Prediction of Biological Targets Using Probabilistic Neural Networks and Atom-Type Descriptors

Tomoko Niwa[†]

*Discovery Research Laboratories, Nippon Shinyaku Co., Ltd. 14,
Nishinosho-Monguchi-cho, Kisshoin, Minami-ku, Kyoto, 601-8550, Japan*

Received June 9, 2003

Prediction of biological targets for molecules from their chemical structures is beneficial for generating focused libraries, selecting compounds for screening, and annotating biological activities for those compounds whose activities are unknown. We studied the ability of a probabilistic neural network (PNN), a variant of normalized radial basis function (RBF) neural networks, to predict biological activities for a set of 799 compounds having activities against seven biological targets. The compounds were taken from the MDDR database, and they were carefully selected to comprise distinct biological activities and diverse structures. The structural characteristics of compounds were represented by a set of 24 atom-type descriptors defined by 2D topological chemical structures. The modeling was done in two ways: (1) compounds having one certain activity were discriminated from those not having that activity and (2) all compounds were classified into seven biological classes. In both cases, around 90% of the compounds were correctly classified. Further validation of the modeled PNNs was done with 26 317 compounds having biological activities against various targets except for the seven targets used for modeling, and 67−98% compounds were correctly classified depending upon the targets. A PNN trains much more quickly than widely used neural networks such as a feed-forward neural network with error back-propagation. Calculation of atom-type descriptors is easy even for a large-size chemical library. Combination of PNN and atom-type descriptors thus provides a powerful way to predict biological activities from structural information.

Introduction

With the increasing use of robotics technologies such as high-throughput screening (HTS) and parallel synthesis, we are now suffering from overflowing data; efficient methods to deal with so much information are strongly demanded. Clustering[1] and partitioning of databases[2] were widely used in generating libraries and selecting screening compounds for a particular target. Artificial intelligence (AI) tools as neural networks are also suitable for analyzing a huge amount of complicated data. For example, to evaluate druglikeness of compounds, Ajay and co-workers employed Bayesian neural networks[3] and Sadowski and co-workers used feed-forward neural networks with error back-propagation.[4]

One problem in neural network modeling based on chemical information is that there are so many kinds of molecular descriptors. Also, the choice of appropriate descriptors is often time-consuming. In addition, smaller numbers of descriptors are preferable because the size of input data largely determines the time needed to model neural networks. We have already found that "atom-type" descriptors assigned by elements, structures, and functional groups work well to characterize molecular properties. For example, only 24 atom-type descriptors could successfully discriminate druglike molecules from nondruglike molecules.[5] These descriptors were also successfully utilized in modeling 1-octanol/water partition coefficients (log *P*), blood−brain barrier partition coefficients (log BB), the highest occupied molecular orbital (HOMO), the lowest unoccupied molecular orbital (LUMO) energy levels, and polar surface area (PSA).[5] They well represented steric, electronic, and hydrophobic effects of molecules. It is thus of interest to examine the performance of these atom-type descriptors in the classification of molecules according to the biological activities.

Apart from the selection of descriptors, problems in neural network modeling include the selection of appropriate architectures of neural networks. We have already found that probabilistic neural networks (PNNs) were powerful in predicting the intestinal absorption of drugs[6] and in classifying compounds according to their druglikeness.[5] PNN is Donald Specht's term for kernel discriminant analysis[7] and a variant of normalized radial basis function (RBF) networks.[8] They perform well in noisy environments and train much more quickly than feed-forward and Bayesian neural networks. Because PNN is a classifier, it can be applicable to multicategory problems. This means that only one PNN is sufficient to model a combined set of compounds against various biological targets. The objectives of this study are to investigate the performance of the atom-type descriptors and PNNs to classify compounds according to their biological activities. These approaches are, of course, useful for generating specialized libraries, selecting compounds for HTS screening, and annotating biological activity for those compounds whose activities are unknown.

