International Journal of Research and Development in Pharmacy and Life Sciences
Available online at http//www.ijrdpl.com
October - November, 2012, Vol. 1, No.4, pp 167-175
ISSN: 2278-0238

Review Article

CoMFA -3D QSAR APPROCH IN DRUG DESIGN

Sandip Sen*[1], N.A.Farooqui[1] , T.S.Easwari[1], Bishwabara Roy[2]

1. IIMT College of Medical Sciences, Department of Pharmaceutical Chemistry, Meerut.
2. IIMR Institute , Meerut .

***Corresponding Author: E-mail** sandipsen2010@gmail.com

(Received: June 06, 2012; **Accepted:** August 07, 2012**)**

ABSTRACT

Progress in medicinal chemistry and in drug design depends on our ability to understand the interactions of drugs with their biological targets. Classical QSAR studies describe biological activity in terms of physicochemical properties of substituents in certain positions of the drug molecules. The detailed discussion of the present state of the art should enable scientists to further develop and improve these powerful new tools. Comparative Molecular Field Analysis (CoMFA) is a mainstream and down-to-earth 3D QSAR technique in the coverage of drug discovery and development. Even though CoMFA is remarkable for high predictive capacity, the intrinsic data-dependent characteristic still makes this methodology certainly be handicapped by noise. It's well known that the default settings in CoMFA can bring about predictive QSAR models, in the meanwhile optimized parameters was proven to provide more predictive results. Accordingly, so far numerous endeavors have been accomplished to ameliorate the CoMFA model's robustness and predictive accuracy by considering various factors, including molecular conformation and alignment, field descriptors and grid spacing. In the present article we are going to discuss the basic approaches of CoMFA in drug design.

Keywords: CoMFA, Conformation, Alignment, Fields, Grid Spacing.

INTRODUCTION

Classical QSAR correlates biological activities of drugs with physicochemical properties or indicator variables which encode certain structural features [1-5]. In addition to lipophilicity, polarizability, and electronic properties, steric parameters are also frequently used to describe the different size of substituents. In some cases, indicator variables have been attributed to differentiate racemates and active enantiomers [2,3]. However, in general, QSAR analyses consider neither the 3D structures of drugs nor their chirality. CoMFA describe 3D structure activity relationships in a quantitative manner. For this purpose, a set of molecules is first selected which will be included in the analysis. As a most important precondition, all molecules have to interact with the same kind of receptor

(or enzyme, ion channel, transporter) in the same manner, i.e., with identical binding . A sufficiently large box is positioned around the molecules sites in the same relative geometry. In the next step, a certain subgroup of molecules is selected which constitutes a training set to derive the CoMFA model. The residual molecules are considered to be a test set which independently proves the validity of the derived models . Atomic partial charges are calculated and (several) low energy conformations are generated. A pharmacophore hypothesis is derived to orient the superposition of all individual molecules and to afford a rational and consistent alignment. Carbon atom, a positively or negatively charged atom, a hydrogen bond donor or acceptor, or a lipophilic probe, are used to calculate field values in each grid point, i.e., the energy values which the probe would experience in the correponding position of the regular 3D lattice . These 'fields' correspond to tables, most often including several thousands of columns, which must be correlated with the binding affinities or with other biological activity values. PLS analysis is the most appropriate method for this purpose. Normally cross-validation is used to check the internal predictivity of the derived model. The result of the analysis corresponds to a regression equation with thousands of coefficients. Most often it is presented as a set of contour maps. These contour maps show favorable and unfavorable steric regions around the molecules as well as favorable and unfavorable regions for electropositive or electronegative substituents in certain positions . Predictions for the test set (the compounds not included in the analysis) and for other compounds can be made, either by a qualitative inspection of these contour maps or, in a quantitative manner, by calculating the fields of these molecules and by inserting the grid values into the PLS model. Despite the straightforward definition of CoMFA, there are a number of serious problems and possible pitfalls [6]. Several CoMFA modifications have been described which solve or avoid some of these problems [7]. In addition, alternatives to CoMFA were developed, e.g., comparative molecular similarity indices analysis (CoMSIA) [8] and other 3D quantitative similarity activity relationship (QSiAR) methods [9,10].

