Neural Network Approach to Structural Feature Recognition from Infrared Spectra

D. Ricard, C. Cachet, and D. Cabrol-Bass*

LARTIC University of Nice Sophia-Antipolis, Parc Valrose, Nice F 06108 Cedex, France

T. P. Forrest

Dalhousie University, Halifax, Nova Scotia, B3H 4J33 Canada

Received May 20, 1992

Neural networks, with and without hidden nodes, have been trained to recognize structural features of compounds from their infrared spectra. The training of the networks was evaluated by a variety of statistical indices using threshold values obtained by simplex optimization and by evaluation of synthetic spectra of structural groups obtained from the connection weights of the single-layer networks. Results indicate that all of the networks can be trained to recognize the structural groups in the compounds used to train the network. The network with a hidden layer, and dedicated to a single structural group, was better able to recognize structural groups in compounds that had not been used in training the network. Although not as efficient, the single-layer networks are particularly useful in that information may be extracted for use in writing more effective rules for an expert system-based infrared interpreter.

INTRODUCTION

In using infrared spectroscopy for structure elucidation, the direct correlation of spectral features with structural features is not as simple as one might wish. The spectra of most organic compounds are quite complex, containing a large number of absorption bands that could be attributed to many different functional groups. An experienced chemist will utilize various aspects of the overall pattern to decide on the presence or absence of a group, but for an automatic computer-based system, the task can be quite difficult. Many programs using the expert system paradigm have been created for interpreting IR spectra.[1-7] Several spectral features are normally used in combination to assign a factor that indicates the degree of certainty for the assignment of a specific structural group. Generally, three types of evidence are used in determining these certainty factors: positive evidence, negative evidence, and prior probability evidence.

In the course of developing a decision support system for inexperienced interpreters of IR spectra,[8,9] we have developed a rule-based expert system, which in combination with a pattern matching algorithm helps the user in the deductive task. In this system we used the concept of positive evidence and negative evidence to rank the groups in order of their degree of support. In this particular project we have reached the limitations of rule-based expert systems in dealing with the uncertainty and the ambiguous information inherent in infrared spectra. The assignment of parameters for the calculation of certainty factors for spectral assignments was to a large extent arbitrary and, therefore, not completely satisfactory.

Several recent studies have shown the suitability of using neutral networks for tackling problems that are characterized by fuzziness in the data to be processed or in the knowledge to be utilized. In a recent review of chemical applications of neural networks, the authors provide a good description of the basic concepts.[10] Specific applications in molecular spectroscopy, proton NMR,[11] ^{13}C NMR,[12,13] IR,[14-16] and mass spectrometry,[17] show that neural networks constitute a promising approach in this field and call for further studies. As another approach to computer-based assistance, we are investigating the use of a neural network as a component of a pluristructural system, that is, a system which makes use of combined approaches: symbolic (expert systems) and analogic (neural networks). The aim of this investigation is twofold: to determine the efficacy of using a neural network to provide assistance in the interpretation of spectra and to provide information that might be used to support the work of an expert system.

METHODOLOGY

The combination of an expert system and neural network can be used in two distinctly different manners; one in which the network is used in cooperation with the expert system to analyze each unknown, and a second where the network is used instead to help in the creation of the rules that the expert system will use.

In the first case, the neural network is used directly and simultaneously in connection with the expert system. After its training has been completed, it is given the spectrum of an unknown compound as an input vector, and the output (real numbers) is categorized to obtain symbolic conclusions (such as definitively absent, probably absent, not classified, probably present, definitively present) relative to the structural components of the unknown. These conclusions are entered into the database of facts which are then used by the expert system. This kind of preprocessing of spectral data by the neural network is well-suited to infrared studies because first, neural networks are not highly perturbed by fuzzy input, and second, the expert system can use symbolic information instead of whole spectra as facts on which to apply its reasoning. For this purpose, the network architecture and parameters must be chosen to optimize its performance so that the expert system can be given the best opportunity to produce reliable conclusions. This is best achieved by using multilayer networks with optimized threshold values for interpreting the output vectors.

In the second case, the neural network can be used indirectly to help the expert system identify spectral features by taking information from the trained network into consideration in writing the rules for the expert system. The type of information that is required from the network for this purpose is not available in the output from the network. Although the output of the neural net can be used directly as an indicator of the

0095-2338/93/1633-0202$04.00/0

Table I. Training and Test Set Populations

	training set (*NT* = 212)		test set (*NS* = 236)	
	NP	*%P*	*NP*	*%P*
OH hydroxyl	59	27.8%	65	27.5%
CH_2OH primary alcohol	13	6.1%	18	7.6%
COOH carboxylic acid	22	10.4%	22	9.3%
COOR ester	40	18.9%	44	18.6%
C–CHO aldehyde	18	8.5%	16	6.8%
C–CO–C ketone	23	10.8%	25	10.6%
RR′–NH amine or amide (I or II)	19	9.0%	20	8.5%
CNH_2 amine or amide (I)	10	4.7%	12	5.1%
C=C alkene	41	19.3%	53	22.5%
C_6 aromatic	90	42.5%	100	42.4%
$C(CH_3)_2$ *gem*-dimethyl	17	8.0%	26	11.0%
CH_2 methylene	130	61.3%	151	64.0%

presence of a structural feature, no information is given as to how the conclusion was reached. For the type of evidence that could be used by a rule-based system, one would have to extract the indicators of positive and negative evidence from the system. For this purpose, the most important aspect of the network is that it allows one to extract the required correlation information from the connection weights obtained by the trained network. This is difficult in a multilayer network, but can be done quite simply in a single-layer network. We have, therefore, used both types of network, with and without hidden layers, in this investigation.

Network Architecture. Three different types of networks have been used in this work. The first two types have only one output, corresponding to a particular structural feature. These "specialized" networks can be integrated into a wider decision support system for structure elucidation, each being activated separately either by the user or as the result of a rule application of a rule-based expert system in order to confirm or reject a specific structural hypothesis. Two types of single output networks were used, one with a hidden layer and another with no hidden layer.

For the networks with a hidden layer, the 'NeuroShell' program from Ward Systems Group, Inc.[8] was used. The number of nodes in the hidden layer was selected on the basis of empirical trials, which indicated that 33 nodes in the hidden layer is a reasonable number to use; using more units does not improve the result significantly. Munk[15] found that a similar number was a convenient size for a neural net using peak positions as input. A total of 284 input nodes were used, resulting in 9405 ($284 \times 33 + 33$) connections between nodes and 9439 adjustable weights including thresholds associated with each node. For the network without a hidden layer, we wrote a program in PASCAL which used the same number of input nodes and, therefore, had a total of 284 connections between nodes and 285 weights including one threshold which, in this case, has been maintained constant.

