Application of a Self-Organizing Map to Quantitative Structure–Activity Relationship Analysis of Carboquinone and Benzodiazepine

Junko KAWAKAMI, Kenji HOSHI, Akiko ISHIYAMA, Sayaka MIYAGISHIMA, and Kenichi SATO*

Information Science Center, Tohoku Pharmaceutical University; 4–4–1 Komatsushima, Aoba-ku, Sendai 981–8558, Japan.
Received February 28, 2004; accepted April 11, 2004; published online April 14, 2004

Self-organizing map (SOM) of Kohonen seems to be a promising approach beyond the standard one to regression for some classification problems encountered in the field of pharmacy. We apply them, therefore, to the quantitative structure–activitity relationship (QSAR) in carboquinones and benzodiazepines, and show their usefulness. Most QSAR analysis using neural networks has been made by adopting neural networks with supervised learning. On the contrary, SOM obeys unsupervised learning and originally does not involve the use of desired target data. If we note that an appreciable fraction of data may be missing without making the similarity comparison impossible in SOM if the number of attributes considered is appreciable, QSAR analysis using SOM is found to be possible as if supervised learning. Similar to target data in supervised learning, we can take into account target data (=observed activity) as one of attributes in addition to other attributes (=structural descriptors). Choice of optimal descriptors as input parameters was found to be essential to generate valuable SOM.

Key words self-organizing map; benzodiazepine; carboquinone; quantitative structure–activitity relationship; leave-one-out; unsupervised learning

Both classification problems and quantitative structure–activitity relationship (QSAR) analyses of drugs have been better studied by the multi-layer feedforward neural networks with back-propagation learning algorithm (BPNN), than the classical methods in pattern recognition such as linear multiple regression (LMR) or adaptive least square method (ALS), since those problems often involve nonlinear relationships between descriptors and the class (/activity).[1,2] BPNN is surely powerful and interesting approach, however, it has the defects such as the problem of local minimum, overfitting *etc*.[3,4] Recently Bayesian regularized neural networks (BRNN),[5,6] has been successfully applied for QSAR studies[7,8] even for massive sample data,[9] but this excellent framework is rather complicated and needs much computation time in general.

Kohonen developed the unsupervised learning algorithm that simplified the mapping mechanism of the relatively homogeneous structures found in mammalian brains associated with the processing of sensory data,[10] and showed it possible to generate self-organizing map (SOM) of data. It is an essential characteristic of Kohonen's SOM that has the ability to project high-dimensional data onto two-dimensional visualized map which is suitable for easy analysis while preserving the most significant information. Therefore, SOM is applied to extensive problems that characteristics of complicated data structure in high-dimensional space can be grasped visually from two-dimensional maps, and plays an important role to analyze inherent data structure.

The advantage of the SOM, compared with some other projection methods is that the algorithm is very simple, straightforward to implement, and fast to compute. In the field of pharmaceutical sciences, the SOM was applied for searching useful drugs. Anzali *at al.* generated SOMs as a two-dimensional representation of molecules to analyze the shape and surface properties of those three-dimensional molecules responsible for biological activity. After visual comparison of SOMs, they discovered the benzothiadiazole group as a surrogate for methylendioxyphenyl.[11,12] Tetko *et al.* used SOMs to compress the so many input CoMFA data in their three-dimensional QSAR studies.[13] In the former study, we obtained the SOM of norbornann derivatives and of carbonyl compounds, and showed that the classification of them into some groups was successfully achieved according to the clustering appeared in the gray map.[14]

In this paper we use the SOM as a method to predict missing activity in the QSAR studies of carboquinone and benzodiazepine, somewhat different usage of SOM rather than the standard compression or visualization tool of data. Results of the calculation indicate that SOM is considered to be one of useful methods in QSAR study.

The Generation of SOM and the Prediction of an Activity Kohonen's neural network consists of two layers, the first layer is the input layer of n-dimension and the second is the competitive layer where every neuron has n-components corresponding to input layer and is arrayed to two dimensions to grasp output visually. The process that generates SOM consists of two steps, that is, the determination of winner neuron according to the competitive learning and self-organization of neurons.

