Using Backpropagation Networks for the Estimation of Aqueous Activity Coefficients of Aromatic Organic Compounds

H. Chow,*,† H. Chen,‡ T. Ng,‡ P. Myrdal,† and S. H. Yalkowsky†

Departments of Pharmacy Practice and Management Information Systems, University of Arizona, Tucson, Arizona 85721

Received January 9, 1995⊗

This research examined the applicability of using a neural network approach to the estimation of aqueous activity coefficients of aromatic organic compounds from fragmented structural information. A set of 95 compounds was used to train the neural network, and the trained network was tested on a set of 31 compounds. A comparison was made between the results and those obtained using multiple linear regression analysis. With the proper selection of neural network parameters, the backpropagation network provided a more accurate prediction of the aqueous activity coefficients for testing data than did regression analysis. This research indicates that neural networks have the potential to become a useful analytical technique for quantitative prediction of structure−activity relationships.

INTRODUCTION

Organic solutes in water generally do not follow ideal behavior. The aqueous activity coefficient of a solute in water is a measurement of the deviation from ideality. Accurate aqueous activity coefficients are essential to predicting and understanding the solubility of a compound in aquatic and biological media. Conventional methods for estimating aqueous activity coefficients include the general solubility equation for nonelectrolytes developed by Yalkowsky and co-workers[1−3] and the UNIFAC method developed by Fredenslund and other researchers.[4,5] Recently, Myrdal et al.[6] introduced the aqueous functional group activity coefficients (AQUAFAC) method, which uses various combinations of molecular fragments for estimating the aqueous activity coefficients of organic compounds. Regression analysis using the AQUAFAC method has been performed on alkanes, polycyclic aromatic hydrocarbons, alkyl aromatics,[6] and halogenated aromatics.[7] Results from these studies have shown great promise for using the AQUAFAC method for aqueous activity coefficient estimation. In an attempt to investigate the applicability of other analytical techniques based on the AQUAFAC method, we resorted to a new and promising neural network computing approach.

Neural networks have been used successfully in various engineering (e.g., character recognition and image processing) and business (e.g., stock performance prediction) applications. Recently they have also been adopted in various biomedical and chemical areas.[8−13] Work reported by Bodor et al.,[11] Aoyama et al.,[12] and Andrea and Kalayeh[13] are most relevant to our research. Aoyoma et al. and Andrea and Kalayeh presented applications of the neural network approach to estimating quantitative structure−activity relationships of carboquinones and benzodiazepines and dihydrofolate reductase inhibitors, respectively. The neural network models performed better than multiple regression analysis. Bodor et al.[11] experimented with a backpropagation network for solubility prediction. Their research showed the backpropagation model to be superior to the regression analysis technique when mean standard deviations for the training set were compared. However, the neural network did not perform as well as the regression technique for an unknown set of organic compounds.

In our reserach we aimed to demonstrate the applicability of using neural network computing based on the AQUAFAC method for estimating aqueous activity coefficients and to compare the performance of this approach with that of conventional regression analysis.

THEORETICAL BACKGROUND

Aqueous Activity Coefficient and Solubility. The aqueous solubility of a solid solute is controlled by the ideal solubility of the crystalline solute and the thermodynamic activity of the solute in water[14]

$$\log S_{obs} = \log S_i - \log \gamma_w \qquad (1)$$

where S_{obs} is the observed solubility of the compound, S_i is the ideal solubility of the compound, and γ_w is the aqueous activity coefficient. The ideal solubility takes into account the crystalline contribution to the solubility and may be expressed, for room temperature conditions, approximately as[1]

$$\log S_i = -\frac{\Delta S_m}{2.303 \cdot 298 \cdot R}(MP - 25) \qquad (2)$$

where ΔS_m is the entropy of melting, MP is the melting point of the compound in centigrade, and R is the universal gas constant. The entropy of melting of rigid organic compounds can be estimated by the method proposed by Yalkowsky and co-workers[15]

$$\Delta S_m = 13.5 - 4.6(\log \sigma) \qquad (3)$$

where σ is the symmetry number. The symmetry number is the number of indistinguishable positions in which a

* Address correspondence to Hsiao-Hui Chow, Ph.D., Department of Pharmacy Practice, College of Pharmacy, University of Arizona, Tucson, AZ 85721. Phone (602) 626-4055. Fax (602) 626-4063. email: chow@tonic.pharm.arizona.edu.
† Department of Pharmacy Practice.
‡ Department of Management Information Systems.
⊗ Abstract published in *Advance ACS Abstracts,* June 1, 1995.

