Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computational Biology and Chemistry
Computational Biology and Chemistry 27 (2003) 381–386
www.elsevier.com/locate/cbac

Multi-way PLS modeling of structure–activity data by incorporating electrostatic and lipophilic potentials on molecular surface

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

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

Received 1 October 2002; received in revised form 29 October 2002; accepted 5 December 2002

Abstract

We devised and elaborated a surface-based three-dimensional-quantitative structure–activity relationship (3D-QSAR) method, which had been proposed in the previous study. This approach can be applied to more general case where both the electrostatic and lipophilic potentials on molecular surface simultaneously change. The 3D coordinates of all sampling points on molecular surface are projected into a 2D map by Kohonen neural network (KNN). Each node in the map is coded by the associated molecular electrostatic potential (MEP) or molecular lipophilic potential (MLP) values. The electrostatic and lipophilic KNN maps are generated for each compound and the four-way array is constructed by collecting two KNN maps of all samples. The correlation between four-way array and biological activity is examined by four-way partial least-squares (PLS). For validation, the structure–activity data of estrogen receptor antagonists was investigated. The four-way PLS model gave the high statistics at calibration and validation stages. The coefficients of the four-way PLS model back-projected on molecular surface had a reasonable 3D distribution and it was nicely consistent with active site of the estrogen receptor which was recently made clear by X-ray crystallography.
© 2002 Elsevier Science Ltd. All rights reserved.

Keywords: 3D-QSAR; CoMFA; PLS; Kohonen neural network; Multi-way PLS; Molecular electrostatic potential; Molecular lipophilic potential

1. Introduction

A relationship between chemical structures and their biological activities has been studied in the field of quantitative structure–activity relationship (QSAR). The main purpose of QSAR is to obtain a reliable model equation with both easiness of interpretation for structure design and high predictability for new chemical structure. Three-dimensional QSAR (3D-QSAR) is a special discipline in QSAR taking into consideration of the 3D structure of molecule (Kubinyi, 1993).

In the late 1980s, a 3D-QSAR technique named comparative molecular field analysis (CoMFA) was introduced by Cramer et al. (1988). CoMFA uses the steric and electrostatic field variables that are calculated at the intersections of 3D grid surrounding molecular structure. The relationship between these 3D structural descriptors and the biological activities is modeled by partial least-squares (PLS; Geladi and Kowalski, 1986). The result of CoMFA can be displayed as 3D contour maps of PLS regression coefficient in computer graphics and the important regions for biological activity can be easily identified. In advanced CoMFA, molecular lipophilicity potential (MLP) is incorporated into CoMFA field that may cover hydrophobic interaction and entropy component (Miyashita et al., 1993). Moreover, new techniques have been invented for solving some difficulties originated from the statistical limits of PLS (Hasegawa et al., 1997; Kimura et al., 1998). Nowadays, CoMFA is widely used in 3D-QSAR studies, and a large number of applications have been reported (Kubinyi, 1998).

Although CoMFA is useful, it does not always reflect real ligand–receptor interaction. Molecular interactions between a ligand and a receptor are mainly occurred near van der Waals surface of the ligand. All grid points surrounding whole molecule in CoMFA are not important as molecular descriptors. If each molecule is

* Corresponding author. Tel.: +81-532-44-6879; fax: +81-532-47-9315.
E-mail address: funatsu@tutkie.tut.ac.jp (K. Funatsu).

1476-9271/02/$ - see front matter © 2002 Elsevier Science Ltd. All rights reserved.
doi:10.1016/S1476-9271(02)00101-9

represented by physicochemical parameters on molecular surface, more precise and realistic 3D-QSAR could be possible.

