Chinese Chemical Letters Vol. 10, No. 9, pp. 759–762, 1999

An Enhanced Comparative Molecular Field Analysis Method Using Genetic Algorithm

Ting Jun HOU, Ning LIAO, Hong Peng LUO, Xiao Jie XU*

Department of Chemistry, Beida-Jiuyuan Molecular Design Laboratory, Peking University, Beijing 100871

Abstract: In this study, an automated conformer selection procedure using genetic algorithm (GA) has been applied in comparative molecular field analysis (CoMFA) method. Using genetic algorithm, the 3D-QSAR model is optimized to an optimal one. From the calculation results, a group of QSAR models with high predictive ability can be obtained, which is superior than using conventional CoMFA; meanwhile, the active conformers for these compounds in data set can be determined from the best model.

Keywords: Comparative molecular field analysis (CoMFA), genetic algorithm (GA), 3D-QSAR.

Comparative molecular field analysis method (CoMFA) has become one of the most powerful tools for three-dimensional quantitative structure activity relationship studies (3D-QSAR) since its advent in 1988 [1]. Over the past decade, it has been widely applied in computer-aided drug design. CoMFA is used mainly to investigate structure-activity relationships and to forecast the potency of new analogues. Moreover, it also has utility in 3D-structure searching and automated design of new ligands, the investigation of the mechanism of organic reactions, and modeling of 3D-structure of protein from sequence. So the further enhancement of CoMFA will attract a great deal of interest of many researchers.

The basic assumption for CoMFA is that the observed biological properties can be well understood or correlated with suitable sampling of the steric and electronic fields surrounding a set of ligand molecules. In conventional CoMFA, the major obstacle to generate a CoMFA model on a diverse set of compounds is to propose the bioactive conformer and superposition rule. Experimental evaluation of the relative energy of the bound conformers of small molecules supports the argument that usually it is low, but often not the lowest-energy structures. To the relatively rigid compounds, the active conformers will correspond to the lowest-energy conformations. But to the relatively flexible compounds, how to select appropriate conformers and perform alignment? This question is very difficult to answer.

For this reason, we introduce genetic algorithm into conventional CoMFA to automatically select the conformer for every compound. The quality of the models from the CoMFA can be defined as the target evaluation function, after the optimization

procedure of the genetic algorithm, the best 3D-QSAR models and corresponding active conformers can be obtained.

Methods

The idea of genetic algorithm is borrowed from genetics and natural selection. A population of "chromosomes" encoding solutions to the problem has been first generated and then it "evolves" through a process similar to biological evolution, including genetic crossover, genetic mutation and natural selection. Chromosomes encoding good partial solutions survive, reproduce and combine to generate new chromosomes which hopefully encode better solutions in the succeeding generations. Chromosomes with small fitness will gradually perish in the succeeding generations [2].

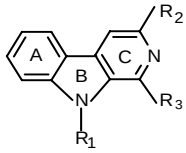
According to the genetic algorithm, the initial populations are generated by randomly selecting one conformer for each compound. Then every individual in the population is estimated according to fitness score using a CoMFA procedure. We define the fitness score to be the leave-one-out cross-validated coefficient. Once all models in the population have been rated using the fitness scores, the crossover, mutation and selection operation are repeatedly performed.

The general steps of CoMFA procedure used to evaluate every individual in the populations also follow the basic steps for conventional CoMFA procedure. First, the conformers selected for every individual in the populations are structurally aligned. In this step, a new simultaneous superposition method proposed by S. J. Kearsley was applied, which uses a quaternion based method to achieve the best fit superposition of two sets of coordinates [3]. Second, evenly-spaced, rectangular grids are generated to enclose the molecule aggregate. A probe atom (sp^3 carbon with +1 charge is used in this study) will be placed at every lattice point to measure the electrostatic (Coulombic, with 1/r dielectric) or steric (van der Waals 6-12) field by using molecular mechanics. In this paper, the parameters for calculating steric field come from Tripos60 forcefield. Third, all these steric and electrostatic values of every lattice grid are analyzed by partial least square (PLS) to get the fitness score for every individual in the population. Our enhanced CoMFA method is written with C language, and now it is embedded in Peking University Drug Design System (PUDDS) as a separate module.

