Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computational Biology and Chemistry
Computational Biology and Chemistry 27 (2003) 211–216
www.elsevier.com/locate/compchem

Simultaneous determination of bioactive conformations and alignment rules by multi-way PLS modeling

Kiyoshi Hasegawa [a], Masamoto Arakawa [b], Kimito Funatsu [b,*]

[a] Nippon Roche Research Center, Kajiwara, Kamakura 247-8530, Japan
[b] Toyohashi University of Technology, Tenpaku, Toyohashi 441-8580, Japan

Received 30 April 2002; received in revised form 17 June 2002; accepted 27 June 2002

Abstract

In this study, we propose a new three-dimensional quantitative structure-activity relationship (3D-QSAR) method for selecting bioactive conformations and alignment rule simultaneously. The possible conformations of all molecules are generated by conformational analysis and they are superimposed on template conformer with possible alignment rules. The field variables are calculated as 3D descriptor of structures. Four-way partial least-squares (PLS) analysis is applied, and the conformations and alignment rule largely contributing to biological activity are selected. In order to demonstrate this method, the data set of benzodiazepine derivatives, antagonists of (CCK-B), was used as a test sample. As a result, appropriate conformers and alignment rule were selected and significant PLS model was obtained. The resulting final model could give the reasonable 3D coefficient contour maps. Moreover, external prediction was carried out by use of external data sample and its prediction was proved to be high enough.

Keywords: Alignment rule; Bioactive conformation; CoMFA; Multi-way PLS; 3D-QSAR

1. Introduction

Three-dimensional quantitative structure-activity relationship (3D-QSAR) and comparative molecular field analysis (CoMFA) have been widely used in the domain of chemistry. However, there are some critical problems that should be solved. A major problem of CoMFA and most other 3D-QSAR methodologies is that the results are dependent on a chosen bioactive conformation and an alignment rule (Todeschini and Gramatica, 1998). Experimentally, bioactive conformation of ligand can be obtained from structural determination of the ligand-receptor complex by X-ray crystallography. In this situation, the field-fit procedure may be useful to define bioactive conformations of other derivatives. When no crystal structures are available, that situation is often encountered in real drug design, bioactive conformations have to be studied theoretically.

Recently, we have proposed a novel method with three-way array data (Multi-way array, 2002) for solving the conformation problem in 3D-QSAR studies (Hasegawa et al., 1999, 2000). In this method, three-way partial least-squares (PLS) was used to select bioactive conformation from theoretically possible conformations. Each dimension of three-way array data corresponds to sample compounds, CoMFA field variables and conformations, respectively. The significant three-way PLS model was obtained and reasonable conformation could be selected from the regression coefficient of the three-way PLS model (Hasegawa et al., 1999, 2000).

In this study, we propose the extended method for selecting bioactive conformations and alignment rule simultaneously. The possible 3D conformations of all molecules are generated by conformational analysis and they are characterized by field variables of CoMFA. Four-way array for four-way PLS analysis is created by

* Corresponding author
E-mail address: funatsu@tutkie.tut.ac.jp (K. Funatsu).

1476-9271/02/$ - see front matter
PII: S0097-8485(02)00055-4

similarity criterion, and conformation/alignment rule largely contributing to biological activity is selected. In order to demonstrate the general utility, the data set of benzodiazepine derivatives, antagonists of cholecystokinin-B (CCK-B), is used as a test sample. As a result, robust PLS model and reasonable 3D coefficient contour maps were obtained. Moreover, external prediction was carried out by use of external data sample and its prediction was proved to be high enough.

2. Materials and methods

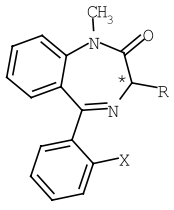
2.1. Data set

The series of 16 CCK-B antagonists reported in the literature (Bureau, 1994) was used as a test data set. CCK is a gastrointestinal hormone peptide, and has been implicated as neuron-transmitter or neuron-modulator. CCK-B is a subtype of CCK receptors and CCK-B antagonists are expected for the treatment of panic disorder and anxiety disorder. Some benzodiazepine derivatives are known as an antagonist of CCK-B receptor. Chemical structures of benzodiazepine derivatives and values of $\log(1/IC_{50})$ are listed in Table 1. Since the structures of benzodiazepine are highly flexible, the determination of bioactive conformations and alignment rules affects the performance of QSAR model.