The third type of network that was used contained one hidden layer and 12 output nodes, each node corresponding to a particular structural feature. A structural group is characterized by its position in the output vector. In this case the network treats the 12 different groups simultaneously. The multi-output network has the advantage of a reduced training time compared to that required to train several separate networks. Nevertheless, for practical reasons related to the technical resources available, the output vector was restricted to 12 components (see Table I).

Learning Method. In all cases the back-propagation algorithm was used as the learning method (for a description of this, and other types of neural network algorithms, the review article in ref 10 is recommended). This choice was made for several reasons, among them being: the large amount of work on both the theoretical and practical aspects of this algorithm; the efficient implementations of the algorithm that are available for most computer systems; the suitability of the input and output for infrared spectra and structural data; and the possibility of maintaining a strong similarity of networks, yet having single or hidden layers, and single or multiple output. The back-propagation algorithm provides the flexibility of network architecture adaptable to the comparative studies needed in this investigation.

For the network with a hidden layer, the training was carried out with the program's defaults parameters, i.e., learning rate = 0.6 and momentum = 0.9. The training was terminated when for each example the error was below a threshold of 0.0001.

For the network without a hidden layer, the modification of the weights is done with the classical DELTA rule, using the activation function $[\exp(kx) - 1]/[\exp(kx) + 1]$ with $k = 0.1$. As this function leads to outputs between −1 and +1, the results were later converted to values in the range of 0–1 for the purpose of comparison with the results of other networks. The training was done by presenting the 212 examples randomly, each example being presented the same number of times. The training was terminated after 2000 complete cycles.

Input Vectors. The choice of input vectors is a critical part of designing an effective neural network. The dilemma lies in representing the spectra as economically as possible without significant loss of information. An infrared spectrum can arise from various sources and be represented in several forms, but for a neural network, the information must be presented in an input vector of consistent dimension and connotation. Munk et al.[14] have tested a network in which they introduced the spectral information as peak positions and have shown that networks can learn to associate functional groups with peak positions. As we wished to use the trained network to help with writing rules for an expert system, we wanted the input vector to include the minimum of pre-processing and human expertise. We, therefore, chose to derive the input vectors directly from frequency/intensity values instead of peak positions. In peak position format, the information provided to the network is dependent upon the peak-picking algorithm, and the information on peak shape is lost. The complete spectrum was presented as a set of intensities, in the form of a list of numerical values, each value representing the absorbance at a frequency determined by its position in the list. In order to have a reasonably fast training speed for the networks, the resolution of the spectra was limited to 12 cm^{-1}. A list of 284 numerical values was used to represent a spectrum; the frequency scale of the spectra was 600–4000 cm^{-1}, with the first numerical value of the input vector representing the intensity at 600 cm^{-1}. The actual resolution of the spectra in the spectral database was 4 cm^{-1}, but taking all points leads to very slow convergence of the learning process.

Output Vectors. The output vectors consist of a list of numbers, each representing the measure of support found by the network for the particular structural feature of the compound. In the case of a network which is used for a single structural feature, the output consists of a single value. The list of structural features that were used in this work is found in Table I. These were chosen as features that one could reasonably expect to be identified in the compounds of our data set by infrared spectroscopy.

Table II. Comparison of Overall Quality of Fit

	training set				test set			
	full set S_r	present S_{rp}	absent S_{ra}	T_d	full set S_r	present S_{rp}	absent S_{ra}	T_d
single output, no hidden layer	0.100	0.156	0.076	42.3	0.240	0.419	0.162	10.62
single output, one hidden layer	0.006	0.007	0.006	887.4	0.233	0.393	0.163	11.28
multioutput, one hidden layer	0.031	0.032	0.028	336.1	0.290	0.455	0.220	8.34

If the network is to be used simply to confirm or reject the presence of a particular structural feature, a binary output may be adequate, but if an uncertainty measure related to the structural feature is needed, a real number output is necessary. The numerical output provided by our network was set to range between 0 and 1, with values close to 1 meaning that the structural feature related to this output node is likely to be present and, conversely, with values on the other end of the scale, near 0, indicating that the associated structural feature is likely to be absent from the compound. The outputs of each were used both as real numbers and also transformed into symbolic conclusions using optimized threshold values. In the case of multidimensional output, the transformation of the output data has to be done for each component of the vector, using threshold values specifically determined for each structural group.

Training Sets. The selection of a training set is a very important part of the development of artificial neural network applications.[10] A good choice would consist of a "reasonable" size set of the examples that constitutes a good representation of the domain.

Several trials with training sets of various sizes have indicated that, for our purposes and with the resources available, a few hundred examples would be adequate. In fact, the choice of training set examples is very much dependent upon the availability of the spectral and structure databases. The spectral data were obtained from the Mattson ICON FTIR spectral library of 500 organic compounds.[19] The compounds for the training set were selected from the library by simply taking all of the odd-numbered samples except for polymeric compounds (the even-numbered samples were used for the test set). The final training set consisted of 212 FTIR spectra of organic compounds whose molecular weights ranged from 50 to 250. These compounds could have one or several, or even sometimes none, of the structural groups that the network was being trained to recognize. For each structural group, the distribution of compounds containing at least one instance of the group in the training set can be seen in Table I.

Test Sets. The estimation of neural network learning quality is usually done by performing a test which consist of presenting a series of input vectors from compounds not used in the training step, for which corresponding output values, although known, are not given to the network, so that the quality of the network predictions can be judged. As a test set in this case, the even-numbered spectra from the spectral library, the 236 compounds that were not used in the training set, were used. Since one important aspect of the use of neural networks is their capability of generalization, the test set was made as large as reasonably possible compared to the training set. The distribution of compounds containing at least one instance of a structural group in the test set is similar to that found in the training set, as seen in Table I.