0095-2338/95/1635-0723$09.00/0

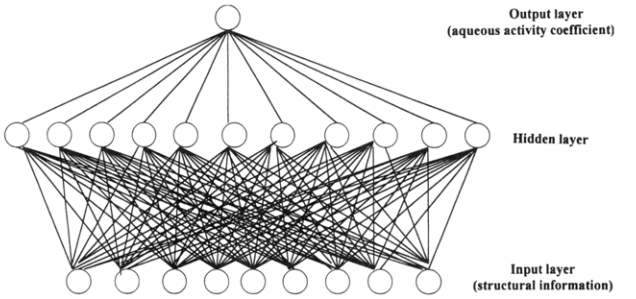

Figure 1. Schematic representation of the neural network used in estimating aqueous activity coefficients from the structural information.

compound may be oriented. For example, anthracene has a symmetry number of four, whereas phenanthrene has a symmetry number of two. The ideal solubility can be calculated from the known MP and R and the estimated ΔS_m. Substituting eq 2 into eq 1 and solving for the activity coefficient provides an equation describing the activity coefficient. That is

$$\log \gamma_\omega = -\log S_{obs} - \frac{\Delta S_m}{1364}(MP - 25) \qquad (4)$$

The AQUAFAC method is an attempt to predict the aqueous activity coefficient from fragmentation of the molecular structure. In past research,[6] the aqueous activity coefficient of a compound has been shown to be well approximated from a summation of constitutive functional groups, i.e.

$$\log \gamma_\omega = \sum n_i q_i$$

where n_i is the number of times group i appears in the compound and q_i is the contribution of group i to the total activity coefficient.

Backpropagation Neural Networks. Neural networks consist of fully interconnected rows of processing units called nodes. Nodes are organized into groups called layers. An input layer receives input (i.e., molecular structural information for this application). An output layer produces output (i.e., activity coefficients for this application). A hidden layer provides interconnection between input and output. A feed-forward network with the backpropagation algorithm (referred in this paper as backpropagation neural networks) may contain multiple hidden layers. The interconnection between each node is associated with a weighting factor. A schematic representation of the neural network used in this research in estimating aqueous activity coefficients from the structural information is shown in Figure 1. This network has nine input nodes (containing the molecular structural information) and one output node (the aqueous activity coefficient). After experimentation, one hidden layer and 11 hidden nodes were selected to provide the interconnections between input and output. Each node in a layer is connected in the forward direction to every node in the next layer. For a backpropagation neural network, activation of network flow is in one direction only, from the input layer through the hidden layer and then on to the output layer. A backpropagation network typically starts out with a random set of connection weights. The network adjusts its weights based on some learning rule (delta rule for this application) each time it sees an input–output (structure–activity) data pair. Each pair of data goes through two stages of activation: a forward pass and a backward pass. The forward pass involves presenting a sample input to the network and letting activation flow until the output layer is reached. During the backward pass, the network's actual output (from the forward pass) is compared with the target output and errors are computed for the output units. The weights connected to the output units can be adjusted in order to reduce those errors. The error estimates of the output units can then be used to derive error estimates for the units in the hidden layers. Finally, errors are propagated back to the connections stemming from the input units. A complete round of forward–backward passes and weight adjustments using all input–output pairs is called an epoch. A backpropagation network needs to "learn" from the same data set through a few hundred—sometimes even thousands—epochs in order to gradually refine its connection weights. Training continues until either the errors for the network stabilize (convergence) or a predetermined number of epochs is exceeded. After training, the network has the ability to recall a previous training data set. A trained backpropagation network should exhibit excellent generalization capability for predicting unknown data.

