Bull. Chem. Soc. Jpn., 73, 1955—1965 (2000)

Headline Articles

Classification and Prediction of Reagents' Roles by FRAU System with Self-Organizing Neural Network Model

Hiroko Satoh,*,# Kimito Funatsu,† Keiko Takano,†† and Tadashi Nakata

RIKEN (The Institute of Physical and Chemical Research), 2-1 Hirosawa, Wako, Saitama 351-0198

†Department of Knowledge-Based Information Engineering, Toyohashi University of Technology, 1-1 Tempaku, Toyohashi, Aichi 441-8580

††Department of Chemistry, Faculty of Science, Ochanomizu University, 2-1-1 Otsuka, Bunkyo-ku, Tokyo 112-8610

(Received March 17, 2000)

The classification and prediction of the roles for reagents in reactions are presented. The same dimensional representation of various reagents independent of the number of atoms was achieved by selecting representative factors by the FRAU (Field-characterization for Reaction Analysis and Understanding) system. Training of a self-organizing model considering both negative and absent data was accomplished by modifying the original counter-propagation (CP) type of Kohonen neural network to treat absent data differently from negative data. The modified CP Kohonen neural network successfully classified the reagents and produced a reagent-roles correlation model that gives good answers predicting roles of the reagents.

Reagents play important roles in chemical reactions, and very various reagents have been developed and used in synthetic studies. The roles of reagents may vary according to the substrate, solvent, and so on. In other words, a reagent has various potential roles in reactions. If the roles of the reagents are numerically predicted on a computer before an experiment, it might be very useful in broad fields of synthetic study. For example, the most preferable reagents for the desired reaction can be chosen before synthesis.

In general, reagents in similar structures and electronic features have similar roles, because the potential roles of a reagent are largely related to the features. If sufficient methods to numerate the structural and electronic features and to explain the roles based on the numeric features are possible, numerical prediction of the roles can be achieved.

Recently, we developed the FRAU (Field-characterization for Reaction Analysis and Understanding) system and demonstrated its usefulness to discriminate similarities and differences in structures as well as the roles of metallic reagents based on FRAU's features for metallic atoms and atoms connecting to the metallic atoms.[1,2]

FRAU estimates the possibilities of occurring reactions for a molecule (e.g., a reagent) based on electrostatic and steric interactions with a pseudoreactant (e.g., a substrate). The good points of FRAU are to take account of three-dimensional field around a molecule, and to measure the molecular features, called FRAU features (FFs), based on interactions with a pseudoreactant to analyze chemical reactions. The input data of FRAU are a molecular structure with atomic charges. An arbitrary computational level to obtain the input data is adaptable according to the purposes; namely, the degrees of accuracy and the calculation time are adjustable to the purposes. Thus, FRAU makes it possible to rapidly obtain more accurate properties for molecules than the properties based on topological relationships between atoms.

The number of the FFs, which are estimated for each atom, depends on the number of atoms in a molecule. The atomic features are necessary to know important sites and directions for occurring reactions. However, in order to compare between reagent molecules, representation by the same number of features for various reagents independent of the number of atoms are needed. The same dimensional representation of molecules is one of the important problems in reaction classification studies, where comparing points are needed to discuss the similarities and differences.

In the meanwhile, a neural network is one of appropriate methods for treating multi-dimensional nonlinear data. We used neural networks for reaction classification studies[1,3] because the factors controlling reactions can be considered as data in multi-dimensional space and a reaction resulting from

PRESTO, JST(Japan Science and Technology Corporation).

complicated interactions among these factors can be considered as nonlinear relationships of these multi-dimensional data. Pioneer studies that used neural networks for reactions were performed by Gasteiger et al.[4—9] The first one was an application of a multilayer back-propagation neural network and an associative memory system to learning and predicting the reactivity for a bond.[4,5]

Recently, a self-organizing neural network introduced by Teuvo Kohonen[10—12] has been applied to reaction classification studies, and its usefulness has been demonstrated.[1,3,7—9] The Kohonen neural network makes it possible to classify input data based on the nature being inherent in the input data. This character is suitable for classifying reactions without any preconception of categories or types of reactions as the templates. This is one of the reasons why we have been using the Kohonen neural network for our reaction classification studies.[1,3]

The classification can be developed for predictions by a counter-propagation type of Kohonen neural network (hereafter, called CP Kohonen neural network),[12,13] that is a Kohonen neural network to which known signals and an output layer are added. The processes of characterization, classification and prediction are *formally* similar to the thinking way of a chemist's brain, where, following the recognition and classifications of reactions, predictions are performed. The CP Kohonen neural network gives an organizing map where input data are classified in the basis of the similarities in the input data with a map, where the known signals are trained according to the similarities in the input data. Good points of the CP Kohonen neural network comparing a back-propagation type is to give a stable interpolated answer and to show us major factors of the input data that concern with the interested similarities because the known signals do not influence the classification. The CP Kohonen neural network has an advantage to treat non-linear data that fits our purposes, although there are many statistical methods used to determine major factors, such as principal component analyses. The CP Kohonen neural network trains values for prediction according to similarities in the input data. Thus, if the similarities in the input data have good correspondence with the similarities in the known signals, the CP Kohonen neural network produces a good model that can give an appropriate answer. However, in the original CP Kohonen neural network, which treats one type of known signal, it is not distinguishable between negative and absent data. In order to treat reaction data, it is needed to treat absent data differently from negative ones, e. g., reactions and reagents that did not occur and work, respectively. The negative data are important to consider chemical reactions.

The purpose of this article is to develop and to show a process for the characterization, classification, and prediction of the potential roles of reagents using chemical information with more accuracy than the properties based on the topological relationships between atoms. For this purpose, we first solved problems about the same dimensional representation of reagents based on FFs independent on the number of atoms. Second, we improved the CP Kohonen neural network to treat absent data differently from negative ones. We then have used these results to construct a reagent-roles correlation model that predicts the potential roles of the reagents.

Method

1. Overview. First, a combination of FFs that has a good correspondence with the similarities in the reagents' roles was selected as the representative set of FFs. The selection was performed by searching for a good classification of thirty kinds of reagents (Fig. 1) using the Kohonen neural network.[14] Then, an original counter-propagation (CP) Kohonen neural network was modified to distinguish between absent and negative data. Using the modified CP Kohonen neural network, a self-organizing model that relates similarities in the representative FFs of the thirty kinds of reagents with the roles was constructed. The reproduction and prediction of roles by the model were examined for the reagents and ten kinds of reagents as test data (Fig. 2). Detailed procedures are described below.