Recently, we developed a new surface-based 3D-QSAR method according to the idea described above (Hasegawa et al., 2002). In the method, the 3D coordinates of all sampling points on molecular surface are projected into a 2D map by Kohonen neural network (KNN; Devillers, 1996). The 2D map has the same number of elements for describing molecule, irrespective of the size of molecule. Each node in the map is coded by the associated molecular electrostatic potential (MEP) value. The three-way array is constructed by collecting all 2D KNN maps. The correlation between three-way array and biological activity is analyzed by three-way PLS (Bro, 1996). The three-way PLS is a robust statistical method against ill noise in data and it can keep the important neighboring relationship between nodes in KNN map. Our new method was applied to 25 dopamine receptor antagonists and the excellent three-way PLS model with the good statistics was obtained (Hasegawa et al., 2002).

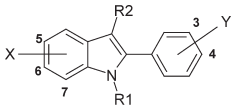
On the basis of this successful result, we reached an idea that our approach could be extended to more general case where both the electrostatic and lipophilic potentials on molecular surface simultaneously change. In this case, the electrostatic and lipophilic KNN maps are generated for each compound and the four-way array is constructed by collecting two KNN maps of all samples. The correlation between four-way array and biological activity is examined by four-way PLS. For validation, the structure–activity data of estrogen receptor antagonists was investigated. The four-way PLS model gave the good statistics at calibration and validation stages. The coefficients of the four-way PLS model back-projected onto molecular surface had a reasonable 3D distribution and it was nicely consistent with active site of the estrogen receptor recently made clear by X-ray crystallography.

2. Materials and methods

2.1. Data set

A series of 36 estrogen receptor antagonists reported in literature was used as test data set (von Angerer et al., 1984). It has been well known that the active site of the estrogen receptor is mainly composed of hydrophobic environments. The lipophilic nature of antagonist is crucial for antagonistic activity and this data set is a good example for validation of our approach. The logarithm value of relative binding affinity (RBA) was used as biological activity. RBA is given as the ratio of the molar concentration of 17β-estradiol and 2-phenyl indole required for decreasing the receptor bound radioactivity by 50%, multiplied by 100. The binding affinities of 2-phenyl indoles for the estrogen receptor were measured by a competitive binding assay with 17β-[^{3}H] estradiol. Chemical structures of 2-phenyl indoles and their values of log(RBA) were listed in Table 1. It has been believed that the estrogen receptor is a good target for treatment of advanced breast cancer. The indole binds to the estrogen receptor competing with a substrate 'estradiol'. This binding prevents the growth of advanced breast cancer because the hormonal function of estardiol is related to the differentiation of cancer cell.

Table 1
Chemical structures of 2-phenyl indoles and their antagonistic activities



X, Y=OH

Number	R1	R2	*X*	*Y*	RBA	log(RBA)
1	H	H	6	4	0.01	−4.61
2	H	CH_3	6	4	0.06	−2.81
3	H	C_2H_5	6	4	0.13	−2.04
4	H	H	5	4	0.01	−4.61
5	H	CH_3	5	4	0.06	−2.81
6	CH_3	H	6	4	3.80	1.34
7	C_2H_5	H	6	4	16.0	2.77
8	C_3H_7	H	6	4	8.50	2.14
9	C_4H_9	H	6	4	4.30	1.46
10	CH_3	CH_3	6	4	10.0	2.30
11	C_2H_5	CH_3	6	4	33.0	3.50
12	C_3H_7	CH_3	6	4	13.0	2.56
13	*i*-C_3H_7	CH_3	6	4	13.0	2.56
14	CH_3	C_2H_5	6	4	5.90	1.78
15	C_2H_5	C_2H_5	6	4	21.0	3.04
16	C_3H_7	C_2H_5	6	4	19.0	2.94
17	CH_3	H	5	4	0.80	−0.22
18	C_2H_5	H	5	4	5.80	1.76
19	C_3H_7	H	5	4	18.0	2.89
20	CH_3	CH_3	5	4	4.60	1.53
21	C_2H_5	CH_3	5	4	9.50	2.25
22	C_3H_7	CH_3	5	4	16.0	2.77
23	*i*-C_3H_7	CH_3	5	4	3.50	1.25
24	C_4H_9	CH_3	5	4	4.60	1.53
25	C_5H_{11}	CH_3	5	4	2.30	0.83
26	C_2H_5	C_2H_5	5	4	23.0	3.14
27	C_3H_7	C_3H_7	5	4	1.70	0.53
28	C_2H_5	CH_3	7	4	0.02	−3.91
29	C_2H_5	H	6	3	1.70	0.53
30	CH_3	CH_3	6	3	0.55	−0.60
31	C_2H_5	CH_3	6	3	3.00	1.10
32	C_3H_7	CH_3	6	3	3.50	1.25
33	C_2H_5	H	5	3	1.70	0.53
34	CH_3	CH_3	5	3	0.60	−0.51
35	C_2H_5	CH_3	5	3	2.20	0.79
36	C_3H_7	CH_3	5	3	7.40	2.00