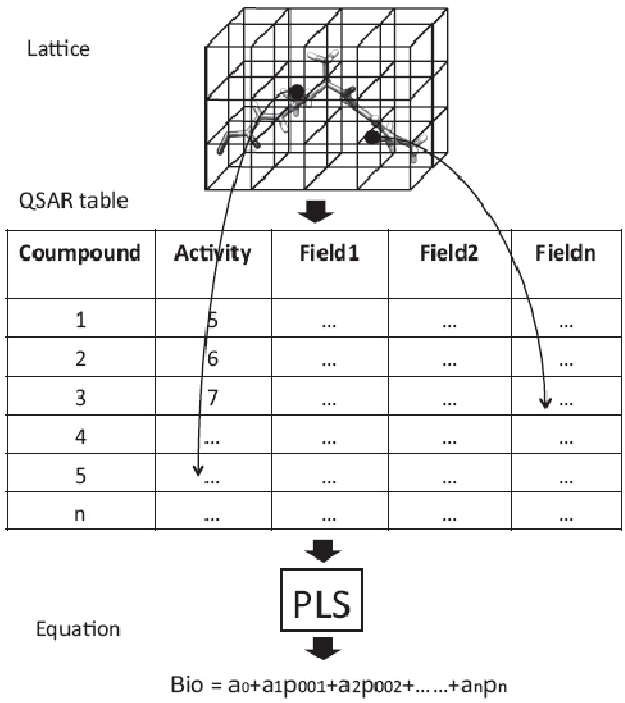

Fig. 1. The standard CoMFA process.

COMPOUND SELECTION AND SERIES OPTIMIZATION

One of the major applications of QSAR is to optimize the existing leads by structural modifications so as to improve their activity and reduce the side-effects. However there are many issues to be taken care of while selecting substituent for the modification of compounds; some of the important ones are given below: [11, 12]

- The compounds/substituent selected should be convincingly different from the existing ones, so as to minimize co linearity among the variables.
- The chosen compounds/substituent should have the properties which behave independent of each other, thereby maximizing dissimilarity and orthogonality.
- The selection should be done in such a manner so as to map the substituent (descriptor) space with minimum number of compounds.
- Synthetic accessibility/feasibility of the selected compounds should also be taken into consideration.