2. Model Structures. For the first purpose that is representation of molecules in various structures by the same number of features, boranes, hydrides, Grignard reagents, and bases shown in Fig. 1 were used as the models. The second purpose is to predict the potential roles of the reagents from the reagents' features by a chemical informational approach. The mechanism analyses of reactions for reagents with specific substrates under specific conditions (e.g., solvents, temperature, concentration) are not the purposes here. The structures for some of the reagents in Fig. 1 have been studied by X-ray, NMR, and theoretical analyses.[15—20] Boranes are generally known as dimeric structures, and it is considered that the active species in hydroborations are monomeric structures. Aluminium hydrides, such as diisobutylaluminium hydride (DIBAL) (**15**), are also considered to be monomeric active species in hydroaluminations. Grignard reagents exist as an equilibrium mixture which may involve various mono-, bi-, and poly-nuclear solvated component in solutions, that are known as Schlenk equilibrium, and active species have not been determined because they are varied by the structures of the Grignard reagents, solvents, concentration, and temperature. In this study, simple stable structures in the ground state were analyzed, because if the features of simple structures are well concerned with the roles in reactions, the simple features can be considered to be essential to determine the roles. If the simple features are not sufficient to explain the roles, this means that more complicated features are needed, which will be added in the next stage. The structures were optimized by ab initio RHF/3-21G-(*) molecular-orbital calculations[21] to obtain the geometries and charges for FRAU calculations. Ab initio calclulations were used here to obtain reliable properties for the consisted metallic atoms.

3. Classification of Reagents. Calculation of FFs. FRAU calculated FFs for the thirty kinds of reagents. FRAU estimates three kinds of FFs: extent of reaction field (FF_{field}), electrostatic feature ($FF_{electro}$), and steric feature (FF_{steric}). $FF_{electro}$ and FF_{steric} were calculated in the basis of interactions with unit charge (+1) and with an sp^3 carbon as probes to detect the features, respectively. The positions of each of the probes are dots that are evenly dispersed on the surface of a sphere drawn around each atom in a molecule. The radius of the sphere is the van der Waals radius in this study, and any radius can be chosen. The detailed procedures of FRAU were described in our previous paper.[1]

Selection of Representative FRAU Features. FFs are calculated to each atom in a molecule. Thus, the number of sets of

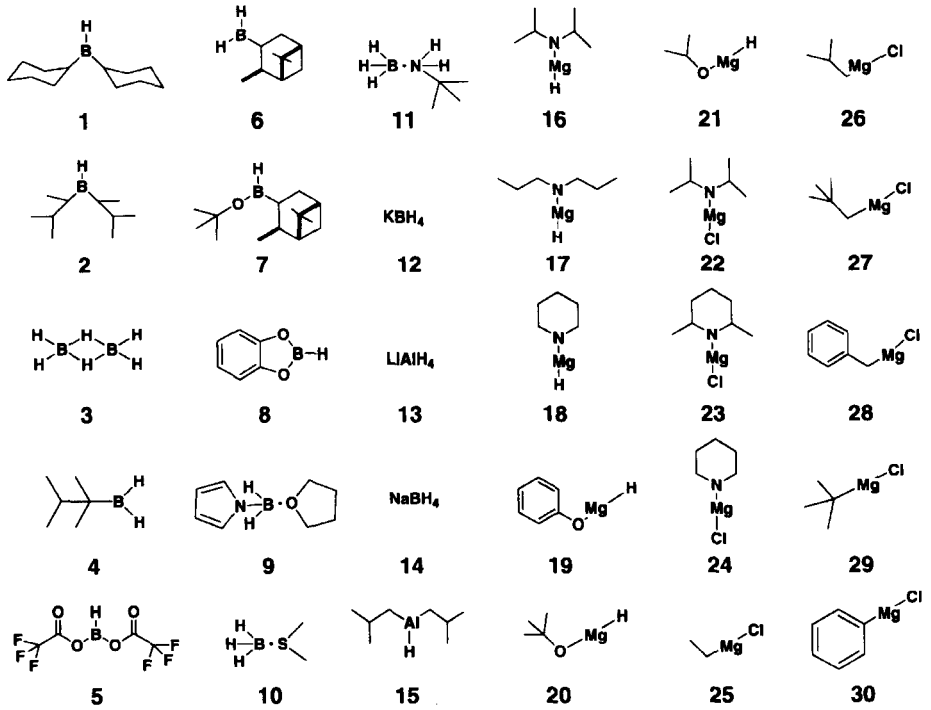


Fig. 1. Model reagents (planar descriptions). Three dimensional structures were optimized by ab initio RHF/3-21G(*) calculations. Optimized geometries for **12**, **13**, and **14** were octahedral type of structures.

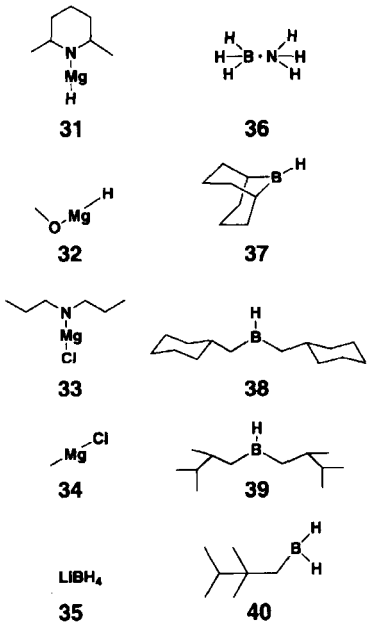


Fig. 2. Reagents for test. They were not used for construction of the model.

the three kinds of FFs is the same as the number of atoms in the molecule. For example, for dicyclohexylborane, which contains 36 atoms ($C_{12}H_{23}B$), 36 sets of the three kinds of FFs are obtained.

In order to compare between reagents in various structures, it is needed to represent the reagents by the same dimensional factors. Thus, the same dimensional representative FFs that have good correspondence with similarities in the structures and roles were selected after investigating various combinations of FFs.

Classification. The selected FFs were input to a Kohonen neural network as discriminators to classify thirty kinds of the reagents (Fig. 1).

The Kohonen neural network outputs a planar map that is called as Kohonen map, where similar input data are set into the same or close neuron, and does not perceive boundaries between clusters of the input data on the Kohonen map. The boundaries were determined by degree of gap between the weight vectors of neighboring neurons by a U-matrix method.[22]

4. Construction of a Reagent-Roles Correlation Model. Counter-Propagation (CP) Kohonen Neural Network. A CP Kohonen neural network is constructed from a Kohonen neural network by the composing of output layers and known answer signals (Fig. 3-(1)). During the training of a CP Kohonen neural network, a modification of the weight vectors between the active and output layers is also performed. In the training, the position of the connection between neurons of active and output layers is determined according to similarities in the input signals, and the dimension and value of the weight vectors between output and active layers are determined according to the signals of a known answer. Weight vectors between output and active layers corresponding to a known answer are trained according to the similarities in the input signals.