Competitive Learning It is assumed that the neuron i on the competitive layer has the synaptic weight $\mathbf{w}_i(t)=(w_{i1}, w_{i2},\ldots, w_{in})$ in time t. When input signal $\boldsymbol{x}(t)$ entered from the outside, the neuron which has the minimum Euclidian distance $[\boldsymbol{x}(t)-\mathbf{w}_c(t)]$ is called "winner neuron c". Synaptic weight $\mathbf{w}_c(t)$ of the winner neuron seems to get closer to an input signal in the next time $t+1$, and, therefore, is updated by the next formula

$$\mathbf{w}_c(t+1)=\mathbf{w}_c(t)+\alpha(t)[\boldsymbol{x}(t)-\mathbf{w}_c(t)]. \quad (0<\alpha(t)<1) \tag{1}$$

Here $\alpha(t)$ is the learning coefficient which decreases monotonically during learning.

Self-Organization In addition to the winner neuron c, the neuron k located around it on the competitive layer is required to cooperate with one another, and then updates synaptic weight $\mathbf{w}_k$ in the following formula

$$\mathbf{w}_k(t+1)=\mathbf{w}_k(t)+\alpha(t)\gamma(d)[\boldsymbol{x}(t)-\mathbf{w}_k(t)] \quad (0<\alpha(t)<1,\ 0<\gamma(d)<1) \tag{2}$$

Here $\gamma(d)$ is called the neighborhood function, and its value decreases monotonically with increasing spatial distance between the neuron k and c, and also decreases monotonically with time t, and becomes narrower.

During these two processes neurons on the competitive layer are self-organized and the two-dimensional SOM is generated. In this work we used the SOM_PAK package developed by Kohonen's group which is available from their web site[15] for non-commercial use.

QSAR Analysis by Missing Value SOM When QSAR studies are performed by applying feed-forward neural networks such as BPNN or BRNN which obeys the supervised learning, it is natural that n structural descriptors are assigned to whole neurons in the input layer and biological activity is assigned to a neuron in the output layer. On the contrary, SOM obeys unsuper-

* To whom correspondence should be addressed. e-mail: ksato@tohoku-pharm.ac.jp

vised learning and does not involve the use of target data, *i.e.*, there is no such a neuron in the output layer. Instead of a target data in the supervised learning, we treat the observed activity as one of attributes as well as the structural descriptors. It should be noted that an appreciable fraction of data may be missing without making the similarity comparison impossible in SOM if the number of attributes taken into account is appreciable. Considering both this fact and the leave-one-out (LOO) cross-validation procedure, first we generate the SOM using N-1 data (N is the total number of compounds). Next, we investigate to which neuron in the SOM this first eliminated data is assigned. To predict a biological activity in QSAR studies using SOM, we extend the input space to include a biological activity in addition to n structural descriptors, having $n+1$ dimensions. Procedures to obtain the predicted activity is summalized as follows;

(1) remove only data of the compound which wants to pursue predicted activity from data of whole compounds (LOO), and generates the SOM by unsupervised learning of these data.

(2) assume the $(n+1)$-th dimension of the removed data as the missing value, and give the data to the SOM already generated in step (1), then find a winner neuron.

(3) Finally regard the $(n+1)$-th component of the winner neuron as the predicted activity.

In this work a usual gaussian type formula is adopted for the neighborhood function. In practical applications of SOM, data pre-processing is well known to be important. Therefore, each attribute (descriptors and activity) scale is normalized as such that its variance taken over all the compounds is unity.

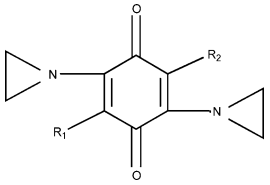

Chart 1

Results and Discussion

Carboquinone Derivatives Carboquinone derivatives are a group of compounds having the configuration shown in Chart 1. They were synthesized by Nakao *et al.*,[16] and were developed to an anticancer drug for the clinical treatment.

As six structural descriptors of data which are given on the input layer, there are the molar refractivity (MR), hydrophobicity (π), substituent constant (F and R), as well as, $MR_{1,2}$ and $\pi_{1,2}$ to estimate the steric effects of R_1 and R_2 and the total hydrophobicity. Using the molar concentration C which is the minimum effective dose (MED) per 1 kg of mouse, the biological activity data $\log(1/C)$ is represented by the column labeled "A" in Table 1. Here MED is the dose giving a 40% prolongation of life compared to the controls. Such refered data in Table 1 is also obtained with the method that administers a dosage in chronic injection by small quantity.[17]