Table 1. AQUAFAC Groups Considered

group	description
CHAR	aromatic CH group
CAR	total aromatic carbons less CHAR
YCH3	methyl group on aromatic ring
YF	fluorine atom attached to an aromatic ring
YCL	chlorine atom attached to an aromatic ring
YI	iodine atom attached to an aromatic ring
YBR	bromine atom attached to an aromatic ring
ORTHO	the number of adjacently positioned halogens
ORTHOBIP	the number of halogens positioned at 2, 2′, 6, and 6′

METHODS

Data Compilation. The solubilities (S_{obs}) and melting points (MP) of 157 substituted aromatic hydrocarbons were extracted from the AQUASOL dATAbASE.[16] The substituents on the aromatics included halogens and methyl groups. The experimental entropies of melting were taken from the literature whenever possible. Otherwise they were estimated using eq 3. Aqueous activity coefficients (γ_w) were computed from the relationship of eq 4. Based on past research,[6,7] nine specific AQUAFAC groups were considered as the contributing structural fragments. These included the aromatic CH group (CHAR), the total aromatic carbons less CHAR (CAR), the methyl group on aromatic ring (YCH_3), the fluorine atom attached to an aromatic ring (YF), the chlorine atom attached to an aromatic ring (YCL), the iodine atom attached to an aromatic ring (YI), the bromine atom attached to an aromatic ring (YBR), the number of adjacently positioned halogens (ORTHO), and the total number of halogens positioned at 2, 2′, 6, or 6′ (ORTHOBIP) of a biphenyl (see Table 1). Each solute was fragmented. The fragmentation scheme employed by AQUAFAC is depicted in Figure 2. The number of times each of the nine specific AQUAFAC groups appeared in the compound was recorded.

Training and Prediction of the Aqueous Activity Coefficients. The data set was randomly divided into three groups: training data (95 compounds), validation data (31 compounds), and testing data (31 compounds). The same training data were used to generate a multiple linear regression equation and to train the neural networks. The

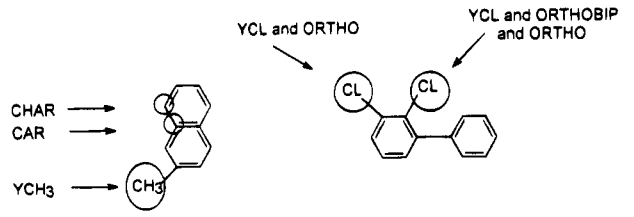

Figure 2. Fragmentation scheme employed by AQUAFAC.

validation data were only used in neural network analysis to determine the optimum number of training epoch.

Multiple Regression Analysis. A multiple regression procedure was performed on the training data, using the SAS statistical analysis program.[17] The dependent variable for regression analysis was the logarithm of the observed aqueous activity coefficients, and the independent variables consisted of combinations of molecular descriptors. The multiple regression model used is as follows

$$\log(\gamma_w) = q_1 n_1 \text{CHAR} + q_2 n_2 \text{CAR} + q_3 n_3 \text{ YCH3} + q_4 n_4 \text{YF} + q_5 n_5 \text{YCL} + q_6 n_6 \text{YI} + q_7 n_7 \text{YBR} + q_8 n_8 \text{ORTHO} + q_9 n_9 \text{ORTHOBIP}$$

where n_i is the number of times the fragment appears in the compound and q_i is the regression coefficient of a particular fragment.

Backpropagation Neural Networks. A backpropagation algorithm implemented in C language (running on a DEC ALPHA 2100, 190 MHz) was developed. The specific algorithm adopted in this research has been described previously.[18] We used the sigmoidal function[19] as the transformation function and the delta rule[19] as the error correction formula. Flexible selection of network parameters such as the number of hidden layers, units within each hidden layer, learing rate, and momentum factor were allowed in our program. The nine molecular fragments listed in Table 1 were used as the neurons in the input layer, and the logarithm of the aqueous activity coefficient was used as the neuron in the output layer. The network was trained with the same training data examined by the multiple regression procedure. During the training phase, extensive experimentation was performed to define the network parameters. The validation data were used in neural network analysis to prevent overtraining of the network. A trained network was used to predict the aqueous activity coefficients of the validation data set. An epoch that provided a stable minimum mean square error for the validation data was selected as the optimum number of training epoch.