2.2. Molecular modeling

Three dimensional structure of each compound was built up from fragment library in SPARTAN (Wavefunction Inc., 2002), and it was fully geometry-optimized using the MMFF94. The partial atomic charges required for calculation of the electrostatic interaction in CoMFA were assigned using the Gasteiger method. The energy-minimized structure was subjected to conformational analysis implemented in SPARTAN. Conformational analysis was accomplished through the stochastic Boltzmann jump procedure. The cycle, in which Boltzmann jump procedure followed by the energy-minimization, was repeated to save many types of conformation into a history file. Cluster analysis was applied to conformations in the history file. The conformations with the root mean square (RMS) deviations greater than a certain value were selected as the unique conformations. They were used for construction of subsequent four-way variable array.

Table 1
Benzodiazepine derivatives and their antagonistic activities



Compounds	X	R	Stereo	Log ($1/IC_{50}$)
1	H	NHCO-2-indolyl	S	0.61
2	F	NHCOPh-*p*-Cl	S	−0.46
3	H	NHCOPh-*p*-*t*-Bu	S	−0.89
4	F	NHCOPh-*m*-I	S	−0.23
5	F	NHCOPh-*m*-I	R	−0.37
6	H	NHCONHPh-*p*-Cl	R	2.25
7	H	NHCONHPh-*p*-Cl	S	0.38
8	H	NHCONHPh-*m*-CH_3	R	2.69
9	H	NHCONHPh-*m*-CH_3	S	0.82
10	F	NHCOPh-*p*-Cl	R	−1.04
11	H	NHCOCH=CHPh(E)	S	0.19
12	F	NHCOPh-*p*-Br	S	−0.59
13	H	NHCO-cyclohexyl	S	−2.00
14	H	NHCO-isopropyl	S	−2.00
15	H	$NHCOCH_2$Ph-*m*-OCH_3	S	−0.32
16	H	$NHCOCH_2$Ph-*m*-OCH_3	R	−0.17

2.3. CoMFA

CoMFA field variables were used as 3D structural descriptor of each conformer (Tripos Inc., 2002). The steric (Lennard–Jones) and electrostatic (Coulombic) field variables were calculated at the grid points surrounding molecule using a sp^3-carbon probe atom with a charge of +1. The steric and electrostatic field values were truncated to 30 kcal mol^{-1} in order to avoid infinity of energy values inside molecule. The grid spacing was 2.0 Å. In all three dimensions within a defined box, which was extended beyond the van der Waals envelopes of all conformations by at least 4.0 Å. From this setup, the total number of 3D grid points is 2548 and then the total number of field variables becomes 5096.

2.4. Four-way variable array

Four-way variable array for four-way PLS was constructed from sample compounds, CoMFA field variables, conformations and alignment rules. In order to generate the four-way variable array, the 14 unique conformers of sample eight that has the best activity value were used as template conformers. All conformers of all other molecules were superimposed onto each template conformer with three alignment rules described in Fig. 1. Then the closest conformer based on the following similarity criterion was used in each case of a template conformer and an alignment rule. The similarity criterion between molecule A and B is defined in next equation.

$$S_{AB} = \frac{A_{st}^{1} B_{st}}{||A_{st}||||B_{st}||} + \frac{A_{el}^{1} B_{el}}{||A_{el}||||B_{el}||} \quad (1)$$

here, A_{st} is the steric part and A_{el} is the electrostatic part of CoMFA field variables of molecule A. B_{st} and B_{el} are the steric and electrostatic parts of molecule B. The symbol $||\cdot||$ is the norm of vector. The resulting size of the four-way array is 16 (number of compounds) *5096 (number of CoMFA field variables) *14 (number of template conformers) *three (number of alignment rules).

2.5. Four-way PLS

PLS is a method for building linear regression models between independent variable matrix X and dependent variables vector y or matrix Y (Geladi and Kowalski, 1986). In PLS algorithm, latent variable t and u are derived from independent and dependent variable matrix, respectively, and covariance between t and u is maximized. Even in the situation that number of variables is greater than number of sample and/or that variables are high correlated each other, PLS can construct a robust model. PLS has become a standard regression method in QSAR analysis including CoMFA. Recently, Bro has proposed a multi-way PLS algorithm as the extension of standard PLS (Bro, 1996). When independent and/or dependent variables are multi-way array, multi-way PLS gives more stable model compared with an unfolding approach.

