Classification of Dopamine Antagonists Using TFS-Based Artificial Neural Network

Satoshi Fujishima and Yoshimasa Takahashi*

Laboratory for Molecular Information Systems, Department of Knowledge-Based Information Engineering, Toyohashi University of Technology, Hibarigaoka 1-1, Tempaku-cho, Toyohashi 441-8580 Japan

Received October 7, 2003

In the former work, the authors proposed the Topological Fragment Spectral (TFS) method as a tool for the description of the topological structure profile of a molecule. This paper describes the TFS-based artificial neural network (TFS/ANN) approach for the classification and the prediction of pharmacological active classes of chemicals. Dopamine antagonists of 1227 that interact with different types of receptors (D1, D2, D3, and D4) were used for the training. The TFS/ANN successfully classified 89% of the drugs into their own active classes. Then, the trained model was used for predicting the class of unknown compounds. For the prediction set of 137 drugs that were not included in the training set, the TFS/ANN model predicted 111 (81%) drugs of them into their own active classes correctly.

INTRODUCTION

We often say that "A is similar to B" or "C is similar to D in terms of xyz". "Similarity" is a very important concept in solving problems in science. This is true in chemistry. The use of molecular similarity methods, especially structural similarity, provides us with a lot of information on structure−activity and structure−property problems.[1,2] And it is still under active development in the area of drug design, for the selection of candidate analogues as new chemicals and for the estimation of molecular properties.[3−5] The basic idea behind them is that structurally similar compounds are likely to possess similar molecular properties and similar biological activities. Most of the approaches for the evaluation are based on finding particular functional atoms or atomic groups defined in advance. However, the result of such a structural similarity analysis depends on the chosen set of substructures defined as the descriptors.[6] In the former work, the authors proposed the Topological Fragment Spectral (TFS) method as a tool for the description of the topological structure profile of a molecule.[7] The method does not require any kind of a priori substructure definition like a dictionary file for substructures to be searched for. The TFS representation method is also useful for the similar structure searching on chemical structure databases[8] and the visualization of similar structure data space.[9] The aim of our these research projects is in establishing a basis of chemical data mining and risk report based on structural similarity without any set of substructures defined in advance.

In the present work, the utility of the TFS representation method will be validated for the classification and the prediction of pharmacological activity classes of drugs using the artificial neural network combined with the input signals of TFS descriptors.

* Corresponding author phone: +81-532-44-6878; fax: +81-532-44-6873; e-mail: taka@mis.tutkie.tut.ac.jp.

METHODS

TFS Representation of Chemical Structure. The TFS (Topological Structural Fragment) is based on the enumeration of all possible substructures from a chemical structure and the numerical characterization of them. For a given structure represented as a chemical graph (hydrogen suppressed graph), all the possible subgraphs embedded in it are enumerated. Subsequently, every subgraph is characterized with a specific numerical quantity. To perform the characterization we have used two methods in the present study as follows: (i) the overall sum of degree of the nodes composing each subgraph and (ii) the overall sum of the mass numbers of the atoms (atomic groups) corresponding to the nodes of the subgraph. With the first method a chemical structure is represented by a simple graph, thus the characterization of the structure depends only on the topology of the structural skeleton. For the second method, attached hydrogen atoms are taken into account as augmented atoms and are represented by weighting correspondingly their respective nodes in the graph. The latter is very similar to that of the mass spectra of chemicals. Thus the TFS can be regarded as a function of the chemical structure, i.e., $TFS = F(\text{chemical structure})$. A schematic flow of the TFS creation from a chemical structure is shown in Figure 1. The TFS of promazine characterized by two different methods is shown in Figure 2.

The computational time required for the exhaustive enumeration of all possible substructures from a chemical structure is often very large especially for the molecules that involve highly fused rings. In addition to this, a large difference in the dimensionality between the fragment spectra to be compared may lead to the unexpected result. To avoid these problems an alternative approach based on the use of the subspectrum may be employed for such a similarity analysis, in which each spectrum can be described with structural fragments up to a specified size in the number of edges (bonds).