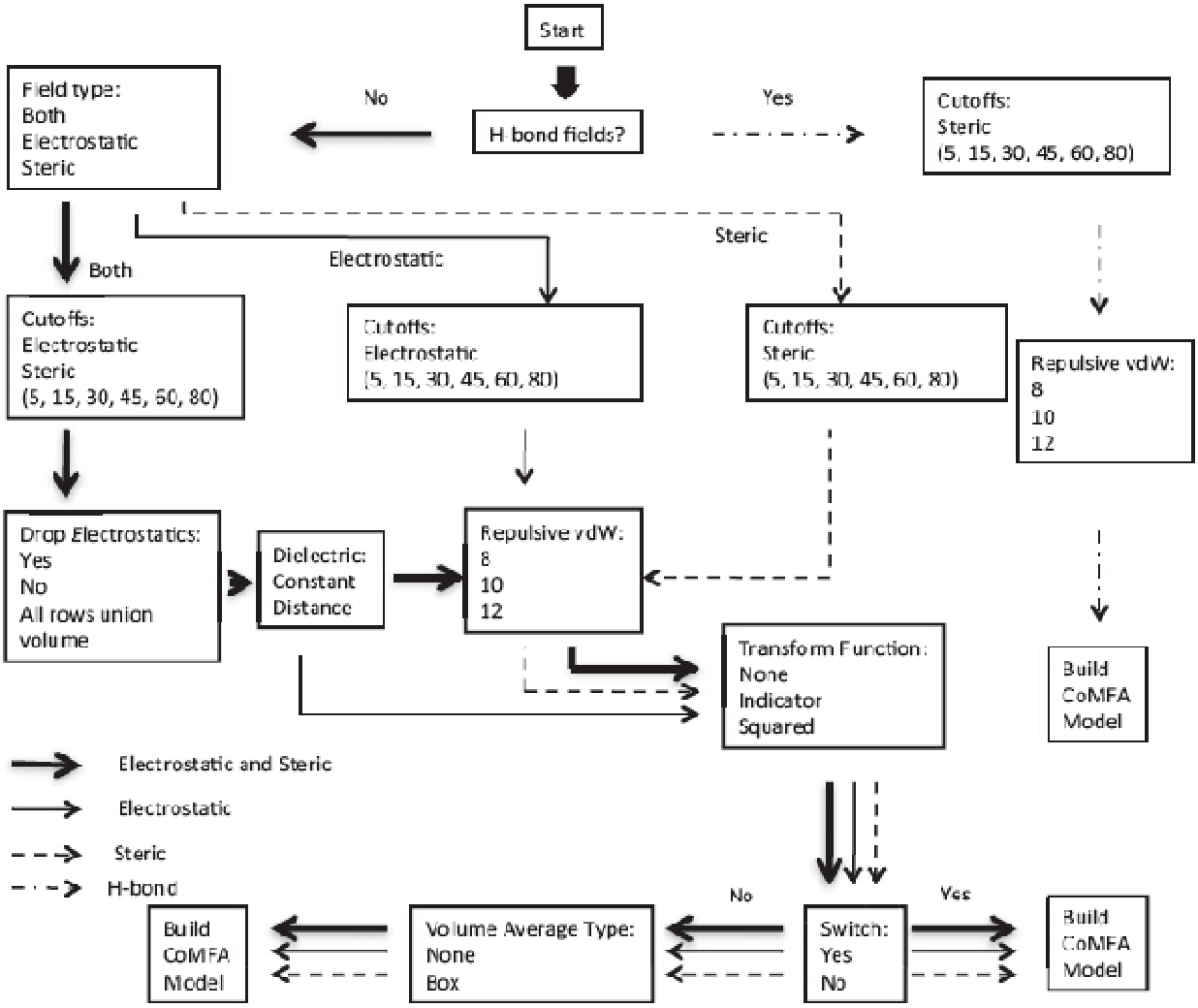

Fig. (2). Decision tree for determining possible combinations of CoMFA settings

OPTIMIZATION OF 3D-STRUCTURE OF THE MOLECULES

An important issue in 3D-QSAR is how to generate and represent the starting molecular structure for analysis. The problem can be resolved both by experimental as well as computational techniques [12]. A large number of well resolved experimentally determined crystal structures are available in databases like Cambridge Structural Database [13] and Protein Data Bank [14]. The crystal structures offer the advantage that some conformational information about the flexible molecule is included. However, molecular modeling methods are particularly useful for compounds that have not been made or cannot even exists under normal conditions. Computationally the 3D-structures can be generated by three methods:

(a) Manually by sketching the structures interactively in a 3D-computer graphics interface or from an existing 3D-structure included in the fragment libraries,

(b) Numerically by using mathematical techniques like distance geometry, quantum or molecular mechanics, and,

(c) By automatic methods that are often used for building 3D-structure databases.

After the generation of starting 3D-molecular structures, their geometries are refined by minimizing their conformational energies using theoretical calculation methods. Commonly used structure optimization techniques include:

(a) Molecular mechanics methods which usually does not explicitly consider the electronic motion, and thus are fast, reasonably accurate and can be used for very large molecules like enzymes,

(b) Quantum mechanics or *ab initio* methods which takes into account the 3D-distribution of electrons around the nuclei, and therefore are extremely accurate but time consuming, computationally intensive and cannot handle large molecules,

(c) Semi-empirical methods which are basically quantum mechanical in nature but employs an extensive use of approximations as in molecular mechanics. Generally, the molecular geometry is optimized by molecular mechanics methods, and its atomic charges are calculated mostly by semi-empirical methods or less frequently by *ab initio* methods.