Modification of the Original CP Kohonen Neural Network. In order to treat reaction data, negative and absent data should be distinguished; namely, it should be distinguished whether the reaction (reagent) did not occur (work) or the data is just absent from the data set. The distinction is hardly adapted to the original CP Kohonen neural network that has only one type of known answer signals. Thus, the original CP Kohonen neural network was modified according to what follows.[14]

The architecture of the modified CP Kohonen neural network is shown in Fig. 3-(2), where known answer signals for the presence of data are added. The additional signals with the other known answer signals control the training rate of weight vectors between

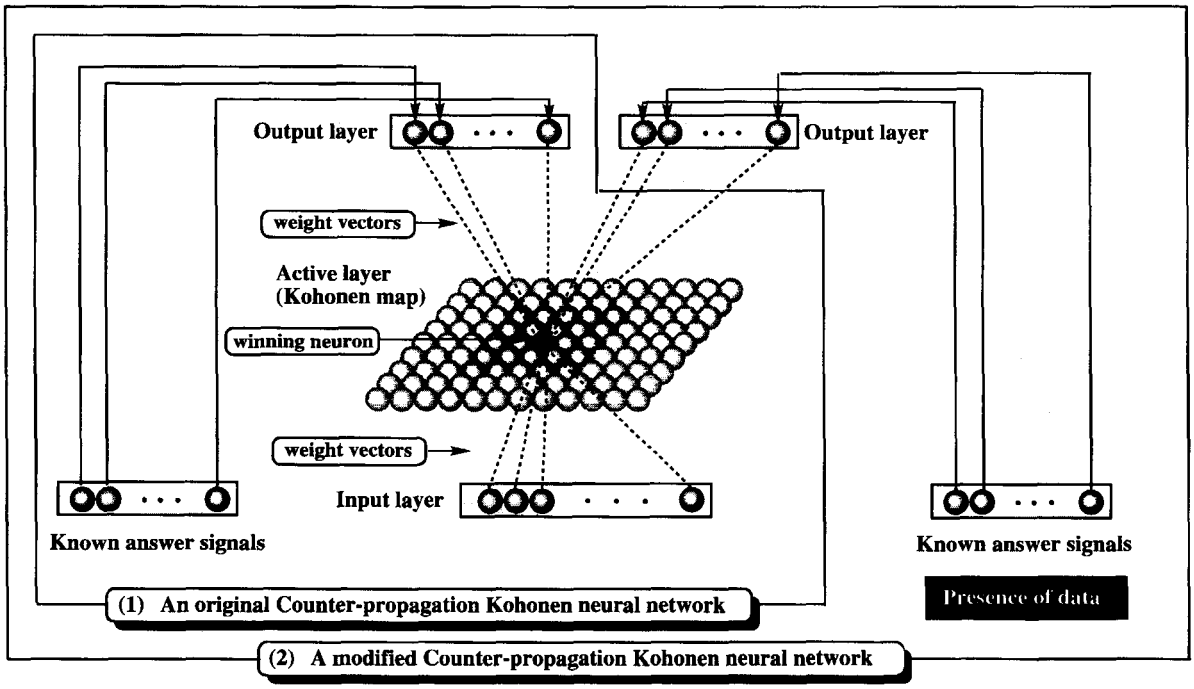


Fig. 3. Counter-propagation Kohonen neural network. (1) An original counter-propagation Kohonen neural network. An output layer and signals of known answer are added to a Kohonen neural network. During the training, modification of weight vectors between the active and output layers is also performed. (2) A counter-propagation Kohonen neural network modified and used, here. A known signal and an output layer for presence of data were added and used in training weight vectors between the active and the output layers.

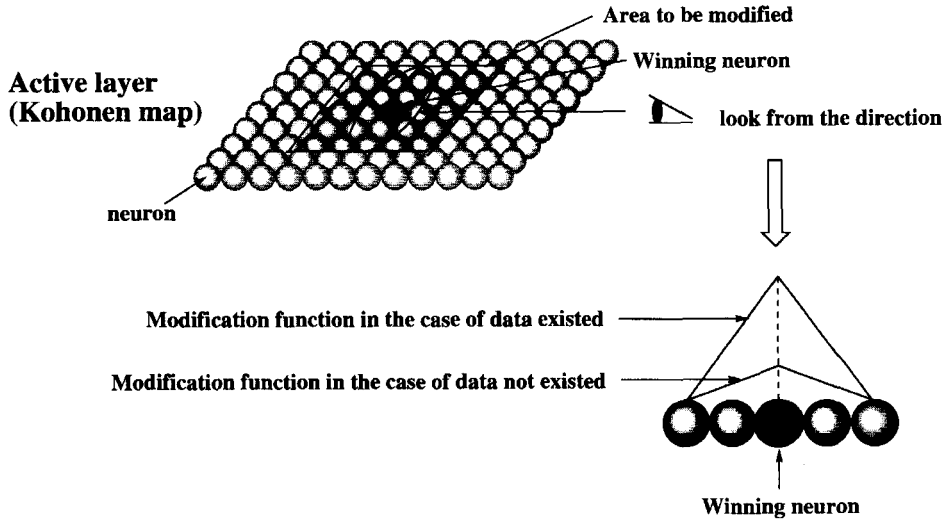


Fig. 4. How to introduce absent data. A rate of the modification function for weight vectors between active and output layers is reduced for the absent data. In this case, triangle modification function was used.

the active and output layers. As shown in Fig. 4, where a triangle function is used as a modification function; when the data is absent, the training rate for weight vectors is reduced to small rather than that for the existed data. Namely, a small voice is given to the absent data to express the possibilities of whether the data might be absent because the reaction (reagent) does not actually occur (work). The volume of the voice for absent data is one of the input parameters.

Construction of the CP Kohonen Neural Network Model. The modified CP Kohonen neural network was used to construct a correlation model between reagents and the roles in reactions. In the construction, the selected representative FFs as the input signals, the roles of the reagents as the first known answer signals, and the information on the presence of the data as the second known answer signals were input. As the known answer signals, nine types of roles of the reagents were set with the information on the presence of the data that were represented by binary notations (1 or 0), as shown in Table 1. All of the binary notations for each reagent are listed in Table 2. The input data for the reagents' roles were determined from some literature and books,[23,24] as well as SYNLIB[25] and ISIS[26] databases. When the working of a reagent against the same functional group varies, a binary notation was determined as follows. If an additional reagent is necessary for

Table 1. Binary Notations Used as Input and Known Signals in the Training of CP Kohonen Neural Network

	There is data where the reagent worked.		There is data where the reagent did not work.		There is no data.	
	Known signal	Presence of data	Known signal	Presence of data	Known signal	Presence of data
Reduction of ketone to alcohol	1	1	0	1	0	0
Reduction of aldehyde to alcohol	1	1	0	1	0	0
Reduction of carboxylic acid to alcohol	1	1	0	1	0	0
Reduction of ester to aldehyde	1	1	0	1	0	0
Reduction of ester to alcohol	1	1	0	1	0	0
Reduction of epoxide	1	1	0	1	0	0
Hydroboration or hydroalumination	1	1	0	1	0	0
Base	1	1	0	1	0	0
Alkylation	1	1	0	1	0	0