Thresholds for Output Interpretation. In using an output value to make a decision on the presence or absence of a structural group, threshold values must be established for both the positive and negative conclusion levels. The determination of the threshold values for the acceptance, AL (level above which the group is accepted), or rejection, RL (level below which the group is rejected), of a structural group is usually done empirically. In this work, we have used an optimization procedure based on the modified simplex method[20] to determine the threshold values which minimize a cost function taking into account correct and incorrect predictions obtained with these thresholds. The cost function used in the optimization procedure was

$$F = K_1 IP_f + K_2 IA_f - K_3 IP_c - K_4 IA_c$$

K_1 and K_2 being the penalty coefficients attributed to the numbers of each type of false assignment (IP_f indicated present when absent and IA_f indicated absent when present), and K_3 and K_4 being the credit coefficients attributed to each type of correct assignment (IP_c indicated present when present and IA_c indicated absent when absent). In our case the factors were set to 1, so as to provide for an undifferentiated measure of the two types of error, but the factors could easily be adjusted, so that for a particular purpose one could favor the acceptance or the elimination of doubtful assignments. For each trained network a set of 100 additional examples, which were not included in the training set, was used to apply the optimization procedure, leading to the values of RL and AL to be used for classification purposes. For each output O_i, in the range [0...1] of real values, the two threshold values RL_i (reject level) and AL_i (accept level) are used to designate the structural feature as IA_i (indicated absent) or IP_i (indicated present), or when the output O_i falls between the two threshold values, as NC_i (not classified).

Performance Indicators Used. Ideally, after completing the learning session, the neural network should be able to classify the spectra of compounds that had been presented to it. Thus, for each example of the training set, the network calculated output should be equal to the target values. Practically, this is not often the case because the adjustment of the network weights is not only done for the particular example but on the basis of the whole training set. Thus, separability between compounds belonging to one category and those which do not may vary for different structural groups.

Various discriminating indexes have been used to measure the quality of a trained network. A single indicator of performance is inadequate to provide a complete appreciation of the various aspects of network performance; we have, therefore, used several different performance indices.

One measure of performance, a goodness of fit measurement, the root mean square of the residuals

$$S_r = \sqrt{\frac{\sum(\text{target value} - \text{neural output})^2}{N}}$$

where N is the number of examples in the set, provides a direct measure of the fit between the network's output and the target values. The values for the results of the training and test sets of the various network configurations were calculated and are given in Table II (column 1 for the training set and column 5 for the test set). These values for the global fit do not indicate the fit for those groups present and those absent, which may be significantly different. In order to evaluate

Table III. Comparison of Overall Quality of Responses

	full set				present			absent	
	R_r	Q_r	GQ	EQ_r	P_f	Q_{pr}	A_{50}	A_f	Q_{ar}
				(A) Training Set					
single output, no hidden layer	0.970	0.997	0.967	0.869	0.859	0.993	0.997	0.992	0.998
single output, one hidden layer	1.000	1.000	1.000	1.000	0.998	1.000	0.997	1.000	1.000
multioutput, one hidden layer	0.997	0.999	0.996	0.984	0.983	0.996	0.997	0.999	1.000
				(B) Test Set					
single output, no hidden layer	0.922	0.945	0.873	0.504	0.625	0.901	0.868	0.933	0.954
single output, one hidden layer	0.966	0.936	0.905	0.629	0.728	0.874	0.871	0.948	0.950
multioutput, one hidden layer	0.929	0.922	0.858	0.447	0.649	0.810	0.738	0.909	0.944

this difference, one has to calculate separately the weighted mean of the root mean square of the residuals for the compounds having the structural group, S_{rp}, and for those that do not have it, S_{ra}. These quantities were calculated for the training and test sets and are presented in Table II (columns 2 and 3 for the training set and 6 and 7 for the test set).

The residual errors in the network gives one measure of the network's state of training, but not necessarily a good indication of the network's ability to discriminate between samples that have and those that do not have a particular structural group. This capability can be related to the difference between the weighted means of the output obtained for compounds having a particular group and the weighted mean of output for compounds without the group. In order to take into account the dispersion about the means, this difference is divided by the common standard deviation, giving an index of discrimination, T_d:

$$T_d = \frac{m_p - m_a}{\sqrt{(v_p + v_a)}}$$

m_p = mean of neural output for presents; m_a = mean of neural output for absents; v_p = variance of neural output for presents; and v_a = variance of neural output for absents. The larger this index is, the better the discrimination. Values are calculated for the training and test sets of each network and are given in Table II (columns 4 and 8).

All of the above measures are directly related to the calculated output values; however, ultimately the output value has to be used for prediction, thus another set of measures has to be defined that is related to the number of correct and incorrect predictions that are made. Using the categorization previously described, several indices are calculated to show various aspects of the quality of the network:

Global response ratio:

R_r = number of samples classified/number of samples
$= \Sigma(IP + IA)/\Sigma(P + A)$

Global response quality:

Q_r = number of correct responses/number of responses
$= \Sigma(IP_c + IA_c)/\Sigma(IP_c + IA_c + IP_f + IA_f)$

Global quality:

GQ = number of correct responses/number of samples
$= \Sigma(IP_c + IA_c)/\Sigma(P + A)$

Although the global quality index is commonly used as an indicator of network performance, it can give an overoptimistic view of network performance when the populations of molecules containing the structure are either very low or very high. In order to assess the quality of the network, this index should be compared to the statistical chance of giving a correct response based on the distribution of molecules containing the structural features in the population. The chance of a correct response for a structural feature can be calculated on the basis of the portion of the sample containing the group, P_i, as follows: $S_t = 1 - 2P_i + 2P_i^2$. For very poorly or very highly represented groups, the probability of a chance correct characterization can be very high. For example, the primary amino group, which is present in only 5.1% of the samples, has an S_t value of 0.903. The improvement in the quality of response over chance can be expressed as the extrastatistical quality response index, EQr, the fraction of the possible improvement that has been actually achieved:

Extrastatistical quality:

EQ_r = difference between experimental and chance correct assignments/maximum possible improvement
$= (GC - S_t)/(1 - S_t)$

These global response indices give an overall view of network performance, but they do not show the differences between the performance of the network in detecting the presence of a group and in detecting the absence of the group. There may be very significant differences here; therefore, separate indices for these were calculated for this purpose as indicated below:

Fraction found of groups present:

P_f = number correctly classified as present/number present
$= \Sigma IP_c/\Sigma P$

Quality of found present response:

Q_{pr} = number correctly classified as present/number correctly classified as present + number incorrectly classified as present
$= \Sigma IP_c/\Sigma(IP_c + IP_f)$

Fraction found absent of groups absent:

A_f = number correctly classified as absent/number absent
$= \Sigma IA_c/\Sigma A$

Quality of found absent response:

Q_{ar} = number correctly classified as absent/number correctly classified as absent + number incorrectly classified as absent
$= \Sigma IA_c/\Sigma(IA_c + IA_f)$

The values are reported in Table IIIA and B, respectively, for the training and test sets.