Prediction of the aqueous activity coefficients was performed on the 31 testing compounds using both the multiple linear regression equation and the validated networks. Mean square error (MSE) was used as an index for comparison of the prediction from the multiple regression with that using the neural network approach. The entire training and testing procedure was repeated 40 times. Each round, the computer randomly assigned data to training, validation, and testing data groups.

RESULTS

Table 2 shows the results of the regression analysis from training data set no. 23. The results from this data set are discussed in detail here, since they resemble the average

Table 2. Results of Regression Analysis from Training Data Set No. 23

Parameter Estimates			
parameter	q value	std error	prob > $\|T\|$
CHAR	0.301	0.012	0.0001
CAR	0.370	0.066	0.0001
YCH3	0.301	0.074	0.0001
YF	−0.069	0.154	0.6539
YCL	0.487	0.072	0.0001
YI	0.995	0.242	0.0001
YBR	0.759	0.092	0.0001
ORTHO	0.029	0.101	0.7742
ORTHOBIP	−0.156	0.061	0.0117

Analysis of Variance

F value 1910
R-square 0.995
root mean square error (RME) 0.528
mean square error (MSE) 0.278
sum of square error (SSE) 23.94

Table 3. Final Network Parameters for Data Set No. 23

topology	9-11-1
learning rate	0.35
momentum factor	0
learning rule	delta rule
transfer function	sigmoidal function
number of epochs	276
training time	4.08 s

results from the 40 test sets. The regression coefficients of CHAR and YCH3 groups are similar (0.301 ± 0.012 vs 0.301 ± 0.074), suggesting that these two groups contribute analogously to the activity coefficient estimation and could be combined in future analysis. For YF, ORTHO, and ORTHOBIP groups, the regression coefficients were not statistically significant (prob > 0.01). This is likely to be due to the fact that the data set did not contain enough compounds with these types of groups. The overall model had an R-square of 0.995, a root mean square error of 0.528, a mean square error of 0.278, and a sum of square error of 23.94. This suggests that the multiple linear regression model was able to describe the relationships between structural information and the aqueous activity coefficient. The mean square error was used later as a benchmark for comparison with the performance of the neural network model.

The neural network parameters for data set no. 23 are summarized in Table 3. Based on our experimentation with various network parameters during the training phase, one hidden layer and 11 hidden nodes were selected to provide interconnections between input and output. A learning rate of 0.35 and no momentum factor provided a stable convergence and a low MSE. The number of training epochs for data set no. 23 determined from the validation set was 276. Total training time was 4.08 s.

Figure 3 illustrates the relationship between the MSE and the number of training epochs for data set no. 23 up to 2000 epochs. As expected, the network improved its performance initially for both training (solid line) and validation (dotted line) data (both errors dropped significantly). However, as the number of training epochs was increased, performance for predicting training compounds continued to improve, but the prediction for the compounds in the validation set deteriorated. The minimal MSE of validation data was achieved when the network was trained for 276 epochs.

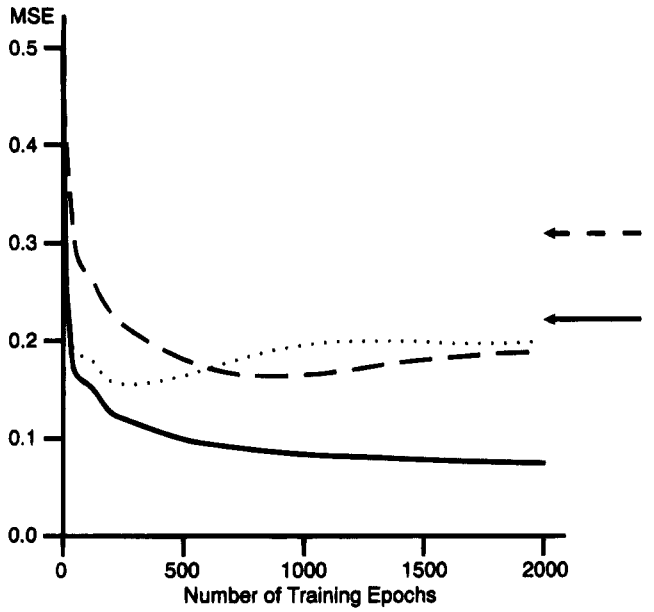

Figure 3. Effects of the number of training epochs on the performance of neural networks. (solid line) training data; (dotted line) validation data; (dashed line) testing data; (←) MSE of the training data from regression analysis; (← —) MSE of the testing data from regression analysis.

