Identifying Functional Groups in IR Spectra Using an Artificial Neural Network

Ralph J. Fessenden*[,a] **and László Györgyi**[b]
[a] *Department of Chemistry, University of Montana, Missoula, Montana 59812, USA*
[b] *Institute of Inorganic and Analytical Chemistry, Eötvös Lóránd University, Budapest, Hungary*

Artificial neural networks are capable of learning and are potentially superior to other computer programs at pattern recognition. We have used a simple two-layer, feed-forward neural network to obtain structural information from IR spectra of organic compounds. The network was taught to recognize the presence and absence of selected functional groups and bond types by simply presenting it with IR spectra of training compounds. The back-propagation algorithm was used to adjust the weights of the network. Spectra of compounds not belonging to the training set were used for testing. The trained network was able to recognize the presence and absence of the functional groups and bond types in the spectra of previously unseen compounds. Percent transmittance *vs.* wavenumber was the most successful input data representation. Using both bond type and functional group identification in the output layer significantly reduced the number of incorrect classifications.

Inspired by the structure and function of biological neural networks, the study of artificial neural networks is an exponentially growing interdisciplinary field of science.[1] Although artificial neural networks have various architectures and modes of operation, they all contain parallel, distributed information processing structures whose elements are interconnected by unidirectional signal channels called connections. Each element or neuron contained in a connection has a local memory and can carry out localized information processing operations.[1d]

Computer simulated neural networks are known to be able to learn and then recognize patterns. A classical pattern recognition problem in chemistry is the correlation of spectra with structure, *i.e.*, the identification of functional groups in a compound on the basis of spectral information. There has been a great deal of work done in this area in past decades.[2–9] The most advanced systems combine the information available from various spectral methods and suggest a complete structural description of the unknown compound. The two common features of these expert systems are that they contain a spectral knowledge base and a built-in reasoning system. The first of these features is a list of spectral characteristics (band position, band width, slope, chemical shift, *etc.*) of substructures of interest. The second is an approximation of the human thought process when a structural analysis is undertaken by an expert.

Without undertaking the ambitious project of building a full structural interpeter, this work addresses a subproblem – the recognition of the functional groups in IR spectra. We were interested in seeing if it would be possible to use an artificial neural network to interpret IR spectra in such a way that neither a database containing the position and characteristics of the key absorption bands nor a set of rules for inferring the structure from the spectrum would be needed. Rather, we anticipated that the network would be able to learn by example, *i.e.* by simply presenting it with spectra of known structures. Since neural networks are capable of generalization, it was hoped that the trained network would be able to interpret the spectrum of a previously unseen compound to the extent it was taught to do so.

There have been some recent applications of neural networks in spectroscopy. Meyer and co-workers[10] used this method to identify carbohydrates on the basis of their ^{1}H NMR spectra. After the work described in this paper was completed a very detailed study of functional group identification in IR spectroscopy by a neural network was published by Robb and Munk.[11] There are, however, some major differences between their work and our approach. While Robb and Munk used a large number of compounds and functional groups, they also used a different network architecture, a one-layer network. This architecture was criticized in the famous book of Minsky and Papert[12] because it can only find a perfect set of weights if the classes of the classification problem (functional groups, in this case) are linearly separable (in this case, in the wavenumber space). The linear activation function used by Robb and Munk imposes further limitations, namely, the output (functional group assignment) is assumed to be a linear function of the input (spectra), which is clearly an oversimplication of the problem.

The approach presented here utilizes a net architecture which was developed[1] to overcome the limitations of the one-layer neural networks. Indeed, our results indicate that a better discrimination in the classification decisions is possible when this more advanced network is used.

Other neural network applications in chemistry include models for knowledge representation (mainly for fault diagnosis) in chemical engineering,[13] several works on protein secondary structure analysis and prediction,[14] a model for DNA promoter site recognition,[15] and for prediction of electrophilic aromatic substitution reactions.[16]

Experimental

Network Description.—For this work we used a two-layer, feed-forward neural network. Fig. 1 is a schematic diagram that shows the architecture of such a neural network. The first layer, called the input layer, is where the information is presented to the network. This layer is used only to present the network with its input. The input layer does not perform any computation and it is not therefore included in the layer count. The last or third layer, called the output layer, is where the response of the network is registered. The layer between the input and output layers is called the hidden layer. Each neuron of the input layer is fully interconnected with each neuron of the hidden layer which in turn is fully interconnected with each neuron of the output layer. There are no connections between the neurons within a layer nor any direct connection between those of the input and output layers. The weights (constant multipliers), or connection strengths, associated with each connection are adjusted during learning.

