 5c *Frequency distribution of q^2 values observed among all placements for test III (horizontal coordinate: q^2 value; vertical coordinate: population)*

Table 1 *Variation of q^2 values observed in AOS and APS for three test sets*

Set	All-orientation search (AOS) Best	Worst	Span [a]	All-placement search (APS) Best	Worst	Span
	Leave-one-out cross-validation R^2 (q^2)					
Set I	0.817	0.551	0.266	0.842	0.463	0.379
Set II	0.612	0.333	0.279	0.654	0.256	0.398
Set III	0.550	0.300	0.250	0.597	0.230	0.367

[a] span = best – worst.

when the orientation or placement of the molecular aggregate is changed, the same molecular field surrounding the aggregate will be mapped differently onto the grid (see Figure 1 and Figure 4). Since the tabulated data on the grid will be processed by the following PLS and yield the final model, the variation of the field sampling eventually results in the variation of q^2 values. Either rotating or translating the molecular aggregate within the grid will affect the q^2 value.

Based on the analysis above, an instant idea is that increasing the grid resolution in CoMFA studies may help to achieve more consistent results. In Figure 3, we have demonstrated the influence of different grid spacings on the results of AOS. Indeed, lowering the grid spacing from 4.0 Å to 1.5 Å narrowed the distribution of q^2 values among all orientations. However, the highest q^2 value tended to be lower when higher resolution was adopted. Similar tendency was observed in the results of APS. This is because the increase in the number of lattice points also increases the noise in PLS analysis and leads to a less statistically significant model. Thus, if not incorporated with a variable selection procedure, increasing the grid resolution in CoMFA studies will generally result in increased computation time and decreased predictivity.

In this study, we have introduced the concept of all-orientation search and all-placement search. AOS and APS are not designed just to demonstrate the variation of q^2 values but rather are straightforward strategies to optimize the field sampling routine in conventional CoMFA. In AOS/APS, all the possible samplings of the molecular field are tested by systematically rotating/translating the molecular aggregate within the grid. Among all the trials, the one yielding the highest q^2 value can be picked out. By performing CoMFA in this way, the arbitrariness in the result can be eliminated. As the data in Table 1 imply, either APS or APS will optimize the result approximately to the same extent.

AOS and APS are implemented entirely within the SYBYL working environment by using SPL scripts. This feature makes the application of these routines convenient for SYBYL users. Trying all the possibilities certainly requires more computation, but generally it is bearable. For instance, APS of test set I at 2.0 Å grid spacing lasted for about 3 hours on a SGI O2/R10000 workstation. This moderate cost is quite worthy considering that the q^2 value of this set of molecules has been improved from the originally reported 0.555 [1] to 0.842 in this study.

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 3D space as important for biological activity. Therefore, some methods, such as GOLPE, give optimized CoMFA model by variable selection procedure. The data in Table 2 and Table 3 indicate that the orientation or placement of the molecular aggregate also influences the results of GOLPE although the influence is less significant. For all test sets, GOLPE will always give the best result by using the "best" orientation/placement of the molecular aggregate. This finding clearly reveals the possibility of incorporating AOS/APS with GOLPE. In our point of view, a CoMFA procedure can be

Table 2 *Variation of q^2 values observed in the combined application of AOS and GOLPE for three test sets*

Set	Orientation	CoMFA [a]	GOLPE [b]	GOLPE/SRD [c]
		Leave-one-out cross-validation R^2 (q^2)		
Set I	best	0.817	0.923	0.914
	random	0.619	0.796	0.774
	worst	0.551	0.865	0.840
Set II	best	0.612	0.920	0.869
	random	0.403	0.823	0.820
	worst	0.333	0.750	0.741
Set III	best	0.550	0.764	0.725
	random	0.386	0.545	0.514
	worst	0.300	0.655	0.561

[a] Given by conventional CoMFA
[b] Given by D-optimal and FFD techniques
[c] Given by SRD and FFD techniques

Table 3 *Variation of q^2 values observed in the combined application of APS and GOLPE for three test sets*

Set	Placement	Leave-one-out cross-validation R^2 (q^2)		
		CoMFA [a]	**GOLPE [b]**	**GOLPE/SRD [c]**
Set I	best	0.842	0.905	0.880
	random	0.600	0.852	0.809
	worst	0.463	0.867	0.832
Set II	best	0.654	0.909	0.857
	random	0.570	0.906	0.809
	worst	0.256	0.791	0.707
Set III	best	0.597	0.728	0.661
	random	0.364	0.619	0.620
	worst	0.230	0.469	0.531

[a] Given by conventional CoMFA;
[b] Given by D-optimal and FFD techniques;
[c] Given by SRD and FFD techniques.

roughly divided into two successive stages: the first stage is to use a grid to map the molecular field (sampling) while the latter one is to process the tabulated data on the grid and derive the CoMFA model (analyzing). AOS/APS can help to find the grid which can represent the molecular field with maximum signal/noise ratio at the sampling stage; while a variable selection procedure like GOLPE can help to find the optimum relationship between the sampled molecular field and the bioactivity at the analyzing stage. Therefore, if AOS/APS is combined with GOLPE, one will get results better than the ones by using AOS/APS or GOLPE alone. This strategy is valuable for future CoMFA studies.