10.1021/ci030035t CCC: $27.50
Published on Web 04/01/2004

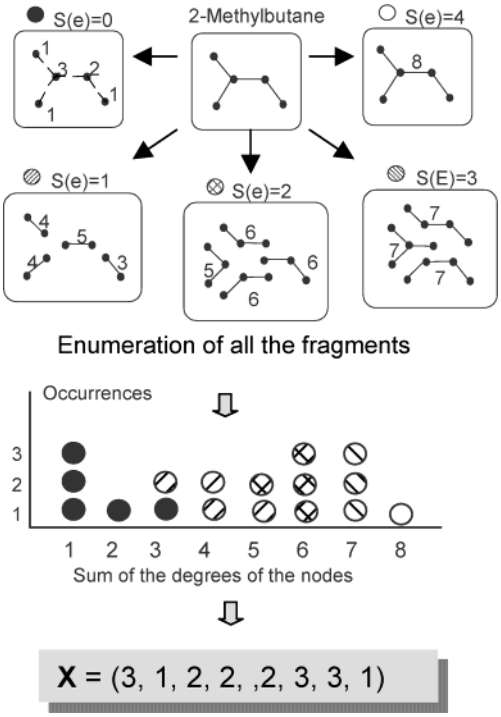

Figure 1. A schematic flow of TFS creation. S(e) is the number of edges (bonds) of fragments to be generated.

Quantitative Evaluation of Structural Similarity Based on the TFS. Obviously, the fragment spectrum obtained by these methods can be described as a kind of multidimensional pattern vector. Consequently, using this pattern representation of a spectrum it is possible to apply various quantitative measures for the evaluation of similarity. In the present work, Euclidean distance measure was used for evaluating the similarity

$$D(X_i,X_j) = \sqrt{\sum (x_{ik} - x_{jk})^2} \qquad (1)$$

where x_{ik} and x_{jk} are pattern vectors which represent the frequency value of peak k of fragment spectra of the ith molecule and the jth molecule, respectively. $D(X_i,X_j)$ is the Euclidean distance between the patterns X_i and X_j. The different dimensionalities of the spectra to be compared are adjusted as follows: If

$$X_i = (x_{i1},x_{i2},...,x_{iq}) \text{ and } X_j = (x_{j1},x_{j2},...,x_{jq},x_{j(q+1)},...,x_{jp}) \quad (q<p)$$

then

$$X_i \text{ is redefined as } X_i = (x_{i1},x_{i2},...,x_{iq},x_{i(q+1)},...,x_{ip})$$

here,

$$x_{i(q+1)} = x_{i(q+2)} = ... = x_{ip} = 0$$

This approach was fully computerized and used in the following data analysis.

Data Set. In this work we employed 1364 dopamine antagonists that interact with four different types of receptors (D1, D2, D3, and D4 receptors of dopamine). The data are taken from the MDDR[10] database that is a structure database of investigative new drugs, and they are all the data of dopamine receptor antagonists that are available on it. The data set was divided into two groups: the training set and the prediction set. The training set consists of 1227 compounds (90% of the total data), and the prediction set consists of 137 compounds (10% of the total data) remained. They were randomly chosen.

Neural Network. Discrimination of pharmacological activity classes of chemicals was investigated using artificial neural network (ANN). A three-layer learning network with a complete connection among layers was used. The TFS was submitted to the ANN as input signals for the input neurons. The number of neurons in the input layer was 165, that is the same as the value of dimensionality of the TFS. The number of neurons in the single hidden layer was determined by trial and error. Training of the ANN was carried out by error back-propagation method. All the neural network