CONFORMATIONAL ANALYSIS OF MOLECULES

It is a well recognized fact that each compound containing one or more single bonds is existing at each moment in many different so-called rotamers or conformers. Although small molecules may have only a single lowest energy conformation but large and flexible molecules do exists in multiple conformations at physiological conditions. Therefore, it becomes necessary to include various such conformations of the molecules in a 3D-QSAR study [12]. Depending upon the type of molecules in the study, any of the following conformational search methods can be adopted:

- Systematic search (or grid search)
- Random search
- Monte Carlo method
- Molecular dynamics method
- Simulated annealing.
- Distance geometry algorithm
- Genetic and evolutionary algorithms

DETERMINING BIOACTIVE CONFORMATIONS OF MOLECULES

Bioactive conformation refers to that conformation of the molecule when it is bound to the receptor. Intrinsic forces between the atoms in the molecule as well as extrinsic forces between the molecule and its surrounding environment significantly influence the bioactive conformation of the molecule. Reliability of any 3D-QSAR methodology depends on the determination of bioactive conformations [12, 15]. Bioactive conformations of the molecules can be obtained both by experimental as well as theoretical techniques Experimental methods for establishing bioactive

Conformations include:

- X-ray crystallography
- NMR spectroscopy

ALIGNMENT OF MOLECULES

One of the most crucial problems in most of the alignment-based 3D-QSAR methods is that their results are highly sensitive to the manner in which the bioactive conformations of all the molecules are superimposed over each other [12, 15]. In cases, where all the molecules in a data set have a common rigid core structure, molecules can be aligned easily using least-square fitting procedure. However in case of structural heterogeneity in the dataset, alignment of highly flexible molecules becomes quite difficult and time consuming. Several approaches have been proposed to superimpose the molecules as accurately as possible, some of which are as follows:

- Atom overlapping based superimposition
- Binding sites based superimposition
- Fields/pseudo fields based superimposition
- Pharmacophore based superimposition
- Multiple conformers based superimposition

CALCULATION OF MOLECULAR INTERACTION ENERGY FIELDS

After superimposition, the overlaid set of molecules is positioned in the center of a lattice or grid box, to calculate interaction energies between the ligands and different probe atoms placed at each intersection of the lattice [12, 16]. Various aspects that are required to be taken care of while calculating the interaction energies in CoMFA methodology are as follows:

- The standard size of the grid spacing is 2 Å. The grid spacing is inversely proportional to the rigorousness of calculations. As the grid spacing decreases to 1Å or less, the calculations becomes more intensive requiring much more computing time and disc storage space. The reduced grid spacing (0.5 Å) is usually employed while

extracting interaction energy fields for a reference (most active) compound during molecular superimposition based on fields, as described earlier.