In their analysis of the results of training a network to use infrared peak positions, Robb and Munk[14] used a discrimination index, "accuracy at 50% information retrieved (A50)", based upon the number of false positive indications obtained at a threshold level equal to the median for those compounds containing the group. This index represents a special case of

Table IV. Fit of Test Set Outputs to Target Values for Different Network Configurations

	single output, no hidden layer				single output, one hidden layer				multioutput, one hidden layer			
	full set S_r	present S_{rp}	absent S_{ra}	T_d	full set S_r	present S_{rp}	absent S_{ra}	T_d	full set S_r	present S_{rp}	absent S_{ra}	T_d
OH	0.201	0.292	0.154	23.52	0.203	0.261	0.177	25.85	0.267	0.323	0.243	18.90
CH_2–OH	0.170	0.531	0.089	7.20	0.170	0.492	0.106	7.47	0.204	0.617	0.116	5.27
COOH	0.112	0.318	0.059	13.83	0.099	0.296	0.043	14.87	0.174	0.514	0.079	7.74
COOR	0.218	0.363	0.169	15.93	0.233	0.386	0.180	14.71	0.241	0.406	0.183	13.61
C–CHO	0.197	0.506	0.152	7.17	0.151	0.391	0.115	10.10	0.211	0.378	0.194	9.22
C–CO–C	0.237	0.550	0.163	7.10	0.216	0.495	0.153	8.26	0.230	0.586	0.136	6.59
RR′–NH	0.226	0.633	0.137	4.93	0.192	0.557	0.108	6.41	0.279	0.663	0.210	4.18
CNH_2	0.172	0.689	0.075	3.69	0.172	0.639	0.095	4.22	0.230	0.645	0.183	3.66
C=C	0.354	0.536	0.280	8.58	0.330	0.510	0.256	9.82	0.449	0.503	0.432	7.01
C6	0.329	0.324	0.333	17.65	0.300	0.313	0.291	20.17	0.421	0.494	0.357	11.17
$C(CH_3)_2$	0.279	0.702	0.163	4.92	0.325	0.792	0.202	3.30	0.345	0.762	0.248	3.19
CH_2	0.386	0.385	0.387	12.91	0.411	0.341	0.512	10.19	0.429	0.348	0.544	9.52
overall	0.240	0.419	0.162	10.62	0.233	0.393	0.163	11.28	0.290	0.455	0.220	8.34

the quality-of-found-present-response, Q_{pr}, in which the threshold value is set to a level such that 50% of the groups are found. This index gave values of 100% for most of the compounds in our training set (only one group, the alkenes, had a value less than 90%); however, the results on the test set were more diverse and are, therefore, given in Table III (the values in this table are presented as fractions rather than percentages).

These indices thus give one the opportunity to assess the quality of the network's ability to recognize the presence or absence of each particular structural feature and give an indication of the reliability of each prediction.

RESULTS AND DISCUSSION

Evaluation of Quality of Fit of the Training. The quality of the fit of the training by the different types of networks can be seen in Table II. The standard residual results for the training set generally indicate that the best fit is found for the networks with a hidden layer and single output node dedicated to one structural group. Removal of the hidden layer or training the network to recognize several groups at the same time caused an increase in the average error. However, for the test set of compounds, the difference in the fit for the various configurations is not as great as in the training set. In these particular training regimes, it would appear that, although the single output network with a hidden layer converges best on the training set, the differences in the residuals of the test set is much less significant.

Evaluation of Discrimination Ability. One measure of the ability of the network to discriminate between those that have and those that do not have the group, the T_d value, is given in Table II. A higher value of this index indicates a greater discrimination ability; as seen from the values in Table II, there is little difference in the global results for the different networks. However, the results for the individual groups show significant differences, as will be seen later. The T_d values give a measure of the discrimination ability without taking into account the use of optimized threshold values; the values of the global discrimination indices that use the thresholds from the simplex optimization are presented in Table IIIA and B.

From the values of the commonly used indices of global performance, R_r, Q_r, and GQ, for the training set (Table IIIA), one gets the impression of very well-trained networks. The values from the test set (Table IIIB) indicate that the networks are not quite as good at classifying previously unseen samples. The values of the global quality ratio (GQ, fraction of responses that are correct) for the test set indicate that the discriminatory ability of the network can be enhanced by the addition of a hidden layer or by training the network for each group individually. This global quality index is composed of correct negative as well as positive answers; the separate indices for the present, Q_{pr} and A_{50}, and absent, Q_{ar}, classes suggest that the network is better at finding the absence of a group than its presence. This is due to the fact that in most cases there are many more samples in which the group is absent rather than present and chance correct responses are higher. The extrastatistical quality response index, E_{Q_r}, which shows the relative improvement over chance, indicates very clearly that the test set is not as well classified as the training set.

Although the indices for the full set of structural groups give a good overall picture of the network's performance, the results for the individual groups show marked differences in the network's ability to detect different groups. The fit of the output values for the test set as well as the T_d values are given in Table IV.

The values of the standard residuals for the separate groups in Table IV show significant differences between the different groups. The differences between the different types of network are not so great. The same trends are observed in all of the network configurations.

The values for the discrimination indices for the individual groups are given in Table V, for the full set of samples, for those containing the group, and for those in which the group is absent, parts A–C, respectively. The global quality measure is quite good for each group, but one must remember that this index measures the quality of both positive and negative responses, and for poorly populated groups it has a high statistical bias. The E_{Q_r} ratio, which measures the improvement over chance predictions, gives the most dramatic indication of the differences. The fraction-found index, P_f, also shows a great variation and is a good test of the network performance.

From the results in Table V, it is clear that some groups are well-characterized and others not so. The two best are the hydroxyl group and the carboxylic acid group. These have distinctive infrared absorption bands, and one might expect the network to find that these groups are easy to recognize. The two poorest characterized groups are the *gem*-dimethyl group and the NH group. It is easy to speculate as to why these groups might be difficult to characterize. The *gem*-dimethyl group has only one very distinctive region (1380 cm^{-1}), and the peak is split into a doublet, which would be very difficult to recognize with a resolution of 12 cm^{-1}. The NH peak may be difficult to characterize because a diverse category of compounds, primary and secondary amines and amides, were used to train the network for this group.

