

A Neural Network Approach to Infrared Spectrum Interpretation*

Ernest W. Robb¹ and Morton E. Munk^{2,**}

¹ Department of Chemistry and Chemical Engineering, Stevens Institute of Technology, Hoboken, NJ 07030, USA

² Department of Chemistry, Arizona State University, Tempe, AZ 85287-1604, USA

Abstract. The simple linear neural network model was investigated as a method for automated interpretation of infrared spectra. The model was trained using a database of infrared spectra of organic compounds of known structure. The model was able to learn, without any prior input of spectrum-structure correlations, to recognize and identify 76 functional groupings with accuracies ranging from fair to excellent. The effect of network input parameters and of training set composition were studied, and several sources of spurious correlations were identified and corrected.

Key words: machine interpretation of spectra, infrared spectrum interpretation, computer assisted structure elucidation, neural networks.

Computer-assisted structure elucidation systems, as exemplified by the CASE system developed at Arizona State University [1], utilize spectroscopic data from a compound of unknown structure to derive a small set of plausible alternative structures consistent with those data. CASE, which is at an early stage of development, now relies chiefly on C-13 NMR spectral data, supplemented by information from 2-D NMR experiments. A two-track spectrum interpretation procedure produces a list of structural building blocks and constraints. Using this dual output, the structure generator builds one or more complete structures compatible with the spectral data.

A major emphasis in our current developmental work is an expansion of spectrum interpretation capabilities of the program. Enrichment of the structural information about an unknown compound serves to further limit the alternative structures produced by the program. With this objective in mind we turned our attention to infrared data. From the infrared spectrum of a compound, the presence or absence of significant structural fragments in the molecule can be deduced, a

* Dedicated to Professor W. Simon on the occasion of his 60th birthday

** To whom correspondence should be addressed

benzene moiety, for instance, or a methyl ester grouping. This information can then be used as a further constraint in the structure generating process.

Of the various forms of spectroscopy from which the organic chemist derives structural information, the infrared spectrum presents the greatest challenge for automated machine interpretation. In contrast to C-13 NMR, for example, where single carbon atoms may be identified from their one-to-one correspondence to spectral peaks, the structural moieties identifiable from the infrared spectrum can range in size from a bonded pair of atoms to a congeries of a dozen or more atoms. Furthermore, the presence or absence of an identifiable grouping is often based not upon the presence or absence of a single spectral band, but upon a complicated Boolean function based on information from several spectral regions.

Because of this challenge, and because of the potentially rich harvest of structural information available from infrared spectra, there has been no shortage of assaults on the problem of their machine interpretation. Many artificial intelligence methods have been tried, including simple search and matching procedures [2, 3]; the use of correlation tables [4–8] and correlation coefficients [9]; expert system approaches, both rule-driven [10–22] and table-driven [23]; the use of Bayesian statistics [24, 25] and other statistical approaches [26, 27]; set theory [28], including fuzzy sets [2, 29, 30]; and binary linear separation [31–36], KNN separation [25, 36, 37], principal component analysis [38], eigenvector projection [39], hierarchical cluster analysis [40–42], factor analysis [39], and other pattern recognition techniques [43–47]. Several reviews are available [30, 48, 49].

Some of these approaches have shown promise, leading to successful applications including a stand-alone infrared interpreter [14–16, 20, 21]. None, however, have demonstrated the accuracy and reliability which we desired for incorporation into the CASE structure elucidation system.

We were therefore led to investigate the use of neural networks, a novel computing structure which has very recently been used effectively in artificial intelligence applications as diverse as speech recognition, military target identification, robot control, image processing, and financial analysis. Recent chemical applications include the prediction of chemical reaction product distributions [50], of drug safety [51], and of protein secondary structures from amino acid sequence information [52–55].

We report here the results of a preliminary study of the use of a simple neural network model as a technique for the automatic recognition of functional groups in organic molecules from infrared spectral information.

Neural Networks: An Introduction

The term “neural network” is based on an analogy with the workings of the nervous systems of the higher animals, in which an electrical signal propagating along a nerve cell (neuron) is transmitted across the synaptic gap by neurotransmitter molecules to another neuron. These intersynaptic connections organize the neurons into networks which are capable of parallel distributed processing of information.

This model is readily simulated in the computer. Data is introduced to a layer of input units. These may be represented in the simulation by the elements of an array. Each unit is considered to be “connected” to the units of an adjoining layer.

Each interconnection is simulated by a transmission coefficient. The interconnection can be made strong, weak, non-existent, or inhibitory by giving the coefficient a large, small, zero, or negative value. A propagation rule combines the coefficients with the outputs from the units of each layer according to some algebraic formula to produce the inputs to the units of the next layer. This process continues until output from the final layer is produced. The overall result is the transformation of a pattern of input data into a pattern of output data.

Speech recognition is an example of the kind of problem amenable to this analysis. The input values in such an application would correspond to the duration and frequencies of the sound components present in a sample of speech; the output units would specify the phonemes present in the speech sample. In an application directed toward the machine interpretation of infrared spectra, the data to the input units should consist of some encoding of the infrared spectrum of a compound; the output data would be some representation of information about the presence or absence of functional groups in the compound.

A feature of simulated neural networks is that correct values of the coefficients required to transform a given input pattern to the desired output pattern need not be known in advance. Instead, the values can be developed by a process known as “training” or “learning”. A “training set” is required for which the output pattern of each element of the set is known. Training begins with arbitrary values for the transmission coefficients. The input pattern for an element of the training set is entered and the output pattern is calculated using an appropriate mathematical relationship. The calculated output pattern is compared to the “target output pattern” (the known output pattern for that element of the training set and the desired outcome of the learning process). A “learning rule” uses the difference between the calculated and target output patterns to improve the values of the transmission coefficients. As this process is repeated with the different input and target patterns, the coefficients converge to values which correctly transform each input pattern in the training set to the correct output pattern.

The attractiveness of this feature for our purpose is obvious. The correlations between infrared bands and structural features do not need to be specified in advance. By training the network on the infrared spectra of a collection of compounds in a training set, with the functional groups present in each compound specified in target patterns, the network should be able to learn the appropriate correlations.

What is termed the “simple linear model” for a simulated network was chosen for the initial study [56]. In this model, there are no intermediate layers of units—the input units are connected directly to the output units. (Intermediate layers of units between the input units and output units are usually referred to as “hidden units”.) The strength of the connection between an input unit i and an output unit j is given by a coefficient c_{ij} . Each input unit is connected to every output unit by such a coefficient. The contribution of an input unit to an output unit is simply the product of the input value x_i and the coefficient connecting the two units. The value y_j at an output unit is the sum of all the contributions reaching it from input units, without any further transformation. This gives the propagation rule

$$y_j = \sum_{i=1}^n x_i c_{ij}. \quad (1)$$

The "delta training rule" was used for coefficient improvement in this study [57]. For each output unit, the correction made to each coefficient connected to that unit is proportional to the difference between the calculated output value and the target value, $t_j - y_j$ and to that coefficient's contribution to the result, $x_i c_{ij}$. The correction formulas are thus

$$\delta c_{ij} = k(t_j - y_j)x_i c_{ij}, \quad (2)$$

$$c_{ij} \leftarrow c_{ij} + \delta c_{ij}. \quad (3)$$

It is useful to think of this in anthropomorphic terms—a coefficient which helps to give a correct answer is rewarded by being made larger, while a coefficient which contributes to a wrong answer is punished by being made smaller; in both cases the magnitude of the change being in proportion to the coefficient's importance in determining the result.

If the input pattern for the simple linear network model is thought of as a vector \mathbf{x} , then the output pattern is a vector \mathbf{y} which is a linear transformation of \mathbf{x} , and is simply the product of multiplication by the matrix of coefficients \mathbf{c} . It has been shown that the use of the delta learning rule with the simple linear model actually constitutes an iterative approach to a least squares solution; that is, the coefficients converge to values which minimize the sum of the squares of the differences between the output values and the target values [58], the summation being over all of the training set.

The simple linear network model has been shown to have some very fundamental limitations [56] and is not expected to give as good a result as more complex models having hidden units and non-linear propagation rules. We thought it more suitable for an initial study, however, because the relationships between input and output values are readily interpreted in terms of the numerical values of the coefficients; the effect on the results obtained of various network parameters could thus be readily determined.

Methods

The Training Set

For training the network coefficients, a set of organic compounds of known structure, along with their machine readable infrared spectra, was required. In the initial stages of the work, 2915 compounds whose spectra had been measured as capillary thin films were used. In the later stages, 3780 compounds whose spectra had been measured as Nujol mulls were added to the training set [59]. The characteristics of the full training set are summarized in the following statistics: Number of compounds—6695; average molecular weight—174.4; average molecular formula— $\text{C}_{8.82}\text{H}_{7.73}\text{Br}_{0.09}\text{Cl}_{0.26}\text{F}_{0.13}\text{I}_{0.02}\text{N}_{0.76}\text{O}_{1.65}\text{P}_{0.02}\text{S}_{0.11}$. Monofunctional compounds comprised 40% of the set, difunctional compounds 38%, compounds having three or more functional groups 11%, and compounds lacking a functional group 11%. The commonest functionalities were: alcohols 20.3%, amines 16.1%, ethers 12.3%, alkenes 11.7%, acid 11.3%, ketones 10.0%, esters 8.8%, amides 8.1%, nitro compounds 6.4%, nitriles 3.6%, and aldehydes 2.7%.