Table 2. Input Signals for Thirty Reagents in the Training of Kohonen Neural Network[a,b]

Reagent No.	Reduction of ketone to alcohol		Reduction of aldehyde to alcohol		Reduction of carboxylic acid to alcohol		Reduction of ester to aldehyde		Reduction of ester to alcohol		Reduction of epoxide		Hydroboration or hydroalumination		Base		Alkylation	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
1	1	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
2	1	1	1	1	0	1	0	1	0	1	0	0	1	1	0	0	0	0
3	1	1	1	1	1	1	0	0	0	0	0	0	1	1	0	0	0	0
4	1	1	1	1	1	1	0	0	0	0	0	0	1	1	0	0	0	0
5	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	1	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
7	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	1	1	1	1	0	0	0	1	0	1	0	0	1	1	0	0	0	0
9	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
10	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	0	0	0
11	1	1	1	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0
12	1	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
13	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0
14	1	1	1	1	0	1	0	0	1	1	0	1	0	0	0	0	0	0
15	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
16	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
17	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
18	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
19	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0
25	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
26	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
28	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
29	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
30	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1

a) Reagents' numbers that correspond to those in Fig. 1. b) Roles in reactions. (1) Known answer signals for the roles of reagents. (2) Known answer signals for presence of data.

an interested reagent to work, such as the reduction of epoxides under $NaBH_4$ with $BF_3 \cdot Et_2O$, it was not determined whether the reagent works, because in such a case the interested reagents are changed to the actual active species. If an interested reagent works, even against a specific substrate, it was determined that the reagent works. The volume of the voice for absent data was set at 30% of that for the existing data.

5. Prediction of Roles of Reagents. Reproduction and prediction by the constructed model were investigated for ten reagents used in the modeling (structures of **4**, **6**, **8**, **10**, **13**, **15**, **16**, **21**, **23**, and **27**), and for ten reagents not used in the modeling as test data (Fig. 2), respectively. Reagents **38**, **39**, and **40** have not been

reported, and for the others, experiments have been reported.

Results and Discussion

1. Results of the Classification. A set of six parameters consisting of maximum and minimum $FF_{electro}$ values in a molecule with values of FF_{field} and FF_{steric} on the sites having the maximum and minimum $FF_{electro}$ successfully distinguished similarities and differences in the reagents and was selected as a representative set of FFs. The selected representative sites and the FFs for each molecule are shown in Fig. 5 and Table 3, respectively.

The selected set of FFs was used as discriminators to classify the reagents by the Kohonen neural network. The resulting Kohonen map is shown in Fig. 6-(I). In Fig. 6-(I), each square is a neuron. Reagents were set on neurons where they were drawn. The black solid lines are boundaries recognized by the U-matrix method. The bold solid lines mean a larger difference between neurons than the fine solid lines, and the fine solid lines mean a larger difference between the neurons than the dotted lines. The actual shape of this map is a torus. For visualization, the torus is cut along two perpendicular lines and the surface is spread into a plane. Thus, the top of this map connects with the bottom, and the left edge connects with the right one.

On the Kohonen map, input data having similar FFs are set into the same or close neuron. Thus, the correlation between the FFs and types of the reagents can be found by labeling the map according to the similarities in the types. Figure 6 is colored according to the similarities in the roles of the reagents. It shows that reagents which play similar roles were successfully set on the same or close neurons forming clusters. Borane-dimethyl sulfide (BMS) (**10**) was isolated from cluster **b**, colored grayish blue. The location of BMS is suitable because the roles of BMS are rather similar to those of boranes (cluster **a** colored light blue), even though BMS belongs to borane complexes. The results show that the selected set of FFs has a good correspondence with the similarities of the roles of the reagents.

The reason why several combinations of maximum and minimum values of FFs were mainly investigated for selecting representative FFs is because it is considered that those extreme values might have a higher influence to determine the characters of the reagents. A good correspondence of the six parameters (the maximum and minimum $FF_{electro}$ values in a molecule and values of FF_{field} and FF_{steric} on the sites having the maximum and minimum $FF_{electro}$) with the roles shows that those sites having the maximum and minimum electrostatic interaction energy are major for determining the reagents' characters. Then, in order to know which are the major parameters to discriminate the similarities and differences in the reagents, neurons are colored according to the size of weight vectors corresponding to the six parameters, as shown in Fig. 7. Referring to the maps in Fig. 7 with the classification results in Fig. 6 let us know that the major factors to determine the boundaries are the maximum and minimum of $FF_{electro}$ values. Various combination of FFs were investigated to select the representative FFs, and a set of the maximum and minimum $FF_{electro}$ values was not sufficient to clearly discriminate the similarities, although it was also investigated. This means that the steric interaction

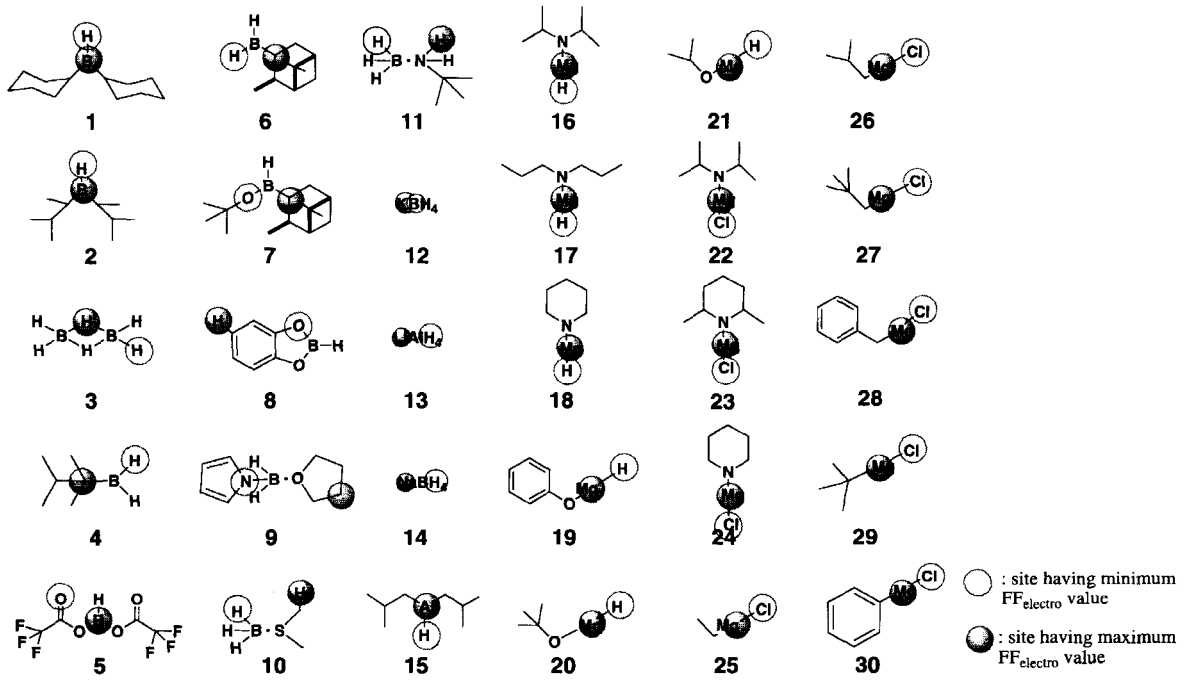


Fig. 5. Selected representative sites. The sites having maximum and minimum $FF_{electro}$ values were important sites to determine the reagents' characters. A set of six parameters of maximum and minimum $FF_{electro}$ values in a molecule with values of FF_{field} and FF_{steric} on the sites having the maximum and minimum $FF_{electro}$ were selected as discriminators that distinguishes similarities and differences in the reagents' roles.