2.2. Molecular modeling

The 3D structure of each compound was built up from the fragment library in SPARTAN (SPARTAN, 2002), and it was fully geometry-optimized at the PM3 level. The energy-minimized structure was subjected to conformational analysis implemented in SPARTAN. Conformational analysis was carried out through systematic conformation option. The systematic conformation option means all torsion bonds in molecule are rotated according to empirical rule (60 increments in the case of sp^3–sp^3 bond). A global energy-minimum conformation of each compound was selected for superimposition. The indole ring was used as the fitting points for superimposition of molecular structures. The MEP value on van der Waals surface was calculated using the electrostatic potential (ESP) suiting method in SPARTAN. The molecular lipophilic potential (MLP) value was calculated on the same point as MEP according to Furet's empirical equation (Furet et al., 1988). The distance between sampling points was set to be 0.5 Å and approximately 40 000 points were sampled on van der Waals surface. The Cartesian coordinates and the associated MEP and MLP values were exported to a text file for the next KNN training.

2.3. KNN

KNN is based on the idea that human brain tends to compress and organize sensory data spontaneously. KNN can be used to generate a projection of objects from a higher dimensional space onto a two dimensional space. In other words, this method enables a decrease in dimension while conserving a topology of the information as much as possible (Devillers, 1996).

KNN is typically made up from two layers (input and output layers). The input layer contains m neurons corresponding to m variables describing objects. The output layer is a two-dimensional geometrical arrangement of n neurons and the topology is usually defined as 'torus'. 'Torus' means that the right or top edge of map is continued to its left and low edge, respectively, and vice versa. The m neurons of the input layer are all connected to each of the n neurons of the output layer as shown in Fig. 1. The network is trained by adjustment of the connection weight in two phases, competitive learning and self-organization phases. Initially, the connection weights are set to random values. Each object in the data set is considered to be a vector $\mathbf{x}$, consisting of m values x_i; each neuron j in the output layer is characterized by a weight vector $\mathbf{w}_j$, consisting of m weights w_{ij}. The Euclidean distance, d_j, is calculated between each input vector $\mathbf{x}$ and each weight vector $\mathbf{w}_j$.

$$d_j = \sum_{i=1}^{m} (x_i - w_{ij}) \quad (1)$$

A node having the shortest distance to the input vector $\mathbf{x}$ is referred to as winner. After the winning neuron, denoted by j^*, is found, the adjustment of weight vector starts. The weight vector of the winner node is modified in order to make this node even closer to the current object:

$$w_{ij^*}(t+1) = w_{ij^*}(t) + \alpha[x_i - w_{ij^*}(t)] \quad (2)$$

where α is learning rate and t is iteration number. The weights of the neighboring nodes on the output layer are also modified to become closer to the winner and, hence, to the current object. The range of the neighboring neurons beyond the winning neuron is determined by learning area (r). Then, the same procedure is repeated for all objects. This process called as epoch is carried out repeatedly while the values of learning rate and learning area (α, r) are monotonically decreased.