Input Units

The input vector for a compound must represent as economically as possible the essential information contained in the infrared spectrum of the compound. In the belief that peak positions and intensities are more informative than raw intensity data, the digitized spectra were converted into a list of peak positions/percent transmission pairs for each compound. For input to the network, these lists were converted into vectors. For the initial stages of this work, the spectral range between 400 and 4000 cm^{-1} was divided into 640 intervals of width 5.625 cm^{-1} , each of which was assigned to one input unit. The serial number of the input unit corresponding to a particular spectral frequency is given by $i = (\nu - 400)/5.625$ (Eq. (4)), rounded to the nearest integer. If a compound had no absorption band in an interval, the value for that input unit was set at zero. If there was a band within an interval, then an input value $0 < x_i \leq 1$ in proportion to the strength of the band was assigned according to the relationship

$$x_i = 1.00 - (\% \text{ transmission})/100.0. \quad (5)$$

The spectra available for the compounds in the training set had for the most part been scaled so that the most intense band in each spectrum had a percent transmission between 0% and 5%; no further adjustment was made to the band intensities to correct for path length or concentration differences.

One of the findings in this work was that too great a number of input units caused spurious results due to statistical anomalies (*vide infra*). Accordingly, the number of input units was reduced to 256 for the later stages of the work. At the same time, the widths of the spectral intervals were adjusted so that they were narrowest in the low frequency portion of the spectrum where the characteristic frequencies that require greater discrimination occur, and broadest at the high-frequency end of the spectrum where the anomalies had been noted. This was achieved by using the following formula to assign the frequency of a band to an input unit number

$$i = 6.0(\nu)^{1/2} - 120.0, \text{ rounded to the nearest integer.} \quad (6)$$

Output Units

The output vector and target vector for each compound should be representations of the functional groups present, or absent, in the compound. A survey of the compounds of the training set revealed some 400 functional groups which were potentially infrared active. Many of these were of infrequent occurrence and were therefore deleted, leaving 128 infrared-active functionalities which occurred in 25 or more of the compounds of the training set. Each of these was associated with an output unit by assigning to it a unique integer in the range $1 \cdots 128$. The list of the functionalities so assigned is given in Table 1. Some of these groupings were intentionally made very general (e.g., "C—O bond" or "aromatic"), some were less general and correspond to the traditional functional groups of the organic chemist, and some consisted of highly specific sub-functionalities (e.g., " $-\text{CH}_2-\text{CHOH}-\text{CH}_2-$ ", " $\text{Ar}-\text{O}-\text{CH}_3$ ").

Table 1

Infrared-active functionality	N^a	y_0^b	y_+^c	A_{50}^d
1. O—H	583	0.109	0.640	92.4
2. Alcohol	426	0.085	0.620	91.6
3. Primary alcohol	260	0.072	0.525	80.6
4. Ar—CH ₂ OH	35	0.018	0.181	24.7
5. Secondary alcohol	157	0.058	0.313	51.3
6. Methyl carbinol	53	0.029	0.213	39.4
7. CH ₃ —CH—CH ₂ —	38	0.023	0.182	34.1
8. —CH ₂ —CHOH—CH ₂ —	44	0.020	0.156	18.7
9. α -branched	25	0.013	0.096	8.0
10. Cyclohexanols	27	0.012	0.151	29.1
11. Tertiary alcohol	30	0.018	0.160	29.5
12. 1,2-Glycol	24	0.014	0.172	10.3
13. Phenol	71	0.027	0.361	62.1
14. N—H	392	0.086	0.572	89.3
15. Primary amine	250	0.062	0.552	84.4
16. Aliphatic	154	0.042	0.495	76.9
17. —CH ₂ —NH ₂	111	0.038	0.365	62.2
18. —CHR—NH ₂	40	0.016	0.215	32.6
19. Aromatic	95	0.033	0.568	88.2
20. Secondary amine	115	0.042	0.326	50.3
21. Aliphatic	97	0.036	0.351	58.9
22. —NHMe	21	0.012	0.216	52.8
23. C—N	564	0.106	0.598	94.5
24. Tertiary amine	138	0.031	0.586	82.6
25. Aliphatic	119	0.028	0.599	85.8
26. —N(Me)—	82	0.020	0.503	74.5
27. S—H	59	0.012	0.287	47.7
28. R—SH	43	0.010	0.236	35.8
29. sp hybridized C or N	209	0.027	0.633	96.5
30. Acetylenes	38	0.018	0.439	75.6
31. RC \equiv CH	13	0.008	0.292	60.9
32. R ₂ CX—C \equiv CH	17	0.009	0.286	47.0
33. Nitrile	125	0.016	0.557	90.6
34. R—CN	73	0.012	0.469	81.5
35. ArCN	20	0.004	0.513	87.0
36. =C=	42	0.004	0.914	100.0
37. Isocyanates	28	0.003	0.926	100.0
38. C=O	925	0.050	0.880	99.7
39. Acid halide	95	0.022	0.612	85.8
40. Acid chloride	88	0.020	0.629	86.9
41. R—COCl	48	0.010	0.683	83.7
42. Ar—COCl	27	0.017	0.299	48.9
43. Carboxylic acid	80	0.027	0.627	97.6
44. R—COOH	73	0.024	0.588	94.8
45. Satd., no α -halogen	45	0.019	0.551	83.6
46. —CO—O—	367	0.068	0.684	89.4
47. Esters	334	0.061	0.690	92.0

Table 1 (continued)

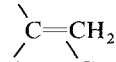
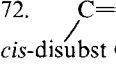
Infrared-active functionality		N^a	y_0^b	y_+^c	A_{50}^d
48.	Methyl esters	106	0.049	0.319	65.4
49.	Ethyl esters	160	0.057	0.445	72.5
50.	Esters, prim. alcohols	59	0.025	0.144	19.5
51.	Acetates	33	0.017	0.174	44.2
52.	Esters, satd. acids	140	0.038	0.447	59.6
53.	Esters, unsat. acids	33	0.016	0.182	41.0
54.	Esters, α -halo acids	25	0.010	0.083	8.1
55.	Esters, aromatic acids	36	0.022	0.185	39.3
56.	β -keto esters	15	0.017	0.125	23.7
57.	Aldehydes	101	0.042	0.366	73.9
58.	R—CHO	28	0.013	0.431	76.7
59.	Ar—CHO	54	0.029	0.289	59.6
60.	Ketones	274	0.077	0.447	67.7
61.	MeCO—, all types	106	0.045	0.287	58.6
62.	Dialkyl ketones	117	0.036	0.353	55.4
63.	No α -halogen	93	0.027	0.362	54.3
64.	Aryl and unsatd. ketones	96	0.033	0.353	50.5
65.	Conj. ketones, one C=C	73	0.025	0.341	48.6
66.	amides	49	0.023	0.173	22.7
67.	Tertiary amides	24	0.011	0.108	10.1
68.	C=C	374	0.112	0.411	61.4
69.	—CH=CH ₂	115	0.044	0.399	80.9
70.	No α —N, O, or halogen	90	0.039	0.45	88.9
71.		57	0.029	0.207	33.2
72.		24	0.016	0.203	51.0
73.	<i>cis</i> -disubst C=C	80	0.033	0.122	15.8
74.	<i>trans</i> -disubst C=C	86	0.040	0.240	37.3
75.	—CH=CH—CO—	30	0.016	0.176	34.7
76.	Trisubst. C=C	72	0.033	0.149	21.0
77.	Aromatic ^e	1207	0.139	0.789	98.9
78.	5 adjacent H	283	0.064	0.621	89.6
79.	4 adjacent H	260	0.088	0.392	60.7
80.	3 adjacent H	259	0.080	0.427	69.4
81.	2 adjacent H	396	0.100	0.547	82.3
82.	1 adjacent H	368	0.121	0.424	68.8
83.	Benzene	1012	0.149	0.741	95.4
84.	Phenyl (—C ₆ H ₅)	283	0.064	0.620	89.5
85.	<i>o</i> -disubst	204	0.070	0.375	60.0
86.	<i>m</i> -disubst	124	0.055	0.304	57.8
87.	<i>p</i> -disubst	200	0.070	0.452	69.7
88.	1,2,3-trisubst.	47	0.023	0.172	21.2
89.	1,2,4-trisubst.	125	0.053	0.290	48.9
90.	Naphthalene	26	0.014	0.192	33.5
91.	Furan	25	0.012	0.208	46.6
92.	Thiophene	31	0.012	0.157	26.5
93.	Pyridine	83	0.039	0.23	41.3

Table 1 (continued)