Table 3. Selected Representative FRAU's Features (FFs) for the Reagents

Reagent	Sites having minimum $FF_{electro}$			Sites having maximum $FF_{electro}$		
No.	$FF_{electro}$	FF_{steric}	FF_{field}	$FF_{electro}$	FF_{steric}	FF_{field}
1	−13.454	256.961	253068	10.919	281.860	309377
2	−9.1724	271.503	252849	12.776	292.972	292352
3	−5.2507	267.347	248989	13.274	313.062	151018
4	−11.480	238.851	251560	14.973	477.965	18533
5	−8.1680	190.646	524012	44.590	203.644	438130
6	−11.526	235.594	251559	7.5380	424.580	48907
7	−33.422	222.098	284007	15.621	285.051	216902
8	−26.279	215.623	326616	16.355	261.958	238697
9	−35.407	430.112	68969	24.953	304.273	206290
10	−19.256	260.357	250266	14.445	262.250	233260
11	−39.802	269.070	256267	35.413	258.575	258254
12	−67.907	275.703	370214	26.285	19.6603	2422772
13	−29.474	239.217	257019	41.683	99.0511	847636
14	−55.343	261.024	253623	35.444	39.2887	1526515
15	−20.992	241.131	328616	29.007	265.351	552630
16	−40.573	231.209	362756	26.975	250.520	678728
17	−40.753	230.966	362876	25.117	250.187	704083
18	−36.978	230.305	362680	33.482	228.767	779961
19	−29.897	230.305	360838	42.365	211.166	773137
20	−35.948	230.312	362154	32.833	211.975	772059
21	−36.011	230.151	361815	32.636	211.224	771967
22	−25.630	209.678	946002	38.978	271.667	594481
23	−25.464	209.524	945928	38.632	271.077	615552
24	−22.381	208.736	945287	46.611	244.503	704713
25	−23.125	208.351	946616	50.338	224.836	703261
26	−24.032	208.623	946875	49.068	236.137	680578
27	−24.554	208.811	946937	48.018	246.892	670890
28	−25.820	209.217	946016	42.793	241.687	642569
29	−23.023	208.623	946772	53.404	237.187	692784
30	−20.703	208.659	945317	51.967	231.063	698610

a) Reagents' numbers that correspond to those in Fig. 1.

energy and field extension around the sites are also important to determine characters of the reagents as well as the electrostatic interaction energies on the sites.

2. Results from the Reproduction and Prediction. Reagents Used in the Modeling. The predicted degrees of the roles for ten reagents used in the training of the CP Kohonen neural network model are given in Table 4 along with the input binary notations that stand for the roles and presence of the data. In Table 4, a value that is described on values in parenthesis is the predicted one; the first and second values in parenthesis are the input signal for the roles and for presence of the data, respectively. The range of predicted values was from 0.0 to 1.0. The values of 0.0 and 1.0 correspond to the lowest and highest possibilities to react, respectively.

Almost all of the input values were successfully reproduced, as shown in Table 4, where the predicted values that are largely different from the input ones are underlined. This shows the good capability of the model to reproduce the input values. The underlined cases are discussed below.

The predicted values for roles of isopinocamphan-3-ylborane (**6**) as a reducing agent from aldehydes to alcohols and as that from carboxylic acids to alcohols were extremely different from the input signal. The input signal for the role of **6** was 0 and the predicted value was 0.8. Namely, the organizing map predicted 80% of possibility for **6** to work as reducing agents from aldehydes to alcohols and from carboxylic acids to alcohols, although the input signal denoted that there were no data. The predicted degrees of the roles were according to the similarities in the input FFs between **6** and the other reagents, such as 1,1,2-trimethylpropylborane (thexylborane) (**4**), gathered closely in the Kohonen map. The predicted values for **6** are suitable to knowledge that chemists possess.

For $LiAlH_4$ (**13**), the predicted possibility as a reducing agent from esters to aldehydes was 0.4, although the input signal denoted that it does not work. This was caused by nearby position of **13** to that of **15**. Incidentally, it is known that **13** with Et_2NH reduces esters to aldehydes.

For DIBAL (**15**), 1.0 was predicted to reduce carboxylic acid as the possibility. It is known that **15** does not reduce carboxylic acids, except pyrazolecarboxylic acids. In this execution, for a reagent that works even against to a specific substrate, it was determined the reagent works. Thus, this result is an instance of an overestimate. Some extension is needed in the future, where information on the majorities of the negative data, namely, reaction data on **15** not reactive to many carboxylic acids, are also considered in training the