Table V. Quality of Responses for Test Set.

	full set				present			absent	
	R_r	Q_r	GQ	E_{Q_r}	P_f	Q_{pr}	A_{50}	A_f	Q_{ar}
(A) Network with Single Output, No Hidden Layer									
OH	0.941	0.968	0.911	0.777	0.831	0.964	1.000	0.942	0.970
CH_2–OH	0.962	0.974	0.936	0.549	0.333	0.857	0.900	0.986	0.977
COOH	0.987	0.991	0.979	0.875	0.864	1.000	1.000	0.991	0.991
COOR	0.945	0.960	0.907	0.693	0.705	0.886	0.960	0.953	0.973
C–CHO	0.924	0.982	0.907	0.263	0.375	0.857	0.890	0.945	0.986
C–CO–C	0.928	0.950	0.881	0.374	0.400	0.714	0.750	0.938	0.966
RR′–NH	0.958	0.956	0.915	0.454	0.400	0.889	0.830	0.963	0.959
CNH_2	0.983	0.970	0.953	0.517	0.167	1.000	0.670	0.996	0.970
C=C	0.843	0.894	0.754	0.294	0.358	0.792	0.684	0.869	0.909
C_6	0.835	0.909	0.758	0.505	0.730	0.869	0.890	0.779	0.938
$C(CH_3)_2$	0.903	0.934	0.843	0.200	0.154	0.800	0.890	0.929	0.938
CH_2	0.856	0.851	0.729	0.412	0.748	0.934	0.950	0.694	0.728
overall	0.922	0.945	0.873	0.504	0.625	0.901	0.868	0.933	0.954
(B) Network with Single Output, One Hidden Layer									
OH	0.970	0.956	0.928	0.820	0.892	0.921	1.000	0.942	0.970
CH2–OH	0.975	0.974	0.949	0.639	0.556	0.833	0.750	0.982	0.982
COOH	0.975	0.991	0.966	0.799	0.864	1.000	1.000	0.977	0.991
COOR	0.966	0.952	0.919	0.735	0.773	0.872	0.880	0.953	0.968
C–CHO	0.983	0.983	0.966	0.732	0.688	0.917	0.889	0.986	0.986
C–CO–C	0.966	0.961	0.928	0.620	0.640	0.842	0.857	0.962	0.971
RR′–NH	1.000	0.962	0.962	0.754	0.650	0.867	1.000	0.991	0.968
CNH2	1.000	0.970	0.970	0.693	0.583	0.778	0.900	0.991	0.978
C=C	0.919	0.889	0.818	0.477	0.585	0.816	0.813	0.885	0.905
C6	0.898	0.915	0.822	0.636	0.810	0.920	0.962	0.831	0.911
C(CH3)2	0.992	0.876	0.869	0.330	0.231	0.400	0.448	0.948	0.909
CH2	0.953	0.800	0.763	0.485	0.768	0.885	0.949	0.753	0.681
overall	0.966	0.936	0.905	0.629	0.728	0.874	0.871	0.948	0.950
(C) Network with Multioutput, One Hidden Layer									
OH	0.932	0.945	0.881	0.703	0.862	0.903	0.970	0.889	0.962
CH_2–OH	0.987	0.957	0.945	0.609	0.444	0.727	0.750	0.986	0.968
COOH	0.945	0.973	0.919	0.524	0.227	0.833	1.000	0.991	0.977
COOR	0.941	0.968	0.911	0.707	0.727	0.970	1.000	0.953	0.968
C–CHO	0.924	0.982	0.907	0.263	0.500	0.800	0.800	0.936	0.990
C–CO–C	0.987	0.948	0.936	0.664	0.600	0.833	0.800	0.976	0.958
RR′–NH	0.987	0.927	0.915	0.454	0.500	0.588	0.588	0.954	0.954
CNH_2	0.992	0.940	0.932	0.298	0.417	0.417	0.294	0.960	0.968
C=C	0.814	0.844	0.686	0.100	0.509	0.659	0.667	0.738	0.894
C_6	0.818	0.845	0.691	0.367	0.600	0.800	0.781	0.757	0.873
$C(CH_3)_2$	0.941	0.896	0.843	0.200	0.231	0.462	0.333	0.919	0.923
CH_2	0.877	0.836	0.733	0.421	0.834	0.875	0.872	0.553	0.746
overall	0.929	0.922	0.858	0.447	0.649	0.810	0.738	0.909	0.944

In almost all cases, the results indicate a better performance of the multilayer networks that were trained to recognize one group at a time. However, in most cases the performance of the network without a hidden layer is almost as good as the network with a hidden layer, although for some cases the hidden layer improves the results significantly.

Basis of Discrimination by the Network. If the results of the trained network are used to characterize new compounds, it is important to determine the spectra–structure correlations that the network has learned to use in classifying the spectra. In the case of infrared spectroscopy the relationships on which the classification should be based are well-known. A well-trained network would be expected to have found correlations similar to those that have been established over the years by infrared spectroscopists. It is possible that a network becomes trained to characterize groups on the basis of correlations that do not have physical meaning, particularly if a large number of input nodes are used relative to the size of the training set. Such a network may appear to be effective on a limited set of samples, but it would not be reliable in extrapolating to new compounds. The verification of the basis of discrimination of the network by analysis of the connection weights is, therefore, an important part of the development of a neural network for infrared interpretation.

During the learning phase, the connection weights are adjusted step-by-step, leading to correlations building up between the spectral data and the defined structural units. For nets with a hidden layer, the input spectral data and output structural data are not directly related, so that the network's knowledge, which is distributed among all network connections, cannot be easily understood. It is not practical to determine the contribution of each input node to the resulting structural information. However, for a network with no hidden layers, the input data and the calculated output are directly connected, so that the connection weights link the spectral data directly to the structural group. An example showing the weights for each input node (wavenumber) for the carboxyl group is shown in Figure 1. Strong positive weights correlate well with the pertinent infrared bands of the carboxyl group (C=O stretching C—O stretching, O—H stretching). The negative weights represent residual bands of other compounds of the training set that do not contain the carboxyl group.

