REVIEW ARTICLE

Accepted 12 June 2001
Revised 3 May 2001
Received 12 October 2000
Indian J. Pharm. Sci., 2001, 63(5) 367-370

Comparative Molecular Field Analysis (CoMFA): A Modern Approach towards Drug Design

L. G. RATHI, SUSHIL K. KASHAW, R. K. AGRAWAL* AND P. MISHRA
Department of Pharmaceutical Sciences,
Dr. H. S. Gour University, Sagar-470 003 M.P.

The Comparative Molecular Field Analysis (CoMFA) approach permits analysis of a large number of quantitative descriptors and uses chemometric methods such as partial least squares (PLS) to correlate changes in biological activity with changes in chemical structure. It is one of the most robust modern tools for quantitative structure activity relationship studies. This review gives the introductory overview and general methodology used to carry out CoMFA.

A primary goal in any drug design strategy is to predict the biological activity of new compounds. Comparative Molecular Field Analysis (CoMFA)[1] is one of the most robust modern tools for quantitative structure activity relationship studies. The CoMFA method of three dimensional quantitative structure activity relationship (3D-QSAR) was introduced by Cramer in 1988, in which an assumption is made that at molecular level, the interaction between an inhibitor and its molecular target, which produces an observed biological effect that is usually non-covalent and the changes in biological activities or binding affinities of sample compounds correlate with changes in the steric and electrostatic fields of these molecules[2].

3D-QSAR method has been used to develop a 3D-model and pharmacophore[3] describing the structure activity relationship for a series of compounds. The CoMFA approach permits analysis of a large number of quantitative descriptors and uses chemometric methods such as partial least squares (PLS) to correlate changes in biological activity with changes in chemical structure. One of the characteristic of 3D-QSAR method is the large number of variables which are generated in order to describe the non-bonded interaction energies between one or more probes and the drug molecule[4].

*For correspondence

Characteristic features of this method are[1]; 1. representation of ligand molecules by their steric and electrostatic fields, sampled at the intersections of a three dimensional lattice. 2. A new "field fit" technique, allowing optimal mutual alignment within a series, by minimizing the root mean square (RMS) field differences between molecules. 3. Data analysis of partial least squares (PLS) using cross-validation to maximize the likelihood that the results have predictive validity and 4. Graphical representation of results as contoured three dimensional coefficient plots.

Conventional CoMFA is performed by any option of QSAR. In a standard CoMFA procedure, all molecules under investigation are first structurally aligned and the steric and electrostatic fields are sampled with probe atoms. Usually a sp^3 carbon atom with a positive unit charge (+1) is moved on a rectangular grid that encompasses the aligned molecules[5,6]. The CoMFA grid spacing should be 2.0 Å in all three dimensions within the defined region which should be extend beyond the Van-der-Waal's envelopes of all molecule by at least 4.0 Å. A CoMFA QSAR table of thousands of column is formed thereafter from the numerical values of the fields at each grid point which is subsequently analysed using special multivariate statistical analysis procedures such as PLS analysis[7] and cross-validation[8]. The optimal number of components

(ONC) in the final PLS model is determined by q^2 value, obtained from the leave-one-out cross-validation technique. A cross validated R^2 (q^2) obtained as a result of this analysis serves as a quantitative measure of the predictability of the final CoMFA model. It should be noted that q^2 is different from the cross-validated correlation coefficient in multilinear regression and a $q^2>0.3$ is already considered significant[9]. For small data sets, in order to maximize the q value and minimize the standard error of prediction, the number of components should be increased only when adding a component raised the q^2 value by 5% or more[10].

In most cases, the molecular field is developed from the quantum-chemically calculated atomic partial charges of the molecule under investigation. MNDO, AM_1, PM_3 calculated Mulliken charges have been used most widely for this purpose. The fields arising from the charge distribution on the Frontier-Molecular Orbitals (FMO-s) have also been suggested for CoMFA analysis[11].

The CoMFA method of 3D-QSAR is carried out on following lines:

Biological data:

A series of compounds with their structure and biological activity values form the training set. The biological activity of the ligands is a critical input for QSAR studies. For 3D field based QSAR methods such as CoMFA, accurate biological data from the enzyme or receptor binding assay is required.

Molecular modeling:

First a group of compounds having a common pharmacophore is selected. Then 3D structures of reasonable conformations are generated. The geometries of the compounds are modeled with the standard bond lengths and bond angles using the molecular modeling package. The conformations of the molecules must be generated from the systematic search of all the rotatable bonds with a uniform increment. All rotatable bonds are searched with uniform increase at 15° from 0 to 360°. Some conformers may be rejected during the search due to non-bonded steric interactions. The rotation of the linker chain produces many low energy conformations. The compounds having lowest energy conformation posses an extended conformation. The compounds having an extended conformation are highly active while those which have less conformation are less active. The lowest energy conformer which has an extended conformation is used for CoMFA study and for superimposition. Those compounds having less rotatable bonds can not attain a proper conformation.