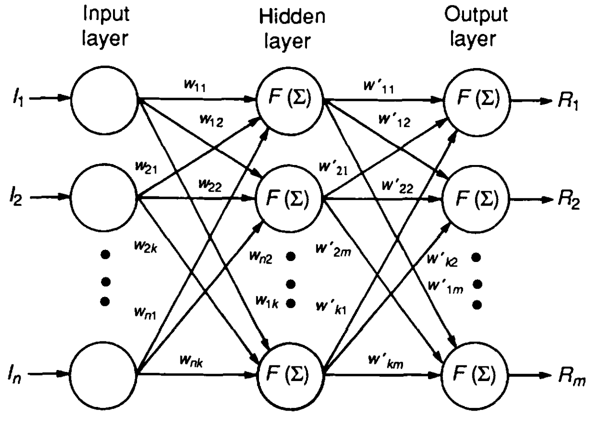

Fig. 1 Schematic diagram of a simple two-layer, feed-forward neural network

After an input is presented to the network (*i.e.* the input layer is activated), the activity propagates forward according to the eqn. (1).

$$a_{j,k} = F(\Sigma w_{ij}a_{i,k-1} + b_j) \qquad (1)$$

In eqn. (1), $a_{j,k}$ represents the activation of the *j*th neuron in *k*th layer (the current layer) except for the input layer; F is the transfer (activation) function; w_{ij} is the weight of the connection from the *i*th neuron of the previous ($k - 1$)th layer (whose activity is $a_{i,k-1}$) to the *j*th neuron on the current layer. The b_j term is the bias term associated with the *j*th neuron of the current layer.

Learning in these networks consists of adjusting the connection strengths to minimize the difference between the response of the network to a particular input and the expected (correct) answer (supervised learning).

This study employed a commercial neural network software, BRAINMAKER, version 1.7, from California Scientific Software.* We used BRAINMAKER to create a fully connected, two-layer, feed-forward network. The maximum number of neurons (input, hidden, and output) the BRAINMAKER can handle is 512. For our study, this value placed a practical upper limit of the number of input neurons of *ca.* 470. This restriction required that each spectrum be sampled rather than inputted as a nearly continous analogue-type format.

Input and Output Neurons. The meaning of the input and output neurons will be described later in this paper in our discussion of the input coding of the spectra and the output coding for functional group identification. The number of input neurons was varied between 13 and 462, while that of the output neurons was either six (one for each functional group under study) or nine (one for each functional group plus the three bond types; see Table 1).

Hidden-layer Neurons.—We used networks with 18 hidden-layer neurons. This value is two times our usual number of output neurons. No systematic study of the effect of the number of hidden-layer neurons on the network performance was carried out.

Transfer Function.—The sigmoid function, $F(\text{net}) = [1 - \exp(-\text{net})]^{-1}$, was used in our network. This function has the limits of 0 in $-\infty$ and 1 in $+\infty$. The steepest slope is at net = 0, where the derivative is 0.25.

Learning.—The standard back-propagation algorithm † was used to adjust the weights of the connections between the neurons. The back-propagation learning rate was set to 1 and the momentum coefficient was 0.9. A bias term was added to each hidden and output layer neuron automatically by BRAINMAKER. The training tolerance was set at 0.1. This means that, for termination of a training session, each output neuron for each of the training spectra had to have a value of >0.9 if the correct answer was 1.0 or a value of <0.1 if the correct answer was 0.0. (See Tables 1 and 2.)

Compounds Studied.—Forty eight liquid compounds were selected for the study. Each compound contained only one of five functional groups: hydrocarbon, alcohol, carboxylic acid, ester and ketone. In addition, each functional group class, except for the carboxylic acids, contained at least four compounds with an unsubstituted phenyl group, which was defined as a functional group for this study. Further, we defined hydrocarbon to be the functional group of those compounds without an oxygen-containing functional group. Cyclohexane was, therefore, classified as a hydrocarbon. Isopropylbenzene was classified as a hydrocarbon with a phenyl functional group. Cyclohexanol was classified as an alcohol, not a hydrocarbon without a phenyl group. Benzyl alcohol was classified as an alcohol, not a hydrocarbon with a phenyl group. The names and structures of the compounds used in this study, with the classification and coding of their functional groups, are listed in Table 2.

Spectra.—The IR spectrum of each compound was obtained as a thin film using a Perkin-Elmer 1600 series FTIR instrument. The compounds were not specially purified for this study and no special care was taken in obtaining the spectra. The spectra were captured in digital format and then converted to ASCII code using a BASIC program supplied to us by the Perkin-Elmer Corporation. The raw spectrum was truncated to 1668 data points, each point being a pair of values (cm^{-1}, %T). The frequency range of the truncated data was from 4000–666 cm^{-1}. The actual data used to train or test a neural network were selected as a subset from these digitized truncated spectra.

Representation of Spectral Data for the Network.—The digitized truncated spectra was used in one of the following four forms: absorbance (A) *vs.* μm; A *vs.* cm^{-1}; percent transmittance (%T) *vs.* μm; or %T *vs.* cm^{-1}. The network input was created by equidistant wavelength (or frequency) sampling of the whole truncated spectrum. The absorption units, such as %T, from the selected wavelength points of the spectra were normalized to values between 0 and 1 using eqn. (2).

$$\text{Normalized } \%T = (\%T \text{ at sample point} - \text{min. } \%T)/(\text{max. } \%T - \text{min. } \%T) \qquad (2)$$

The number of samples was determined solely by the number of neurons in the input layer. In each case the whole truncated spectrum was sampled. No interpolation was performed between two neighbouring digital values. We took the first point beyond the required μm or cm^{-1} value. Since the original digital spectra are densely sampled, this procedure was satisfactory for our purposes.