Table 1. Physico-Chemical Parameters ($MR_{1,2}$—R), Observed Biological Activities (A) and Normalized Observed Biological Activities (NA) for 37 Carboquinone Derivatives

No	R^1	R^2	$MR_{1,2}$	$\pi_{1,2}$	π_2	MR_1	F	R	A	NA
1	C_6H_5	C_6H_5	5.08	3.92	1.96	2.54	0.16	−0.16	4.33	0
2	CH_3	$(CH_2)_3C_6H_5$	4.5	3.66	3.16	0.57	−0.08	−0.26	4.47	0.054
3	C_5H_{11}	C_5H_{11}	4.86	5	2.5	2.43	−0.08	−0.26	4.63	0.117
4	$CH(CH_3)_2$	$CH(CH_3)_2$	3	2.6	1.3	1.5	−0.08	−0.26	4.77	0.171
5	CH_3	$CH_2C_6H_5$	3.57	2.51	2.01	0.57	−0.12	−0.14	4.85	0.202
6	C_3H_7	C_3H_7	3	3	1.5	1.5	−0.08	−0.26	4.92	0.230
7	CH_3	$CH_2OC_6H_5$	3.79	2.16	1.66	0.57	−0.04	−0.13	5.15	0.319
8	$R^1=R^2=CH_2CH_2OCON(CH_3)_2$		6.14	0.72	0.36	3.07	−0.08	−0.26	5.16	0.323
9	C_2H_5	C_2H_5	2.06	2	1	1.03	−0.08	−0.26	5.46	0.440
10	CH_3	$CH_2CH_2OCH_3$	2.28	1.03	0.53	0.57	−0.08	−0.26	5.57	0.482
11	OCH_3	OCH_3	1.58	−0.04	−0.02	0.79	0.52	−1.02	5.59	0.490
12	CH_3	$CH(CH_3)_2$	2.07	1.8	1.3	0.57	−0.08	−0.26	5.6	0.494
13	C_3H_7	$CH(OCH_3)CH_2OCONH_2$	4.24	0.98	−0.52	1.5	−0.04	−0.13	5.63	0.506
14	CH_3	CH_3	1.14	1	0.5	0.57	−0.08	−0.26	5.66	0.518
15	H	$CH(CH_3)_2$	1.6	1.3	1.3	0.1	−0.04	−0.13	5.68	0.525
16	CH_3	$CH(OCH_3)C_2H_5$	2.75	1.53	1.03	0.57	−0.04	−0.13	5.68	0.525
17	C_3H_7	$CH_2CH_2OCONH_2$	3.56	1.45	−0.05	1.5	−0.08	−0.26	5.68	0.525
18	$R^1=R^2=CH_2CH_2OCH_3$		3.42	1.03	0.53	1.71	−0.08	−0.26	5.69	0.529
19	C_2H_5	$CH(OC_2H_5)CH_2OCONH_2$	4.23	0.98	−0.02	1.03	−0.04	−0.13	5.76	0.556
20	CH_3	$CH_2CH_2OCOCH_3$	2.78	1.23	0.73	0.57	−0.08	−0.26	5.78	0.564
21	CH_3	$(CH_2)_3$-dimer	1.96	2	1.5	0.57	−0.08	−0.26	5.82	0.580
22	CH_3	C_2H_5	1.6	1.5	1	0.57	−0.08	−0.26	5.86	0.595
23	CH_3	$CH(OCH_2CH_2OCH_3)^-$	4.45	0.01	−0.49	0.57	−0.04	−0.13	6.03	0.661
24	CH_3	$CH_2CH(CH_3)OCONH_2$	3.09	0.75	0.25	0.57	−0.08	−0.26	6.14	0.704
25	C_2H_5	$CH(OCH_3)CH_2OCONH_2$	3.77	0.48	−0.52	1.03	−0.04	−0.13	6.16	0.712
26	CH_3	$CH(C_2H_5)CH_2OCONH_2$	3.55	1.25	0.75	0.57	−0.08	−0.26	6.18	0.720
27	CH_3	$CH(OC_2H_5)CH_2OCONH_2$	3.77	0.48	−0.02	0.57	−0.04	−0.13	6.18	0.720
28	CH_3	$(CH_2)_3OCONH_2$	3.09	0.95	0.45	0.57	−0.08	−0.26	6.18	0.720
29	CH_3	$(CH_2)_2OCONH_2$	2.63	0.45	−0.05	0.57	−0.08	−0.26	6.21	0.732
30	C_2H_5	$(CH_2)_2OCONH_2$	3.09	0.95	−0.05	1.03	−0.08	−0.26	6.25	0.747
31	CH_3	CH_2CH_2OH	1.78	0.34	−0.16	0.57	−0.08	−0.26	6.39	0.802
32	CH_3	$CH(CH_3)CH_2OCONH_2$	3.09	0.75	0.25	0.57	−0.08	−0.26	6.41	0.809
33	CH_3	$CH(OCH_3)CH_2OCONH_2$	3.31	−0.02	−0.52	0.57	−0.04	−0.13	6.41	0.809
34	H	$N(CH_2)_2$	1.66	0.18	0.18	0.1	0.1	−0.92	6.45	0.825
35	$R^1=R^2=CH_2CH_2OH$		2.42	−0.32	−0.16	1.21	−0.08	−0.26	6.54	0.860
36	CH_3	$N(CH_2)_2$	2.13	0.68	0.18	0.57	0.06	−1.05	6.77	0.949
37	CH_3	$CH(OCH_3)CH_2OH$	2.47	−0.13	−0.63	0.57	−0.04	−0.13	6.90	1