In this study, KNN was trained by all points sampled on van der Waals surface. Therefore, each input neuron has three weights corresponding to three Cartesian coordinates ($m = 3$). According to the finally established weight vectors determined by KNN training, each sampling point was placed on a specific node in the output layer. After that, each output neuron was coded by the associated MEP or MLP values of the occupying sampling point. All coding nodes were collected together to define chemical descriptors that were used to correlate the estrogen receptor antagonist activities (see Section 3 and Fig. 2). Moreover, in order to compare KNN maps each other the template approach was employed (Anzali et al., 1996). A reference KNN is trained with the coordinates of sampling points of the most active compound (compound **11** in Table 1). Then, other antagonists are filtered through this network and thus produce comparative KNN maps.

The parameters of KNN experiments were defined as follows: map size $(n) = 50 \times 50$, initial learning rate $(\alpha) = 0.01$, initial learning area $(r) = 20$, training steps $= 50$. These values were determined according to the previous 3D-QSAR study (Hasegawa et al., 2002). The KNN experiments were simulated using TUT_-SOM (Toyohashi University of Technology Program for Self-Organization Map Satoh et al., 1998) on SGI workstation.

3. Four-way PLS

PLS is a method for building linear regression model between independent variable matrix X and dependent variables vector y (Geladi and Kowalski, 1986). In PLS algorithm, a latent variable t is derived from independent variable matrix and the covariance between t and y

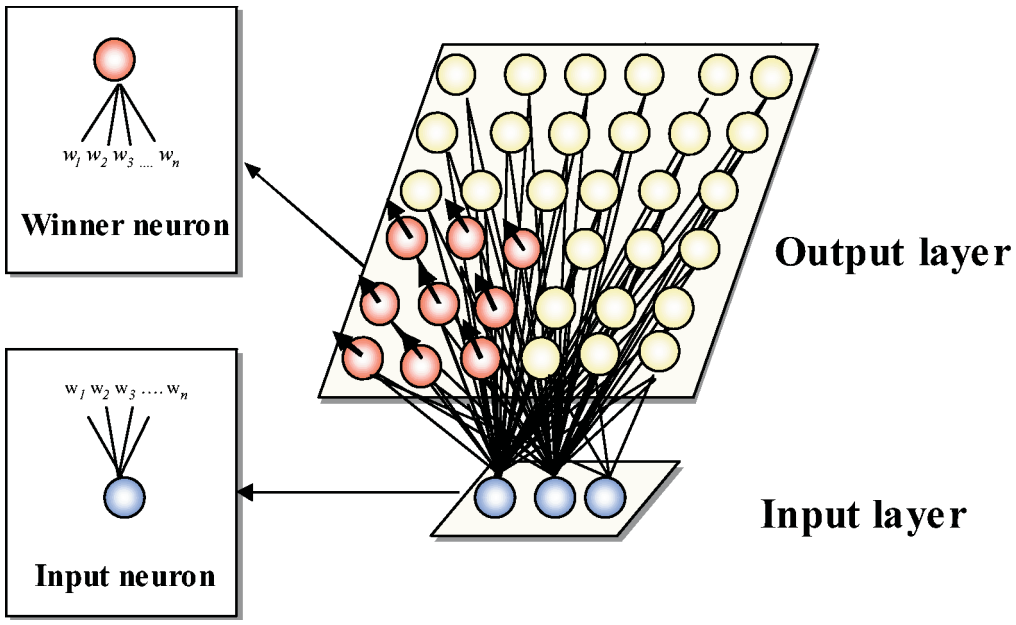


Fig. 1. Architecture of Kohonen neural network.

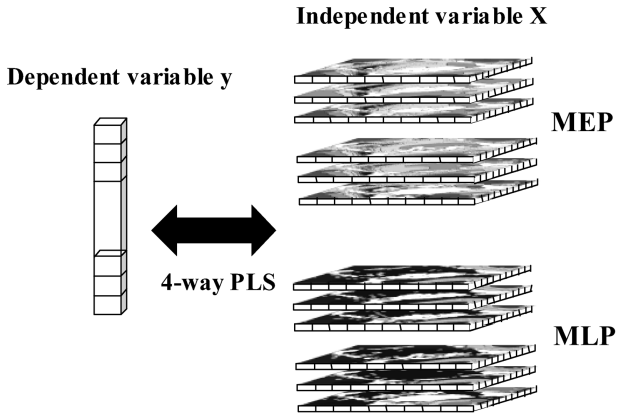


Fig. 2. Correlation between antagonistic activity and four-way array.