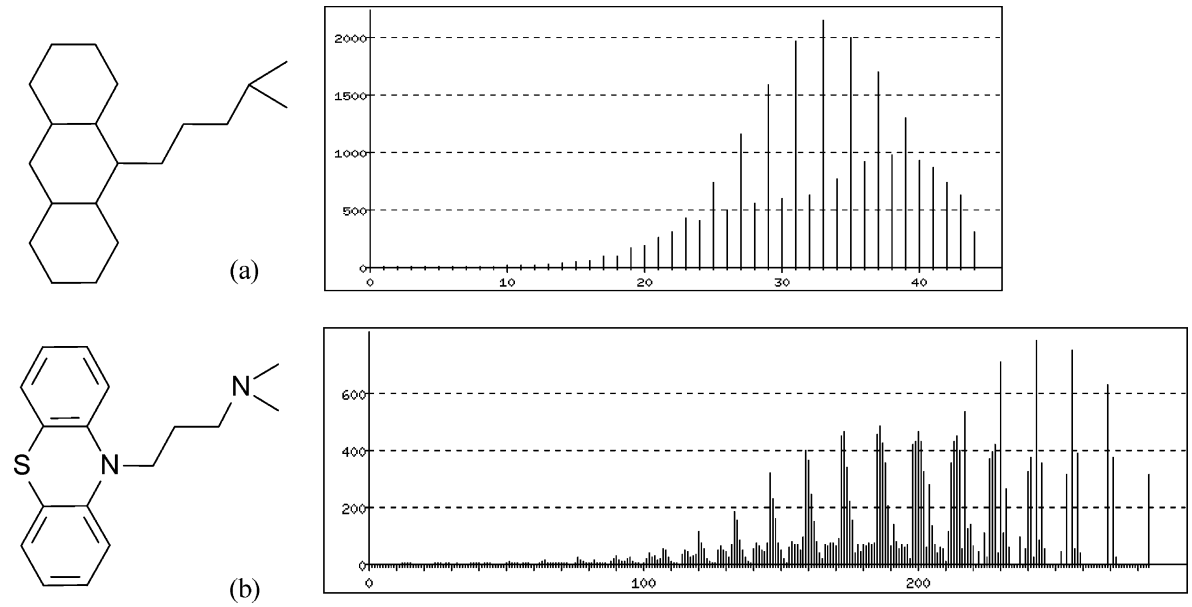

Figure 2. TFS of promazine generated by the different characterization methods. (a) is characterized by the sum of degrees of nodes on the fragments. (b) is characterized by the sum of atomic mass numbers in fragments.

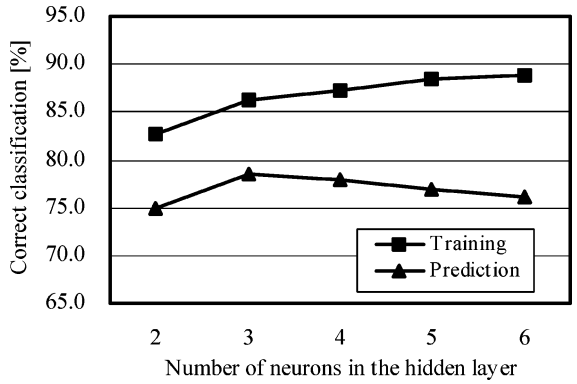

Figure 3. Results of the training and the prediction by the TFS-based artificial neural network with the different number of neurons in the hidden layer. For the training, 1227 of the dopamine antagonists were used, and 137 of them for the prediction. Every plot was based on the average value of five trials with the different initial weights randomly generated.

Table 1. Results of TFS-Based Neural Network Analysis for 1364 Dopamine Receptor Antagonists

	training		prediction	
class	no. of samples	correct (%)	no. of samples	correct (%)
all	1227	1087 (88.6)	137	111 (81.0)
D1	155	112 (72.3)	18	11 (61.1)
D2	356	312 (87.6)	39	27 (69.2)
D3	216	193 (89.4)	24	23 (95.8)
D4	500	470 (94.0)	56	50 (89.3)

analyses in this work were carried out using a computer program, NNQSAR, developed by the authors.[11]

RESULTS AND DISCUSSION

Classification of Dopamine Receptor Antagonists by TFS/ANN. Here, dopamine antagonists of 1227 that interact with different type of receptors (D1: 155 compounds, D2: 356 compounds, D3: 216 compounds, and D4: 500 compounds) were used for training an artificial neural network (ANN) with their TFS to classify the type of action. For the first, we tried to determine the suitable number of neurons in the hidden layer of the TFS/ANN model to give the correct predictions as much as possible. Due to the number of input neurons of 165, two to six neurons for the hidden layer were tested in the computational trials. The results of the preliminary tests suggested that the TFS/ANN model with the hidden layer of three neurons would be better than those with the others for the predictive ability and the stability (Figure 3).

Thus, the ANN model of 165 × 3 × 4 neurons was employed for the present work. The TFS/ANN model that gave us the best result for the present data set correctly classified 88.6% (1087/1227) of the training set into their own classes. For 137 compounds in the prediction set, the TFS/ANN correctly predicted the active classes of 111 compounds. The total prediction rate was 81.0%. The details of the results are summarized in Table 1.

In the comparison between the results in Table 1, it is shown that the TFS/ANN gave us well-trained models and a better prediction for D3 and D4 antagonists. However the table shows the results for D1 receptor antagonists was poorer than those for other classes in both cases of training and prediction. It is considered that the TFS/ANN model could not learn enough for the training set because the number of samples was relatively smaller than those of the other classes. While the prediction result for D2 antagonists was not better than that for D3 antagonists, nonetheless the number of D2 antagonists in the training was larger than that of D3 antagonists. It is considered that the D2 antagonists in the database may contain the data tested before the current classification of dopamine receptors. But the table shows that the total prediction rate was still good. These results show that the TFS is a very powerful tool to describe the structural features of chemicals, and it would be suitable to use an input signal along with artificial neural network modeling for the classification and the prediction of a pharmacological active class of them.