Alignment rules:

The alignment, i.e. molecular conformation and orientation is one of the most sensitive inputs for CoMFA study. The alignments define the putative pharmacophore for the series of ligands. One of the active molecules serves as a template structure to derive all alignments . The molecule with the smallest number of possible conformations forms the template. There are two different approaches to superimpose the molecules.

1. Atom based alignments: In these alignments, atoms of the molecules are used for RMS fitting onto the corresponding atoms of the template structure, and 2. Shape based alignment : In these alignments, centroids rather than the exact superimposition of the atoms of the rings are used for RMS-fitting to the template structure[12].

The conformation of the compounds identified by systematic search is used for superimposition. DISCO[13] uses a clique-detection method to find superimposition's of the molecules that contain at least one conformation of each compound in the user-defined three dimensional arrangement of site points. The exact superimposition of the atoms is essential to exhibit good predictivity. Slight variations in alignment rules lead to dramatic differences in the external predictions. Many factors such as entropy, solvation and desolvation etc. can have an effect on the binding modes. To improve the alignment, a field fit procedure has been proposed. The objective of the field fit is to minimize the residual mean square differences between a fixed template field (consisting of a steric and an electrostatic field) and the fields of the molecules to be aligned. It is claimed that most often better results are obtained by using this field fit after a first, preliminary alignment of the molecules. ALADDIN[14] calculates the location of points which may be considered for the superimposition of the molecules for all low energy conformers of a series of compounds (only conformers separated by a certain distance are considered for the same molecule); such points are e.g. atoms, ring centers and projection from the molecule to H-bond, donor, acceptor and charged groups at the binding site.

Grid establishment:

Once the molecules are aligned, a grid or lattice is established which surrounds the set of analogs in potential

receptor space; although 0.2 nm (=2 Å) is the default values for the distance between the grid points, other values may be chosen; smaller distances seem desirable, but they lead to unreasonably large number of grid points. The field which a certain probe atom would experience at every grid point is calculated for each molecule. For the steric field a ($a/r^{1/2}$-b/r^6) Lennard-Jones potential is calculated, the electrostatic field is a 1/r coulomb potential the use of r^{12}-r^6 Van-der-Waal's potentials has been criticised to produce unrealistic results in case of steric overlap of the ligand and the receptor. Large positive energy values, e.g. grid points inside the molecules, are set constant at certain cut-off values to avoid unreasonable large energy values.

Normally the steric and electrostatic fields are kept separate for the ease of interpretation of performed. Other fields than those implemented in the CoMFA program have been proposed for 3D-QSAR analysis, e.g. different interaction fields calculated by the program GRID[15-18] or hydrophobic fields derived from HINT[19,20].

Partial least square (PLS) analysis:

The last step in a CoMFA study is partial least square (PLS) analysis to determine the minimal set of grid points which is necessary to explain the biological activities of the compounds. Most often good to excellent results are obtained. The predictive value of the model must be checked by cross-validation. In cross validation many PLS runs are performed in which one (leave-one out technique) or several objects are eliminated from the data sets either randomly or in a systematic manner and a model is generated from the remaining compound. This model is then used to predict the activity of the dropped compound. This process is repeated until the corresponding model predicts all the compounds. If necessary, the model is refined and the analysis is repeated until a model of high predictive ability is obtained. The PLS analysis gives the optimum number of component that are used to generate the final models without cross-validation. The results from cross-validation analysis is expressed as r^2_{cv} value which is defined as -

$$r^2_{cv} = 1 - \text{PRESS}/(\Sigma \text{ y-y mean})^2$$

where, PRESS (Predictive Residual Sum of Squares) = $(\Sigma\ y\text{-}y_{mean})^2$

The r^2_{cv} can take up values in the range from 1, suggesting a perfect model, to less than 0 where errors of prediction are greater than the error from assigning each compound mean activity of the model. The cross-validation analysis can also be performed by setting the number of cross-validation groups to 2 (leave-half-out, LHO). In this case, cross-validation groups are randomly selected and a model is derived. This is then used to predict the activity of the compounds from other groups.

The PLS variant GOLPE[2] seems to be better suited than ordinary PLS analysis because it eliminates variables not contributing to prediction in a stepwise procedure. Some recent applications in CoMFA studies confirm that the predictive power of the CoMFA model increases after reduction of the number of variables according to the GOLPE procedure. Biological activities of new compounds can be predicted by transforming the PLS results into a multiple regression equation[22].