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. In 1996, Bro proposed a multi-way PLS algorithm as the extension of standard PLS (Bro, 1996). When independent variable is multi-way array, multi-way PLS gives more stable model compared with standard PLS.

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

$$\max_{w^J w^K w^L} \left[\sum_{i=1}^{I} t_i y_i \middle| t_i = \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{i=1}^{L} X_{ijkl} w_j^J w_k^K w_l^L \right.$$

$$\left. \text{and } \|w^j\| = \|w^k\| = \|w^L\| = 1 \right] \quad (3)$$

where w^J, w^K, w^L are weight vectors of second, third and fourth dimensions, respectively. X_{ijkl} is four-way independent variable array and y_i is dependent variable vector. As standard PLS, the covariance between y and latent variable t is maximized at each component, and this model is subtracted from X_{ijkl} and y_i, then the following component is built from residues. After determination of weight vectors in all components, the regression equation can be calculated from weights and latent variable.

$$y_i = \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{l=1}^{L} X_{ijkl} b_{jkl} + e_i \quad (4)$$

where b_{jkl} is regression coefficient array and e is residue error. The algorithm for obtaining regression coefficients from weights and latent variable was fully described in literature (Smilde, 1997). The related programs for multi-way PLS modeling were written in MATLAB 5.3 on Windows 2000 operating system.

In this study, four-way array for four-way PLS was constructed from compounds, two-dimensional KNN maps and two potentials. The resulting size of the four-way array is 36 (number of compounds) × 50 (number of first dimension in KNN maps) × 50 (number of second dimension in KNN map) × 2 (MEP and MLP). The procedure for construction of four-way array was graphically illustrated in Fig. 2.

4. Results and discussion

4.1. Four-way PLS analysis

Mean centering was applied to four-way independent variable array (36 compounds × 50 × 50 Kohonen

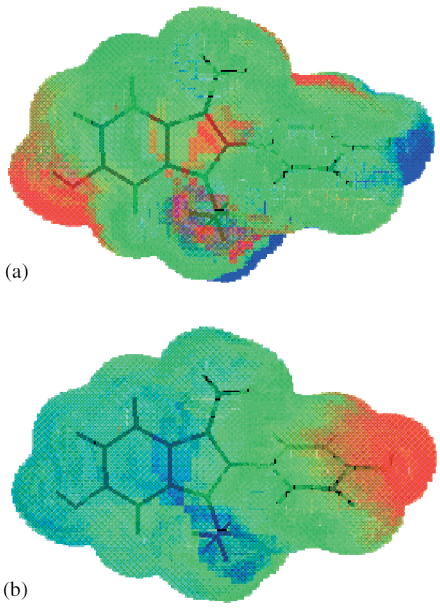


Fig. 3. Coefficient map derived from four-way PLS analysis. (a) MEP, (b) MLP. Blue and red colors represent plus and minus coefficients, respectively. Coefficient map was drawn at 0.01 level.

maps × 2 molecular potentials) and dependent variable vector (36 antagonist activities) as a preprocessing of modeling. The six-component four-way 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 0.89 and 0.80, respectively. The contributions of MEP and MLP to antagonist activity were estimated 80 and 20% from the sum of squares of coefficients.

The four-way PLS model was converted into the regression-like model and the regression coefficients were back-projected onto molecular surface. The resulting back-projection map of coefficients was shown in Fig. 3. Panel (a) and (b) show the projection maps of MEP and MLP, respectively. Coefficient map was drawn at the 0.01 level. Compound **11** with the highest antagonist activity was taken as reference for specifying 3D space. The blue and red colors represent the plus and minus coefficients, respectively.