Infrared-active functionality	N^a	y_0^b	y_+^c	A_{50}^d
94. $-\text{NO}_2$	80	0.022	0.610	92.2
95. $\text{Ar}-\text{NO}_2$	64	0.020	0.642	95.0
96. $\text{C}-\text{O}$	1318	0.272	0.702	95.8
97. Ether	410	0.134	0.423	68.9
98. Dialkyl ether	134	0.059	0.299	44.8
99. $-\text{CH}_2-\text{O}-\text{CH}_3$	43	0.022	0.149	22.5
100. $-\text{CH}_2-\text{O}-\text{CH}_2-$	92	0.044	0.222	38.3
101. Aryl alkyl ether	178	0.071	0.405	68.2
102. $\text{Ar}-\text{O}-\text{CH}_3$	135	0.058	0.372	65.7
103. Ketal/acetal	74	0.037	0.220	52.6
104. Diethyl ketal	27	0.019	0.222	48.6
105. Epoxide	19	0.009	0.077	5.1
106. $\text{P}=\text{O}$	38	0.023	0.198	53.6
107. $\text{P}-\text{O}$	47	0.026	0.191	56.1
108. $\text{P}-\text{Cl}$	19	0.013	0.204	55.3
109. $\text{S}=\text{O}$	36	0.020	0.144	27.2
110. $\text{S}-\text{O}$	16	0.009	0.125	10.7
111. $\text{R}-\text{S}-\text{R}$	32	0.013	0.067	4.7
112. $\text{C}-\text{X}$	804	0.199	0.488	70.0
113. $\text{C}-\text{F}$	213	0.081	0.283	43.6
114. $-\text{CF}_3$	77	0.042	0.286	82.6
115. $\text{Ar}-\text{F}$	12	0.050	0.230	33.2
116. $\text{C}-\text{Cl}$	417	0.120	0.399	59.9
117. $-\text{CH}_2-\text{Cl}$	123	0.051	0.199	23.6
118. $\text{Ar}-\text{Cl}$	147	0.057	0.265	36.4
119. $-\text{CCl}_3$	24	0.010	0.106	8.2
120. $\text{C}-\text{Br}$	206	0.072	0.172	20.7
121. $-\text{CH}_2-\text{Br}$	88	0.034	0.216	34.0
122. $\text{Ar}-\text{Br}$	65	0.028	0.122	14.6
123. $\text{C}-\text{I}$	43	0.012	0.037	2.8
124. CMe_2	223	0.085	0.300	51.5
125. <i>t</i> -butyl ($-\text{CMe}_3$)	68	0.030	0.176	24.5
126. <i>gem</i> -diMethyl (R_2CMe_2)	61	0.028	0.096	11.9
127. Isopropyl ($-\text{CHMe}_2$)	47	0.040	0.172	23.6
128. $-(\text{H}_2)_6-$	167	0.058	0.403	65.7

^a Number of compounds in the initial training set (2915 compounds) having the functionality^b Mean output value for compounds lacking the functionality^c Mean output value for compounds having the functionality^d See text^e Includes substituted benzenes, pyridines, pyrazines, pyrimidines, etc., thiophenes, furans, pyrroles, oxazoles, thiazoles, pyrazoles, etc., and benzo-analogs

The target vector for a compound, then, consisted of a sequence of 128 real numbers corresponding to the serially numbered output units, each number either

having the value 1.0 if the functionality assigned to that output unit was present in the compound, or the value 0.0 if the functionality was absent from the compound.

The Training Program

The training program, written in standard Pascal, went through the compounds of the training set in order. For each compound, the list of infrared peak positions and intensities was converted to an input vector, and a target vector was constructed from a list of the compound's functionalities. From the input vector, an output vector was computed by Eq. (1). This output vector and the target vector were then used to compute new values for the appropriate coefficients according to Eqs. (2) and (3). A training session utilizing the full training set of 6695 compounds required 40 min on a Prime 450 minicomputer or 18 min on a VAX 11-785.

A value of 0.12 was used for the proportionality constant k in Eq. (2). When greater values were used, some evidence for oscillatory behavior was observed, while smaller values would have required more training sessions to reach convergence. An initial value of 0.10 was used for all of the coefficients prior to the first training session. Values obtained for the coefficients in a training session were stored and utilized in the subsequent training session.

The coefficients converged smoothly as successive training sessions were carried out. We found that 32 training sessions were sufficient for convergence; further sessions caused no significant change in the values of any coefficients. It is of interest that approximately 80% of the coefficients converged to zero.

The Test Set

A test set of organic compounds, separate from the training set, was necessary for rigorous evaluation of the accuracy of functional group identifications by the trained network. This test set consisted of 541 compounds chosen from the Sadtler collection [60]. The compounds chosen approximated those of the training set in molecular size and complexity; their average molecular weight was 171.9, the average molecular formula was $C_{9.03}H_{12.34}Br_{0.08}Cl_{0.15}F_{0.04}I_{0.02}N_{0.58}O_{1.50}S_{0.09}$ and the representation of different functional groups was similar to that in the training set. Some of the spectra of the test set compounds were measured as capillary films, others as Nujol mulls, and some as KBr pellets. These published spectra were digitized and the digitized spectra were converted to lists of peak positions and peak intensities.

The Test Program

The test program converted the list of infrared peak positions and peak intensities for each compound into an input vector. From the input vector and the set of coefficients generated by the final training session, an output vector was computed using Eq. (1). The elements of the output vector which corresponded to the 24 functionalities tested for (Table 2) were then compared with pre-set cutoff values, as described below, yielding for each functionality a prediction that the functional group was "definitely absent", "possibly present", "probably present", or "definitely present".

Table 2

Group	N^a	y_0^b	y_+^b	A_{50}^b	Cutoff output values			
					c	d	e	f
1. OH/NH bond	3611	0.263	0.679	98.4	0.167	0.433	0.500	0.633
2. Triple bond/cumulene	343	0.016	0.572	98.0	0.100	0.233	0.300	0.500
3. C=O	2706	0.141	0.813	97.6	0.100	0.433	0.567	0.700
4. C=C	781	0.096	0.198	32.3	0.033	0.433	0.500	0.633
5. Aromatic	3973	0.340	0.721	96.2	0.233	0.433	0.567	0.700
6. C—O single bond	3213	0.331	0.640	84.6	0.033	0.500	0.700	0.967
7. C—halogen bond	1531	0.131	0.234	42.4	0.000	0.433	0.500	0.567
8. Alcohol/phenol	1356	0.159	0.346	57.4	0.033	0.433	0.567	0.767
9. Amine	1033	0.128	0.381	57.0	0.033	0.433	0.700	0.900
10. Ammonium ($N^+—H$)	505	0.057	0.365	72.4	0.033	0.433	0.633	0.900
11. Acetylene	48	0.008	0.136	45.1	0.033	0.167	0.500	0.567
12. Nitrile	243	0.009	0.494	83.9	0.033	0.167	0.300	0.500
13. Cumulene	54	0.006	0.679	98.2	0.233	0.500	0.567	0.700
14. Acid halide/anhydride	148	0.012	0.498	86.4	0.100	0.367	0.567	0.700
15. Carboxylic acid	750	0.111	0.440	60.8	0.100	0.433	0.633	0.767
16. Ester	589	0.064	0.421	69.4	0.033	0.433	0.567	0.833
17. Aldehyde	179	0.020	0.074	9.6	0.033	0.167	0.300	0.367
18. Ketone	674	0.064	0.141	27.6	0.033	0.300	0.433	0.567
19. Amide	542	0.103	0.338	41.1	0.100	0.433	0.567	0.767
20. $—CH=CH_2$ (vinyl group)	132	0.011	0.115	41.5	0.033	0.167	0.233	0.300
21. Phenyl group	812	0.079	0.489	75.9	0.033	0.433	0.633	0.900
22. Carboxylate ion	211	0.034	0.164	36.8	0.033	0.300	0.367	0.567
23. Nitro group	427	0.058	0.387	76.6	0.033	0.433	0.500	0.567
24. Ether	822	0.112	0.269	45.3	0.033	0.433	0.500	0.567

^a Number of compounds in the augmented training set (6695 compounds) having the functionality

^b See footnotes, Table 1

^c Cutoff between "definitely absent" and the unreported category

^d Cutoff between the unreported category and "possibly present"

^e Cutoff between "possibly present" and "probably present"

^f Cutoff between "probably present" and "definitely present"

Results

Distribution of Output Values for Various Functional Groups

During the training of the network, the target values at an output unit corresponding to a functional group were set at 1.00 when the functional group was present and 0.00 when the group was absent. If, after training, the network was able to discriminate perfectly for the presence of a group, then the output values generated from a particular compound should be either 1.00 or 0.00 depending on whether the functional group in question was present or absent. In practice, one would not expect to achieve perfect discrimination; however, one would hope that the network could discriminate for the presence of a functional group at least to the extent that

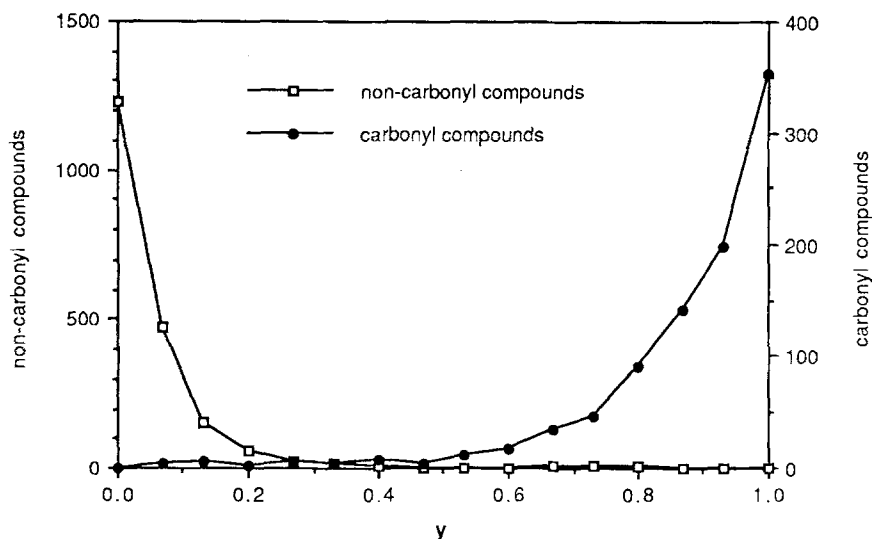


Fig. 1 a

Fig. 1. Distribution of output values y for training set compounds; **a** at output unit 38 ($\text{C}=\text{O}$), **b** at output unit 111 ($\text{R}-\text{S}-\text{R}$), **c** at output unit 3 ($-\text{CH}_2\text{OH}$)