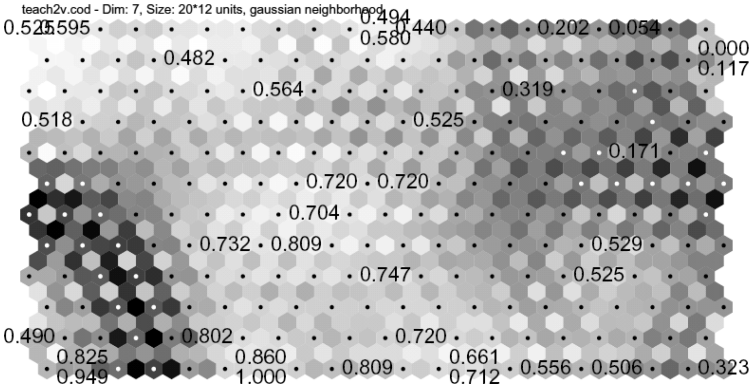

Fig. 1. Self-Organizing Map of Kohonen for 36 Carboquinones with Leave-No. 6-Out of all 37 Carboquinones
Normalized observed biological activities are labeled for every winner neuron.

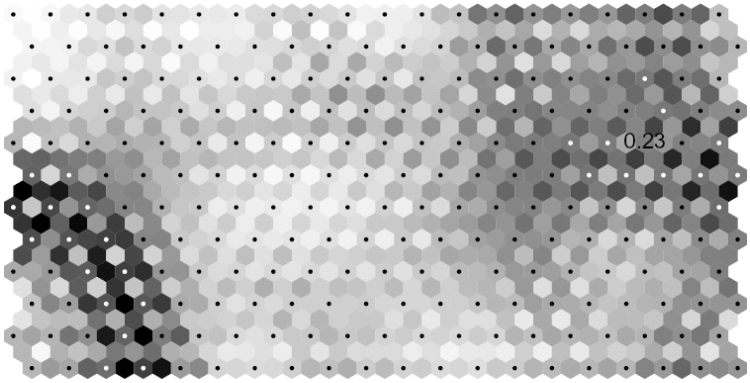

Fig. 2. Self-Organizing Map of Kohonen for the No. 6 Carboquinone Whose Activity Is Missing, and the Winner Neuron Is Settled
Predicted biological activity for the No. 6 is obtained from the value of the 7-th dimensional component of its codebook vector.

QSAR studies for carboquinone derivatives are performed based on a missing value SOM method, and seven neurons on the input layer, 20×12 neurons in the competitive layer are placed, respectively. We set the total number of neurons in the competitive layer to size of several times larger than the number of data set in order to avoid overlapping assignment of many derivatives to a neuron. It was checked that the calculated results remain almost unchanged when the number of neurons in the competitive layer are changed to some extent. Initial conditions are set as the learning coefficient α=0.15, the radius of neighborhood function r=20, and 50000 times of learning cycles are processed.

Results of the calculation indicate that it can predict biological activity with an average of 4.2% errors, and the cross-validation correlation coefficient between them is 0.87. We also checked the fitting ability of the present SOM method by calculating the predicted activity for every carboquinone derivative after neural network training for whole data set, and obtained the high correlation coefficient value 0.96.