We attempted to estimate the structural requirements for estrogen receptor antagonist from the coefficient map in Fig. 3. At first, the MEP character is discussed from panel (a). The plus coefficient region is around the *Y* position. It means that the hydrogen atom of hydroxyl group having the plus MEP value enhances the antagonist activity. The hydrogen atom of hydroxyl group may act as hydrogen-bonding donor. On the other hand, the negative coefficient region is around the *X* position. The oxygen atom of hydroxyl group having the negative MEP value at *X* improves the antagonist activity. The oxygen atom of hydroxyl group may act as hydrogen-bonding acceptor. Next, the MLP character is discussed from panel (b). The plus coefficient region is around the R1 position. The hydrophobic substituent having the plus MLP value at R1 enhances the antagonist activity. The negative coefficient region is around the *Y* position. *Y* has two opposite coloring maps derived from MEP and MLP. This suggests that the *Y* region has the higher contribution to antagonistic activity. The clear coloring region is not observed around the R2 position. It may be considered that the contribution of the R2 position is lower than that of R1.

For comparative study, standard (two-way) PLS was applied to the same data set. The comparative KNN map was transformed to the vector by the unfolding method (Polanski and Walczak, 2000). The values of R^2 and Q^2 were 0.98 and 0.79, respectively. The discrepancy between R^2 and Q^2 is higher than that of the two-way PLS model (0.19 vs. 0.09). This indicates the two-way PLS model is over-fitting and the prediction of new chemical structure is expected to be difficult. The back-projection map of coefficients was shown in Fig. 4 (0.005 level). The visual inspection of Fig. 4 demonstrates that the 3D distribution of the two-way PLS model is more messy and complex than that of the four-way PLS model due to the small fraction of coefficient values (Fig. 4 vs. Fig. 3).

4.2. Validation of model

The 3D structure of complex between 17β-estradiol and estrogen receptor was made clear by X-ray crystallography (PDB code: 1A52; Tanenbaum et al., 1998). From the X-ray crystallography, it was found that there are at least four key ligand–receptor interactions: The 3-hydroxyl group interacts with the carboxylate of Glu-

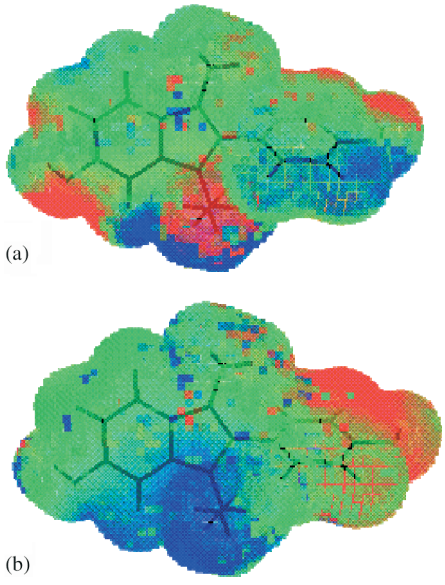


Fig. 4. Coefficient map derived from two-way PLS analysis. (a) MEP, (b) MLP. Blue and red colors represent plus and minus coefficients, respectively. Coefficient map was drawn at 0.005 level.

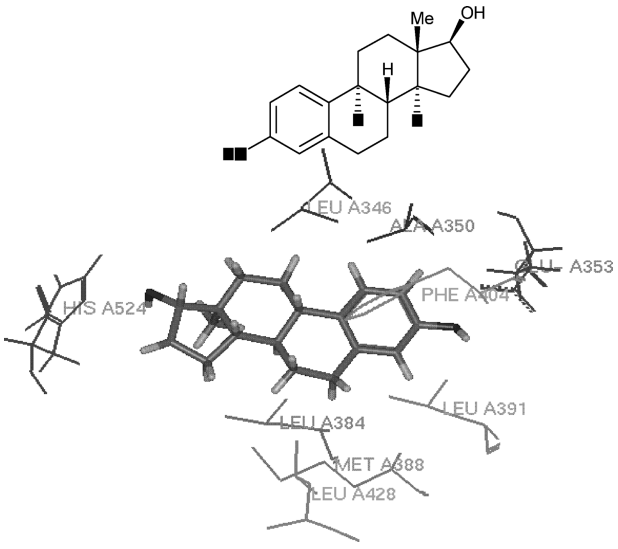


Fig. 5. 17β-Estradiol in active site of estrogen receptor.