The visual presentation of the network's perception of a particular group may be shown in the form of a synthetic spectrum by applying the weights to the spectra of those compounds of the training set that contain the structural group. The relative contribution C_{ij} of each absorption frequency i to the synthetic spectrum of the functionality j is expressed

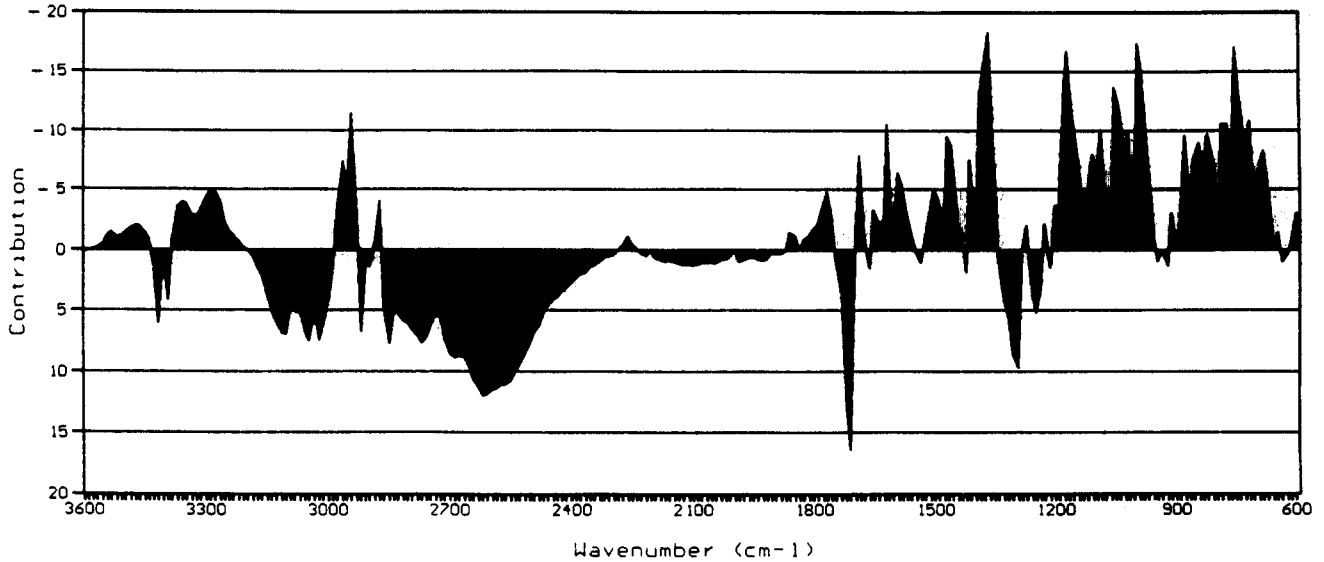

Figure 1. Plot of connection weights vs wavenumber for the carboxyl group, COOH. The plot is presented with the positive weights below the axis in accord with the normal manner of presenting an infrared spectrum.

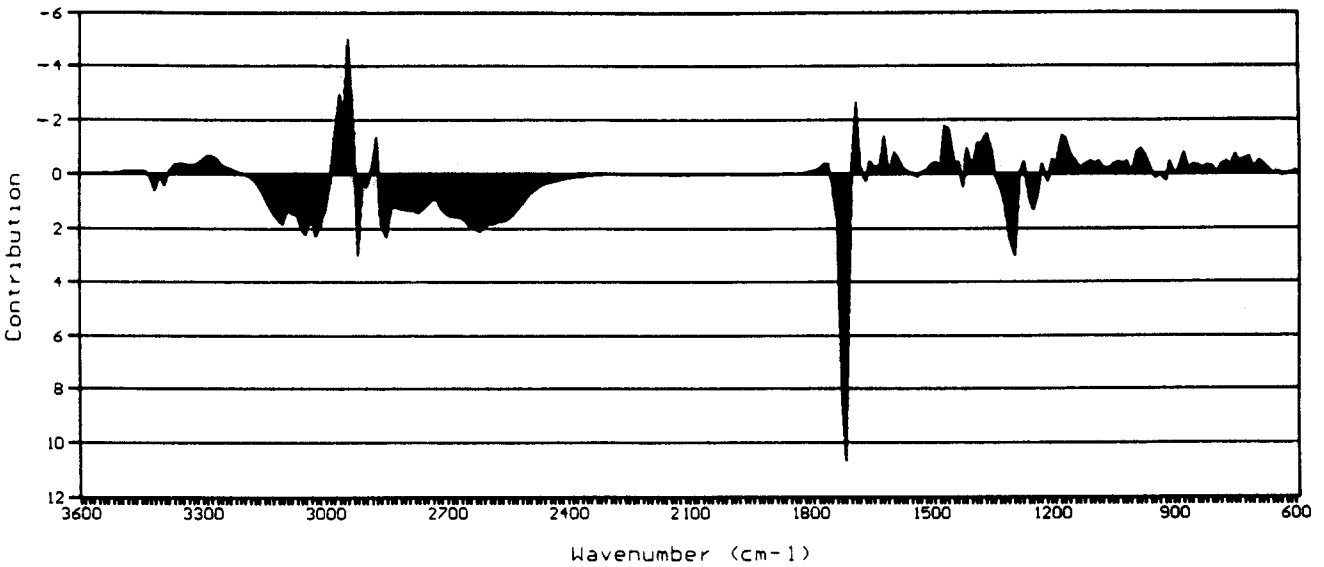

Figure 2. Synthetic spectrum of the carboxylic group, COOH, showing positive contribution from the C—O, C=O, and O—H stretching vibrations and negative contributions from C—H stretching vibrations.

by the relation:

$$C_{ij} = W_{ij}E_{i}\text{av}$$

where W_{ij} is the connection weight between cell i of the input layer to the cell j of the output vector and E_{i}av is the average input of the cell i; this average being done for all compounds of the training set that have the structural group j.

An example of such a plot (for the carboxylic group) is shown in Figure 2. The more positive the contributions, the more the corresponding absorption frequencies are indicative to the net of the presence of the structural feature j. The negative contributions occur in those spectral regions where adsorption would appear to diminish the possibility that the group is present. These peaks may be interpreted as an indication of negative evidence or of an effect of prior probability discrimination. Explicit negative evidence may be expected to be learned by the network; for example, a strong O—H peak might be expected to exclude the possibility of an acid chloride or the presence of an N—H peak might be expected to decrease the validity of assigning a C—N absorption to a tertiary amine. The sample distribution may also be considered to contribute to the positive or negative aspect of the synthetic spectrum by the way of providing prior probability evidence. One might expect that a group that occurs rarely will not be well-learned by the network and that significant negative absorption will exist in its profile, providing an effective prior probability index for the group. If a training set is selected to represent the expected distribution of the samples to be analyzed, then prior probability evidence is legitimate evidence to be used in predicting the likelihood of the presence of a particular group. Indeed, this has been specifically incorporated into the information base of several expert systems. This representation of the networks knowledge, therefore, incorporates all of the evidence categories that the rule-based systems have attempted to incorporate into their knowledge base.[7,21]