Table 2. Results for the Training in 10-Fold Cross-Validation Test

training set (recognition rate %)					
class	D1	D2	D3	D4	all
max	80.0	87.6	92.6	97.2	88.7
min	72.3	74.1	87.5	93.0	85.0
average	76.5	80.8	91.1	94.4	87.6

Table 3. Results for the Prediction in 10-Fold Cross-Validation Test

prediction set (prediction rate %)					
class	D1	D2	D3	D4	all
max	88.2	82.5	95.8	93.4	83.1
min	61.1	67.5	75.5	85.5	79.6
average	73.2	72.2	88.9	90.5	81.0

Cross-Validation Test. In the previous section, the TFS-based artificial neural network successfully classified and predicted many of the dopamine antagonists into their own active classes. We also tested the validity of the results using a cross-validation technique. For the same data set of 1364 compounds that contain both the training set and the prediction set, 10 different trial data sets were generated. For making these trial data sets the whole data were randomly divided into 10 subsets that had almost the same number of samples (137 or 136 compounds for each). Each trial data set was used as the test set for the prediction, and the remaining data sets were employed for the training of the TFS/ANN. The results of these computational trials were summarized in Tables 2 and 3.

In the cross-validation test, the total recognition rates for 10 trials of the training were 85.0−88.7%. The total average of recognition rates in the training was 87.7%. Then, for the predictions the total recognition rates for 10 trials were 79.6−83.1% in the correct prediction rate. The total average of the prediction rates for the cross-validation test was 81.0%. These values are quite similar to those described in the previous section. It is considered that the present results strongly validate the utility of the TFS-based neural network approach to the problems for classification and prediction of pharmacological active compounds.

CONCLUSION AND FUTURE WORK

It is concluded that the topological fragment spectra is a very powerful tool to describe the structural features of chemicals, and it would be suitable to use an input signal along with an artificial neural network for the classification of pharmacologically active compounds and the prediction

of their active class. The authors would like to extend the TFS/ANN model for other active class compounds in future work. An additional tool that can be used for identification and interpretation of the TFS peaks our interest and will be also required for knowledge discovery of the related area.

ACKNOWLEDGMENT

The authors thank Prof. Takashi Okada and Dr. Masumi Yamakawa for the fruitful discussion and comments. This work was supported by Grant-In-Aid for Scientific Research on Priority Areas (B) 13131210, Ministry of Education, Culture, Sports, Science and Technology, Japan.

REFERENCES AND NOTES

(1) *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; Wiley: New York, 1990.
(2) Takahashi, Y. Identification of Structural Similarity of Organic Molecules. *Topics Curr. Chem.* **1995**, *174*, 105−133.
(3) Rarey, M.; Stahl, M. Similarity Searching in Large Combinatorial Chemistry Spaces. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 497−520.
(4) Raymond, J. W.; Willett P. Effectiveness of Graph-Based and Fingerprint-Based Similarity Measures for Virtual Screening of 2D Chemical Structure Databases. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 59−71.
(5) Wilton, D.; Willett, P. Comparison of Ranking Methods for Virtual Screening in Discovery Program. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 469−474.
(6) Flower, D. On the Properties of Bit String-Based measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379−386.
(7) Takahashi, Y.; Ohoka H.; Ishiyama Y. Structural Similarity Analysis Based on Topological Fragment Spectra. In *Advances in Molecular Similarity*; Carbo, R., Mezey, P., Eds.; JAI Press: Greenwich, CT, 1998; Vol. 2, pp 93−104.
(8) Takahashi, Y.; Fujishima, S.; Kato, H. Chemical Data Mining Based on Structural Similarity. *J. Comput. Chem. Jpn.* **2003**, *2*, 119−126.
(9) Takahashi, Y.; Konji, M.; Fujishima, S. MolSpace: A Computer Desktop Tool for Visualization of Massive Molecular Data. *J. Mol. Graph. Model.* **2003**, *21*, 333−339.
(10) MDL Drug Data Report, MDL, Ver. 2001.1, 2001. The list of compounds used here is available by requesting to the authors.
(11) Ando, H.; Takahashi, Y. Artificial Neural Network Tool (NNQSAR) for Structure−Activity Studies, *Proceedings of the 24th Symposium on Chemical Information Sciences*, 2000, pp 117−118.

CI030035T