[†] E-mail: t.niwa@po.nippon-shinyaku.co.jp. Phone: +81-75-321-9171. Fax: +81-75-321-9038. E-mail: t.niwa@po.nippon-shinyaku.co.jp.

10.1021/jm0302795 CCC: $27.50 © 2004 American Chemical Society
Published on Web 04/06/2004

Table 1. Biological Activities and the Number of Compounds Used in the Analysis

biological activities[a]	no. of compds[b]
histamine H3 antagonist	31 (3.9%)
carbonic anhydrase inhibitor	54 (6.8%)
HIV-1 protease inhibitor	216 (27.0%)
5-HT2A antagonist	120 (15.0%)
tyrosine-specific protein kinase inhibitor	200 (25.0%)
ACE inhibitor	140 (17.5%)
progesterone antagonist	38 (4.8%)

[a] The same expressions as those used in the MDDR database 2000.2 (22,11) are shown. [b] The figures in parentheses are the percentage values.

Table 2. Atom-Type Descriptors Used in the Analysis

hydrogen	H
carbon	C[sp], C[sp^2], C[sp^3], C[aromatic]
nitrogen	N[sp], N[sp^2], N[sp^3], N[aromatic], N[amide], N[sp^3 planar], N[ammonium]
oxygen	O[sp^2], O[sp^3], O[carboxyl]
sulfur	S[sp^2], S[sp^3], S[SO], S[SO_2]
phosphorus	P
halogens	F, Cl, Br, I

Method

Test Compounds. The compounds used for analysis were taken from MDL Drug Data Report (MDDR 2000.2)[9] containing 118 309 compounds. MDDR covers the patent, literature, journals, meetings, and congresses. For each patent or literature in this database, one representative compound has biological descriptions in the activity field whereas the rest of compounds have nothing in this field. To exclude compounds with similar structures, we first selected 27 116 compounds having biological descriptions in the activity field and valid chemical structures. From these 27 116 compounds, 799 compounds were then selected for modeling. The activity classes and the expression of biological activities used for selection are shown in Table 1. All compounds thus retrieved were employed in the neural network modeling. They were similar to those used by Xue and co-workers[10,11] to study informative fingerprints representing the structural properties of molecules with distinct biological activities and diverse structures. Among the 27 116 compounds, 799 compounds were used for modeling and the remainder of 26 317 compounds was used for validation.

Availability of Informtion on the Test Compounds. The names of the compounds (extreg) are available from the author.

Descriptors. There are two ways to prepare descriptors. The first is to calculate a large number of descriptors and then select meaningful descriptors, and the second is to identify a preferred set of molecular descriptors directed not for a particular case but for universal cases. Our selection was the latter, and we defined 24 descriptors based on the following considerations. (1) Global descriptors such as log *P* and molecular weight were avoided because they are so rough for discriminating compounds based on biological activities. (2) Particular descriptors were also avoided because compounds belonging to one biological class have structural diversities to some extent. (3) For atoms that rarely appear, such as phosphorus, detailed assignment was avoided, since sparse descriptors do not generally work well in neural network modeling. (4) The atoms in biologically important functional groups, such as amide and caroboxyl groups, were defined in detail.