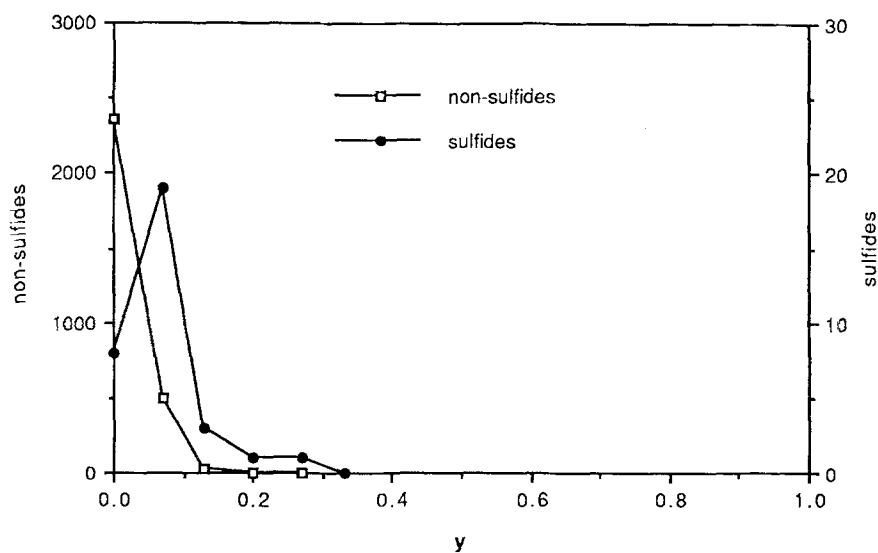


Fig. 1 b

the output value would be substantially greater when the group was present than when it was absent.

The discrimination actually achieved was measured by collecting, during the final training session, the output values at each of the 128 output units for each compound of the training set. These values were tabulated separately for compounds having the functional group and for those lacking it, and the tabulations were plotted as distributions. In Fig. 1, three distributions typifying the results obtained are shown.

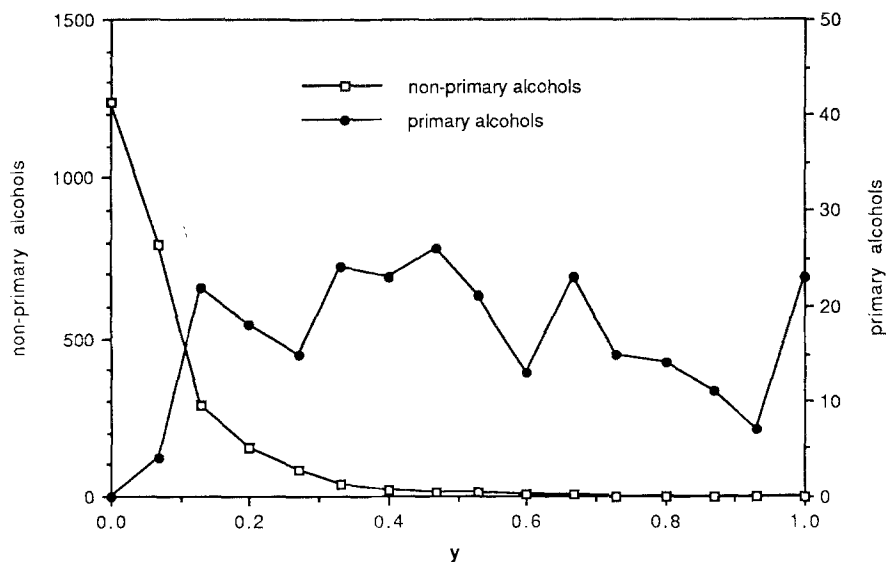


Fig. 1 c

Fig. 1 a shows the output values obtained at the output unit assigned to the carbonyl group. For compounds actually having a carbonyl group, many gave output values of 1.00 or greater, and nearly all gave output values greater than 0.500; the mean output value was 0.880. For training set compounds that lacked a carbonyl group, many gave output values of 0.000, and nearly all gave output values less than 0.500; the mean value was 0.050. There was almost no overlap between the two distributions. Thus the neural network model is able to discriminate almost perfectly for the presence of a carbonyl group. A number of other groupings showed perfect or almost perfect discrimination. These include isocyanates, aromatic compounds, carboxylic acids, compounds having C—O single bonds, compounds having benzene rings, and aromatic nitro compounds.

Fig. 1 b shows the results at the output unit assigned to the sulfide grouping, typifying those cases where there is almost no discrimination. Compounds containing the sulfide grouping gave output value at this unit ranging from 0.00 to 0.36, with a mean value of 0.067; while compounds lacking the sulfide grouping gave values in the range 0.000–0.180, with a mean value of 0.013. The two distributions overlap extensively, hence the output value for an individual compound gives essentially no information about the presence of absence of the sulfide grouping. This result is not surprising, since the non-polar C—S bond does not give rise to a strong characteristic band in the infrared spectrum. Other functional groupings lacking characteristic bands in the 4000–400 cm^{-1} range, e.g., the C—I and C—Br groupings, gave similar results.

The results at the output unit assigned to the primary alcohol grouping, $-\text{CH}_2-\text{OH}$, are shown in Fig. 1 c. These results are typical of those obtained for most of the functional groups surveyed, in that the network model was able to discriminate, but only imperfectly, between compounds having the grouping and those lacking it. Primary alcohols, for example, had a mean output score of 0.524 at the output unit for that group, compared with 0.071 for compounds without any

—CH₂OH group. However, the two distributions overlap for output values in the range 0.03–0.57, so that a score in this range for an individual compound is ambiguous.

Such a result could be applicable to the CASE program. A cutoff value of, say, 0.75 for the primary alcohol grouping could be set. Compounds scoring higher than this could be assumed to be primary alcohols with little chance of error, so that the requirement could be passed on to a structure generator as a positive constraint. Of course, many primary alcohols would score lower than 0.75 and would thus not be detected, in which case the program would have to rely solely on NMR data, as it presently does for structural information.

For other functional groups where there was imperfect discrimination, we found no overlap between the two distributions at the low end of the output scale. In these cases, one could set a very low cutoff value, and could say with very little chance of error that compounds scoring below the cutoff lacked the functional group. This information could be given to a structure generator as a negative constraint, requiring that the functional group be absent from all final structures. Again, many cases where the group was absent would be passed over, forcing reliance on other spectral data to exclude the group from the final structures generated.

A single numerical parameter, accuracy at 50% information retrieved, (A_{50}), was chosen to characterize these pairs of partially overlapping distributions in an easily interpreted manner. If the median output value for a group of compounds possessing a functional group is taken as the lowest acceptable value for a “functional-group-present” assignment, then only 50% of the valid information is retrieved; the other half is lost. If the distributions of output values for compounds with and without a functional group overlap, i.e., if output values for compounds without the functional group equal or exceed the median output value for compounds with the functional group, then the use of that median output value as the “cutoff” will result in false positive identifications, the number of which will depend on the extent of the overlap. The accuracy at 50% information retrieved is calculated as the percentage of correct identifications relative to all positive identifications including false ones. In the case of the primary alcohol grouping, the median cutoff score is 0.51 and the accuracy at this value is 80.6%. For comparison with Figs. 1 a and 1 b, the A_{50} value for the carbonyl group is 99.7% and for the sulfide group is 4.7%.

Table 1 gives the A_{50} values for all of the functional groups considered in the initial training regimen (640 input units, 2915 training compounds). Discrimination was excellent ($A_{50} > 90\%$) for 17 groups, good (A_{50} 75%–90%) for 22 groups, fair (A_{50} 50%–75%) for 37 groups, and poor ($A_{50} < 50\%$) for 53 groups. As expected, those functionalities which have characteristic bands unique to them show excellent discrimination. Examples include the nitriles, isocyanates, and nitro compounds. Groups which have strong characteristic bands in spectral regions common to other functionalities as well provided a better indication of the inherent discriminating power of the method. With this group of functionalities the results are mixed, e.g., alcohols, esters, and carboxylic acids show excellent discrimination, but only fair to good discrimination is observed for aldehydes, ketones, anides, and compounds containing the S=O linkage. It is interesting to note that this difference in discrimination correlates with the frequency of occurrence of the functional group in the training set. Groups which are well represented in the training set show good

discrimination, while groups which occur infrequently show only fair or poor discrimination (Table 1). A similar correlation is observed in comparing performance of general classes with that of subclasses within them, e.g., alcohols, amines and esters show better discrimination than tertiary alcohols, secondary amines, and methyl esters, respectively.

This observed correlation is a consequence, at least in part, of a training method in which all of the compounds of the training set are presented serially to the neural network. Consider, for example, the aldehyde grouping, one of whose characteristic infrared bands is in the region near 1725 cm^{-1} , and which only occurs 101 times in the initial training set. During a training session, the coefficient connecting the input unit for this frequency and the output unit for the aldehyde group will be strengthened 101 times. But for each instance in which the carbonyl band of one of the more numerous acids, esters, or ketones falls inside the aldehyde range, this connection will be weakened; and there are considerably more than 101 such instances in the training set. As a result, the final value for the coefficient connecting absorption at 1725 cm^{-1} with the aldehyde group will be small, leading to small output values for compounds containing the group. In general, functional groups with lower output values display less discrimination. Although an increase in the population of such functional groups may improve discrimination, it is also likely that with the simple linear model, discriminating ability will decline as the extent of overlap between the characteristic spectral regions of different functional groups increases. Decreasing discrimination can also be expected as structural similarities between groups increase, e.g., between primary, secondary, and tertiary alcohols.

It follows that other training regimens that do not involve the presentation of each individual training compound might result in better recognition of uncommon functionalities. We plan to pursue this avenue in future work. The effect of training set composition on performance has also been noted in infrared interpretation systems based on other pattern recognition techniques [31, 35].