For a given training-testing session, all 48 spectra were processed in an identical manner. Therefore, each input neuron represented absorption of energy at a specific wavelength regardless of which compound's spectrum was input into the network.

* California Scientific Software, 160 E. Montecito #E, Sierra Madre, CA 91024, USA.

† See, *e.g.*, ref. 1(*c*), ch. 3.

Table 1 Results of testing a trained network

	Test compound																	
	Isopropyl-benzene		Heptane		Propanoic acid		Ethyl benzoate		Ethyl ethanoate		Cyclo-hexanone		Phenyl-propan-2-one		Butan-1-ol		1-Phenyl-ethanol	
Functional group or bond	Correct answer	Result	Correct answer	Result	Correct answer	Result	Correct answer	Result	Correct answer	Result	Correct answer	Result	Correct answer	Result	Correct answer	Result	Correct answer	Result
Hydrocarbon	1.000	0.836	1.000	0.912	0.000	0.000	0.000	0.011	0.000	0.024	0.000	0.031	0.000	0.052	0.000	0.069	0.000	0.042
Carboxylic acid	0.000	0.000	0.000	0.000	1.000	0.951	0.000	0.000	0.000	0.007	0.000	0.051	0.000	0.000	0.000	0.031	0.000	0.001
Ester	0.000	0.015	0.000	0.000	0.000	0.002	1.000	0.994	1.000	0.979	0.000	0.078	0.000	0.177	0.000	0.008	0.000	0.027
Ketone	0.000	0.015	0.000	0.098	0.000	0.017	0.000	0.008	0.000	0.030	1.000	0.911	1.000	0.702	0.000	0.001	0.000	0.001
Alcohol	0.000	0.052	0.000	0.006	0.000	0.008	0.000	0.007	0.000	0.011	0.000	0.003	0.000	0.003	1.000	0.929	1.000	0.952
Phenyl	1.000	0.999	0.000	0.032	0.000	0.000	1.000	0.742	0.000	0.059	0.000	0.007	1.000	0.997	0.000	0.020	1.000	0.967
Oxygen–hydrogen bond	0.000	0.056	0.000	0.074	1.000	0.966	0.000	0.048	0.000	0.040	0.000	0.064	0.000	0.030	1.000	0.980	1.000	0.963
Carbon–oxygen bond	0.000	0.054	0.000	0.004	1.000	0.974	1.000	0.990	1.000	0.953	0.000	0.079	0.000	0.099	1.000	0.958	1.000	0.977
Carbonyl group	0.000	0.007	0.000	0.044	1.000	0.999	1.000	0.986	1.000	0.975	1.000	0.971	1.000	0.860	0.000	0.006	0.000	0.004