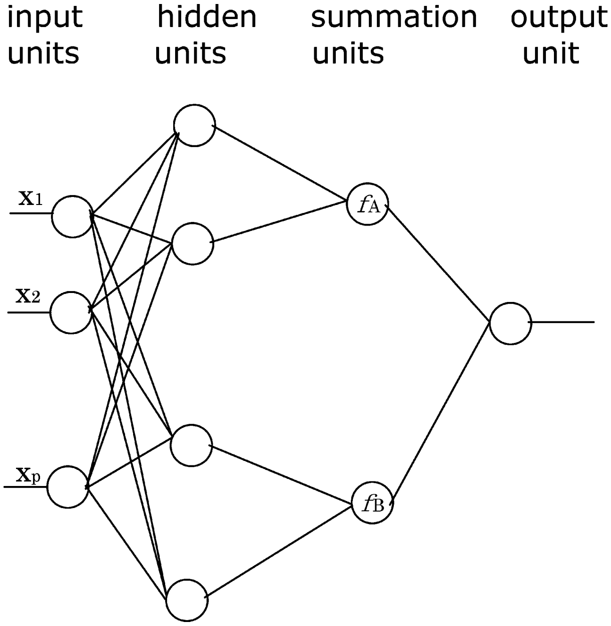

Figure 1. Architecture of PNN.

The descriptors we defined are listed in Table 2. For hydrogen, halogen, and phosphorus atoms, only one type of descriptor was assigned. For carbon, nitrogen, oxygen, and sulfur atoms, their descriptors were defined by using hybridization types (sp, sp^2, and sp^3) and structural information. For example, C[aromatic] and N[aromatic] are defined for the carbon and nitrogen atoms in aromatic rings, respectively. N[amide] represents the nitrogen atom in an amide group. N[sp^3 planar] is the sp^3 nitrogen in a planar structure such as the nitrogen atom in an indole moiety. N[ammonium] is an ammonium nitrogen atom. O[carboxyl] represents an oxygen atom in a carboxyl group. S[SO] and S[SO_2] stand for the sulfur atoms in the functional groups SO and SO_2, respectively. For the carbon, nitrogen, and sulfur atoms not defined above, the hybridization types (sp, sp^2, and sp^3) were used for definitions. C[sp], C[sp^2], and C[sp^3] represent sp, sp^2, and sp^3 carbons, respectively. Similarly, O[sp^2], O[sp^3], S[sp^2], and S[sp^3] were defined for oxygen and sulfur atoms. Note that we used the occurrence of each descriptor. Therefore, our descriptors can be regarded as extended molecular formulas. These “atom-type” descriptors were computed with in-house programs. We used all 24 descriptors in every model.

PNN Modeling. Probabilistic neural network is Donald Specht’s term for kernel discriminant analysis and is regarded as a normalized RBF network.[7] PNN has been used to analyze biological, spectral, and analytical data.[6,12,13] The details of PNN have been described elsewhere.[6,13] Briefly, PNN is a memory-based feed-forward network, consisting of four layers: input, hidden, summation, and output layers. PNN replaces the sigmoid activation function often used in neural networks with a radial basis function, and probability density functions are evaluated using the Parzen’s nonparametric estimator.[9]

PNN utilizes one probability density function for each category, as shown by

$$f_A(\mathbf{x}) = \frac{1}{m(2\pi)^{p/2}\sigma^{p}} \sum_{i=1}^{m} \exp\left[-\frac{(\mathbf{x}-\mathbf{x}_{Ai})^T(\mathbf{x}-\mathbf{x}_{Ai})}{2\sigma^2}\right] \quad (1)$$

where $f_A(\mathbf{x})$ represents the probability density function for category θ_A with a vector random variable $\mathbf{x}$, m is the number of training patterns, p is the number of independent features (descriptors), and $\mathbf{x}_{Ai}$ is ith training pattern from category θ_A. σ is the width of the Gaussian-shaped kernels. $f_B(\mathbf{x})$ is similarly defined by eq 1 for category θ_B and so on.