Practical execution of the procedure (1) and (2), based on the missing data SOM presented in the preceding section, generated SOMs shown in Figs. 1 and 2, respectively, as an example, for the case to predict the No. 6 carboquinone derivative in the Table 1. Figure 1 shows the gray map of SOM for 36 carboquinone derivatives with leave-No. 6-out of all 37 ones. Each winner neuron is labeled by the normalized biological activity as the term "NA" in Table 1. Next, the procedure (2) generates the SOM for the No. 6 carboquinone derivative which has a missing activity term is shown in Fig. 2, where a winner neuron is settled. Finally, we obtain the predicted activity for the No. 6 derivative from the seventh component of the codebook vector of the allocated neuron labeled by normalized value 0.230. A black circles pictured as scatter diagram in Fig. 3 are the predicted activity values, where measured value agrees with predicted one along the straight line from (4, 4) to (7.5, 7.5). The cross-validation correlation coefficient between the predicted activity and the observed one provided by this calculation is the value 0.874, and is nearly equal to the calculated results using BPNN[2] or Bayesian regularized neural networks.[7] Such calculated results suggest that Kohonen's self-organizing neural network

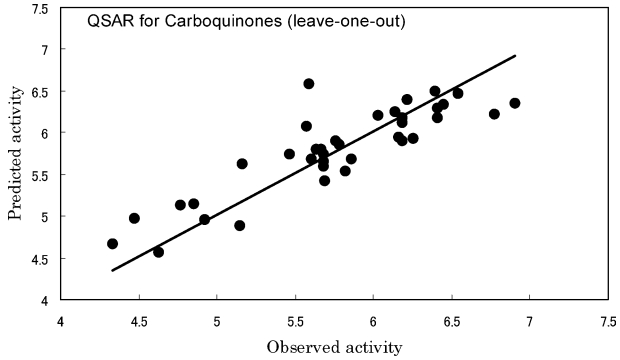

Fig. 3. Scatter Plot of Predicted Activities from Leave-One-Out Cross-Validation *versus* Observed Ones for Carboquinones

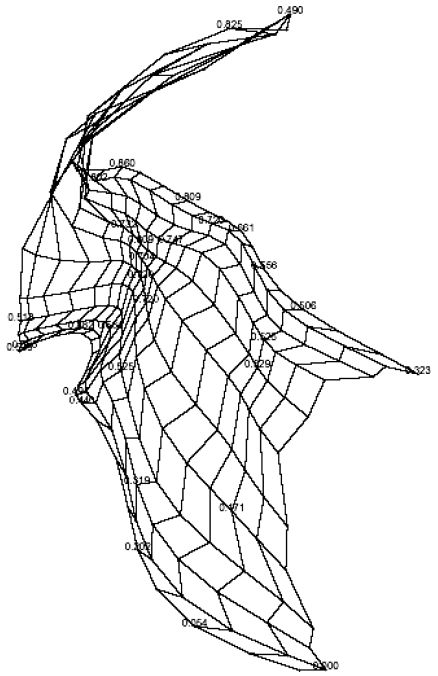

Fig. 4. Sammon Map of the SOM That Is Shown in Fig. 1

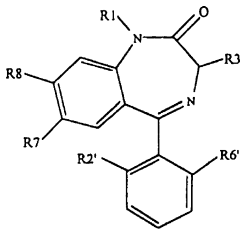

Chart 2

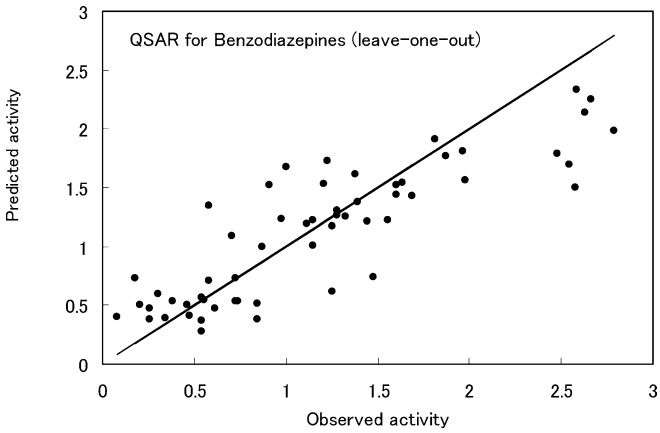

Fig. 5. Scatter Plot of Predicted Biological Activities *versus* Observed Ones for Benzodiazepine