The Basis for Functional Group Discrimination

The neural network establishes its own correlations between infrared absorption frequencies and functional groups during the training procedure. We wished to compare these correlations with those traditionally used by chemists and infrared spectroscopists in assigning structural features. For the simple network model used in this work, this could be done by examination of the final numerical values of the network coefficients for each functional group. The total contribution to a particular output unit j from each input unit i is given by

$$B_{ij} = \sum c_{ij}x_i,$$

where the summation is carried out over all of the occurrences in the training set of the functional group corresponding to j . Fig. 2 a, for example, shows the contributions from all of the input units to the output unit corresponding to the nitrile group. (In the figure, increasing values of B_{ij} are plotted downward so that the plot resembles an infrared spectrum.) It can be seen that the only contribution to this output unit is from the input units corresponding to absorption between 2220 cm^{-1} and 2256 cm^{-1} ; the coefficients connecting all of the other input units to the nitrile

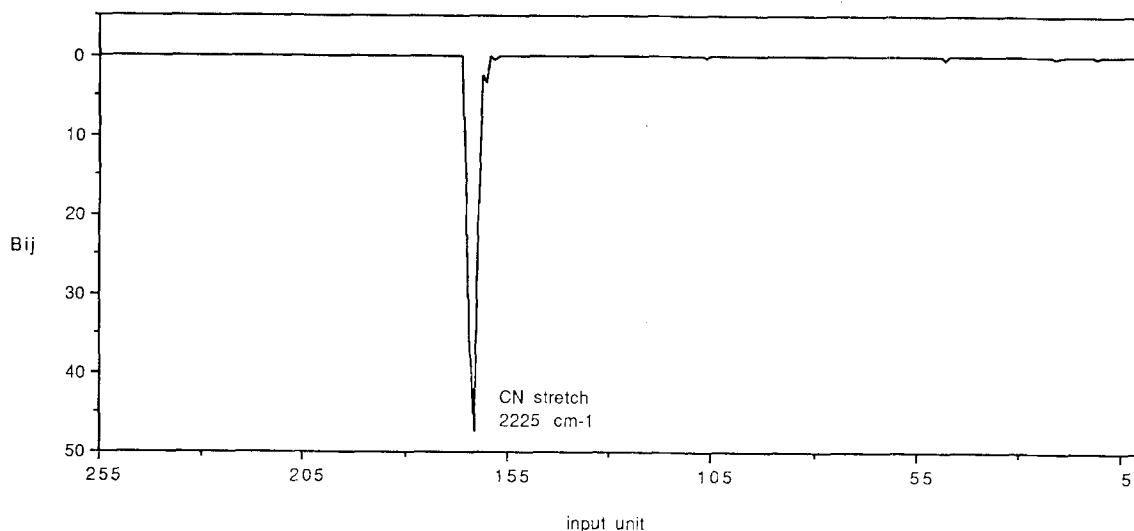
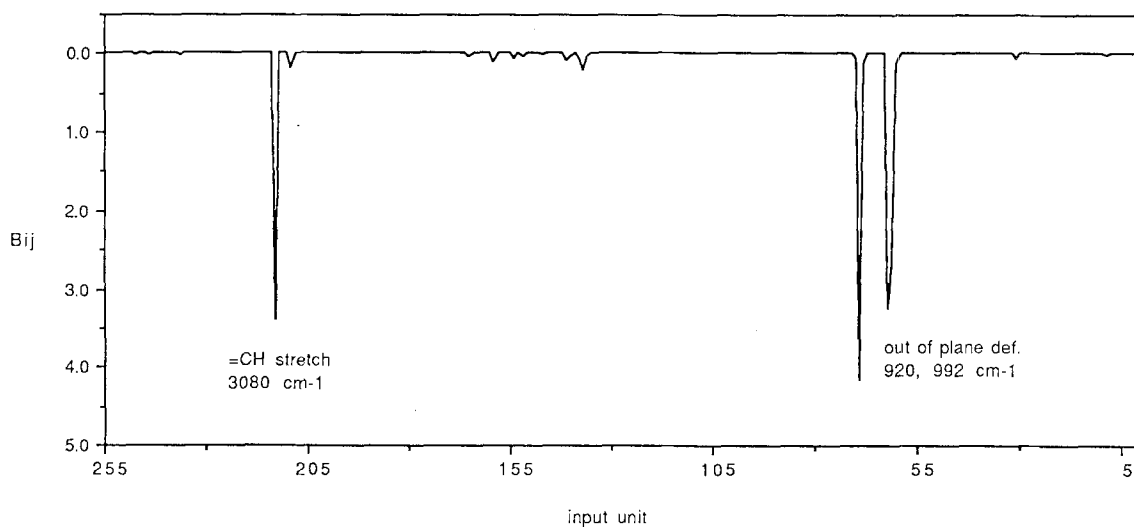
**Fig. 2 a**

Fig. 2. Contributions B_{ij} (see text) as a function of input unit i ; **a** for $j = 33$ (nitrile), **b** for $j = 69$ ($-\text{CH}=\text{CH}_2$), **c** for $j = 31$ ($\text{RC}\equiv\text{CH}$), **d** for $j = 24$ (tertiary amine), **e** for $j = 23$ ($\text{C}-\text{N}$ bond)

**Fig. 2 b**

output unit have a value of zero. The network thus uses the $\text{C}\equiv\text{N}$ stretching band to detect nitriles, ignoring all other infrared absorption bands that a compound may have in reaching this decision.

Similar plots of contribution as a function of the input units have been prepared for the other functional groups surveyed. These plots are of interest in their own right, since they provide new independent information on infrared structural correlations. In most cases, the network model bases its positive response on well-known characteristic group frequencies. For example, in Figs. 2 b and 2 c, the contributions to the output units corresponding to the vinyl group and to the terminal acetylene

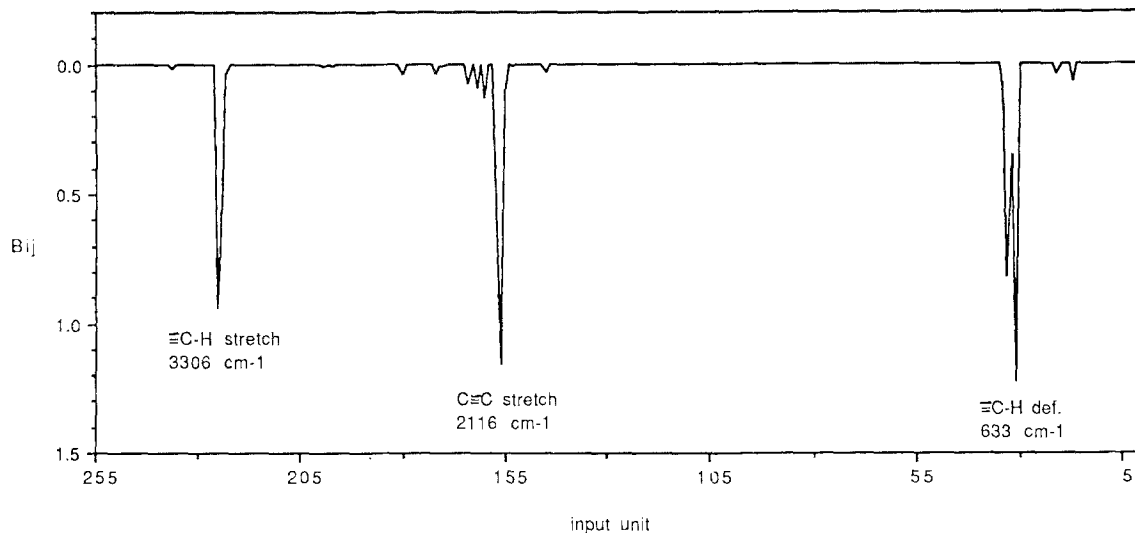


Fig. 2 c

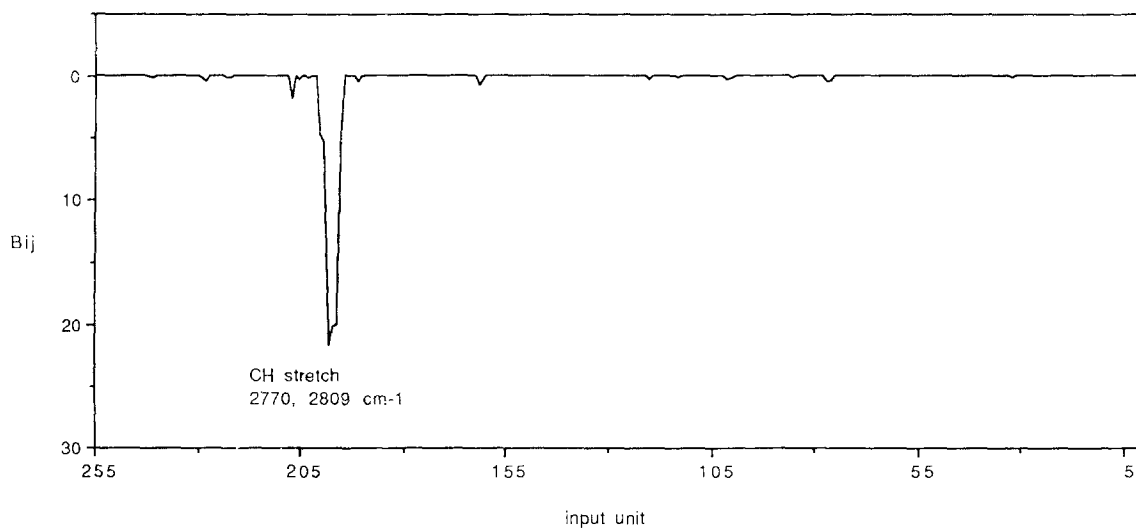


Fig. 2 d

groups are shown. The frequencies displayed are those familiar to chemists as diagnostic for these groups.