Figure 1 illustrates the architecture of a PNN having four layers. For simplicity, PNN with two categories is shown. The input units are merely distribution units that supply the same input values to all of the hidden units. The number of input units is equal to the number of descriptors. There is one hidden unit for each training pattern, and the number of hidden units is equal to the number of compounds in the training set. The

Table 3. Classification Performance of the Trained Neural Network with Two Output Categories

	train[a]		test[a]		predict[a]		all[a]		
biological activities	C+ [b]	C− [b]	C+ [b]	C− [b]	C+ [b]	C− [b]	C+ [b]	C− [b]	smoothing factor[c]
histamine H3 antagonist	92.5	98.1	90.5	98.2	90.5	98.2	91.6	98.2	0.461
carbonic anhydrase inhibitor	96.3	99.8	96.4	99.9	94.5	99.3	95.9	99.7	0.450
HIV-1 protease inhibitor	96.9	95.3	91.7	93.1	94.0	92.5	95.3	94.3	0.298
5 HT2A antagonist	94.4	98.8	91.7	97.8	95.0	96.9	94.0	98.2	0.284
tyrosine-specific protein kinase inhibitor	91.5	93.9	84.5	90.2	86.5	93.2	89.1	93.0	0.432
ACE inhibitor	95.0	96.8	93.6	93.8	90.7	95.2	93.9	95.9	0.265
progesterone antagonist	95.7	99.2	90.0	98.4	90.0	98.2	93.2	98.8	0.378

[a] Train, test, and predict mean the training, test, and prediction sets, respectively. All is the combined set of train, test, and prediction sets. The percentages of correctly classification compounds are shown. [b] For H3 antagonists, C+ means the percentage of the compounds predicted to have H3 antagonist activity. C− is the percentage of the compounds predicted not to have H3 antagonist activity. [c] The value of the trained smoothing factors.

number of summation units is the number of the categories. The hidden-to-output weights are usually 1 or 0; for each hidden unit, a weight of 1 is used for the connection going to the output to which that pattern belongs, while all other connections are given weights of 0. The network produces activations in the output layer corresponding to the probability density function estimate. The highest output represents the most probable category. A probability density function is defined for each category and is composed of Gaussian shaped kernels, as shown by eq 1. Since the numbers of input, hidden, summation, and output units do not change during model developments, the only weights to be learned are the widths of the Gaussian shaped kernels, σ. These widths are called “smoothing factors” or “bandwidths”.[7,8] In this study, only one smoothing factor is applied to all the input descriptors. This means that all of the input descriptors have the same impact on predicting the output.

From 799 compounds, a training set (60%), a test set (20%), and a prediction set (20%) were selected randomly. A PNN was trained on the training data set. To prevent the PNN from overfitting the training data, the PNN was evaluated on its ability to make correct predictions of the test data set. The best smoothing factor was iteratively optimized, and an upper limit used to test a smoothing factor was set at 0.8. Thus, a smoothing factor varies from 0 to 0.8. The external prediction set was used to study the predictive power of the trained PNN. PNN modeling was done with the Neuroshell 2 program[14] run on a Pentium II desktop computer.

Results

Classification into Two Categories. First, we studied whether PNNs could discriminate between compounds having a certain biological activity from those not having that activity by dividing the data set comprising 799 compounds into two categories. For example, in the case of H3 antagonists, one category consisted of all H3 antagonists (38 compounds) whereas the other included the rest of the compounds (761 compounds). The training (60%), test (20%), and prediction (20%) sets are randomly selected five times, and the average values of five runs are summarized in Table 3. The differences among the five runs were generally small. Modeling a PNN was easy, and it took less than 1 min to obtain a PNN. After training of a PNN, its classification ability was checked by calculating the percentage of the compounds correctly classified. As can be seen in Table 3, about 93% of the H3 antagonists in the training set were correctly classified and about 98% of the rest of the compounds were predicted not to have H3 antagonist activity. More than 90% of the compounds were successfully classified not only for the test set but also for the prediction set. Similar results were obtained for the other biological targets. The predictive abilities of the trained PNNs proved to be remarkably high.