Similar to the validation data, the initial MSE of the testing data (dashed line) decreased significantly. However, an increase in MSE of the testing data was observed as training continued. The minimum MSE of the testing data occurred at approximately 750 epochs. For this particular data set, the optimal training epoch selected from the validation data set was smaller than that for the testing data.

Comparison of the prediction of aqueous activity coefficients between the neural network approach and regression analysis on the testing data set no. 23 is summarized in Table 4. Boldface indicates the analysis giving the more accurate prediction. The activity coefficient values of this test data set ranged from 2 to 9.5 log unit. The neural network predicted more accurate for 22 of the 31 testing data, with 77% of the predicted activity coefficients within one-half a log unit of the experimental data. When multiple linear regression analysis was used, 65% of the predicted activity coefficient was within one-half a log unit of the experimental data.

Table 5 summarizes the MSE from multiple regression analysis and from neural networks for 40 randomly selected test data sets. Boldface indicates the analysis having the lower MSE. The neural network performed better than regression in 28 of the 40 multiple tests. Statistical comparison of the two techniques is summarized in Table 6. The average mean square error for the multiple tests was 0.250 ± 0.064 vs 0.210 ± 0.072 ($p < 0.01$) for multiple regression analysis and for the backpropagation neural network analysis, respectively.

DISCUSSION

The solutes used in our data set consisted of substituted aromatic hydrocarbons. The substituents on the aromatics included halogens and methyl groups. The data set used in this study contained a large number of polychlorinated biphenyls (PCBs), which were of special interest to us because of their abundance in the environment and their well-documented toxicity. Accurate prediction of their aqueous

Table 4. Comparison of the Prediction of Aqueous Activity Coefficients between the Neural Network Approach (BNN) and Regression Analysis (REG) on Testing Data Set No. 23[a]

obs γ_w	pred (BNN)	diff (obs-pred)	pred (REG)	diff (obs-pred)	chemical name
4.61	**4.73**	**−0.12**	4.43	0.18	1,2,3,5-tetrachlorobenzene
4.70	**4.22**	**0.47**	4.14	0.56	1,4-diiodobenzene
6.52	**6.78**	**−0.25**	7.08	−0.56	2′,3,4,5,5′-pentachlorobiphenyl
5.94	**5.94**	**0.00**	5.76	0.18	2′,3,4-trichlorobiphenyl
7.08	**7.26**	**−0.18**	7.42	−0.34	2,2′,3,3′,5,6-hexachlorobiphenyl
7.01	7.30	−0.29	**7.27**	**−0.25**	2,2′,3,3′,6,6′-hexachlorobiphenyl
7.16	**7.67**	**−0.51**	8.08	−0.92	2,2′,3,4′,5,5′,6-heptachlorobiphenyl
6.25	6.44	−0.19	**6.27**	**−0.02**	2,2′,3,4′-tetrachlorobiphenyl
7.15	**7.61**	**−0.46**	8.24	−1.09	2,2′,3,4,4′,5,5′-heptachlorobiphenyl
7.00	**7.26**	**−0.26**	7.42	−0.42	2,2′,3,4,4′,6-hexachlorobiphenyl
7.10	**7.67**	**−0.57**	8.08	−0.98	2,2′,3,4,5,5′,6-heptachlorobiphenyl
6.36	**6.83**	**−0.47**	6.92	−0.57	2,2′,3,4,5-pentachlorobiphenyl
6.94	**6.44**	**0.50**	6.27	0.67	2,3′,4,6-tetrachlorobiphenyl
5.98	**5.94**	**0.04**	5.76	0.22	2,3′,5-trichlorobiphenyl
6.84	**7.21**	**−0.37**	7.58	−0.74	2,3,3′,4,4′,6-hexachlorobiphenyl
6.26	**6.38**	**−0.12**	6.42	−0.16	2,3,4′,5-tetrachlorobiphenyl
5.72	**6.38**	**−0.66**	6.42	−0.70	2,3,4,4′-tetrachlorobiphenyl
6.05	**6.00**	**0.05**	5.61	0.44	2,3,6-trichlorobiphenyl
6.02	**6.38**	**−0.36**	6.42	−0.40	2,4,4′,5-tetrachlorobiphenyl
5.46	6.48	−1.02	**6.11**	**−0.65**	2,2′,4,6-tetrachlorobiphenyl
5.70	6.00	−0.30	**5.61**	**0.09**	2,4,6-trichlorobiphenyl
5.37	**5.43**	**−0.06**	5.11	0.26	2,4-dichlorobiphenyl
5.69	**5.86**	**−0.17**	5.92	−0.23	2,4′,5-trichlorobiphenyl
5.39	**5.43**	**−0.04**	5.11	0.28	2,5-dichlorobiphenyl
7.39	6.31	1.07	**6.58**	**0.81**	3,3′,4,4′-tetrachlorobiphenyl
7.65	6.31	1.33	**6.58**	**1.07**	3,3′,5,5′-tetrachlorobiphenyl
2.61	**2.57**	**0.04**	2.73	−0.12	bromobenzene
9.50	8.68	0.82	**9.89**	**−0.39**	decachlorobiphenyl
2.00	1.95	0.05	**2.01**	**−0.01**	1,3-difluorobenzene
4.46	**4.22**	**0.23**	4.14	0.32	1,3-diiodobenzene
9.22	8.46	0.76	9.24	−0.01	2,2′,3,3′,4,5,5′,6,6′-nonachlorobiphenyl