- The typical size of the grid box is 3 - 4 Å larger than the union surface of the overlaid molecules. Since the electrostatic/Coulombic interactions are long-rang in nature, a larger grid box may be needed. Due to inherent correlation between electrostatic energies among lattice points in close proximity, a similar size grid box can be used for steric/van der Waals interactions.
- Many times the position of the grid box considerably influences the statistics particularly the number of components in the final CoMFA model. Generally, the initial models are developed at various locations to spot the best grid position. Two approaches have been proposed to reduce the instability. The first one suggests rotating the set of overlaid molecules in a manner that they are not parallel to any of the grid edges. The second strategy recommends substituting the field value at a lattice point by average of the field values at the vertices of a cube centered on the grid point, whose side length is two-thirds of the grid spacing.
- In CoMFA, the interaction energies are calculated using probes. The probe may be a small molecule like water, or a chemical fragment such as a methyl group. The electrostatic energies are calculated with H+ probe, whereas a sp3 hybridized carbon atom with an effective radius of 1.53 Å and a +1.0 charge is used as probe for including the steric energies. Each probe is positioned in turn at every intersection point of the lattice, and the interaction energies between the probe and each of the compounds are calculated using different molecular force fields.
- A force field is a mathematical equation, which using a combination of bond lengths, bond angles, dihedral angles, interatomic distances along with coordinates and other parameters, empirically fit the potential energy surface. Major forces encountered in the drug-receptor intermolecular interactions include electrostatic/ Coulombic, hydrogen bonding, steric/van der Waals and hydrophobic. The electrostatic and hydrogen bonding interactions are responsible for ligand-receptor specificity, whereas hydrophobic interactions generally provide the strength for binding. The most commonly employed fields in CoMFA are steric and electrostatic, which are mainly enthalpic in nature. However, many times the entropic effects, in the form of hydrophobic interactions, are also included in the CoMFA analysis. Creativity of the research and the validity of the underlying theory are the major parameters deciding the type of field to be generated and included in a CoMFA model.
- In CoMFA, the standard Lennard-Jones function is used to model the van der Waals interactions whereas electrostatic interactions are determined by the Coulomb's law. The slope of the Lennard-Jones potentials is very steep close to the van der Waals surface, as a result of which the potential energy at lattice points in the proximity of the surface changes significantly. This implies that a trivial difference in the mutual shift or conformational changes of the compounds may result in very large differences in energy values. Moreover, the Lennard-Jones and Coulombic potentials show singularities (unacceptably large values) at the atomic positions. Therefore to avoid all these problems in CoMFA, the cut-off values (± 30 kcal/mol) for steric and electrostatic energy are defined.

DATA PRETREATMENT AND SCALING

Before performing the actual chemo metric analysis in 3D-QSAR, the raw data is usually pretreated to minimize redundancy. [12] One of the common reduction methods is based on the standard deviation cut-off, in which all the energy columns with a low standard deviation are eliminated from the data, since they require longer computing time without contributing significantly to the results. Similarly several variable selection methods are available, which can be used to reduce co linearity among the descriptors thereby preventing data over-fitting and improving the prediction performance of the model. Also, in CoMFA the steric and electrostatic values are amended by using cut-offs (± 30 kcal/mol, as mentioned earlier), depending upon the position of the lattice point. Many times after pretreatment, the data

is subjected to scaling which assigns equal weight to all the descriptors and places them on a common platform for a meaningful statistical analysis. Scaling significantly improves the signal to noise ratio and also allows ranking the relative importance of individual variables. Different scaling techniques are available and can be used effectively in 3D-QSAR approaches. For example: auto scaling scales the variables to zero mean and a unit standard deviation by dividing each column with its standard deviation, block-scaling provides each category of variables with the same weight by dividing the initial auto scaling weights of descriptors in one class by the square root of the number of descriptors in that class (CoMFA standard scaling), and block-adjusted scaling which is particularly useful when other variables are included along with the energy values in the analysis. This scaling gives other variables a comparable weight to the total variables. Sometimes the pretreated data is subjected to centering by subtracting the column means from all the data. This does not change any coefficient values or comparative weights of the descriptors, but the number of significant components from PLS may be one less than from the data without centering. The method is supposed to improve the ease of interpretation and numerical stability.

MODEL GENERATION AND VALIDATION

After pretreatment and scaling of the descriptors (interaction energies and other variables, if necessary), they are correlated to the biological activities of the molecules, assuming a linear relationship between them [12,16,17]. Since the number of independent (x) variables in CoMFA is much larger than the number of compounds in the data set, the traditional linear regression analysis cannot be used to perform the fitting process. Therefore to extract a stable and best QSAR model from a range of possible solutions, the partial least-squares (PLS) technique is used. Other methods to model linear relationships include MLR, PCA, and PCR *etc.* However many times the relationship between the dependent (y) and independent (x) variables is not linear or it can't be predicted, in such cases non-linear chemo metric methods like neural networks are employed; these methods make no assumption about the relationship between the variables during training and model development. Most of these chemo metric techniques for QSAR modeling are discussed in the later sections. The most important criterion for judging the quality of a QSAR model is its ability to predict accurately not only the activities of molecules that form part of training set (internal prediction), but also of molecules not included in the development of the model (external prediction) [17]. The internal predictive capability of the model can be judged from cross-validated by techniques like leave-one-out and leave-group-out, whereas its external productivity can be evaluated by using a separate set of molecules (the test set) not included in the model development. To further assess the robustness and statistical confidence of the derived models Fischer statistics, randomization (y-scrambling) and bootstrapping analysis are also performed. All these cross validation methods have been explained in the later sections.