Table 2 Names and structures of compounds used in this study

	Functional group and bond type coding[a,b]								
	1	2	3	4	5	6	7	8	9
Hydrocarbons									
Cyclohexane, C_6H_{12}	1	0	0	0	0	0	0	0	0
Cyclopentane, C_5H_{10}	1	0	0	0	0	0	0	0	0
Ethylbenzene, $C_6H_5CH_2CH_3$	1	0	0	0	0	1	0	0	0
Heptane, $CH_3(CH_2)_5CH_3$	1	0	0	0	0	0	0	0	0
Hexane, $CH_3(CH_2)_4CH_3$	1	0	0	0	0	0	0	0	0
Isopropylbenzene, $C_6H_5CH(CH_3)_2$	1	0	0	0	0	1	0	0	0
Octane, $CH_3(CH_2)_6CH_3$	1	0	0	0	0	0	0	0	0
Alkane-mixture $CH_3(CH_2)_xCH_3$	1	0	0	0	0	0	0	0	0
Polystyrene, $-[CH_2CH(C_6H_5)]_x-$	1	0	0	0	0	1	0	0	0
Toluene, $C_6H_5CH_3$	1	0	0	0	0	1	0	0	0
Ketones									
Acetophenone, $C_6H_5COCH_3$	0	0	0	1	0	1	0	0	1
Acetone, CH_3COCH_3	0	0	0	1	0	0	0	0	1
Butan-2-one, $CH_3COCH_2CH_3$	0	0	0	1	0	0	0	0	1
1-Phenylbutan-1-one, $C_6H_5COCH_2CH_2CH_3$	0	0	0	1	0	1	0	0	1
Cyclohexanone, $C_6H_{10}O$	0	0	0	1	0	0	0	0	1
Cyclopentanone, C_5H_8O	0	0	0	1	0	0	0	0	1
Heptan-2-one, $CH_3CO(CH_2)_4CH_3$	0	0	0	1	0	0	0	0	1
Octan-2-one, $CH_3CO(CH_2)_5CH_3$	0	0	0	1	0	0	0	0	1
Phenylpropan-2-one, $C_6H_5CH_2COCH_3$	0	0	0	1	0	1	0	0	1
1-Phenylpropan-1-one, $C_6H_5COCH_2CH_3$	0	0	0	1	0	1	0	0	1
Carboxylic acids									
Ethanoic acid, CH_3CO_2H	0	1	0	0	0	0	1	1	1
Heptanoic acid, $CH_3(CH_2)_5CO_2H$	0	1	0	0	0	0	1	1	1
Hexanoic acid, $CH_3(CH_2)_4CO_2H$	0	1	0	0	0	0	1	1	1
2-Methylpropanoic acid, $(CH_3)_2CHCO_2H$	0	1	0	0	0	0	1	1	1
Octanoic acid, $CH_3(CH_2)_6CO_2H$	0	1	0	0	0	0	1	1	1
Oleic acid, $CH_3(CH_2)_7CH{=}CH(CH_2)_7CO_2H$	0	1	0	0	0	0	1	1	1
Pentanoic acid, $CH_3(CH_2)_3CO_2H$	0	1	0	0	0	0	1	1	1
Propanoic acid, $CH_3CH_2CO_2H$	0	1	0	0	0	0	1	1	1
Esters									
Benzyl benzoate, $C_6H_5CO_2CH_2C_6H_5$	0	0	1	0	0	1	0	1	1
Butyl ethanoate, $CH_3CO_2CH_2CH_2CH_2CH_3$	0	0	1	0	0	0	0	1	1
Ethyl ethanoate, $CH_3CO_2CH_2CH_3$	0	0	1	0	0	0	0	1	1
Ethyl benzoate, $C_6H_5CO_2CH_2CH_3$	0	0	1	0	0	1	0	1	1
Ethyl propanoate, $CH_3CH_2CO_2CH_2CH_3$	0	0	1	0	0	0	0	1	1
3-Methylbutyl ethanoate, $CH_3CO_2CH_2CH_2CH(CH_3)_2$	0	0	1	0	0	0	0	1	1
Isopropyl ethanoate, $CH_3CO_2CH(CH_3)_2$	0	0	1	0	0	0	0	1	1
Methyl benzoate, $C_6H_5CO_2CH_3$	0	0	1	0	0	1	0	1	1
2-Phenylethyl ethanoate, $CH_3CO_2CH_2CH_2C_6H_5$	0	0	1	0	0	1	0	1	1
Propyl ethanoate, $CH_3CO_2CH_2CH_2CH_3$	0	0	1	0	0	0	0	1	1
Alcohols									
Benzyl alcohol, $C_6H_5CH_2OH$	0	0	0	0	1	1	1	1	0
Butan-1-ol, $CH_3CH_2CH_2CH_2OH$	0	0	0	0	1	0	1	1	0
Butan-2-ol, $CH_3CH_2CH(OH)CH_3$	0	0	0	0	1	0	1	1	0
Cyclohexanol, $C_6H_{11}OH$	0	0	0	0	1	0	1	1	0
Hexan-1-ol, $CH_3(CH_2)_4CH_2OH$	0	0	0	0	1	0	1	1	0
3-Methylbutan-1-ol, $(CH_3)_2CHCH_2CH_2OH$	0	0	0	0	1	0	1	1	0
Isobutyl alcohol, $(CH_3)_2CHCH_2OH$	0	0	0	0	1	0	1	1	0
1-Phenylethanol, $C_6H_5CH(OH)CH_3$	0	0	0	0	1	1	1	1	0
2-Phenoxyethanol, $C_6H_5OCH_2CH_2OH$	0	0	0	0	1	1	1	1	0
3-Phenylpropan-1-ol, $C_6H_5CH_2CH_2CH_2OH$	0	0	0	0	1	1	1	1	0

[a] A value of 1 means that the functional group or bond type is present; a value of 0 means that it is absent. [b] Coding for function groups and bond types: 1, hydrocarbon; 2, carboxylic acid; 3, ester; 4, ketone; 5, alcohol; 6, phenyl; 7, O–H bond; 8, C–O bond; and 9, C=O bond.