[a] Boldface indicates the analysis with a Dower MSE.

Table 5. Summary of MSE from Multiple Regression Analysis (REG) and from Neural Networks (BNN) for 40 Randomly Selected Test Data Sets[a]

data set no.	BNN	REG	data set no.	BNN	REG
1	**0.141**	0.194	21	**0.251**	0.405
2	**0.304**	0.379	22	0.272	**0.197**
3	**0.133**	0.256	23	**0.213**	0.302
4	**0.251**	0.350	24	**0.146**	0.274
5	**0.193**	0.314	25	**0.117**	0.239
6	**0.144**	0.270	26	**0.266**	0.306
7	**0.215**	0.278	27	0.185	**0.156**
8	**0.168**	0.223	28	0.198	**0.157**
9	**0.175**	0.186	29	**0.172**	0.319
10	0.285	**0.213**	30	**0.196**	0.254
11	0.153	**0.144**	31	**0.137**	0.155
12	0.494	**0.221**	32	0.216	**0.211**
13	0.178	**0.149**	33	**0.203**	0.318
14	0.267	**0.199**	34	0.232	**0.206**
15	**0.176**	0.294	35	**0.218**	0.266
16	**0.162**	0.290	36	0.191	**0.262**
17	**0.145**	0.192	37	0.360	**0.217**
18	**0.154**	0.229	38	**0.229**	0.351
19	**0.178**	0.234	39	**0.217**	0.228
20	**0.129**	0.288	40	0.309	**0.262**

[a] Boldface indicates the analysis with a lower MSE.

Table 6. Summary of Mean Square Error (MSE) from 40 Randomly Selected Data Sets

	MSE	no. of data set with lower MSE
linear regression	0.250 ± 0.064	12
neural network	0.210 ± 0.072[a]	28

[a] Satistically significantly lower than that from multiple linear regression analysis ($p < 0.01$).

activity coefficients would improve the assessment of aquatic toxicity, leaching, soil and sediment transport and mobility, adsorption, and volatization of these compounds. We have shown in our studies that, based on the fragmented structural information, neural network analysis gave more accurate predictions of the activity coefficients than regression analysis.