Group Characteristics Learned by the Network. It might be expected that those structural features that show very distinctive characteristic infrared absorptions would be those for which network discrimination power is greatest. Functional groups having infrared bands in common with other groups but whose shape or intensity is distinctive, e.g., carboxylic acids, esters, or alcohols, could also be well-classified. Some other structural features are less likely to be well-categorized

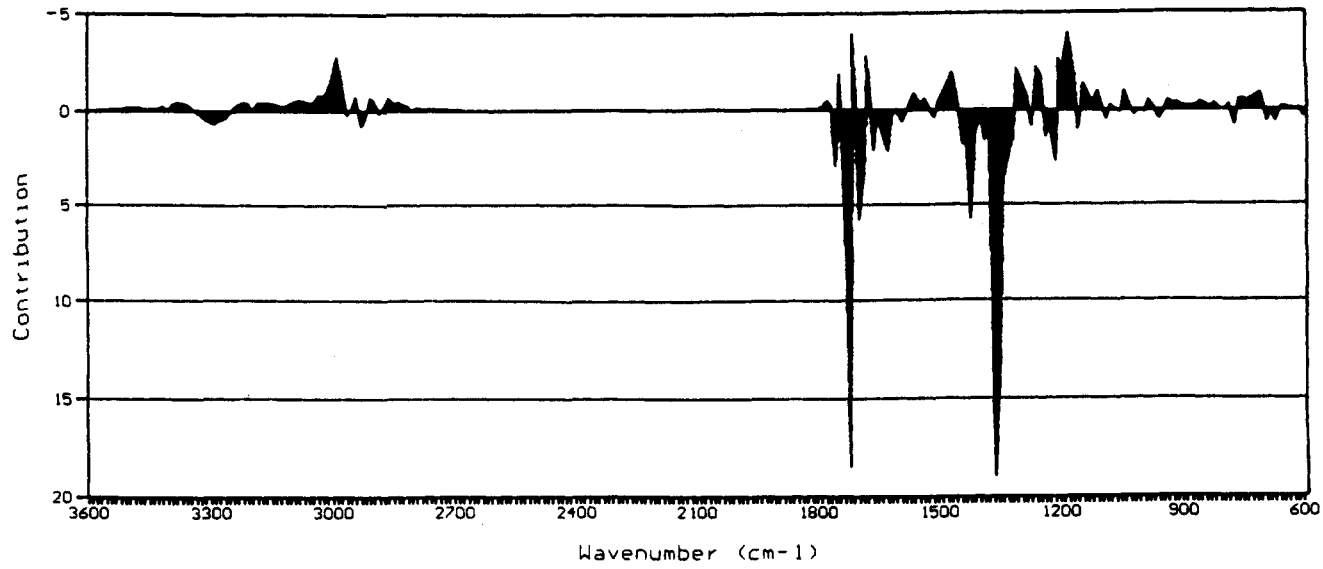

Figure 3. Synthetic spectrum of the keto group, RCOR, showing positive contributions from the C=O stretching band, as well as from the C-H bending band of the methyl group, which is also present in most of the compounds in the training set that contains the keto group.

(e.g., tertiary amines, isolated double bonds). However, some interesting observations can be made on the results, and these can be taken into account both qualitatively and quantitatively in the input to the expert system. As may be seen in the synthetic spectrum of the keto group (Figure 3), the network has associated the presence of the CH bending bands (1360 cm^{-1}) as well as the CO stretching band (1700 cm^{-1}) with the group. Two-thirds of the compounds that were used as representative of the keto group in the training were methyl ketones, and all but one of the others contained some other type of methyl group. It can, therefore, be easily understood why the network learned to associate the CH stretching of the methyl group with the presence of the keto group.

In the aldehyde case, two major zones of positive contributions can be seen in the range 1676–1752 cm and 2676–2880 cm^{-1}, which can be identified respectively to C=O and C-H stretching vibrations. Close examination of the carbonyl region shows a separation of the conjugated and aliphatic aldehydes. In the carboxylic acid case, three positive contributions zones in the ranges 1280–1330, 1690–1730, and 2400–3180 cm^{-1} can be associated respectively to C-O, C=O, and O-H stretching vibrations. The strong negative contribution near 3000 cm^{-1} indicates that the presence of an absorption at that frequency would be used by the network to reduce the value of the output for the carboxylic acid group.

The network topological representation resulting from the learning process is strongly dependent on the composition of the training set, such that it can sometimes lead to unexpected results. If the network is to be used to classify compounds that are very diverse, it is important to maximize the diversity of the training compounds used for each structural group. If the domain of the compounds for which the network is to be used is restricted, and known, it would be advantageous to train the network on a set of compounds with a similar distribution of functionality.

In spite of indications of good global fit, the network can give poor results in particular situations. Examination of the connection weight distributions can give an explanation for erroneous results. For example, in the case of the ketone group, it can easily be seen that peaks from the methyl group have been recognized as important indicators of the keto group. As another example, 2-hydroxybenzaldehyde is not properly classified with respect to the OH or the aldehyde group. In this case, hydrogen bonding between the two adjacent groups gives an atypical experimental spectrum, and thus the false conclusion of the network is not surprising. This exemplifies one of the limits of the neural network approach, i.e., the difficulty in handling special cases which occur infrequently in the training set. In fact, this difficulty is frequently encountered when applying pattern matching methods which rely upon correlation tables. Explicit knowledge may be used to overcome this limitation, something which can be provided by cooperation with a rule-based expert system.

CONCLUSION

The results show that it is practical to use spectral data directly to train a network to recognize structural features. The several indices that we used show that the performance of the network should be evaluated very carefully as it is easy to obtain very optimistic evaluations of the network's capabilities from certain indices. This is particularly true of the global indices for both highly and thinly populated categories. It is easy to obtain what appears to be good results when there are few examples and a large number of input nodes. The usefulness of the network for the prediction of the structures of compounds should be checked by using a large set of previously unseen test compounds and by determining the basis of the correlation by analysis of the connection weights.

The results from the network without a hidden layer indicate that the categorization follows the expected correlations in most cases. The generation of a synthetic spectrum from the connection weights provides an invaluable tool for the diagnosis of the network training. For a particular training set, it is possible to see the spectral information that was used for the correlations and to see any anomalies that might have become incorporated. It is then possible to modify the composition of the training set to produce a network trained more closely to meet the users requirements. Since the network with a hidden layer gives slightly better results, it can be used as the active network, while the one without the hidden layer can be used for the verification of the validity of the correlations and to obtain information to use by the expert system.

4The input to the expert system that we have created for assisting with the interpretation of infrared spectra currently requires that the user "read" the spectrum and make a verbal description to the expert system. With an interface between the neural network and the expert system, the need for this step can be removed.