Table 4. Classification Performance of the Trained Neural Networks with Seven Output Categories

biological activities	train[a]	test[a]	predict[a]	all[a]
histamine H3 antagonist	97.8	81.4	76.7	90.3
carbonic anhydrase inhibitor	100.0	98.0	100.0	99.6
HIV-1 protease inhibitor	90.3	84.7	84.7	88.1
5 HT2A antagonist	95.6	90.8	90.0	93.5
tyrosine-specific protein kinase inhibitor	90.7	80.0	81.0	86.6
ACE inhibitor	97.9	93.6	92.1	95.9
progesterone antagonist	96.5	97.5	90.0	95.3

[a] Train, test, and predict mean the training, test, and prediction sets, respectively. All is the combined set of train, test, and prediction sets. The percentages of correctly classified compounds are shown.

The smoothing factors in Table 3 varied from 0.265 to 0.461. Except for tyrosine kinase inhibitors, there are tendencies that when the number of compounds comprising a category was large, the smoothing factor for that category was small. It is reasonable because larger smoothing factors give more relaxed surface fits through the data and provide good interpolative abilities for prediction, whereas smaller smoothing factors produce sharp surface fits and reduce the overlap between categories. Namely, excellent classification performances were achieved by adjusting smoothing factors to best discriminate the compounds.

Classification into Seven Categories. PNNs having seven output categories were trained successively, and their results are summarized in Table 4. The training (60%), test (20%), and prediction (20%) sets were randomly selected, and the averages of five runs are listed in this table. It is striking that only one PNN could successfully classify 799 compounds into seven biological categories. Classification performance for the external prediction set is an excellent metric to validate the quality of a trained PNN. As can be seen in Table 4, the differences between the classification performances for the prediction and test sets were small for all biological targets. This fact shows the high predictive power of the trained PNNs. The percentages of compounds correctly classified were larger than 80% except for histamine H3 antagonists, and about half of them are around 90%. A possible reason for the relatively worse performance for histamine H3 antagonists is that only 38 histamine H3 antagonists were included in the total of 799 compounds.

The averaged smoothing factor of five runs was 0.273, and the standard deviation of the five smoothing factors was 0.048. As mentioned above, the smoothing factor was large for a category comprising a relatively small

Table 5. Effects of Fixed Smoothing Factors on the Classification Performance of the Trained Neural Networks with Two Output Categories

	smoothing factors					
	0.2[a]		0.3[a]		0.4[a]	
biological activities	Cn+ [b]	Cn− [b]	Cn+ [b]	Cn− [b]	Cn+ [b]	Cn− [b]
histamine H3 antagonist	98.1	98.5	99.4	96.8	99.4	97.3
carbonic anhydrase II inhibitor	99.6	99.5	100.0	99.0	100.0	98.7
HIV-1 protease inhibitor	96.2	96.0	95.2	94.1	92.7	89.3
5 HT2A antagonist	97.0	97.8	96.5	96.4	96.0	95.5
tyrosine kinase inhibitor	93.6	97.1	92.3	95.6	89.2	92.8
ACE inhibitor	97.3	95.0	97.9	88.7	98.0	83.9
progesterone antagonist	97.4	98.3	97.4	97.3	97.4	97.7

[a] The smoothing factors used for classifications are shown. [b] For H3 antagonists, Cn+ means the percentage of the compounds predicted to have H3 antagonist activity and Cn− is the ratio of the compounds predicted not to have H3 antagonist activity.

Table 6. Effects of Fixed Smoothing Factors on the Classification Performance of the Trained Neural Networks with Seven Output Categories

	smoothing factors		
biological activities	0.2[a]	0.3[a]	0.4[a]
histamine H3 antagonist	89.7	91.6	69.7
carbonic anhydrase II inhibitor	98.5	99.6	100.0
HIV-1 protease inhibitor	91.7	85.5	82.5
5 HT2A antagonist	95.5	91.8	89.5
tyrosine kinase inhibitor	89.7	86.6	80.5
ACE inhibitor	95.9	95.9	94.6
progesterone antagonist	95.8	95.3	87.4

[a] The smoothing factors used for classifications are shown.

number of compounds; the smoothing factors for histamine H3 antagonists, carbonic anhydrase inhibitors, and progesterone antagonists were 0.461, 0.450, and 0.387, respectively (Table 3). However, for these three categories, the percentages of compounds correctly classified were near those for such categories that comprise more than 100 compounds (Table 4).