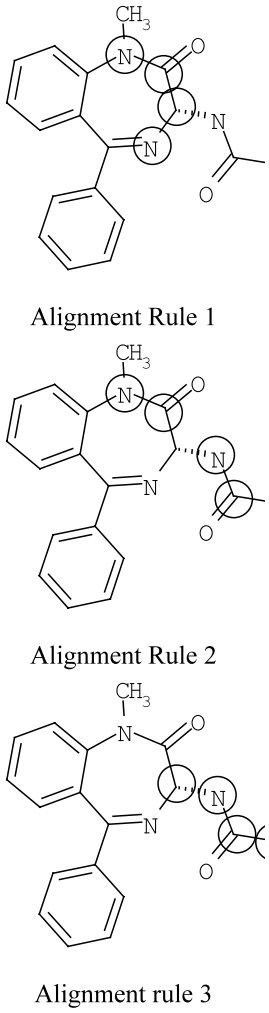


Fig. 1. Alignment rules.

The essence of four-way PLS algorithm can be expressed as follows:

$$\max_{w^J w^K w^L} [\sum_{j=1}^{J} t_i y_i | t_i = \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{l=1}^{L} x_{jkl}\, w_j^J w_k^K w_l^L \quad \text{and} \quad |w^J| = ||w^K|| = ||w^L|| = 1|] \quad (2)$$

where w^J, w^K, w^L are weight vectors of second, third and fourth dimensions, respectively. X (I*J*K*L) is four-way independent variable array and y (I*1) is dependent variable vector, that must be centered in the direction of I. Like a traditional PLS algorithm, covariance between y and score vector t is maximized

at each component, and these models are subtracted from X and y, then the following component is built from residues. After the determination of weight vectors in all components, the regression equation can be calculated from weight and score vectors.

$$y_i = \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{l=1}^{L} x_{ijkl} b_{jkl} + e_i \quad (3)$$

where B (J^*K^*L) is a regression coefficient array and e is a residue error of regression. The algorithm for obtaining regression coefficients of multi-way PLS was described in literatures (Smilde, 1997; de Jong, 1998; Smilde et al., 2000; Faber and Bro, 2002). The related programs for calculation of multi-way PLS modeling were written in MATLAB-5.3 for WINDOWS. All simulations were executed on IBM compatible PC running WINDOWS 2000.

3. Results and discussion

3.1. Four-way PLS analysis

Mean centering was applied to four-way independent variable array and dependent variable vector as a preprocessing of PLS modeling. As a result, three-component PLS model was obtained by the leave-one-out cross-validation experiment. The values of R^2 and Q^2, cross-validated R^2, were 0.900 and 0.596, respectively. WA and WC were calculated for selection of conformation and alignment rule.

$$\mathrm{WA}(l) = \sum_{i}^{\mathrm{field}} \sum_{j}^{\mathrm{conf}} (B_{ijl}^2) \Big/ \sqrt{\sum_{k}^{\mathrm{align}} \Big[\sum_{i}^{\mathrm{field}} \sum_{j}^{\mathrm{conf}} (B_{ijk}^2) \Big]^2} \quad (4)$$

$$\mathrm{WC}(l) = \sum_{i}^{\mathrm{field}} \sum_{k}^{\mathrm{align}} (B_{ilk}^2) \Big/ \sqrt{\sum_{j}^{\mathrm{conf}} \Big[\sum_{i}^{\mathrm{field}} \sum_{k}^{\mathrm{align}} (B_{ilk}^2) \Big]^2} \quad (5)$$

here, B is a regression coefficient array of four-way PLS model. WA and WC indicate a ratio of contribution to PLS model in each alignment rule and conformation. Values of WA and WC are given in Table 2. Alignment rule 2 and conformation 9 were selected for further examination (Conformation, 2002). Since alignment rule 2 provides a superposition based on both benzodiazepine core ring and peptide bond, it is suggested that their inter relationship has an important significance to describe the biological activities of derivatives.

3.2. Two-way PLS analysis

Traditional two-way PLS method was applied to the data set constructed from conformation 9 and alignment rule 2. The field variable matrix was mean-centered and then block-scaled to give the same weight for the steric and electrostatic fields. A three-component PLS model was obtained and the values of R^2 and Q^2 of the model were 0.973 and 0.738, respectively. Since the Q^2-value was high enough and it was also improved against that of the four-way PLS model (0.596 vs. 0.738), this model was regarded as the final model. The plot of calculated values by the final PLS model versus observed values is given in Fig. 2. It can be recognized that the model is robust and it has no apparent outliers.