DISPLAY OF RESULTS

CoMFA generates an equation correlating the biological activity with the contribution of interaction energy fields at every grid point. To allow simple and easy visual interpretation, results are generally shown as coefficient (or scalar product of coefficients and standard deviation) contour plots, depicting important regions in space around the molecules where specific structural modifications significantly alters the activity [12,18]. Generally two types of contours are shown for each interaction energy field: the positive and negative contours. The contours for steric fields are shown in green (positive contours, more bulk favored) and yellow (negative contours, less bulk favored), while the electrostatic field contours are displayed in red (positive contours, electronegative substituent favored) and blue (negative contours, electropositive substituent favored) colors. In addition of contour plots, CoMFA also provides two types of plots from PLS models: score plots and loading/ weight plots. Score plots between biological activity (Yscores) and latent variables (X-scores) show relationship between the activity and the structures, whereas plots of latent variables (X-scores) display the similarity/dissimilarity between the molecules, and their clustering propensities.

DRAWBACKS AND LIMITATIONS OF CoMFA

Despite of offering many advantages over classical QSAR and good performance in various practical applications,

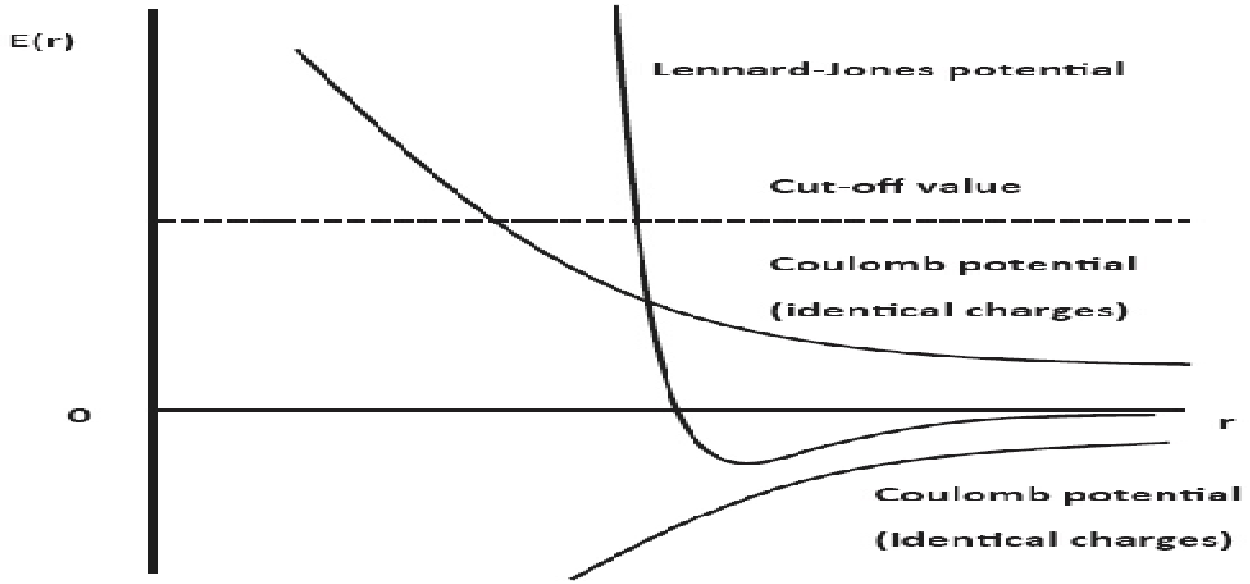

Fig. (3). Steric and electrostatic fields in CoMFA studies

CoMFA has several pitfalls and imperfections as given below [12,18,19]:

- Too many adjustable parameters like overall orientation, lattice placement, step size, probe atom type *etc.*
- Uncertainty in selection of compounds and variables.
- Fragmented contour maps with variable selection procedures.
- Hydrophobicity not well-quantified.
- Cut-off limits used.
- Low signal to noise ratio due to many useless field Variables.
- Imperfections in potential energy functions.
- Various practical problems with PLS.
- Applicable only to *in vitro* data.