Effects of Fixed Smoothing Factors on Classification Performances. The optimized smoothing factors varied from 0.265 to 0.461 in Table 3. According to Specht, the misclassification rate does not change dramatically with small changes in smoothing factors.[7] It is interesting to examine how fixed smoothing factors affect the classification performances. The numbers of the compounds, descriptors, and output categories uniquely define the architecture of a PNN. Therefore, all that was needed to examine the effects of fixed smoothing factors was to change the smoothing factors in the previously modeled PNNs. Tables 5 and 6 list the classification performances for all 799 compounds using the values of 0.2, 0.3, and 0.4 as smoothing factors. As can be seen in these tables, the changes in the smoothing factors only weakly affected the classification performances. Generally, values of 0.2 and 0.3 gave better results than 0.4, but the differences were small. These results clearly show that without optimizing smoothing facts, it is possible to obtain PNNs with sufficient classification abilities. This fact is particularly valuable in real-world drug discovery and development where prompt modeling is often demanded.

In this study, only one smoothing factor was used for all descriptors. This means that all of the input descriptors have the same impact on predicting the output. Of course, use of multiple smoothing factors (one smoothing factor for each input descriptor) is possible. In our preliminary modeling, we employed multiple descriptors. Unexpectedly, using multiple smoothing factors gave worse results than employing a single smoothing factor. A possible reason for this is that the number of compounds used in the present modeling was relatively small. Multiple smoothing factors might give better results where a large number of compounds are available.

Table 7. Validation Test Using External Data (26 317 Compounds)

biological activities[a]	classification performance (C−)[b]
histamine H3 antagonist	90.6
carbonic anhydrase inhibitor	98.0
HIV-1 protease inhibitor	73.3
5-HT2A antagonist	78.4
tyrosine-specific protein kinase inhibitor	66.9
ACE inhibitor	78.2
progesterone antagonist	94.0

[a] The same expressions as those used in the MDDR database 2000.2 (22,11) are shown. [b] The percentages of correctly classified compounds are shown. C+ is 100.0 minus C−.

Validations Using External Data. Among 27 116 compounds having biological descriptions in the activity field and valid chemical structures in the MDDR database, only 799 compounds were used for modeling and the remainder of 26 317 compounds was left unused. The remainder has biological activities against various targets except for the seven targets used in selecting 799 compounds. Further validation of the modeled PNNs was done with these 26 317 compounds to examine generalization capabilities. We used the modeled PNNs having two output categories (Table 3), and the average values of five runs are summarized in Table 7. When all compounds are perfectly classified, the values of C+ and C− are 0% and 100%, respectively. As can be seen in Table 7, about 67−98% compounds were correctly classified depending on the targets. Although the PNNs were built with only 799 compounds, fairly good results were given for the remainder of 26 317 compounds. It should be noted that the seven targets cover the major drug targets as GPCRs, kinases, enzymes, nuclear hormone receptors, and Zn peptidases. These facts clearly show that the only 799 compounds well represent the structural properties of the drug world and the validity of using 799 compounds in the present modeling.

Classification performances were high for histamine H3 antagonists, carbonic anhydrase inhibitors, and progesterone antagonists. The structural variations for these targets were small, which explains the excellent results for these targets. On the other hand, the structural variations in HIV-1 protease inhibitors and tyrosine-specific protein kinase inhibitors are large. For example, tyrosine-specific protein kinase inhibitors comprise various compounds that are ATP-competitive and not ATP-competitive. Their classification performances are still practically good. The biological activity of 5TH2A antagonist is very specific, while tyrosine-specific kinase inhibitors have various biological activities such as antiinflammatory and anticancer activities. From these results, our modeling procedures proved to be widely applicable.