In a few cases, correlations not well known or widely used by spectroscopists are uncovered. For example, the network model bases its identification of ethers and amines in part on the position of the C-H stretching frequency, which occurs at an unusually low frequency for CH_3 and CH_2 groups adjacent to ether oxygen or amino nitrogen atoms. Fig. 2 d, which shows the contribution of input units to the output unit assigned to the tertiary amine group, illustrates this correlation.

The neural network model used was able to identify with high accuracy some very general structural features such as "aromatic" or " C-N bond", structural features for which specific characteristic infrared bands have not been recorded. The

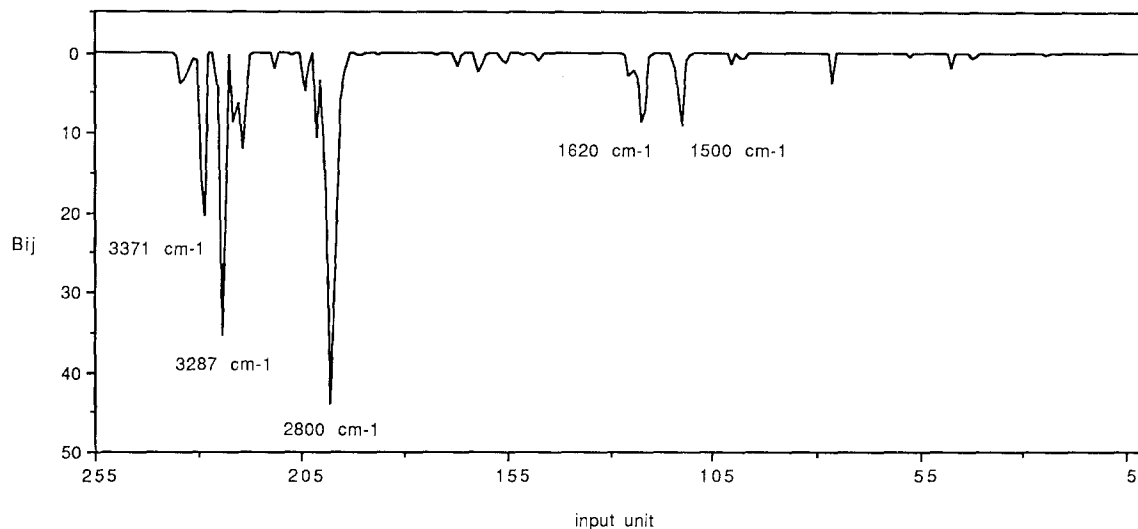


Fig. 2 e

input unit contributions to the “C—N bond” output unit shown in Fig. 2 e reveal that the network is in fact using stretching and deformation N—H absorptions to detect primary and secondary amides and primary and secondary amines. If all such compounds are given high output scores for the “C—N bond” group, then a fairly high percentage of compounds actually having a C—N bond will be identified as having that group. Analysis of other very general group designations verified that the ability of the network to detect them rests on its ability to carry out a Boolean OR operation on those specific sub-classes within the general category which do have distinctive characteristic absorptions of their own.

Interfering Groups

For many of the functional groups investigated, it was noted that there were a few compounds which gave persistent false positive identifications for the presence of the group. For example, there were 11 compounds in the training set which, although lacking a phenyl group, gave output values at the unit assigned to the phenyl group that were greater than 0.90 and that did not decrease with additional training sessions. Investigation showed that these compounds were all either 2-substituted thiophenes or sulfite esters. In either case, they had strong bands at 690 cm^{-1} which mimicked the out-of-plane bending band that the network was using to identify phenyl groups. There were also a number of cases where the absorptions characteristic of a group were absent, resulting in persistent false negative identifications. Although not directly pertinent to neural network performance, these interferences are of general interest, and are summarized in the *Appendix* to this paper for the 24 groups used in the final test of the network model.

Effect of the Order of Presentation of the Training Compounds

In each training session, the compounds of the training set were presented to the network in the same order. It was of interest to learn the effect that changing this

order would have on the trained network. A new version of the input files was prepared, in which the order of the compounds had been randomly scrambled. Thirty-two training sessions were carried out with the scrambled files. Very nearly the same distributions of output values were obtained for the functional groups. The largest difference between A_{50} values obtained for any group was less than 2.0%. Thus the order in which the compounds stand in the training set has only an insignificant effect on the results obtained.

Is Information on Line Widths Necessary?

In most prior work on automated interpretation of infrared spectra, information on the width of spectral bands, as well as on their position and intensity, has been considered necessary. The case most often cited in support of this has been that of the carboxylic acids, whose O—H stretching band is very broad. We noted that superimposed on this broad band are a number of weak subsidiary maxima, the most prominent of which comes near 2650 cm^{-1} . These weak maxima form a pattern which is unique to the carboxylic acids. The broadness of the underlying envelope, while a useful mnemonic for the human interpreter of spectra, is thus unnecessary for the machine recognition of the acids. We believe that the decision to exclude band-width information from the input, which greatly simplifies the representation of spectra to the neural network, is justified by the generally satisfactory results obtained.

Testing the Network—Spurious Correlations

The training procedure enables a neural network to map input patterns to the desired output patterns by utilizing any correlations present. Some of these correlations may not be valid, if they are based on extraneous characteristics of the training set that will not necessarily be found in the real-world input data of the network's intended application. The only way such spurious correlations can be found is by testing the trained network with an independent set of input data not used in its training. Such a test is therefore necessary in the development of any neural network.

In preliminary trials on the test set described above, the trained network did perform less well than the data on the training compounds had led us to expect. This was most notably the case with polar functional groups such as amide groups. These were under-represented in the original training set, which included only liquid and low-melting compounds whose spectra had been measured as capillary films. The training set was therefore augmented with an additional 3780 compounds whose spectra had been measured as Nujol mulls.

In addition, these preliminary tests revealed two types of spurious correlation which were of interest. In the first, the network was unable to distinguish between O—H and N—H containing compounds in the test set, even though it was apparently able to make this distinction for compounds in the training set. We found that this was a consequence of the very large number of input units (640) used, resulting in a very narrow spectral region (5.625 cm^{-1}) associated with each input unit. Both O—H groups and N—H groups have characteristic absorption in the same region of the spectrum, and are present in the training set in large and approximately equal numbers. However, only a few compounds will be found in each narrow input unit

interval and statistically it is likely that there will be a preponderance of one type. Thus, seven amines and two alcohols in the training set had absorption bands in the interval 3389.7 cm^{-1} – 3395.3 cm^{-1} and the network interpreted a band in this interval as evidence for an N—H group; while in the next interval, 3395.3 cm^{-1} – 3400.9 cm^{-1} , one amine and four alcohols absorbed, and the network interpreted bands in this interval as being due to O—H groups. These interpretations, having no basis in fact, lead to incorrect predictions on test set compounds. The problem was solved by decreasing the number of input units to 256, thereby increasing the width of the spectral interval associated with each unit. With more compounds absorbing in a given interval, such statistical imbalances were eliminated.

A second spurious correlation, noted after training with the full training set, was the interpretation by the network of the Nujol band at 2924 cm^{-1} as evidence for polar groups such as “O—H” and “C—O”, as well as the “aromatic” grouping. While technically correct—70.3% of the spectra for aromatic compounds were taken as mulls, for example—these correlations lead to incorrect predictions on the test set. The solution was to modify both the training and test programs so that the 2924 cm^{-1} band in Nujol spectra was not presented to the network.

Testing the Final Network Model

For a final test of the simple linear neural network model, 24 functional groups of a general nature, shown in Table 2, were selected. Some of these had shown excellent and some only fair discrimination in earlier training. The network coefficients for these groups were retrained, in 32 additional training sessions, using the augmented training set, the decreased number of input units, and suppression of Nujol bands, as discussed above. The results, in terms of the mean output values, and A_{50} values obtained for the training set compounds, are given in Table 2. The changes made in the training procedure lowered the A_{50} values for some of the groups (compare similar groups in Tables 1 and 2), an effect attributable to the elimination of the spurious correlations. It was anticipated, however, that this would be counterbalanced by better performance in the test program.

For each compound tested, the network yields a numerical output value for each functional group tested for. Comparison of this output value with the distributions of output values obtained from the training set gave an estimate of the probability that the functional group is present in the compound. For ease of interpretation of the test results, the range of probabilities was converted into assertions, specifying four degrees of certainty about the presence or absence of the functional group:

Assertion	Probability
“definitely present”	= 90% probability of being present
“probably present”	= 75% probability of being present
“possibly present”	= 50% probability of being present
“definitely absent”	= 99% probability of being absent.

In order for the correct assertion to be assigned by the program, output values were compared with cutoff points; the output value cutoff points were taken from the

Compound 25: cinnamyl alcohol

functional group	y_i	prediction	based on band at	prediction was
1. O-H/N-H bond	0.596	probably present	3345 cm ⁻¹	correct
2. triple bond/cumulene	0.098	definitely absent		correct
5. aromatic	1.193	definitely present	744 cm ⁻¹	correct
6. C-O single bond	0.923	definitely present	1010 cm ⁻¹	correct
8. alcohol/phenol	0.470	may be present	1068 cm ⁻¹	correct
11. acetylene	0.003	definitely absent		correct
12. nitrile	0.026	definitely absent		correct
13. cumulene (=C=/=N+=)	0.101	definitely absent		correct
14. acid halide/anhydride	0.016	definitely absent		correct
16. ester	0.030	definitely absent		correct
17. aldehyde	0.020	definitely absent		correct
21. phenyl group	1.036	definitely present	692 cm ⁻¹	correct
23. nitro group	0.000	definitely absent		correct

predictions for this compound: 13 correct, 0 incorrect; 100% correct

Fig. 3. Sample page of output from the final test program

training set distribution of output values for each of the 24 functionalities and are shown in Table 2. Note that there is a category between "possibly present" and "definitely absent" about which no assertion is made; this corresponds to the range of output values where the two distributions overlap.