3.3. Contour map

Figs. 3 and 4 shows the steric and electrostatic coefficient contour maps, respectively. Compound 8 having conformation 9 was used as reference for specifying 3D space. The light and dark contours

Table 2
Values of WA and WC

Alignment	WA	Conformation	WC
1	0.107	1	0.182
2	0.841	2	0.268
3	0.530	3	0.314
		4	0.187
		5	0.329
		6	0.197
		7	0.290
		8	0.177
		9	0.356
		10	0.208
		11	0.307
		12	0.310
		13	0.221
		14	0.302

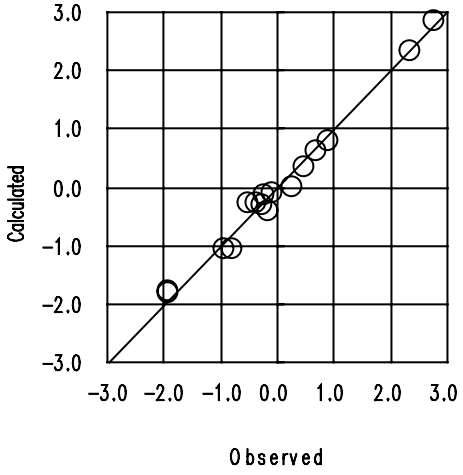


Fig. 2. Plot of observed and calculated values ($R^2 = 0.973$).

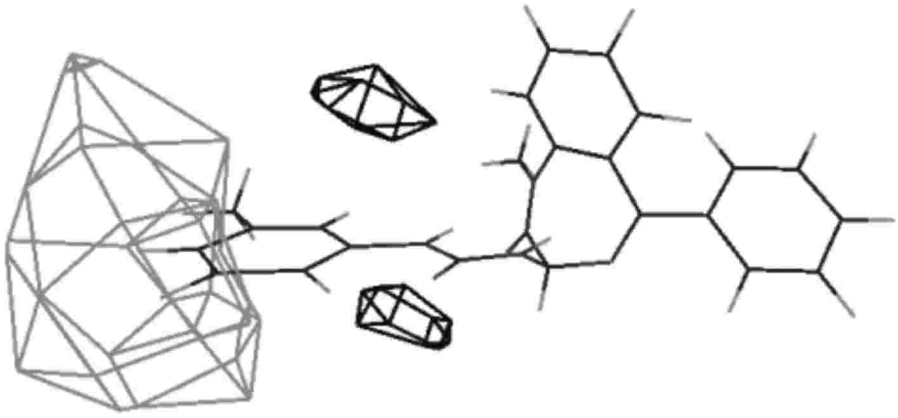
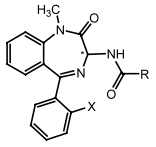


Fig. 3. Steric contour map of final CoMFA model.

Table 3
Prediction set for external validation



Compounds	X	R	Stereo	Log(1/IC$_{50}$)	Predicted
1	F	Ph-*o*-I	R	−0.58	−0.55
2	F	Ph-*o*-Br	R	−0.77	−0.55
3	F	Ph-*o*-Cl	R	−1.20	−0.82
4	H	Ph-*p*-CF_3	R	−0.52	−0.41
5	H	Ph-*p*-Br	R	−0.46	−0.48

$R^2 = 0.899$.

indicate that the signs of coefficient values are plus and minus, respectively.

In Fig. 3, the large light region is around the tip of side chain and two dark regions are in both sides of side chain. It is suggested that there is a long narrow pocket in the active site of receptor. Thus, bulky substituents in side chain decrease biological activity. On the other hand, the long side chain is favorable for potent compounds. In Fig. 4, the dark region is located near the nitrogen atom of urea bond. It is indicated that urea or peptide bond has an important role such as hydrogen-bonding donor. In addition, the light region is located on *ortho*-position of pendent benzene ring. It can be understood from Table 1 that X has to be hydrogen atom for high biological activity. Indeed almost compound that has high activity value, e.g. Number 8, 6, 9 and 1 in Table 1, has hydrogen atom on this position. From these considerations, it can be concluded that the final PLS model can give the reasonable 3D coefficient contour maps.