Coding for the Functional Groups. The Output Neurons.—If a functional group was present in a compound, its code was 1 and, if it was absent, its code was 0. Each compound was assigned a coded answer that reflected its structure. This coded answer was used for training and for evaluation of the results of the testing. The functional group codes were for hydrocarbon, carboxylic acid, ester, ketone, alcohol and phenyl. The bond-type codes were for the O–H, C–O and C=O bonds. These bond-type codes allowed all the functional groups, except phenyl, to be double coded, *i.e.*, each functional group was independently represented by a unique output pattern in the first five as well as in the last three output neurons. Table 2

Table 3 % Correct identification as a function of data presentation

Percent transmittance				Absorbance			
μm		cm^{-1}		μm		cm^{-1}	
Number of points[b]	% Correct	Number of points[c]	% Correct	Number of points	% Correct	Number of points	% Correct
13	60	14	47	13	54	14	33
26	84	27	49	26	53	27	31
51	83	51	86	51	74	51	76
126	83	129	89	126	69	129	74
251	83	278	91	251	54	278	62
462	96	417	87	462	67	417	60

[a] The % correct is the average of nine separate training sessions starting with a different initial randomized network. The value was determined using a 0.2 tolerance for each test compound. [b] The complete spectrum was recorded from 2.5–15 μm: 13 points gave a separation of 1.0 μm between data samples; 26 points, 0.5 μm; 51 points, 0.25 μm; 126 points, 0.10 μm; 251 points, 0.05 μm; and 462 points, 0.025 μm. [c] The complete spectrum was recorded from 4000–666 cm^{-1}: 14 points gave a separation of 256 cm^{-1} between data samples; 27 points, 128 cm^{-1}; 51 points, 66 cm^{-1}; 129 points, 26 cm^{-1}, 278 points, 12 cm^{-1}; and 417 points, 8 cm^{-1}.

shows the functional group and bond-type codes used for each compound studied.

A Typical Training and Testing Session.—A typical set of compounds used to train a network and their input order in the training set (chosen by a random selection) are listed in the Appendix. Scrambling the order of the compounds so that compounds with the same functional group were not presented sequentially to the network significantly reduced the number of passes needed to train a network. With this set of 39 training compounds, the network required 68 passes to be trained to a tolerance of 0.1.

None of the nine compounds (isopropylbenzene, heptane, propanoic acid, ethyl benzoate, ethyl ethanoate, cyclohexanone, phenylpropan-2-one, butan-1-ol, 1-phenylethanol) used to test the network belonged to the training set. The test results are given in Table 1.

Results and Discussion

Two different sessions of training and testing were carried out. The first was aimed at determining the best data representation for this classification problem. The second was aimed at determining the accuracy of a trained network.

Effect of Representation of Spectral Information on Network Performance.—This portion of the study attempted to answer two interrelated questions: (*a*) how many data points are needed to train a network to correctly identify functional groups in related spectra; and (*b*) what is the best format for the presentation of the spectral data to the network?

The compounds used for these runs are the same as those used to describe a typical training–testing session (Table 1 and Appendix). In this portion of the study, the compounds were not changed from one training–testing session to another. In addition, the order in which the training spectra were entered into the networks was held constant. This procedure produced identically trained networks when all other variables were held constant and the network was initialized using the same seed for the random number generator.

The parameters that were varied were: (*a*) the number of data entry points; (*b*) the use of wavelength or frequency units in the sampling procedure (μm or cm^{-1}); and (*c*) the units of absorption (%*T* or *A*). Nine separate training–testing sessions were carried out with each set of variables using different randomized initial states and the results were averaged. This procedure minimized the effect of the initial random state of the network. The results in this portion of the study can be quantitatively compared.

Each input neuron represents absorption of energy at a specific wavelength (or frequency). Therefore, the amount of spectral detail presented to the network is a function of the number of input neurons used. In addition, each input neuron can be used to sample the absorption of energy at a specific wavelength (μm) or wavenumber (cm^{-1}). The difference between the two scales determines which portion of the spectrum is emphasized. When equidistant wavenumber values are used, the functional group region is emphasized; when equidistant micrometer values are used, the fingerprint region is emphasized. The results of the variation of these parameters is given in Table 3. The use of wavenumber-spaced data gives a slightly higher percentage of correct functional group and bond type identifications than that of micrometer-spaced data but the difference is not great. Either type of input can be used if a sufficient number of data points are presented to the network.

The importance of the number of data points can be seen in the spectra of benzyl alcohol given in Fig. 2. It is apparent that over 100 points (input neurons) are needed to adequately represent the details of the spectrum. However, merely increasing the number of input neurons is not necessarily beneficial. Oversized networks tend to be poor at generalization and for our network with 18 hidden neurons, every additional input neuron increases the size of the network by eighteen connections (weights). The effect of oversizing the network can be seen in Table 3, where, in some cases, the number of correct classifications passes through a maximum as the number of data points is increased.

Absorption of energy can be expressed in units of absorbance (*A*) or percent transmittance (%*T*). The results of a study of these two variables are also summarized in Table 3. Percent transmittance data give better results than do absorbance data. It is not clear why this is the case. A possibility is that the use of absorbance units suppresses the medium and weak absorption bands that play an important role in the global recognition of functional groups by the network.

The number of passes through the training data needed to train the network is also a function of the number of input values. In general, the greater the number of points, the fewer the number of passes required to train a network to a tolerance of 0.1 (Table 4).