Fig. 3 shows a page from the output from the final test program. For each assertion made about a functional group, the output value y_i is printed out, along with the assertion and whether the assertion is correct. The program also prints out, for positive assertions, the frequency corresponding to the input unit which made the greatest contribution to the output value. This corresponds to the infrared band which played the largest role in the network's decision that a functional group is present.

The test set results for the 24 functional groups are summarized in Table 3. Note that many assertions were made about groups for which discrimination by the network was excellent (e.g., "C=O" or "aromatic") and few assertions were made about groups for which discrimination was only fair or poor (e.g., "C=C" or "aldehyde"). This is a consequence of the use of cutoffs that were keyed to the probability of correctness—no assertions at all would have been made about a group for which there was no discrimination at all ($A_{50} = 0.0\%$). For the 541 compounds in the test set, a total of 6915 assertions were made, out of a possible total of 12984 (541 compounds \times 24 groups tested for). The detection level was thus $6915/12984 = 53.3\%$. Of these assertions, 6326 were correct for an overall accuracy of functional group identification of 91.5%.

Broken down by categories, the accuracy of the assertions was 99.5% for "definitely absent", 48.0% for "possibly present", 69.2% for "probably present", and 87.6% for "definitely present". These values are quite close to the target values of 99.0%, 50.0%, 75.0%, and 90.0% which the choice of cutoff values aimed for. The

Table 3

Functional group	Assertions											
	Definitely absent			Possibly present			Probably present			Definitely present		
	Pr. ^a	Abs. ^b	% ^c	Pr. ^a	Abs. ^b	% ^c	Pr. ^a	Abs. ^b	% ^c	Pr. ^a	Abs. ^b	% ^c
1. OH/NH	2	99	98.0	26	14	65.0	55	18	75.3	103	6	94.5
2. Triple bond/cumulene	4	459	99.1	1	2	33.3	5	2	71.4	25	0	100.0
3. C=O	0	149	100.0	4	6	40.0	34	3	91.9	195	3	98.5
4. C=C	3	39	92.9	1	4	20.0	4	2	66.7	1	0	100.0
5. Aromatic	1	95	99.0	17	34	33.3	30	26	53.6	215	24	90.0
6. C—O	0	5	100.0	57	64	47.1	92	38	70.8	104	29	78.2
7. C-halogen bond	0	0	—	11	21	34.4	8	16	33.3	16	29	35.6
8. Alcohol/phenol	0	12	100.0	25	32	43.9	11	7	61.1	2	1	66.7
9. Amine	1	71	98.6	26	25	51.0	11	4	73.3	1	0	100.0
10. Ammonium (NH ₄ ⁺)	0	184	100.0	0	5	0.0	1	0	100.0	0	0	—
11. Acetylene	3	399	99.3	2	3	40.0	0	0	—	0	0	—
12. Nitrile	0	361	100.0	2	5	28.6	8	1	88.9	14	1	93.3
13. Cumulene	0	529	100.0	0	0	—	1	1	50.0	2	1	66.7
14. Acid halide/anhydride	0	502	100.0	3	0	100.0	1	0	100.0	1	0	100.0
15. Carboxylic acid	1	196	99.5	28	32	46.7	12	8	60.0	8	0	100.0
16. Ester	0	155	100.0	29	8	78.4	9	7	56.3	0	1	0.0
17. Aldehyde	1	273	99.6	2	4	33.3	0	0	—	0	0	—
18. Ketone	0	108	100.0	5	8	38.5	1	0	100.0	0	0	—
19. Amide	1	249	99.6	19	10	65.5	4	1	80.0	1	0	100.0
20. —CH=CH ₂	6	356	98.3	3	3	50.0	0	0	—	2	0	100.0
21. Phenyl (C ₆ H ₅)	0	188	100.0	19	21	47.5	37	6	86.0	22	2	91.7
22. —CO ₂ ⁻ ion	1	251	99.6	0	0	—	0	0	—	0	0	—
23. Nitro	2	269	99.3	0	4	0.0	0	1	0.0	3	0	100.0
24. Ether	1	31	96.9	8	7	53.3	6	6	50.0	15	6	71.4
Total	27	4978	99.5	288	312	48.0	330	147	69.2	730	103	87.6

^a Of the total assertions of this category made, the number of compounds in which the functional group is actually present^b Of the total assertions of this category made, the number of compounds in which the functional group is actually absent^c Accuracy of the assertion as a percentage

close agreement between the accuracies found and the target values means that the network performed almost as well in functional group identification on a new set of compounds that it had never seen before as it did on the compounds of its training set. This result establishes the validity of the functional group discriminations made by the network, and rules out the presence of any further types of spurious correlations.

Conclusions

These results show that the neural network paradigm is a practical method for the automated interpretation of infrared spectral data of organic compounds. The network model investigated in this preliminary study proved capable of detecting the presence or absence of a wide range of functional groups from the infrared spectra of organic compounds of moderate structural complexity, with accuracies that ranged from fair to excellent.

The use of the simple linear network model made it possible to detect and determine the cause of several types of spurious correlations. These could then be eliminated by appropriate alterations to the input architecture of the network, the composition of the training set, and the training regimen used. Information of this type is readily extensible to the construction of more complicated network models.

Despite the limited performance anticipated for the simple linear model, an overall accuracy of 91.5% in functional group identification was achieved at a detection level of 53.3%. This performance compares well with previously reported infrared interpretation systems using other artificial intelligence techniques.

Network models having an additional layer of hidden units and using a non-linear learning rule should show substantial improvements on this performance. The training regimen employed in this study resulted in a bias against functional groups that are under-represented in the training set. A search for other regimens that lack this bias offers another promising possibility for further improvements in performance.

The neural network approach offers several advantages over other artificial intelligence techniques for spectral interpretation. Among these may be cited:

- (1) The network architecture is easily programmed. The total programming effort required is much less than that required for the construction of an expert system, for example.
- (2) Compounds can be identified as belonging to broad general categories as well as to more narrowly defined functional group classifications. The categories need not be structurally defined; it is sufficient to be able to specify that each compound in the training set either belongs to or does not belong to the category.
- (3) Because of the learning ability of the network, the correlations between spectral features and structural categories need not be specified, or even known, in advance.
- (4) The addition of new functional groups to a network's repertoire requires no alteration of the network and no additional programming. It is necessary only to define the new group and to rerun the training sessions.

Acknowledgements. The authors gratefully acknowledge the financial support of this work by the National Institutes of Health (NIGMS) and by The Upjohn Company. We wish to thank S. R. Lowry and The Nicolet Company for making available the database of infrared spectra.

Appendix: Interferences for Functional Group Recognition

1. OH/NH bond. False positives: adventitious water in salts, e.g., sodium carboxylates and sulfonates, amine hydrochlorides, etc. False negatives: none.
2. Triple bond/cumulene. False positives: none. False negatives: some isothiocyanates; some $X-CH-CN$, e.g., cyanoacetate esters. An electronegative X makes the CN stretch very weak.
3. $C=O$. False positives: enol ethers, oxazolidines, guanidinium salts. The $C=C$ or $C=N$ stretching frequency is moved into the carbonyl range by electronegative substitution. False negatives: some enols of β -dicarbonyl compounds, some ureas, where the carbonyl band is at very low frequency.
4. $C=C$. False positives: none. False negatives: tetrasubstituted, many trisubstituted, and some cis-disubstituted double bonds. The $C=C$ stretch is weak or nonexistent in these cases.
5. Aromatic. False positives: some conjugated dienes, some halides, some amino acids, some secondary amines. Bands characteristic of these groups in the out-of-plane bending region and in the $1600-1450\text{ cm}^{-1}$ region mimic aromatic bands. False negatives: none.
6. $C-O$ single bond. False positives: many α -fluorocarbonyl compounds, some hydantoins. Reasons unknown. False negatives: long-chain aliphatic primary alcohols, crown ethers, pyrones and coumarins, some hydroxy amino acids.
7. C -halogen bond. False positives: none. False negatives: polyhalo aromatics, e.g., tetrabromotoluene; aryl halides with one or more nitro groups on the ring; long-chain alkyl halides.
8. Alcohol/phenol. False positives: see 1. above. False negatives: some enols, some *o*-hydroxy aromatic carbonyl compounds.
9. Amine. False positives: none. False negatives: many tertiary amines; some long-chain aliphatic amines, where the $N-H$ stretch can be very weak.
10. Ammonium (N^+-H). False positives: imidazoles, hydroxypyridines. The $N-H$ stretch is broad and at low frequency, mimicking the $N(+)-H$ stretch. False negatives: a few miscellaneous compounds, no pattern observed.
11. Acetylene. False positives: isothiocyanates, carbodiimides. False negatives: nearly symmetrical acetylenes, e.g., $Ph-C\equiv C-R$; $R_2CX-C\equiv CH$. When X is electronegative, the triple bond stretch is weak or non-existent.
12. Nitrile. False positives: $X-N=C=O$ where X is very electronegative, e.g., tosyl or pentafluorophenyl. False negatives: cyanohydrins, see 2. above.
13. Cumulene. False positives: none. False negatives: $X-N=C=S$ where X is electronegative, e.g., tosyl, benzenesulfonyl.
14. Acid halide/anhydride. False positive: strained lactones, cyclic carbonates. False negatives: some aromatic acid chlorides.
15. Carboxylic acid. False positives: none. False negatives: some amino acid hydrochlorides.
16. Ester. False positives: some barbiturates. False negatives: pyrones and coumarins; salicylates, aromatic esters with *o*-amino, *o*- or *p*-hydroxyl groups.
17. Aldehyde. False positives: none. False negatives: aromatic aldehydes with *o*-hydroxyl groups; aromatic aldehydes with nitro groups on the same ring.
18. Ketone. False positives: none. False negatives: quinones, some enols, poly- α -haloketones (e.g., trichloroacetone).
19. Amide. False positives: none. False negatives: miscellaneous compounds, no pattern observed.
20. Vinyl. False positives: none. False negatives: vinyl ethers and esters, acrylate esters.
21. Phenyl. False positives: nearly all 1-substituted thiophenes; most sulfite esters. False negatives: many benzoyl compounds (e.g., benzoate esters, benzamides, phenyl ketones, etc.).