353 via charge–charge interaction. The 17β-hydroxyl group forms the hydrogen bond to the imidazole hydrogen atom of His-524. The B ring of estradiol is located in the large lipophilic pocket made from Leu-384, Met-388, Leu-428 and Leu-391. The C ring of estradiol interacts with lipophilic amino acid residues comprised from Ala-350, Leu-346 (Fig. 5). From molecular modeling, these four interactions of estradiol correspond to *X*, *Y*, R1 and R2 of 2-phenyl indole, respectively. This nice complementary relationship between coefficient maps and active site of estrogen receptor shows the powerful ability of our method for reproducing a real and meaningful 3D-QSAR model.

5. Conclusion

In the present study, we devised and elaborated a surface-based 3D-QSAR method, which had been proposed in the previous study. Our approach is divided to KNN and four-way PLS. KNN is used for projecting MEP or MLP on molecular surface to a 2D map. KNN can overcome the technical problem when dealing with molecules with the different size. KNN maps are collected to construct four-way array. four-way PLS is used for correlating four-way array with biological activity. The four-way PLS can give more straightforward and dense contour map than that of two-way PLS. The result can be visualized by contour map as CoMFA and it is easier to interpret structure–activity data.

References

Anzali, S., Barnickel, G., Krug, M., Sadowski, J., Wagener, M., Gasteiger, J., Polanski, J., 1996. J. Comput. Aided Mol. Des. 10, 521.
Bro, R., 1996. J. Chemom. 10, 47.
Cramer, R.D., Patterson, D.E., Bunce, J.D., 1988. J. Am. Chem. Soc. 110, 5959.
Devillers, J., 1996. Neural Networks in QSAR and Drug Design (Principles of QSAR and drug design). Academic Press.
Furet, P., Sele, A., Cohen, N.C., 1988. J. Mol. Graphics 6, 182.
Geladi, P., Kowalski, B.R., 1986. Anal. Chim. Acta 185, 1.
Hasegawa, K., Kimura, T., Funatsu, K., 1997. Quant. Struct. Act. Relat. 16, 219.
Hasegawa, K., Matsuoka, S., Arakawa, M., Funatsu, K., 2002. Comput. Chem. 26, 583.
Kimura, T., Hasegawa, K., Funatsu, K., 1998. J. Chem. Inf. Comput. Sci. 38, 276.
Kubinyi, H., 1993. 3D QSAR in Drug Design. ESCOM Science, The Netherlands.
Kubinyi, H., 1998. 3D QSAR in Drug Design, vol. 2. Kluwer Academic, Leiden, Oxford.
Miyashita Y., Shiraishi Y., Hasegawa K., Sasaki S., 1993. Proceedings of CAMSE'92, p. 879.
Polanski, J., Walczak, B., 2000. Comput. Chem. 24, 615.
Satoh, H., Sacher, O., Nakata, T., Chen, I., Gasteiger, J., Funatsu, K., 1998. J. Chem. Inf. Comput. Sci. 38, 210.
Smilde, A.K., 1997. J. Chemom. 11, 367.
SPARTAN, Wavefunction Inc., 18401 Von Karman Avenue, Suite 370, Irvine, CA, USA.
Tanenbaum, D.M., Wang, Y., Williams, S.P., Sigler, P.B., 1998. Proc. Natl. Acad. Sci. USA 95, 5998.
von Angerer, E., Prekajac, J., Strohmeier, J., 1984. J. Med. Chem. 27, 1439.