Accuracy of the Trained Neural Networks.—As before, in this portion of the study nine compounds from the set of 48 were selected for testing and the remaining 39 were used for training

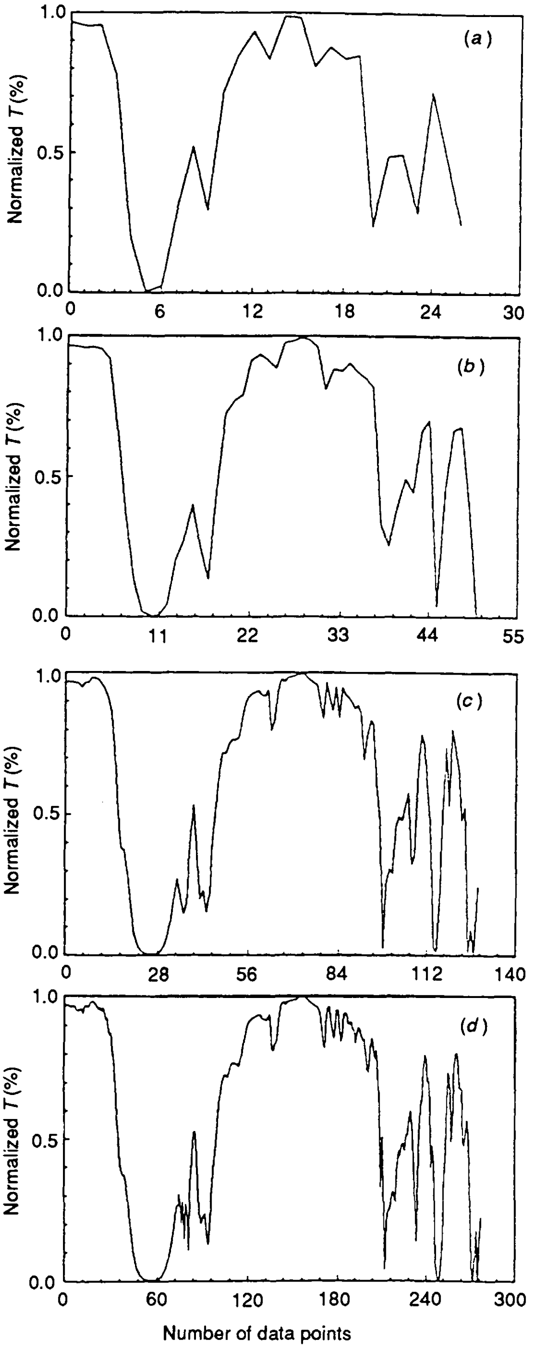

Fig. 2 IR spectra of benzyl alcohol, $C_6H_5CH_2OH$ from 4000–666 cm^{-1}. Spectrum (*a*), 27 data points separated by 128 cm^{-1}; spectrum (*b*), 51 data points separated by 66 cm^{-1}; spectrum (*c*), 129 data points separated by 26 cm^{-1} and spectrum (*d*), the first 240 data points of 278 selected points separated by 12 cm^{-1}.

the network. The test compounds were selected randomly from a group of candidates which fulfilled the following two criteria: (*a*) only two compounds from each functional group class could be selected (except for carboxylic acids, where only one compound could be selected); and (*b*) one of the two selected compounds must have a phenyl group. Fifteen separate selections of 9 testing and 39 training compounds were made. Each selection was used to train three separate networks starting at different initial random settings (45 different training–testing sessions; 405 separate tests).

The input data for each spectrum consisted of 250 points separated by 0.05 μm. The absorption units were normalized %T values. The network was trained to a tolerance of 0.1 and then the network was tested with the test compounds. A testing tolerance of 0.4 was used. To be classified as good, the output results from a test spectrum had to be within 0.4 of the correct values for all output neurons. If at least one output neuron had a value between 0.4 and 0.6 the result was classified as uncertain. A typical group of compounds selected for training and testing are listed in Table 1 and the Appendix.

First the network was trained to recognize only the six functional groups: hydrocarbon, alcohol, carboxylic acid, ketone, ester and phenyl. With this level of training, 25 test spectra of 405 were classified as bad (6.2%) and 28 as uncertain (6.9%). This number gives a correct identification rate of 86.9%.

Subsequently each compound was also coded with the three bond types. This double coding allowed all of the functional groups except phenyl to be identified by a set of rules. A BASIC program was written to evaluate the output results from each trained network. The rules used in the program are listed in the Appendix.

With this rule-based evaluation of the neural network output, good, bad and uncertain classifications are defined as follows. The classification is good when the network selects both the correct functional group and bond types. The classification is bad when the network selects a functional group and bond type that are in agreement with each other, but are incorrect. The classification is uncertain when the selected functional group is not in agreement with the selected bond type. An example of an uncertain classification is the case in which the network selects an ester and an O–H bond. The accuracy of the network with rule-based evaluation is: good, 90.3%; bad, 3% and uncertain, 6.7%.