22. Carboxylate anion. False positives: pyridine N-oxides. False negatives: many amino acids.
23. Nitro group. False positives: none. False negatives: most aromatic nitro compounds having *p*-hydroxy or amino groups.
24. Ether. False positives: none. False negatives: nearly all epoxides, crown ethers.

References

- [1] M. E. Munk, B. D. Christie, *Anal. Chim. Acta* **1989**, 216, 57.
- [2] T. Blaffert, *Anal. Chim. Acta* **1986**, 191, 161.
- [3] S. Moldoveanu, C. A. Rapson, *Anal. Chem.* **1987**, 59, 1207.
- [4] B. Debska, J. Duliban, B. Guzowska-Swider, Z. Hippe, *Anal. Chim. Acta* **1981**, 133, 303.
- [5] M. Farkas, J. Markos, P. Szepesvary, I. Barthas, G. Szalontai, Z. Simon, *Anal. Chim. Acta* **1981**, 133, 19.
- [6] G. Szalontai, G. Simon, Z. Csapo, M. Farkas, Gy. Pfeiffer, *Anal. Chim. Acta* **1984**, 133, 31.
- [7] T. Visser, J. H. van der Maas, *Anal. Chim. Acta* **1980**, 122, 363.
- [8] T. Visser, J. H. van der Maas, *Anal. Chim. Acta* **1981**, 133, 451.
- [9] C. D. Baer, W. W. Brown, *Appl. Spectrosc.* **1977**, 31, 524.
- [10] C. G. A. v. Eijk, J. H. van der Maas, *Fresenius' Z. Anal. Chem.* **1977**, 286, 80.
- [11] N. A. B. Gray, *Anal. Chem.* **1975**, 47, 2426.
- [12] H. J. Luinge, G. J. Kleywest, H. A. Van't Klooster, J. H. van der Maas, *J. Chem. Inf. Comput. Sci.* **1987**, 27, 95.
- [13] D. D. Saperstein, *Appl. Spectrosc.* **1986**, 40, 344.
- [14] G. M. Smith, H. B. Woodruff, *J. Chem. Inf. Comput. Sci.* **1984**, 24, 33.
- [15] S. A. Tomellini, R. A. Hartwick, H. B. Woodruff, *Appl. Spectrosc.* **1985**, 39, 330.
- [16] S. A. Tomellini, R. A. Hartwick, J. M. Stevenson, H. B. Woodruff, *Anal. Chim. Acta* **1984**, 162, 227.
- [17] S. A. Tomellini, D. D. Saperstein, J. M. Stevenson, G. M. Smith, H. B. Woodruff, P. F. Seelig, *Anal. Chem.* **1981**, 53, 2367.
- [18] S. A. Tomellini, J. M. Stevenson, H. B. Woodruff, *Anal. Chem.* **1984**, 56, 67.
- [19] H. B. Woodruff, M. E. Munk, *J. Org. Chem.* **1977**, 42, 1761.
- [20] H. B. Woodruff, G. M. Smith, *Anal. Chem.* **1980**, 52, 2321.
- [21] H. B. Woodruff, G. M. Smith, *Anal. Chim. Acta* **1981**, 133, 545.
- [22] L.-S. Ying, S. P. Levine, S. A. Tomellini, S. R. Lowry, *Anal. Chem.* **1987**, 59, 2197.
- [23] M. O. Trulson, M. E. Munk, *Anal. Chem.* **1983**, 55, 2137.
- [24] S. R. Lowry, T. L. Isenhour, *J. Chem. Inf. Comput. Sci.* **1975**, 15, 212.
- [25] H. B. Woodruff, S. R. Lowry, T. L. Isenhour, *Anal. Chem.* **1974**, 46, 2150.
- [26] S. R. Lowry, H. B. Woodruff, G. L. Ritter, T. L. Isenhour, *Anal. Chem.* **1975**, 47, 1126.
- [27] J. Seil, I. Kohler, C. W. v. d. Lieth, H. J. Opferkuch, *Anal. Chim. Acta* **1986**, 188, 219.
- [28] Y. Mivashita, S. Ochiai, S. Sasaki, *J. Chem. Inf. Comput. Sci.* **1977**, 17, 228.
- [29] T. Blaffert, *Anal. Chim. Acta* **1984**, 161, 135.
- [30] M. E. Elyashberg, V. V. Serov, L. A. Gribov, *Talanta* **1987**, 34, 21.
- [31] B. R. Kowalski, P. C. Jurs, T. L. Isenhour, C. N. Reilly, *Anal. Chem.* **1969**, 41, 1945.
- [32] R. W. Liddell, P. C. Jurs, *Appl. Spectrosc.* **1973**, 27, 371.
- [33] R. W. Liddell, P. C. Jurs, *Anal. Chem.* **1974**, 46, 2126.
- [34] D. R. Preuss, P. C. Jurs, *Anal. Chem.* **1974**, 46, 520.
- [35] E. K. Whalen-Pedersen, P. C. Jurs, *J. Chem. Inf. Comput. Sci.* **1979**, 19, 264.
- [36] H. B. Woodruff, M. E. Munk, *Anal. Chim. Acta* **1977**, 95, 13.
- [37] H. B. Woodruff, S. R. Lowry, T. L. Isenhour, *Appl. Spectrosc.* **1975**, 29, 226.
- [38] D. S. Frankel, *Anal. Chem.* **1984**, 56, 1011.
- [39] S. Williams, R. B. Lam, T. L. Isenhour, *Anal. Chem.* **1983**, 55, 1117.
- [40] J. Zupan, *Anal. Chim. Acta* **1982**, 139, 143.
- [41] J. Zupan, M. E. Munk, *Anal. Chem.* **1985**, 57, 1609.

- [42] J. Zupan, M. E. Munk, *Anal. Chem.* **1986**, 58, 3219.
- [43] J. C. W. G. Bink, H. A. Van't Klooster, *Anal. Chim. Acta* **1983**, 150, 53.
- [44] J. Comerford, P. G. Anderson, W. H. Snyder, H. S. Kimmel, *Spectrochim. Acta* **1977**, 33A, 651.
- [45] L. Domokos, I. Frank, G. Matolcsy, G. Jalsovszky, *Anal. Chim. Acta* **1983**, 154, 181.
- [46] G. Hangac, R. C. Weibolt, R. B. Lam, R. T. Isenhour, *Appl. Spectrosc.* **1982**, 36, 40.
- [47] H. B. Woodruff, G. L. Ritter, S. R. Lowry, T. L. Isenhour, *Appl. Spectrosc.* **1976**, 30, 213.
- [48] N. A. Gray, *Prog. NMR Spectrosc.* **1982**, 15, 20.
- [49] Z. Hippe, R. Hippe, *Appl. Spectrosc. Rev.* **1980**, 16, 135.
- [50] D. W. Elrod, G. M. Maggiore, R. Trenary, *197th Natl. ACS Meeting*, Dallas, Texas, April 10–14, 1989.
- [51] D. F. Stubbs, *197th Natl. ACS Meeting*, Dallas, Texas, April 10–14, 1989.
- [52] N. Qian, T. J. Sejnowski, *J. Mol. Biol.* **1988**, 202(4), 865.
- [53] L. H. Holley, M. Karplus, *Proc. Natl. Acad. Sci.* **1989**, 86(1), 152.
- [54] J. D. Bryngelson, J. J. Hopfield, *197th Natl. ACS Meeting*, Dallas, Texas, April 10–14, 1989.
- [55] M. N. Liebman, *197th Natl. ACS Meeting*, Dallas, Texas, April 10–14, 1989.
- [56] D. E. Rumelhart, G. E. Hinton, J. L. McClelland, in: *Parallel Distributed Processing, Vol. I*, MIT Press, Cambridge, MA, 1987, p. 45.
- [57] D. E. Rumelhart, G. E. Hinton, R. J. Williams, in: *Parallel Distributed Processing, Vol. I*, MIT Press, Cambridge, MA, 1987, p. 318.
- [58] M. I. Jordan, in: *Parallel Distributed Processing, Vol. I*, MIT Press, Cambridge, MA, 1987, p. 365.
- [59] Both the thin film and Nujol spectra were taken from a collection made available by The Nicolet Company.
- [60] *Sadtler Grating Infrared Spectrum Collection*, Sadtler Research Laboratories, Philadelphia, PA.

Received July 3, 1989.