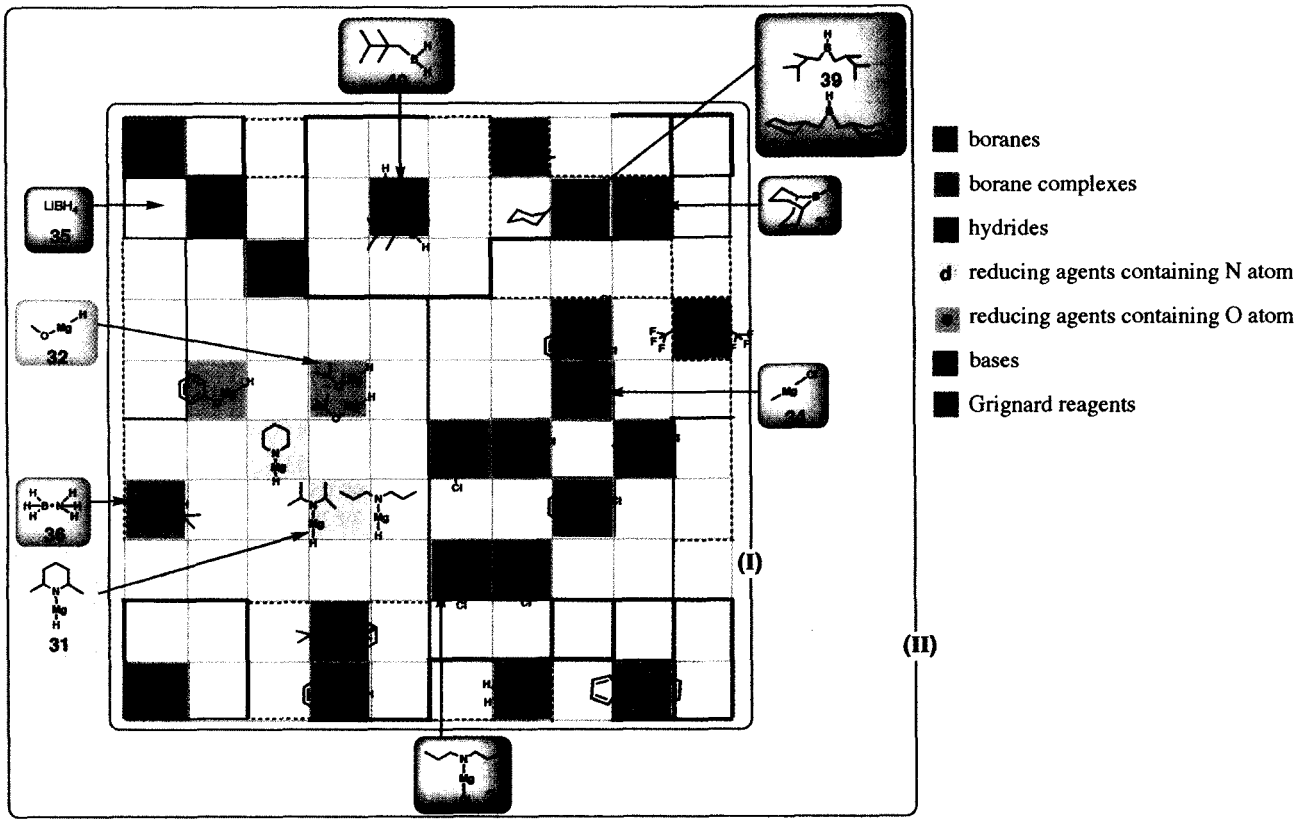


Fig. 6. A Kohonen map. (I) A Kohonen map trained by the representative FFs for the thirty reagents. Each small square in the map is a neuron. Neurons obtaining the input data are colored and mapped by the reagent structures. Black solid lines are boundaries recognized by U-matrix method. Bold solid lines mean larger difference between neurons than the fine solid lines, and the fine solid lines mean larger difference between neurons than the dotted lines. Actual shape of this map is a torus. Thus, the top of this map connects with the bottom, and the left edge connects with the right one. The map is labeled by colors (a—g) according to the roles of the reagents in reactions. (II) The results from mapping of reagents not used in training and constructing of the model. Reagents of **31**—**40** were mapped on the neurons pointed by arrows.

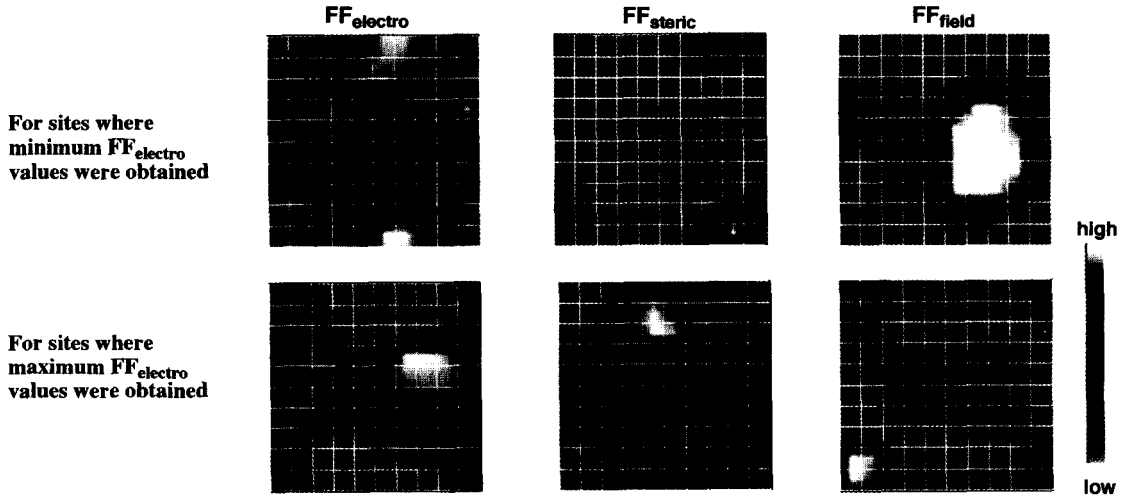


Fig. 7. Kohonen maps colored according to the size of weight vectors corresponding to the six parameters of FFs. Correspondence of gradient between the value and the color on the map are shown in color bar in the right side. Referring of these maps with the classification results in Fig. 6 shows that the major factors to determine the boundaries are maximum and minimum of $FF_{electro}$ values.

model. A predicted possibility as a base for **15** was 0.5, but the input signal denoted there was no data. This is suitable to the nature of **15**, which has the basisity.

The predicted value for the roles of neopentylmagnesium chloride (**27**) as base was rather different from the input signal. The input signal for the role for **27** was 0 and the predicted value was 0.6. The predicted degree of the roles was according to the similarities in the input FFs between **27**

Table 4. Results from Prediction of Roles of Reagents Used in Construction of the Model

	Reduction of ketone to alcohol	Reduction of aldehyde to alcohol	Reduction of carboxylic acid to alcohol	Reduction of ester to aldehyde	Reduction of ester to alcohol	Reduction of epoxide	Hydroboration or hydro-alumination	Base	Alkylation
 **4**	1.0 (1,1)	0.8 (1,1)	0.8 (1,1)	0.1 (0,0)	0.1 (0,0)	0.1 (0,0)	1.0 (1,1)	0.0 (0,0)	0.0 (0,0)
 **6**	1.0 (1,1)	0.8 (0,0)	0.8 (0,0)	0.1 (0,0)	0.1 (0,0)	0.1 (0,0)	1.0 (1,1)	0.0 (0,0)	0.0 (0,0)
 **8**	1.0 (1,1)	0.9 (1,1)	0.1 (0,0)	0.0 (0,1)	0.0 (0,1)	0.1 (0,0)	0.9 (1,1)	0.1 (0,0)	0.0 (0,0)
 **10**	1.0 (1,1)	0.9 (1,1)	0.7 (1,1)	0.0 (0,1)	0.7 (1,1)	0.7 (1,1)	1.0 (1,1)	0.0 (0,0)	0.0 (0,0)
$LiAlH_4$ **13**	1.0 (1,1)	1.0 (1,1)	0.8 (1,1)	0.4 (0,1)	1.0 (1,1)	0.8 (1,1)	0.9 (1,1)	0.8 (1,1)	0.0 (0,0)
 **15**	1.0 (1,1)	1.0 (1,1)	1.0 (1,1)	0.8 (1,1)	0.9 (1,1)	0.9 (1,1)	0.9 (1,1)	0.5 (0,0)	0.0 (0,0)
 **16**	1.0 (1,1)	0.0 (0,0)	0.0 (0,0)	0.0 (0,0)	0.0 (0,0)	0.0 (0,0)	0.0 (0,1)	0.0 (0,0)	0.0 (0,0)
 **21**	1.0 (1,1)	0.1 (0,0)	0.1 (0,0)	0.1 (0,0)	0.1 (0,0)	0.1 (0,0)	0.0 (0,0)	0.0 (0,0)	0.0 (0,0)
 **23**	0.1 (0,0)	0.2 (0,0)	0.0 (0,0)	0.0 (0,0)	0.0 (0,0)	0.0 (0,0)	0.0 (0,1)	0.9 (1,1)	0.4 (0,0)
 **27**	0.0 (0,0)	0.0 (0,0)	0.0 (0,0)	0.0 (0,0)	0.0 (0,0)	0.0 (0,0)	0.0 (0,0)	0.6 (0,0)	0.9 (1,1)