With the rule-based evaluation procedure, we observed nine bad classifications out of 405 spectra tested. Six of these were due to two compounds, toluene and benzyl benzoate. Of 27 uncertain classifications ten were also due to benzyl benzoate. When only six functional groups were used as output neurons, 14 of the 53 bad or uncertain classifications were also due to benzyl benzoate. The trained networks tended to misclassify this ester as a ketone, a misclassification that implies the network could not recognize the C–O bond of the ester. The position of the C–O bond in the spectrum of benzyl benzoate is 1274 cm^{-1}, the same location as for methyl and ethyl benzoate (1276 cm^{-1}), but a considerably different location from those of the other esters (1180 to 1240 cm^{-1}). The carbonyl group in the spectrum of benzyl benzoate absorbs at 1720 cm^{-1}, considerably different than the carbonyl absorption of the other benzoates (1602 cm^{-1}) but similar to that of the other esters (1742 cm^{-1}). Why the network has difficulty with benzyl benzoate is not entirely understood, but must in some way be related to the manner in which a network globally interprets the combination of the C–O and C=O bands.

Conclusions

All training–testing sessions carried out in this study show that a neural network can be trained to recognize several different functional groups and bond types by simply being presented with spectra of the compounds. No specific identification of the bands in the spectrum need be given to the network. Once a network has been trained, it is able to recognize these functional groups and bond types in related spectra. Either micrometer or wavenumber data can be used to train neural networks for functional group identification, but percent transmittance data are superior to absorbance data.

Comparison of our results with those of Robb and Munk[11] indicate that a multilayer network with nonlinear activation function can achieve better classification than a simple one-layer, linear architecture (50% with a linear, one layer *vs.* 90%

Table 4 Average number of passes needed to train a network to a tolerance of 0.1 as a function of the number of data entry points

Micrometers			Wavenumbers		
Number of points	Passes required[a] %T	Passes required[a] A	Number of points	Passes required[a] %T	Passes required[a] A
13	360	1553	14	1114	3773
26	122	331	27	277	705
51	63	197	51	81	173
126	52	46	129	59	64
251	49	38	278	78	50
462	70	30	417	90	41

[a] The number of passes required are the average of nine separate training sessions with different initial randomized networks.

with the nonlinear, multilayer). We also found that normalization of the absorption and equidistant wavelength (or frequency) sampling was all that was needed for the processing of a spectrum. There was no need to locate the major absorption bands.

One crucial aspect of the performance of a neural network used to solve classification problems like this is the percentage of bad classifications. When the network produces an uncertain classification, the problem can still be shown to a human expert or presented again to the network in a modified form. There is no way, however, to detect a misclassification in a real-world situation. We attempted to address this problem and found that double coding of the molecular structure in the output layer followed by a rule-based evaluation procedure is superior to a simple functional group identification.

Our results indicate that neural networks are potentially useful for building structure elucidation systems and further studies are underway to test the limits of this approach.

Appendix

The Compounds Used for Training a Network.—

(1) Benzyl alcohol
(2) Butan-2-one
(3) Butyl ethanoate
(4) Cyclopentane
(5) 2-Methylpropanoic acid
(6) Butan-2-ol
(7) 1-Phenylbutan-1-one
(8) Methyl benzoate
(9) Cyclohexane
(10) Hexan-1-ol
(11) Acetophenone
(12) Alkane mixture
(13) Oleic acid
(14) 3-Methylbutan-1-ol
(15) 2-Heptanone
(16) Ethyl propanoate
(17) Ethylbenzene
(18) Cyclohexanol
(19) Propyl ethanoate
(20) Ethanoic acid
(21) Cyclopentanone
(22) Polystyrene
(23) Octanoic acid
(24) Isobutyl alcohol
(25) Octan-2-one
(26) 2-Phenylethanol
(27) Octane
(28) Pentanoic acid
(29) 3-Phenylpropan-2-ol
(30) Acetone
(31) 3-Methylbutyl ethanoate
(32) Toluene
(33) Hexanoic acid
(34) 2-Phenoxyethanol
(35) 1-Phenylpropan-1-one
(36) Isopropyl ethanoate
(37) 1-Phenylethanol
(38) Benzyl benzoate
(39) Hexane

Criteria for Classification of Functional Groups Using Neural Network Output Results When Nine Output Neurons are Used.—Notation (see Tables 1 and 2). Result (R_i) is the output result from the neural network, $0 < R_i < 1$.

R_1 = hydrocarbon
R_2 = carboxylic acid
R_3 = ester
R_4 = ketone
R_5 = alcohol
R_6 = phenyl
R_7 = O–H bond
R_8 = C–O bond
R_9 = C=O bond

Rules in order of execution.

1 For all output results:
If $R_i \leqslant 0.4$ set $R_i = 0$
If $0.4 < R_i < 0.6$ set $R_i = 0.5$
If $R_i \geqslant 0.6$ set $R_i = 1.0$

2 If all R_i for the set ($R_1, R_2, \ldots, R_5$) are equal to 0 then the compound is classified as uncertain.

3 If two R_i from the set ($R_1, R_2, \ldots, R_5$) are equal to 1.0, then the functional group classification of the compound is uncertain.