Predictive r^2 values:

The predictive ability of each analysis is determined from a set of test compounds that are not used in the training set. The activities of these compounds are predicted from each PLS analysis. The predictive r^2 (r^2_{pred}) will be based on molecules of the test set only and is defined as -

$$r^2_{pred} = \text{SD} - \text{PRESS/SD}$$

were, SD is the sum of the squared deviation between the biological activities of the test set and mean activity of the training set molecules. PRESS is the sum of the squared deviation between predicted and actual activity values from every molecule in the test set.

CoMFA has been used for the quantitative description of effect of compounds on enzymes i.e. receptor-antagonist and agonist activities, an antiviral activities and carcinogenic and toxicological properties of compounds. CoMFA approach has been most widely used in biomedical QSAR studies; however, it has also been applied for the description of the chemical reactivity of compounds. Notably CoMFA has been used to correlate log K for the SN_2 reaction of benzyl benzene-sulfonate and p-methoxybenzylamines.

ACKNOWLEDGEMENTS

Two of the authors, (LGR and SKK) would like to express sincere thanks to the U.G.C. for awarding the Junior Research Fellowship.

REFERENCES

1. Cramer, R.D. III, Patterson, D.E. and Bunce, J.D., **J. Amer. Chem. Soc.,** 1988, 110, 5959.
2. Cho, S.J. and Tropsha, A., **J. Med. Chem.,** 1995, 38, 1060.
3. Hariprasad, V. and Kulkarni, V.M., **J. Comput. Aid. Molec. Design,** 1996, 10, 284.
4. Cruciani, G. and Watson, K.A. **J. Med. Chem.,** 1994, 37, 2589.
5. Kim, K.H., **J. Comput. Aid. Molec. Design,** 1993, 7, 71.
6. Debnath, A.K., Hansch, C., Kim, K.H. and Martin Y.C., **J. Med. Chem.,** 1993, 36, 1007.
7. Wold, S., Ruhe, A., Wold, H. and Dunn, J.W. SIAM, **J. Sci. Stat. Comput.,** 1984, 5(B), 735.
8. Cramer, R.D., Bunce, J.D. and Patterson, D.E. **Quant. Struct. Act. Relat.,** 1988, 7, 18.
9. Agarwal, A., Pearson, P.P., Taylor, W.E., Li, H.B., Dahlgren, T., Herslof, M., Yong, Y., Lambert, G., Welson, D.L., Regan, J.W. and Martin, A.R., **J. Med. Chem.,** 1993, 36, 4006.
10. Personal communication with Dr. David E. Patterson, in Cross-Validated R^2-Guided region selection for CoMFA: A simple method to achieve consistent results, Cho, S.J. and Tropsha, A., **J. Med. Chem.,** 1995, 38, 1060.
11. Waller, C.L. and Marshall, G.R., **J. Med. Chem.,** 1993, 36, 2390.
12. Kulkarni, S.S., Gediya, L.K. and Kulkarni, V.M., **Bioorgan. Med. Chem.,** 1999, 7, 1475.
13. Martin, Y.C., Bures, M.G., Danaher, E.A., Delarrer, J., Lico, I. and Paylik, P.A., **J. Comput. Aid. Molec. Design,** 1993, 7, 83.
14. Van Drie, J.H., Weininger, D. and Martin, Y.C., **J. Comput. Aid. Molec. Design,** 1989, 3, 225.
15. Goodford, P.J., **J. Med. Chem.,** 1985, 28, 849.
16. Boobbyer, D.N.A., Geodford, P.J., Meldhirmie, P.M. and Wade, R.C., **J. Med. Chem.,** 1989, 32, 1083.
17. Wade, R.C., Clark, K.J. and Goodford, P.I., **J. Med. Chem.,** 1993, 36, 140.
18. Wade, R.C. and Goodford, P.J., **J. Med. Chem.,** 1993, 36, 148.
19. Wireko, F.C., Kellog, G.E. and Abraham, D.I., **J. Med. Chem.,** 1991, 34, 758.
20. Kellog, C.E., Joshi, G.S., Abraham, D.J., **J. Med. Chem.,** 1992, 1, 444.
21. Baroni, M., Costantino, G., Crucioani, G, Rioganellit, D., Valigi, R. and Clementi, S., **Quant. Struct.-Act. Relat.,** 1993, 12, 9.
22. Norden, B., Elund, V., Johnels, D. and Wold, S., **Quant. Struct.-Act. Relat.,** 1983, 2, 73.