Discussion

Data Sets. Among 27 116 compounds in the MDDR database, only 799 compounds were used in modeling and the remainder of 26 317 compounds was utilized in validation. We considered that use of the remainder in modeling was problematic for the following reasons. First, it is possible to get compounds having a certain activity, but it is impossible to retrieve those compounds that do not have a certain activity from the MDDR database because the MDDR database has activity information only but is devoid of inactivity information. It is critical to use well-validated data sets in classifications studies. Second, it is not rare that one drug has several biological activities, and so the remainder may comprise some actives. Third, the seven targets cover major drug targets such as GPCRs, kinases, enzymes, nuclear hormone receptors, and Zn peptidases. Fourth, the number of hidden units is equal to the number of the compounds in the training set, and inclusion of all compounds in the modeling results in a huge network. To get meaningful and practically useful PNN models, small- or medium-size and high-content data sets are demanded. Fairly good results obtained for the remainder of 26 317 compounds in Table 7 demonstrate the validity of our selection procedure.

Atom-Type Descriptors. Ghose and Crippen defined atom-type descriptors to calculate the octanol/water partition coefficient, log *P*.[15] Sadowski and coworkers used Ghose–Crippen atom-type descriptors to discriminate drugs from nondrugs by using neural networks.[4] Also, molecular fingerprints such as MDL substructural keys are widely used in clustering.[1] Despite the usefulness of these descriptors and fingerprints, there exist some redundancies. Recently, a reduced set of descriptors has been reported by Xue and co-workers.[10] They proposed the minifingerprints (MFPs) comprising a few selected two-dimensional (2D) descriptors and a number of structural keys. Their MFPs were picked from a pool of descriptors using a genetic algorithm and a set of drug compounds with distinct biological activities. They also claimed that MFPs were specifically designed to recognize compounds with similar activity.[11] Frimurer and co-workers[16] reported a subset of atom-type descriptors assigned by using CONCORD,[17] software to generate 3D structures from 2D chemical presentations.

The reduced sets proposed by Xue and co-workers and Frimurer and co-workers were selected by using the modeling results. There are possibilities that different data sets lead to different reduced sets of descriptors. On the other hand, our atom-type descriptors were defined prior to the network modeling; they were defined without the information about biological activities. Hence, the descriptors used here are more universal and more generalized.

We used not bit strings as MDL fingerprints but the counts of descriptors. The fingerprints are useful for treating detailed structural information. However, our experience is that they are weak for expressing the size and lipophilicity of molecules, which are very important factors for biological activities. The strengths of our descriptors are as follows. (1) They are not so global and not so detailed. (2) They include information about pharmacological important functional groups such as amide and carboxyl groups. (3) They can deal with size and lipophilicity information to some extent. (4) They are easy to compute. After defining the appropriate atom types, we found that the descriptors used in this study are very similar to those used in the docking program "GOLD" by Jones and co-workers.[18] It is not surprising that the descriptors expressing receptor–ligand interactions perform well to represent the physical and structural properties of drugs.

Another thing to be noted is that bond information expression such as connectivity and bond order is lacking in our descriptor set. Bond information is indispensable to substructural and similarity searching and is useful when dealing with a single scaffold or very similar scaffolds. However, there are exists multiple scaffolds in a category, and with too much precise information, it is difficult to handle multiple scaffolds. In the present modeling, only the atom information was employed, but all compounds in a training set were included in a PNN. It is probable that the combination of atom-type descriptors and PNN contributes to the high performance of our models.

Probabilistic Neural Networks. The advantages of PNNs are as follows. (1) It is unnecessary to optimize the architecture of a PNN, since the architecture is uniquely defined by the numbers of compounds, descriptors, and output categories. (2) A PNN trains much more quickly than multilayer feed-forward and Bayesian neural networks, which is beneficial in drug discovery and development. (3) For a multilayered feedforward neural network, a threshold value must be given to divide a set of data into two categories, and a different threshold value leads to different classification results. The outputs of a PNN are categories, and PNNs are free from threshold problems. (4) A PNN can deal with multicategory problems.