4 If $R_6 = 0.5$, the functional group classification of the compound is uncertain.

5 If the functional group result ($R_1, R_2, \ldots, R_5$) and the bond-type result (R_7, R_8, R_9) both contain an 0.5 output, the functional group classification of the compound is uncertain.

6 If there is not 0.5 output for the bond type (R_7, R_8, R_9) and these three outputs do not code a valid functional group, then the functional group classification of the compound is uncertain.

7 If there is no 0.5 output in the set ($R_1, R_2, \ldots, R_9$) then the functional group of the compound can be classified. (At this point the classification was judged to be good or bad by comparing it to the expected answer.)

8 If there is 0.5 output among the first five functional group answers ($R_1, R_2, \ldots, R_5$), then use the bond type information to decide if it is 0 or 1. Replace it and restart from Rule 2.

9 If there is 0.5 output among the bond type answers (R_7, R_8, R_9), then use the functional group information ($R_1, R_2, \ldots, R_5$) to decide if it is 0 or 1. Replace it and restart from Rule 2.

Acknowledgements

Very useful discussions with Alden Wright, Department of Computer Science, University of Montana and Roger Barga, Battelle, Pacific Northwest Laboratories, Richland Washington, are gratefully acknowledged. We also wish to thank Dennis Schaff of the Perkin-Elmer Corporation for the BASIC program that allowed us to convert the digitized FTIR spectra to ASCII format.

References

1 (*a*) *Neurocomputing. Foundations of Research*, eds. J. A. Anderson and E. Rosenfeld, MIT Press, Cambridge, MA, 1988; (*b*) D. E. Rumelhart and J. L. McClelland, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, I & II, MIT Press, Cambridge, MA, 1986; (*c*) P. D. Wasserman, *Neural Computing. Theory and Practice*, Van Nostrand Reinhold, New York, 1989; (*d*) R. Hecht-Nielsen, *Neurocomputing*, Addison-Wesley, Reading, MA, 1990.
2 L. A. Gribov, *Anal. Chim. Acta*, 1980, **122**, 249.
3 H. Abe, T. Yamasaki, I. Fujiwara and S. Sasaki, *Anal. Chim. Acta*, 1981, **133**, 499.
4 R. E. Carhart, D. H. Smith, H. Brown and C. Djerassi, *J. Am. Chem. Soc.*, 1975, **97**, 5755.
5 N. A. B. Gray, *Anal. Chem.*, 1975, **47**, 2426.
6 H. B. Woodruff and M. E. Munk, *J. Org. Chem.*, 1977, **42**, 1761.
7 B. Debska, J. Duliban, B. Guzowska-Swider and Z. Hippe, *Z. Anal. Chim. Acta*, 1981, **133**, 303.
8 M. E. Munk, M. Farkas, A. H. Lipkis and B. D. Christie, *Mikrochim. Acta*, 1986, **2**, 199.
9 H. Huixiao and X. Xinquan, *J. Chem. Inf. Comput.Sci.*, 1990, **30**, 203.
10 (*a*) B. Meyer, T. Hansen, D. Nute, P. Albersheim, A. Darvill, W. York and J. Sellers, *Science*, 1991, **251**, 542; (*b*) J. Thomsen and B. Meyer, *J. Magn. Reson.*, 1989, **84**, 212.

11 E. W. Robb and M. E. Munk, *Mikrochim. Acta*, 1990, **1**, 131.

12 M. Minsky and S. Papert, *Perceptrons*, MIT Press, Cambridge, MA, 1969.

13 (*a*) V. Venkatasubramanian and K. Chan, *AIChE J.*, 1989, **35**, 1993; (*b*) K. Watanabe, I. Matsuura, M. Abe, M. Kubota and D. M. Himmelblau, *AIChE J.*, 1989, **35**, 1803; (*c*) J. C. Hoskins and M. D. Himmelblau, *Comput. Chem. Eng.*, 1989, **12**, 881.

14 (*a*) H. Bohr, J. Bohr, S. Brunak, R. M. Cotteril, B. Lautrup, L. Noershov, O. H. Olsen and S. B. Petersen, *FEBS Lett.*, 1988, **241**, 223; (*b*) N. Qian and T. J. Sejnowski, *J. Mol. Biol.*, 1988, **202**, 865; (*c*) L. H. Holley and M. Karplus, *Proc. Natl. Acad. Sci. USA*, 1989, **86**, 152; (*d*) M. J. McGregor, T. P. Flores and M. J. E. Sternberg, *Protein Eng.*, 1989, **2**, 521.

15 A. V. Lusashin, A. I. Gragerov and M. D. Frank-Kamenetskii, *J. Biomol. Struct. Dyn.*, 1989, **6**, 1123.

16 D. W. Elrod, G. M. Maggiora and R. G. Trenary, *J. Chem. Inf. Comput. Sci.*, 1990, **30**, 477.

Paper 1/02316K
Received 16*th May* 1991
Accepted 8*th July* 1991