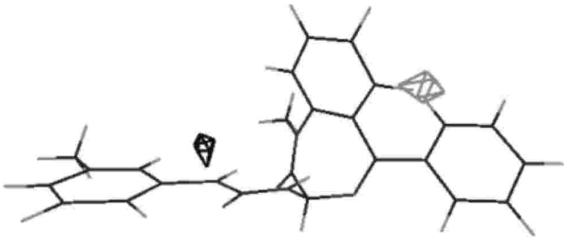


Fig. 4. Electrostatic contour map of final CoMFA model.

3.4. External validation

In order to evaluate the prediction ability of the final PLS model, external validation was carried out with test data set. The chemical structures of the test compounds and their biological activities are given in Table 3. Conformation analysis of these compounds was done in same way as described in the section of molecular modeling, and all the generated conformers were super-

imposed on the template conformer 9 of compound 8 with alignment rule 2. The biological activities of these compounds were predicted by final PLS model. The predicted values are given in last column of. For all compounds, satisfactory prediction value was obtained, and value of RMSEP (RMS error of prediction) was 0.203. Since value of RMS in training set is 0.201, it can be said that the QSAR model constructed in this study can give high prediction for benzodiazepine CCK-B antagonists.

4. Conclusions

In this paper, the new method (four-way PLS) for simultaneous selection of bioactive conformer and alignment rule was proposed. A series of benzodiazepine derivatives was used as test set for demonstration. The four-way PLS successfully indicated bioactive conformations and alignment rule. Then traditional two-way PLS analysis was carried out with the presumed conformers and alignment rule. The reasonable PLS (CoMFA) model was obtained and it could give the meaningful 3D coefficient contour maps. Moreover, the final model could give satisfactory prediction for external data set.

References

Bureau, R., Rault, S., Robba, M., 1994. Eur. J. Med. Chem. 29, 487.
Bro, R., 1996. J. Chemom. 10, 47.
Conformation, 2002. From the WC values in Table 2, conformations 3, 5, 11, 12 may become another candidates for bioactive conformations. So, we compared the statistical performance derived from conformations 3, 5, 9, 11, 12 while fixing alignment 2. Alignment 2, conformation 3: $A=4$, $R^2=0.979$, $Q^2=0.537$, RMSEP = 0.734, Alignment 2, conformation 5: $A=3$, $R^2=0.878$, $Q^2=0.456$, RMSEP = 0.400, Alignment 2, conformation 9: $A=3$, $R^2=0.973$, $Q^2=0.738$, RMSEP = 0.203, Alignment 2, conformation 11: $A=4$, $R^2=0.990$, $Q^2=0.660$, RMSEP = 0.624, Alignment 2, conformation 12: $A=3$, $R^2=0.916$, $Q^2=0.564$, RMSEP = 0.391, The result clearly indicated that alignment 2 and conformation 9 (i.e. final model) is the best combination from the statistical point of view.
de Jong, S., 1998. J. Chemom. 12, 77.
Faber, N.M., Bro, R., 2002. Chemom. Intell. Lab. Syst., in press.
Geladi, P., Kowalski, B.R., 1986. Anal. Chim. Acta 185, 1.
Hasegawa, K., Arakawa, M., Funatsu, K., 1999. Chemom. Intell. Lab. Syst. 47, 33.
Hasegawa, K., Arakawa, M., Funatsu, K., 2000. Chemom. Intell. Lab. Syst. 50, 253.
Multi-way array, 2002. The definition of multi-way data is as follows: any set of data for which the elements can be arranged as, Xijk $\ldots$ $i=1\ldots I$, $j=1\ldots J$, $k=1\ldots K$, $\ldots$ where the number of indices may vary, is a multi-way array. With only one index the array will be a one-way or first-order array (vector), and with two indices a two-way or second-order array (matrix). With three indices the data can be geometrically arranged in a box (third-order array).
Smilde, A.K., 1997. J. Chemom. 11, 367.
Smilde, A.K., Westerhuis, J.A., Boque, R., 2000. J. Chemom. 14, 301.
Todeschini, R., Gramatica, P., 1998. A new method based on the G-WHIM descriptors has been proposed and it is completely independent from the alignment. In: Kubinyi, H., Folkers, G., Martin, Y.C. (Eds.), Three Dimensional QSAR in Drug Design, vol. 2. Kluwer/ESCOM, Dordrecht, The Netherlands, pp. 355–380.
Tripos Inc., 2002. St. Louis, MO 63144, USA.
Wavefunction Inc., 2002. California, MO 92715, USA.