can be applicable to handle QSAR problem with generalization ability. If the No. 11 derivative with massive error of 17.73% is excluded in Fig. 3, the correlation coefficient would improve to give the value 0.916. The No. 11 derivative has a maximal error even in the calculation by the BPNN and by the Bayesian regularized neural networks, and the prediction error is almost the same with our SOM calculation. Such large discrepancy seems to be mainly come from the outstanding large F-value of No. 11 derivative. Figure 4 shows a Sammon map of the SOM whose gray map was already given in Fig. 1, and indicates that carboquinone derivatives which have high similarity ratio locate in the neighboring distance, much clearer than in gray map. It is seen that the No. 11 data of normalized observed biological activity value 0.49 is isolated on the Sammon map in Fig. 4.

Benzodiazepine Derivatives The 57 1,4-benzodiazepin-2-ones used in the study of Maddalena and Johnston[18] are examined in the present work. Such benzodiazepine derivatives are a group of compounds having the configuration shown in Chart 2.

Benzodiazepine has been broadly used as antianxiety agent. It is known that the action of benzodiazepine depends on augmentation of GABA operation in a brain-related communication, that is, GABA and binding affinity between the receptors rise when benzodiazepine connects it to the receptor. Benzodiazepine is the compound whose QSAR analysis has been broadly performed by various methods, and some QSAR studies by BPNN has been done.[18,19] The set of 57 benzodiazepine derivatives used in this study arranged various functional groups at the position of six substituents R_1, $R_{2'}$, R_3, $R_{6'}$, R_7, R_8. As the physicochemical parameter characterizing each substituent, seven substituent constants: molar refractivity (MR), lipophilicity (π), polar constant (F), resonance constant (R), aromatic group dipole (μ), Hammet meta constant (σ_m), Hammet para constant (σ_p) are well used. We performed QSAR analysis using a missing data SOM method about sample data already stated. When we adopt full 42 physicochemical parameters as structural descriptors, only the LOO cross-validated correlation coefficient of about 0.5 was given by the calculation, and then the QSAR analysis by the SOM seemed to be difficult. It is well known that optimum choice of variables is significant to achieve generalization performance of BPNN, and the information that is not so valuable is eliminated by removing the redundant variables.

As for benzodiazepine, important input parameters were investigated in many works. Using selective pruning method an optimum set of 10 input parameters were obtained by Maddalena and Johnston. Then much better optimum set of various six input parameters (T6-2, C6-2 set), such as π_7, σ_{7m}, MR_1, $\sigma_{2'm}$, $\sigma_{2'p}$, $\sigma_{6'p}$, were found by So and Karplus using genetic algorithm as preprocessing in their study using BPNN.[19]

Results of a missing value SOM under the conditions of learning coefficient α=0.15, 20×15 neurons in the competitive layer, initial neighborhood radius r=20 and 50000 learning cycles, show that T6-2 and C6-2 parameter sets successfully provides cross-validation correlation coefficient of Rcv≧0.8, and especially C6-2#4 parameter set gives Rcv=0.86, though it is less than the case of the best performance Rcv=0.93 obtained by So and Karplus. We also checked the fitting ability of the present SOM method by calculating the predicted activity for every benzodiazepine derivative after neural network training for whole data set, and obtained the high correlation coefficient value 0.94.

Calculated Rcv for C6-2#4 parameter set are plotted as scatter diagram in Fig. 5, where the observed activity agrees with the predicted one along the straight line from (0,0) to (3, 3).

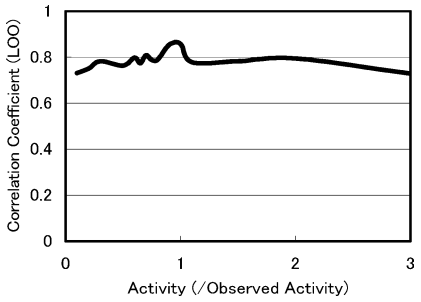

Fig. 6. Results of a Missing Activity SOM Calculation When the Activity Component of the Input Data Is Scaled While All the Descriptors Are Kept Unchanged

Our calculation showed that changes of the conditions above to some extent produce only minor change of the result.