The test data used in our study of 18 compounds are gained directly from literature [4]. All these compounds belong to β-carboline ligands, which possess apparent binding affinities with the benzodiazepine receptor (BzR). Before the CoMFA calculations, all these compounds were modeled using INSIGHT II package. The initial structures were firstly minimized using molecular mechanics with CVFF forcefield, the terminal condition was RMS gradient smaller than 0.001 Kcal/(Å.mol). The conformational analysis was performed for every compound, only those conformers with energy values less than the minimum energy for the most stable conformer plus 10 kcal/mol were kept. Moreover, in order to simplify the calculations, the maximum remaining conformers for each compound were defined as 20. The partial charges for every conformer were computed by CVFF forcefield. During the CoMFA calculations, eight atoms were used in fit. These included the six aromatic carbon atoms of the A ring, the indole nitrogen

moiety (B ring), and the nitrogen atom (C ring) of the β-carbolines and diindoles. The grid used in the CoMFA had a resolution of 2.0Å, and the extent of the grid border is about 2Å. For this data set, 50 populations were used. The genetic operator was applied until the total fitness score of the populations could not be improved over a period of 30 evolution operations. The convergence criterion was met after 600 operations for 2 components.

Table 1. The structures, experimental and predicted biological data for these compounds in the tested dataset.



No.	R_1	R_2	R_3	Nc [a]	Conf [b]	PIC_{50} (Actual)	PIC_{50}_1 [c] (Predicted)	PIC_{50}_2 [d] (Predicted)
1	CO_2CH_3	H	H	2	1	0.70	1.28	0.98
2	$CO_2CH_2CH_3$	H	H	6	3	0.70	1.42	1.20
3	N=C=S	H	H	1	1	0.90	1.65	1.50
4	OCH_2CH_3	H	H	6	5	1.38	1.53	1.74
5	OCH_3	H	H	6	1	2.70	2.08	1.87
6	$O(CH_2)_3CH_3$	H	H	20	16	1.99	1.81	1.67
7	OCH_3	H	H	2	2	2.09	2.05	1.91
8	$O(CH_2)_2CH_3$	H	H	20	12	1.04	1.57	1.74
9	$CO(CH_2)_2CH_3$	H	H	14	12	0.45	0.53	0.89
10	$(CH_2)_3CH_3$	H	H	18	18	2.39	2.38	2.23
11	H	H	H	1	1	3.21	2.49	2.76
12	$CO_2C(CH_3)_3$	H	H	18	1	1.00	1.37	1.20
13	Cl	H	H	1	1	1.65	1.90	2.10
14	NO_2	H	H	1	1	2.10	1.45	1.74
15	$CO_2CH_2C(CH_3)_3$	H	H	8	1	2.88	2.24	2.04
16	NCO_2CH_3	H	CH_2CH_3	4	3	3.88	4.52	4.33
17	H	H	CH_2CH_3	4	2	5.40	5.73	5.66
18	H	H	CH_3	1	1	4.09	3.44	3.75

[a]Nc represents the total conformers for every compounds, in the calculations, the conformers are ranked by their total energy.
[b]Conf represents the number of conformer used in the best model.
[c]PIC_{50}_1 represents the predicted value by the model using the lowest-energy conformers.
[d]PIC_{50}_2 represents the predicted value by the best model from the GA optimization.

Results and Discussion

Before formal calculations, a CoMFA procedure should firstly be performed to determine the optimal components used in the generation of the QSAR models. In this stage, the lowest-energy conformers for the 18 studied compounds were selected, and a conventional CoMFA calculation combined with leave-one-out crossvalidation was preformed from 1 component to 10 components. The results revealed that 2 components possessed the best predictive ability. Its q^2 is equal to 0.63, larger than 0.52 using 1 component and –0.05 using 3 components.

Table 1 showed the best model after GA optimization, it is very evident that CoMFA combined with genetic algorithm allows the construction of models superior to standard techniques. The q^2 of the best model is 0.73, but that value of the model using the lowest-energy conformers is only 0.63. It means that the best model after GA optimization is more predictive than that model using the lowest-energy conformers. After analysis for these models from populations, it can be found that for a given set of molecules, the q^2 value may vary within a very large region. So, in conventional CoMFA, it is possible that low q^2 which often frustrates the researcher may be only caused by the inappropriate conformer selections. In most cases, the CoMFA model with the lowest-energy conformers is not the best model, because the real active conformers maybe do not correspond to the lowest-energy structures. From **Table 1**, it is obvious that to this best CoMFA model, only 9 compounds prefer to the lowest-energy conformers; the other compounds adopt higher energetic structures. So from the best model, we may get the active conformers for the studied compounds.

Acknowledgment

This work is supported by the National Natural Science Foundation of China.

References

1. R. D. Cramer, D. E. Patterson, Bunce, J. D., *J. Am. Chem. Soc.*, **1988**, *110*, 5959.
2. Z. Michalewicz, *Genetic Algorithm+Data structure=Evolution Programs*, **1994**.
3. S. J. Kearsley, *J. Compt. Chem.*, **1990**,*11*,1187.
4. S. A. Mechael, *et. al., J. Med. Chem.*, **1990**, *33*, 2343.

Received 4 may 1999