The quantification of the significance of negative evidence is very difficult as we know from our own work on the expert system and as indicated by the very extensive efforts along these lines by others. The connection weights from the single-layer network provide a way to give measured values to the pertinence factors rather than subjective guesses.

In this initial phase of the training, all spectra from the database were used regardless of quality, except for polymers. It is clear that the training set should be scrutinized to eliminate poor quality spectra as the effect on the training may be quite significant. The next step in this project will be to increase the training database significantly and to develop interfaces to use the results of the trained database in a synergistic relationship with expert systems.

Note Added in Proof. Exploiting the potential of a neural network for structural elucidation using spectral information is very promising, and this work should be considered as a preliminary investigation; in particular, no attempt was made at tackling the problem of finding relationships between infrared data and global structural features like the skeleton of a molecule. Since this paper was submitted, other contributions have been published. Wiegel and Herges[22] used a specialized network to recognize aromatic substitution patterns. Tanabe et al.[23] have developed a hierarchical neural network system for fast identification of a spectrum from a database of 1129 spectra. Meyer and Weigelt[24] used an approach similar to ours to identify 32 different functional groups from digitized infrared spectra, although with a lower resolution and a rather small number of compounds in the training and test sets (100 and 50, respectively). Obviously, using the back-propagation algorithm for adjusting the connection weights with training sets of such a limited size, as also in our case (212), one can only expect to reach a local minima. However, as we have shown, the trained networks are able to generalize fairly well to examples which were not included in the training set.

REFERENCES AND NOTES

(1) Tomellini, S. A.; Hartwick, R. A.; Woodruff, H. B. Automatic Tracing and Presentation of Interpretation Rules Used by PAIRS: Program for the Analysis of IR Spectra. *Appl. Spectrosc.* **1985**, *39*, 331–3.

(2) Moldoveanu, S.; Rapson, C. A. Spectral Interpretation for Organic Analysis Using Expert System. *Anal. Chem.* **1987**, *59*, 1207–12.

(3) Luinge, H. J.; Van der Maas, J. H. Artificial Intelligence for the Interpretation of Combined Spectral Data. Design and Development of a Spectrum Interpreter. *Anal. Chim. Acta* **1989**, *223*, 135–47.

(4) Funatsu, K. T.; Sasaki, S. The automated structure elucidation system CHEMICS. *Chem. Inf. Proc. Int. Conf.* **1989**, 271–81.

(5) Huixiao, H.; Xinquan, X. ESSESA: An Expert System for Elucidation of Structure from Spectra. 1. Knowledge Base of Infrared Spectra and Analysis and Interpretation Programs. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 203–10.

(6) Elyashberg, M. E.; Serov, V. V.; Martirosyan, E. R.; Zlatina, L. A.; Karasev, Y. Z.; Koldashov, V. Y.; Yampol'skii, Y. An expert system for molecular structure elucidation based on spectral data. *THEOCHEM* **1991**, *70*, 191–203.

(7) Tomellini, S. A.; Wythoff, B. J.; Levine, S. P. Developing Knowledge-Based Systems for Interpreting Infrared Spectra. In *Computer-Enhanced Analytical Spectroscopy*; Jurs, P. C., Ed.; Plenum Press: New York, 1992; Vol. 3, Chapter 8, pp 215–38.

(8) Cabrol-Bass, D.; Rabine, J. P.; Forrest, T. P. An Educational Problem Solving Partner in Prolog for Learning Infrared Spectroscopic Analysis. *Comput. Educ.* **1988**, *12*, 241–6.

(9) Cabrol-Bass, D.; Forrest, T. P.; Rabine, J. P.; Ricard, D.; Rouillard, M. IRexpert, an Infrared Interpretation Assistant. *J. Chem. Educ.*, in press.

(10) Zupan, J.; Gasteiger, J. Neural networks: A new method for solving chemical problems or just a passing phase? *Anal. Chim. Acta* **1991**, *246*, 1–30.

(11) Thomsen, J. U.; Meyer, B. Pattern Recognition of the 1H NMR Spectra of Sugar Alditols Using Neural Network. *J. Magn. Reson.* **1989**, *84*, 212–7.

(12) Kvasnicka, V. An application of neural networks in chemistry. *Chem. Pap.* **1990**, *44*, 775–92.

(13) Kvasnicka, V. An application of Neural Networks in Chemistry. Prediction of 13C NMR Chemical Shifts. *J. Math. Chem.* **1991**, *6*, 63–76.

(14) Robb, E. W.; Munk, M. E. A Neural Network Approach to Infrared Spectrum Interpretation. *Mikrochim. Acta (Wien)* **1990**, *1*, 131–55.

(15) Munk, M. E.; Madison, M. S.; Robb, E. W. Neural Network Models for Infrared Spectrum Interpretation. *Mikrochim. Acta (Wien)* **1991**, *2*, 505–14.

(16) Fessenden, R. J.; Györgyi, L. Identifying Functional Groups in IR Spectra Using an Artificial Neural Network. *J. Chem. Soc. Perkin Trans. 2* **1991**, 1755–62.

(17) Curry, B.; Rumelhart, D. E. MSnet: A Neural Network That Classifies Mass Spectra. *Tetrahedron Comput. Methodol.* **1990**, *3*, 213–37.

(18) NeuroShell from Ward Systems Group, Inc., 245 W. Patrick St., Frederick, MD 21701.

(19) ICON FTIR Spectral Library, Mattson Instruments Inc., 1001 Fourier Dr., Madison, WI 53717.

(20) Nelder, J. A.; Mead, R. A simplex method for function minimization. *Comput. J.* **1965**, *7*, 308.

(21) Curry, B. A Distributed Expert System for Interpretation of GC/IR/MS Data. In *Computer-Enhanced Analytical Spectroscopy*; Meuzelaar, H. L. C., Ed.; Plenum Press: New York, 1990; Vol. 2, Chapter 8; pp 183–209.

(22) Weigel, U.-M.; Herges, R. Automatic Interpretation of Infrared Spectra: Recognition of Aromatic Substitution Patterns Using Neural Networks. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 723–31.

(23) Tanabe, K.; Tamura, T.; Uesaka, H. Neural Network System for the Identification of Infrared Spectra. *Appl. Spectrosc.* **1992**, *46* (5), 807–10.

(24) Meyer, M.; Wiegelt, T. Interpretation of Infrared Spectra by Artificial Neural Networks. *Anal. Chim. Acta.* **1992**, *265*, 183–90.