During the training phase for neural network analysis, we performed extensive experimentation to define appropriate network parameters. Although the selection of the number of hidden units in the hidden layers as well as of the number of hidden layers in a backpropagation network was empirical, it is recognized that each of these choices can have a significant effect on network performance. A large number of hidden units and hidden layers can provide more predicting power, but the network will require more computation time and may also suffer in ability to generalize for an unknown data set. In our experiment we attempted to identify a simple topology that is able to predict satisfactorily for both training and testing compounds. We started out with one layer and increased the number of hidden units from 1 to 60 for different learning rates (from 0.25 to 0.35) and momentum factors (from 0 to 1). The selected range of network parameters had been adopted successfully in several applications.[20,21]

Based on our experimentation with various parameters, we found that learning rates did not have a noticeable effect on training, while different momentum factors caused the network to behave differently. For one hidden layer using a learning rate of 0.35 and a momentum factor of 0.9, errors decreased as the number of hidden units increased. This effect was most evident when the number of hidden units was small. The performance of the network stabilized after inclusion of an adequate number of hidden units (more than 5). When the network was configured with the same learning rate but with no momentum term, the network exhibited a similar characteristic of stabilizing after a few hidden units were introduced, but its overall behavior was more stable (had fewer variations) and more accurate (smaller MSE). The best topology determined from the training phase was 9-11-1, with a 0.35 learning rate and no momentum factor. This also satisfied our criterion of network simplicity.

Over-training of neural networks has often created an undesirable effect on the generalization process. The system may memorize the training set and eventually lose its capability to predict unknown data. We have adopted a method suggested by Rumelhart[22] to incorporate a validation data set to prevent the generalization problems. This cross-validation approach has also been used widely in statistics for error estimation (leave-one-out or "jacknifing" is a special case of the cross-validation method). For each of the 40 randomly generated test sets, we first trained the network with the training data up to 10 000 epochs. At each epoch the trained network was evaluated with the validation data. MSEs for the training and validation data were computed and drawn in a graph. The epoch that gave the lowest MSE for the validation data was selected as the "optimum" epoch (i.e., the trained network at that epoch had the best predictive power). We drew graphs for all 40 test sets to make sure that the MSEs for the training and validation data were stable (i.e., no oscillation), and the optimum epoch indeed produced the lowest MSE within 10 000 epochs. The optimum epochs for the 40 test sets varied, from a few hundreds (e.g., test set no. 23, presented as an example in the paper) to several thousands. The trained network at the optimum epoch was then used to compute the MSE for the testing data. By using validation data to select the optimum epoch for the trained network, this approach prevented the network from over-training. We believe research in defining a heuristic approach in selecting the optimum training epochs deserves further attention.

In conclusion, the research here showed the applicability of the neural network approach to aqueous activity coefficient prediction using the AQUAFAC method. With a simple network topology of 9-11-1 and empirically selected network parameters (learning rate of 0.35, no momentum factor, and epochs determined from validation data), the multiple-layered, fully connected backpropagation network prediction was better than that of regression for unknown data. However, as the technique currently stands, developing a good neural network is not a trivial task. The basic AQUAFAC method based on multiple regression seems easier to use than our backpropagation model. Nevertheless, for researchers constructing quantitative structure–activity relationships for a new application, developing a proper and yet simple model (i.e., one that can be solved through standard statistical methods) can require more painstaking effort than applying the neural network approach. The major strength of neural network techniques is that they do not assume a specific model. Instead, they learn from the data to establish the input and output relationship. With the incorporation of a hidden layer and fully connected nodes between layers, neural networks are capable of capturing

complex polynomial surfaces, including third or higher order terms as well as cross-products terms corresponding to interactions between physicochemical properties.[13] As more researchers investigate and refine the newer neural network techniques and as more robust and user-friendly neural network-based software is developed, the neural network approach promises to be a useful analytical tool to establish the global structure−activity relationships.

ACKNOWLEDGMENT

This project was supported in part by BRSG S07RR07002 awarded by the Biomedical Research Support Grant Program, Division of Research Resources, National Institutes of Health and NSF IRI-9211418 awarded by the Division of Information, Robotics, and Intelligent Systems, National Science Foundation.

REFERENCES AND NOTES

(1) Yalkowsky, S. H., Valvani, S. C. Solubility and Partitioning. 1. Solubility of non-electrolytes in water. *J. Pharm. Sci.* **1980**, *69*, 912−922.

(2) Yalkowsky, S. H., Valvani, S. C., Mackay, D. Estimation of the aqueous solubility of some aromatic-compounds. *Res. Rev.* **1983**, *85*, 43−45.