QSAR study by a missing value SOM is more sensitive to the selection of descriptor set compared with BPNN, and only Rcv *ca.* 0.7 is provided if MJ6-1 parameter set (BPNN result Rcv=0.78) or with T10-3 parameter set (BPNN result Rcv=0.94) are adopted.

These results indicate that QSAR study using a missing value SOM depends greatly on the selection of important descriptors, and optimal set of descriptors have to be selected to raise generalization performance of the neural networks also in the application of SOM. It is interesting that a missing activity SOM may be helpful in evaluating whether a set of descriptors are good or not.

In Fig. 6 we show the change of cross-validation correlation coefficient as the seventh attribute (=activity) of input data is scaled down or up by the factor on a x-axis, while six selected C6-2#4 structural descriptors are kept unchanged. It should be noted that a correlation function itself is independent of the absolute value of biological activity. Calculated results show that the cross-validation correlation coefficient takes the maximum value at the scale-factor 1 and it changes very slowly, which supports the stability of a missing activity SOM method.

The fact that the selection of structural descriptors as preprocessing is essential to generate the excellent SOM found in the present research may be applicable for other SOM applications such as data-compression, data-visualization. Since self-organizing neural network itself does not have a pruning function, it would be better to use the optimal parameter set which are selected by other methods as preprocessing like Backward stepwise selection, Bayesian neural networks which comprises automatic variable choice function called "Automatic Relevance Determination".[5,6]

Conclusion

We applied the self-organizing map (SOM) to QSAR studies in a similar way as the supervised neural networks. Results of our calculation based on a missing value SOM could predict the observed activities for carboquinone and benzodiazepine well. Considering the excellent properties of SOM like simplicity of the algorithm, straightforward implementation and fast computation, a missing value SOM should be one of the valuable methods for QSAR study, though its prediction ability seems to be somewhat less than BPNN in general. Another finding that the prediction ability of a missing value SOM depends more severely on the set of selected input descriptors than BPNN, may suggest that it is helpful to know whether some pruning methods to select an optimum descriptors is good or not.

References

1) Aoyama T., Suzuki Y., Ichikawa H., *J. Med. Chem.*, **33**, 905—908 (1990).
2) Ichikawa H., "Kaisougata Nyurarunettowa-ku," Kyouritsu Co., Tokyo, 1993.
3) Bishop C. M., "Neural Networks for Pattern Recognition," Oxford University Press, New York, 1995.
4) Sato K., Nakagawa J., *Chem. Pharm. Bull.*, **45**, 107—115 (1997).
5) MacKay D. J. C., Network: Computation in Neural Systems, **6**, 469 (1995).
6) Neal R. M., "Bayesian Learning for Neural Networks," Springer, New York, 1996.
7) Sato K., Nakagawa J., Matuzaki H., *J. Tohoku Pharmaceutical University*, **44**, 187—193 (1997).
8) Burden F. R., Winkler D. A., *J. Med. Chem.*, **42**, 3183—3187 (1999).
9) Ajay Walters W. P., Murcko M. A., *J. Med. Chem.*, **41**, 3314—3324 (1998).
10) Kohonen T., "Self-Organizing Maps," Springer, Berlin, 2000.
11) Anzali S., Mederski W. W. K. R., Osswald M., Dorsch D., *Bioorg. Med. Chem. Lett.*, **8**, 11—16 (1998).
12) Devillers J., "Neural Networks in QSAR and Drug Design," Academic Press, London, 1996.
13) Tetko I. V., Kovalishyn V. V., Livingstone D. J., *J. Med. Chem.*, **44**, 2411—2420 (2001).
14) Sato K., Kawakami J., Matuzaki H., *J. Tohoku Pharmaceutical University*, **48**,125—131 (2001).
15) Kohonen's group of Helsinki University of Technology, ⟨http://www.cis.hut.fi/research/som-research/nnrc-programs.shtml⟩
16) Nakao H., Arakawa M., Nakamura T., Fukushima M., *Chem. Pharm. Bull.*, **20**, 1968—1974 (1972).
17) Yoshimoto M., Miyazawa H., Nakao H., Shinkai K., Arakawa M., *J. Med. Chem.*, **22**, 491—496 (1979).
18) Maddalena D. J., Johnston G. A., *J. Med. Chem.*, **38**, 715—724 (1995).
19) So S. S., Karplus M., *J. Med. Chem.*, **39**, 5246—5256 (1996).