APPLICATION

Since the time of its origin in 1988, numerous applications of the CoMFA method in different fields have been published. Several data sets have been investigated; the first being the binding affinity of the steroid data set [20,21] for human corticosteroid-binding globulins (CBG) and testosterone-binding globulins (TBG). Many successful endeavors of CoMFA approach in the areas of enzyme inhibition, agrochemistry (pesticides, insecticides or herbicides), physicochemistry (partition coefficients, capacity factors, enantio-separation and 13C chemical shifts), ADME and toxicity, thermodynamics and kinetics have also been exhaustively appraised in several reviews [22, 23, 24, 25].

CONCLUSION

The CoMFA technique has been developed for more than one couple of decades. Thus far a great number of CoMFA studies were performed based on this state-of-the-art approach. Scientists have also contributed everlasting and booming endeavors to im- prove the predictive quality of the CoMFA model. Herein, the practicable CoMFA descriptors, including molecular conformation, structural alignment, molecular fields, grid spacing and additional physical chemical properties, were well presented as a tutorial re- view to provide possible guidance to the further CoMFA studies. Among these crucial determinants, bioactive conformation and molecular superposition engage an essential portrayal in the CoMFA procedure, while different combination of fields and physical chemical properties results in diverse predictable levels. High predictive

models can also be realized by adjusting settings, such as energy cutoff values, lattice size and probe types. In sum, suggestions for future CoMFA studies are outlined below:

1. The initial geometries of the molecules should be in bioactive or theoretical active framework;
2. Different charge methods should be carefully considered to establish a muscular CoMFA model;
3. A reasonable molecular alignment is mandatory for a trust- worthy CoMFA model;
4. Cut-off values are needed both for the steric and electro- static energy calculation and for the PLS analysis to reduce unwanted variance;
5. Other descriptors, such as Clog P, can substantially improve the reliability of the CoMFA model. In the absence of statistic significance in CoMFA generation, those descriptors can be taken into consideration;
6. Different probe atoms could be attentively considered to ameliorate the credibility of CoMFA model;
7. The lattice location and size should be unanimously deliberated.