(3) Pinal, R., Yalkowsky, S. H. Solubility and Partitioning. 7. Solubility of barbiturates in water. *J. Pharm. Sci.* **1987**, *76*, 75−85.

(4) Fredenslund, A., Jones, R. L., Prausnitz, J. Group-contribution estimation of activity-coefficients in nonideal liquid-mixtures. *AIChE J.* **1975**, *21*, 1086−1099.

(5) Arbuckle. W. B. Using UNIFAC to calculate aqueous solubilities. *Environ. Sci. Technol.* **1986**, *20*, 1060−1064.

(6) Myrdal, P., Ward, G. H., Dannenfelser, R.-M., Mishra, D., Yalkowsky, S. H. AQUAFAC 1: Acqueous functional group activity coefficients: Application to hydrocarbons. *Chemosphere* **1992**, *24*, 1047−1061.

(7) Myrdal, P., Ward, G. H., Simamora, P., Yalkowsky, S. H. AQUAFAC: Aqueous functional group activity coefficients. *SAR QSAR Envir. Res.* **1993**, *1*, 53−61.

(8) Moallemi, C. Classifying cells for cancer diagnosis using neural networks. *IEEE EXPERT* **1991**, 8-12, December.

(9) Gonzalez, L. P., Arnaldo, C. M. Classification of drug-induced behaviors using multi-layer feed-forward neural network. *Comput. Methods Programs Biomed.* **1993**, *40*, 167−173.

(10) Shadmehr, R., D'Argenio, D. Z. A neural network for nonlinear bayesian estimation in drug therapy. *Neural Comput.* **1990**, *2*, 216−225.

(11) Bodor, N., Harget, A., Huang, M. Neural network studies. 1. Estimation of the acqueous solubility of organic compounds. *J. Am. Chem. Soc.* **1991**, *113*, 9480−9483.

(12) Aoyama, T., Suzuki, Y., Ichikawa, H. Neural networks applied to pharmaceutical problems. 3. neural networks applied to quantitative structure activity relationships analysis. *J. Med. Chem.* **1990**, *33*, 2583−2590.

(13) Andrea, T. A., Kalayeh, H. Applications of neural networks in quantitative structure-activity relationships of dihydrofolate reductase inhibitors. *J. Med. Chem.* **1991**, *34*, 2824−2836.

(14) Yalkowsky, S. H. Techniques of solubilization of drugs; Marcel Dekker: New York, NY, 1981.

(15) Dannenfelser, R.-M., Surendran, N., Yalkowsky, S. H. Molecular symmetry and related properties. *SAR QSAR Environ. Res.* **1993**, *1*, 273−292.

(16) Yalkowsky, S. H., Dannenfelser, R.-M. *ARIZONA dATAbASE of Aqueous Solubility*; 5th ed.; Tucson, AZ, 1990.

(17) *SAS/STAT User's Guide, Version 6*, 4th ed.; SAS Institute Inc.: Cary, NC, 1990.

(18) Rumelhart, D. E., Hinton, G. E., Williams, R. J. Learning internal representations by error propagation. In *Parallel Distributed Processing*; Rumelhart, D. E., McClelland, J. L. the PDP Research Group, Eds.; The MIT Press: Cambridge, MA, 1986; pp 45−76.

(19) Smolensky, P. Natural and conceptual interpretation of models. In *Parallel Distributed Processing*; Rumelhart, D. E., McClelland, J. L. the PDP Research Group, Eds.; THE MIT Press: Cambridge, MA, 1986; pp 390−431.

(20) Mooney, R., Shavlik, J., Towell, G., Gove, A. An experimental comparison of symbolic and connectionist learning algorithms. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence (IJCAI-89)*; Detroit, MI, August 20−25, 1989; pp 775−780.

(21) Lippmann, R. P. An introduction to computing with neural networks. *IEEE ASSP Magazine*, **1987**, 4 (April), 4−22.

(22) Rumelhart, D. E., Widrow, B., Lehr, M. A. The basic ideas in neural networks. *Commun. ACM* **1994**, *37*, 87−92.

CI950263Q