and the other base reagents gathered closely in the Kohonen map. This result is also suitable to the nature of **27**, which has the strong basisity, although it is not actually usually used as base.

These results demonstrate that the degrees of similarities in the input FFs have a good correspondence with the degree of similarities in the roles of the reagents, and that the CP Kohonen neural network model was successfully trained.

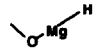
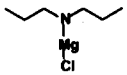
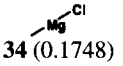
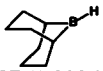
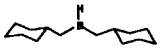
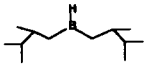
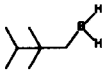
Reagents Not Used in the Modeling. In the CP Kohonen neural network model, ten reagents not used in the modeling (Fig. 2) were allocated as shown in Fig. 6-(II), and the predicted degrees of the roles are listed in Table 5. In the first column of Table 5, the reagent structures and the structural nos. with q-error values in parenthesis are shown. The q-error values indicate a gap between a reagent's FFs and the weight vectors of a neuron where the reagent was mapped. Thus, lower values of the q-error denote a higher reliability of the predicted values.

In Fig. 6, reagents were allocated to the neurons indicated by arrows. The reported borane of 9-borabicyclo-[3.3.1]nonane (9-BBN) (**37**) and ammonia-borane (**36**) are allocated into groups of boranes and borane complexes, respectively. Unreported boranes **38**, **39**, and **40** are allocated into groups of boranes. $LiBH_4$ (**35**) was allocated to a neuron neighbor to $LiAlH_4$. MeOMgH (**32**), 2,6-Me_2-*c*-C_5H_9NMgH (**31**), *n*-Pr_2NMgCl (**33**), and MeMgCl (**34**) were allocated to groups of alkoxymagnesium hydrides, dialkylaminomagnesium hydrides, dialkylaminomagnesium chlorides, and Grignard reagents, respectively. The results show that each of ten reagents was successfully mapped into a suitable neuron.

As shown in Table 5, for **31** and **32**, the possibilities of reducing agents from ketone to alcohol were predicted as 1.0. They were reported as reducing agents from ketone to alcohol.[27] **34**, which is well known as a Grignard reagent, the possibilities of a reagent for alkylation was predicted to be 1.0. Low values of the q-error for them are consistent with these correct results.

For $LiBH_4$ (**35**), all of the predicted values, except that for reducing carboxylic acids, are suitable. The predicted values for reducing carboxylic acids as 0.8, underlined in Table 5, is not good, because it is known that **35** does not reduce carboxylic acids. For 9-BBN (**37**), all of the predicted values, except that for reducing carboxylic acids and esters,

Table 5. Results from Prediction of Roles of Reagents Not Used in Construction of the Model

	Reduction of ketone to alcohol	Reduction of aldehyde to alcohol	Reduction of carboxylic acid to alcohol	Reduction of ester to aldehyde	Reduction of ester to alcohol	Reduction of epoxide	Hydroboration or hydroalumination	Base	Alkylation
 **31** (0.1369)	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
 **32** (0.1242)	1.0	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0
 **33** (0.1534)	0.1	0.2	0.0	0.0	0.0	0.0	0.0	0.9	0.4
 **34** (0.1748)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	1.0
$LiBH_4$ **35** (1.0442)	1.0	1.0	0.8	0.0	1.0	0.8	0.9	0.9	0.0
 **36** (0.7299)	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
 **37** (1.0336)	1.0	0.9	0.7	0.0	0.7	0.7	1.0	0.0	0.0
 **38** (0.3696)	1.0	0.8	0.5	0.0	0.4	0.6	1.0	0.0	0.0
 **39** (0.3497)	1.0	0.8	0.5	0.0	0.4	0.6	1.0	0.0	0.0
 **40** (1.0548)	1.0	0.8	0.8	0.1	0.1	0.1	1.0	0.0	0.0

are suitable. The predicted values for reducing carboxylic acids and esters as 0.7, underlined in Table 5, are not good, because it is known that **37** does not reduce them. The q-error values for **35** and **37** were higher. The q-error values denote a lower reliability of the predicted values.

These results show that the predicted values are almost suitable to the potential of the roles, and that the reliability of the predicted values is given as q-error values. Thus, the CP Kohonen neural network model constructed here can predict the roles of those reagents not used in the modeling.

Stability of the CP Kohonen Neural Network Model. Training of the CP Kohonen network model was repeated to the know stability of the model by an examination of the weight vectors between active and output layers. During the training, the weight vectors from about 0.3 to 0.7 were rather unstable, while those from about 0.0 to 0.3 and from about 0.7 to 1.0 were rather stable. The results mean that the predicted values by the model can be used as three or four classes of possibility measures.

A good point of the CP Kohonen neural network is to know the major factors related to the roles of reagents. This is a different good point from a back-propagation (BP) neural network, and the point is the reason to use the CP neural network, here. A weak point of the CP compared with BP neural networks is that the model can not predict the roles for a reagent that is not similar to reagents used in the modeling. Thus, a combination of the good points of CP and BP neural networks will produce a better model for prediction, which is what we are planing to do.

Conclusion

The same dimensional representative features that were calculated by a FRAU system represented similarities in the roles of reagents in various structures. The representative FRAU's features shown that sites where the maximum and minimum electrostatic interaction energies were calculated are important to determine the reagents' characters. The roles of the reagents were successfully classified in the basis of the representative FRAU's features by a Kohonen neural network. The results were applied to a neural network

model construction for predicting the potential roles of the reagents by a CP Kohonen neural network that has been modified to distinguish between negative and absent data. Thus, a reagent-roles correlation model that can predict the possibilities of the reagents' roles was constructed.