REFERENCES

1. Ramsden C. A., (ed.), 'Quantitative Drug Design', Vol. 4, of Comprehensive Medicinal Chemistry', eds. Hansch, C., Sammes P. G., and J. B. Taylor, Pergamon, Oxford, 1990.
2. Kubinyi H., 'QSAR: Hansch Analysis and Related Approaches', VCH, Weinheim, 1993.
3. Kubinyi H., in 'Burger's Medicinal Chemistry', Vol. I, ed. M. E. Wolff, Wiley, New York, 5th edn., 1995, pp. 497 571.
4. Hansch C. and Leo A., 'Exploring QSAR. Fundamentals and Applications in Chemistry and Biology', American Chemical Society, Washington, DC, 1995.
5. Van De H. Waterbeemd, (ed.), 'Structure Property Correlations in Drug Research', Academic Press, Austin, TX, 1996.
6. Kubinyi, H., (ed.), '3D QSAR in Drug Design. Theory, Methods and Applications', ESCOM, Leiden, 1993.
7. Kubinyi H. , Folkers G. , and Y. C. Martin, (eds.), '3D QSAR in Drug Design', Vols. 2 and 3, Kluwer, Dordrecht, 1998.
8. Hansch, C., Gao, H., Comparative QSAR: Radical reactions of benzene derivatives in chemistry and biology Chem. Rev., 1997, 97, 2995-3060.
9. Good A. C., So S.S., and Richards W.G., J. *Med. Chem.*1993, **36**, 433 438.
10. Kubinyi, H. in 'Computer-Assisted Lead Finding and Opti- mization' Proc. 11thEuropean Symp. on Quantitative Structure Activity Relationships, Lausanne, 1996, eds.
11. .Hopfinger A.J., Tokarski J.S, Three-Dimensional Quantitative Structure-Activity Relationship Analysis, In: Practical Application of Computer-Aided Drug Design, Charifson P.S., Ed., Marcel Dekker, Inc.: New York, USA; 1997, pp. 105-164.
12. Kim K.H., Comparative molecular field analysis (CoMFA). In: Molecular Similarity in Drug Design; Dean, P.M., Ed.; Blackie Academic & Professional: Glasgow, UK; 1995, pp. 291-331.
13. Allen F., The Cambridge Structural Database: a quarter of a million crystal structures and rising. Acta Crystallogr. , B, 2002, 58, 380- 388.
14. Berman H.M., Westbrook J.; Feng Z.; Gilliland G.; Bhat T.N.; Weissig H.; Shindyalov I.N.; Bourne P.E., The protein data bank. Nucleic Acids Res., 2000, 28, 235-242.
15. Akamatsu M.; Current state and perspectives of 3D-QSAR; Curr. Top. Med. Chem.; 2002, 2, 1381-1394.
16. Norinder U; Recent progress in CoMFA Methodology and Related Techniques. In: 3D QSAR in Drug Design - Recent Advances; Kubinyi H.; Folkers G.; Martin Y.C., Eds. Kluwer Academic Publishers: New York, USA, 1998, Vol. 3, pp. 24-39.
17. Richard D., Cramer III, R.D., Bunce J.D., Patterson D.E., Frank I.E., Crossvalidation, bootstrapping, and partial least squares compared with multiple regression in conventional QSAR studies. Quant. Struct.-Act. Relat. , 1988, 7, 18-25.
18. Bultinck, P., Winter H.D., Langenaeker, W.; Tollenaere J.P., Eds.; Marcel Dekker, Inc.: New York, USA, 2004, pp. 571-616.
19. Hopfinger A.J., Tokarski J.S, Three-Dimensional Quantitative Structure-Activity Relationship Analysis;, In: Practical Application of Computer-Aided Drug Design; Charifson P.S., Ed.; Marcel Dekker, Inc.: New York, USA; 1997; pp. 105-164
20. Kim K.H., List of CoMFA References. In: 3D QSAR in Drug Design - Recent Advances; Kubinyi, H., Folkers, G., Martin, Y.C.; Eds., Kluwer Academic Publishers: New York, USA; 1998, Vol. 3, pp. 316-338.
21. Coats E.A.,Kubinyi, H., Folkers, G., Martin, Y.C., Eds.; Kluwer Academic Publishers: New York, USA; 1998, Vol. 3, pp. 199-213.
22. Kim K.H., Greco G., Novellino E.; A Critical Review of Recent CoMFA Applications. In: 3D QSAR in Drug Design – Recent Advances; Kubinyi, H., Folkers, G., Martin, Y.C., Eds.; Kluwer Academic Publishers: New York, USA; 1998, Vol. 3, pp. 257-315.
23. Bordas B., Komives T. Lopata A.; Ligand-based computer-aided pesticide design; A review of applications of the CoMFA and CoMSIA methodologies, Pest Manag. Sci.; 2003, 59, 393-400.

24. Akamatsu M., Current state and perspectives of 3D-QSAR; Curr. Top. Med. Chem.; 2002, 2, 1381-1394.
25. Martin Y.C.; 3D QSAR: Current State, Scope, and Limitations. In: 3D QSAR in Drug Design - Recent Advances; Kubinyi, H., Folkers, G., Martin, Y.C., Eds.; Kluwer Academic Publishers: New York, USA; 1998, Vol. 3, pp. 3-23.