There are of course some disadvantages. One disadvantage is that PNNs suffer from the curse of dimensionality. To avoid this problem, it is critically important to select informative and nonredundant descriptors to reduce the number of the input descriptors. Namely, the qualities of the descriptors largely determine the classification power of PNNs. In fact, we paid extensive effort to consider informative and nonredundant descriptors. Another disadvantage is that PNN is a memory-based method. This was a serious problem in past years, but large memories are inexpensively available today. We have successfully applied PNN and the atom-type descriptors to the discrimination between "druglike" and "nondruglike" compounds. Applications to other problems are now in progress. While PNN has been rarely used in the field of chemometrics, we believe the combination of PNN and the atom-type descriptors has potential for modeling complicated problems.

References

(1) Brown, R. D.; Martin, Y. C. Use of Structure–Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572–584.

(2) Lewis, R. A.; Mason, J. S.; McLay, I. M. Similarity Measures for Rational Set Selection and Analysis of Combinatorial Libraries: The Diverse Property-Derived (DPD) Approach. *J. Chem. Inf. Comput. Sci.* **1997**, *3*, 5599–5614.

(3) Ajay; Walters, W. P.; Murcko, M. A. Can We Learn To Distinguish between "Drug-like" and "Nondrug-like" Molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324.

(4) Sadowski, J.; Kubinyi, H. A Scoring Scheme for Discriminating between Drugs and Nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.

(5) Unpublished work.

(6) Niwa, T. Using General Regression and Probabilistic Neural Networks To Predict Human Intestinal Absorption with Topological Descriptors Derived from Two-Dimensional Chemical Structures. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 113−119.

(7) Specht, D. F. Probabilistic Neural Networks. *Neural Networks* **1990**, *3*, 109−118.

(8) Bishop, C. M. In *Neural Networks for Pattern Recognition*; Oxford University Press: New York, 1995; pp 164−193.

(9) The MDL Drug Data Report is available from MDL Information Systems Inc., San Leandro, CA, 94577.

(10) Xue, L.; Godden, J.; Gao, H.; Bajorath, J. Identification of a Preferred Set of Molecular Descriptors for Compound Classification Based on Principal Component Analysis. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 699−704.

(11) Xue, L.; Godden, J. W.; Bajorath, J. Evaluation of Descriptors and Mini-Fingerprints for the Identification of Molecules with Similar Activity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1227−1234.

(12) Shaffer, R. E.; Rose-Pehrsson, S. L. Improved Probabilistic Neural Network Algorithm for Chemical Sensor Array Pattern Recognition. *Anal. Chem.* **1999**, *71*, 4263−4271.

(13) Mosier, P. D.; Jurs, P. C. QSAR/QSPR Studies Using Probabilistic Neural Networks and Generalized Regression Neural Networks. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1460−1470.

(14) *NeuroShell 2*; Ward Systems Group, Inc., Executive Park West, 5 Hillcrest Drive Frederick, MD 21703.

(15) Ghose, A. K.; Crippen, G. M. Atomic Physicochemical Parameters for Three-Dimensional-Structure-Directed Quantitative Structure−Activity Relationships. 2. Modeling Dispersive and Hydrophobic Interactions. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 21−35.

(16) Frimurer, T. M.; Bywater, R.; Naerum, L.; Lauritsen, L. N.; Brunk, S. Improving the Odds in Discriminating "Drug-like" from "Non Drug-like" Compounds. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1315−1324.

(17) Tripos Associates Inc., St. Louis, MO.

(18) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* **1997**, *267*, 727−748.

(19) Specht, D. F. A General Regression Neural Network. *IEEE Trans. Neural Networks* **1991**, *2*, 568−576.

JM0302795