The FRAU features represent more detailed similarities and differences in molecules and the FRAU features as discriminators for close structures and roles are now in progress and will be reported in the future. Furthermore, verifications of the predicted values by experiments are also in progress and will be described elsewhere.

The benefits of a CP Kohonen neural network is to treat non-linear data as well as to know factors that principally contribute to the classification. The CP Kohonen neural network produced a reagent-roles correlation model that gives good reproduced and stable interpolated answers. The predicted values can be evaluated in three to four classes. For more accuracy, an improvement of neural network methods will be needed. The reliability of extrapolated answers and the construction of an extended model will be investigated by the CP neural network model with other methods, such as a back-propagation neural network.

The procedures for constructing the reagent-roles correlation model presented here are planned to apply to our reaction prediction studies in the future.

We thank Dr. Kazuo Nagasawa of Synthetic Organic Chemistry Laboratory in RIKEN (The Institute of Physical and Chemical Research) for useful discussions and advice on the chemical reactivity of reagents.

References

1 H. Satoh, S. Itono, K. Funatsu, K. Takano, and T. Nakata, *J. Chem. Inf. Comput. Sci.*, **39**, 671 (1999).

2 Some of the reagents analyzed in the Ref. 1 and here contains boron atoms that do not clearly defined as metallic atoms, however, we call them as metallic atoms, here, to distinguish them from other atoms such as carbon, hydrogen, oxygen, and nitrogen.

3 H. Satoh, O. Sacher, T. Nakata, L. Chen, J. Gasteiger, and K. Funatsu, *J. Chem. Inf. Comput. Sci.*, **38**, 210 (1998).

4 V. Simon, J. Gasteiger, and J. Zupan, *J. Am. Chem. Soc.*, **115**, 9148 (1993).

5 K. Schulz and J. Gasteiger, *J. Chem. Inf. Comput. Sci.*, **33**, 395 (1993).

6 J. Gasteiger and J. Zupan, *Angew. Chem., Int. Ed. Engl.*, **32**, 503 (1993).

7 L. Chen and J. Gasteiger, *Angew. Chem.*, **108**, 844 (1996).

8 L. Chen and J. Gasteiger, *Angew. Chem., Int. Ed. Engl.*, **35**, 763 (1996).

9 L. Chen and J. Gasteiger, *J. Am. Chem. Soc.*, **119**, 4033 (1997).

10 T. Kohonen, *Biol. Cybern.*, **43**, 59 (1982).

11 T. Kohonen, "Self-Organizing Maps," Springer-Verlag, Berlin and Heidelberg (1995).

12 J. Zupan, and J. Gasteiger, "Neural Networks for Chemists. An Introduction," VCH, Weinheim (1993).

13 R. Hecht-Nielsen, *Appl. Optics*, **26**, 4979 (1987).

14 Kohonen neural network and counter-propagation Kohonen neural network programs that were used in this study were modules of TUT_SOM, which is a program package developed by K. Funatsu in Toyohashi University of Technology. TUT_SOM includes two neural network programs (Kohonen neural network and counter-propagation neural network); two programs for recognition of boundaries among weight vectors on Kohonen map (U-matrix and Potential function methods), and several convenient graphic viewers of the results given by the above programs. The modification for treating reaction data was performed for the counter-propagation neural network module of TUT_SOM.

15 a) W. E. Lindsell, in "Comprehensive Organometallic Chemistry," ed by G. Wilkinson, F. G. A. Stone, and E. W. Abel, Pergamon Press, Oxford (1982), Vol. 1, p. 155. b) J. D. Odom, in "Comprehensive Organometallic Chemistry," ed by G. Wilkinson, F. G. A. Stone, and E. W. Abel, Pergamon Press, Oxford (1982), Vol. 1, p. 253. c) W. E. Lindsell, in "Comprehensive Organometallic Chemistry II," ed by G. Wilkinson, F. G. A. Stone, and E. W. Abel, Pergamon Press, Oxford (1995), Vol. 1, p. 57. d) C. E. Housecroft, in "Comprehensive Organometallic Chemistry II," ed by G. Wilkinson, F. G. A. Stone, and E. W. Abel, Pergamon Press, Oxford (1995), Vol. 1, p. 129.

16 a) K. Smith and A. Pelter, in "Comprehensive Organic Synthesis," ed by B. M. Trost, B. I. Fleming, and S. L. Schreiber, Pergamon Press, Oxford (1991), Vol. 8, p. 703. b) J. J. Eisch, in "Comprehensive Organic Synthesis," ed by B. M. Trost, B. I. Fleming, and S. L. Schreiber, Pergamon Press, Oxford (1991), Vol. 8, p. 733.

17 a) P. Chabot, in "Handbook of Grignard Reagents," ed by G. S. Silverman and P. E. Rakita, Marcel Dekker Inc., New York (1996), p. 93. b) J. K. Bonesteel, in "Handbook of Grignard Reagents," ed by G. S. Silverman and P. E. Rakita, Marcel Dekker Inc., New York (1996), p. 103. c) H. L. Uhm, in "Handbook of Grignard Reagents," ed by G. S. Silverman and P. E. Rakita, Marcel Dekker Inc., New York (1996), p. 117.

18 G. H. Penner, Y. C. P. Chang, and J. Hutzal, *Inorg. Chem.*, **38**, 2868 (1999).

19 B. Wrackmeyer, in "Annual Reports on NMR Spectroscopy," ed by G. A. Webb, Academic Press Ltd., London (1988), p. 61.

20 S. Tomoda and T. Senju, *Chem. Commun.*, **1999**, 423.

21 Molecular orbital calculations were performed by SPARTAN system (version 5.0 for Unix) from Wave Function Co., Ltd. Arbitrary computational level of theory is adaptable for geometry optimization and estimation of atomic charges according to purposes of FRAU users, namely, FRAU users can choice a computational level of theory that more rapidly optimizes geometry and estimates atomic charges. We performed ab initio MO calculations, here, to get accurate properties especially for metallic atoms.

22 A. Ultsch, G. Guimaraes, D. Korus, and H. Li, in "Proc. Transputer Anvender Treffen/World Transputer Congress TAT/WTC 93 Aachen," Springer-Verlag, New York (1993), p. 194.

23 L. A. Paquette, in "Encyclopedia of Reagents for Organic Synthesis," John Wiley & Sons, Chichester (1995).

24 A. Pelter, K. Smith, and H. C. Brown, in "Borane Reagents," Academic Press, London (1988).

25 Distributed Chemical Graphics, Inc.

26 Distributed MDL Information Systems, Inc.

27 a) E. C. Ashby and J. J. Lin, *J. Org. Chem.*, **43**, 1557 (1978). b) E. C. Ashby, J. J. Lin, and A. B. Goel, *J. Org. Chem.*, **43**, 1560 (1978). c) E. C. Ashby, J. J. Lin, and A. B. Goel, *J. Org. Chem.*, **43**, 1564 (1978).