Conclusion

By using three sets of molecules, we have demonstrated that the result of a conventional CoMFA is sensitive to the overall orientation/placement of the aligned molecules. The q^2 value of a given set of molecules may vary as much as 0.4 units merely due to the change in orientation or placement. The reason comes from the sampling routine in which a discrete grid has to be used to represent the continuous molecular field. We have introduced all-orientation search and all-placement search to optimize the field sampling routine in CoMFA approach. By performing AOS and APS, the molecular aggregate is rotated/translated systematically and accordingly the one with the highest q^2 value can be picked out among all the individual trials. We have also combined AOS/APS with GOLPE in the CoMFA studies. It is shown that the combined application of AOS/APS and GOLPE gives better results than using them respectively. This strategy could be used routinely in the future CoMFA studies.

Acknowledgments This work was financially supported by Chinese “863” High Technology Project. We also thank Mr. Weizhong Li for the support of computing resources and the reviewers for their precious comments.

Supplementary Material Available Tables of the three test sets used in this study as well as the 3D structures of all the involved molecules in SYBYL MOL2 and PDB format. The SPL scripts for performing AOS and APS are available from the author.

References

1. Cramer, R.D.; Patterson, D.E.; Bunce, J.D. *J.Am. Chem.Soc.* **1988**, *110*, 5959-5967.
2. Cramer, R.D.; DePriest, S.A.; Patterson, D.E. In *3D QSAR in drug design: Theory, Methods, and Applications*; Kubinyi, H., Ed.; ESCOM: Leiden, 1993; pp. 443-485.
3. 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. *J.Med.Chem.* **1993**, *36*, 4006-4014.
4. Cho, S.J.; Tropsha, A. *J.Med.Chem.* **1995**, *38*, 1060-1066.
5. Kroemer, R.T.; Hecht, P. *J.Comput.-Aided Mol.Des.* **1995**, *9*, 205-212.
6. Baroni, M. *Quant. Struct.-Act. Relat.* **1993**, *12*, 9-20.
7. SYBYL ver. 6.3, Tripos Associates, St. Louis, MO, U.S.A. 1997. http://www.tripos.com/.
8. Schevitz, R.W.; Bach, N.J.; Carlson, D.G.; Chirgadze, N.Y.; Clawson, D.K.; Dillard, R.D.; Draheim, S.E.; Hartley, L.W.; Jones, N.D.; Mihelich, E.D. *Nature Struct. Biol.* **1995**, *2*, 458-465.
9. (a) Schoen, W.R.; Ok, P.; DeVita, R.J.; Pisano, J.M.; Hodges, P.; Cheng, K.; Chan, W.W.S.; Butler, B.S.; Smith, R.G.; Wyvratt, M.J.; Fisher, M.H. *Bioorgan. & Med. Chem. Lett*, **1994**, *4*, 1117-1122. (b) DeVita, R.J.; Schoen, W.R.; Ok, P.; Pisano, J.M.; Cheng, K.; Butler, B.S.; Chan, W.W.S.; Smith, R.G.; Hodges, P.; Wyvratt, M.J. *Bioorgan. & Med. Chem. Lett*, **1994**, *4*, 1807-1812. (c) DeVita, R.J.; Schoen W.R.; Fisher, M.H.; Frontier, A.J.; Pisano, J.M.; Wyvratt, M.J.; Cheng, K.; Chan, W.W.S.; Butler, B.S.; Hickey, G.J.; Jacks, T.M.; Smith, R.G. *Bioorgan. & Med. Chem. Lett*, **1994**, *4*, 2249-2254. (d) Ok,P.; Schoen, W.R.; Hodges, P.; Chan, W.W.S.; Wyvratt, M.J.; DeVita, R.J.;

Butler, B.S.; Smith, R.G.; Pisano, J.M.; Fisher, M.H. *Bioorgan. & Med. Chem. Lett*, **1994**, *4*, 2709-2714.

10. Cruciani, G.; Clementi, S.; Baroni, M. In *3D QSAR in drug design: Theory, Methods, and Applications;* Kubinyi, H., Ed.; ESCOM: Leiden, 1993; pp. 551-564.
11. Pastor, M.; Cruciani, G.; Clementi, S. *J.Med.Chem.* **1997**, *40*, 1455-1464.