The Neural Network as a Tool for Multispectral Interpretation†

Morton E. Munk* and Mark S. Madison

Department of Chemistry and Biochemistry, Arizona State University, Tempe, Arizona 85287-1604

Ernest W. Robb*

Department of Chemical Sciences and Engineering, Stevens Institute of Technology, Hoboken, New Jersey, 07030

Received August 18, 1995⊗

A neural network which utilized data from the infrared spectra, carbon-13 NMR spectra, and molecular formulas of organic compounds was developed. The network, which had one layer of hidden units, was trained by backpropagation; network parameters were determined by a simplex optimization procedure. A database of 1560 compounds was used for training and testing. The trained network was able to identify with high accuracy the presence of a broad range of substructural features present in the compounds. The number of features identified and the accuracy were significantly greater as compared with networks using data from a single form of spectroscopy. The results have significance for the SESAMI computer-enhanced structure elucidation system.

INTRODUCTION

Spectrum Interpretation in Computer-Enhanced Structure Elucidation. The interpretation by the chemist of the collective spectral data derived from a compound of unknown structure generally forms the basis for elucidating its molecular structure. Given a series of unknowns, modern laboratory spectrometers can already produce such data faster than even experienced chemists can interpret them, and the next generation of instruments will be even more efficient. Therefore, it should come as no surprise that the computer has become the centerpiece of numerous efforts to facilitate the interpretation of spectral data, thereby enhancing productivity.

Computer-based procedures for the interpretation of spectral data can be useful stand-alone tools for the chemist. However, greater productivity can be achieved if they serve as components of a more comprehensive system which links them to programs for structure generation and spectrum simulation. Our own interest in spectrum interpreters is related to the development of SESAMI, an interactive comprehensive system of computer-enhanced structure elucidation.[1] SESAMI's role is to *directly* reduce the collective spectral properties of an unknown to a set of compatible molecular structures which is small in number and exhaustive in scope. This set of structures is the point of departure for the chemist's involvement.

At its current stage of development, SESAMI consists of a number of spectrum interpreters[1] linked to a powerful structure generator.[2] In one of their roles, these interpreters give rise to substructures predicted to be present in the structure of the unknown. These substructures serve as constraints on the structure generation process and comprise one of the mechanisms used to limit the number of compatible molecular structures finally produced by SESAMI.

The more information-rich the constraints, the smaller the set of compatible molecular structures and therefore the closer the chemist to the correct assignment of structure. However, it is important to recognize that if even one of the substructures predicted to be present in the unknown is in error, *every* structure produced by the structure generator will be invalid. Thus, the substructural inferences made by spectrum interpreters and used as constraints in SESAMI should meet two requirements: high information content and high accuracy.

In the hands of an experienced chemist, conventional structure elucidation of compounds of considerable complexity, e.g., natural products, generally requires the breadth of structural information that can only be derived from multiple spectral sources. A computer-based system such as SESAMI should not be expected to solve structure problems of comparable complexity with less information. If such collective data are needed by SESAMI, why not determine if there is an advantage to integrating their interpretation? The output of such a procedure may be viewed as the intersection of the predictions derived from single spectroscopic methods. The set of predictions so obtained not only will undoubtedly be smaller than that produced by any one spectroscopic method but may also correspond to a broader range of substructural inferences of high accuracy than those obtained by taking the best of the inferences derived from the individual spectroscopic methods. This study was undertaken with such a goal in mind.

Neural Networks. Backpropagation neural networks have proved to be effective computational tools for pattern recognition and pattern classification in many areas of science and engineering.[3] Their application to chemical problems has been recently reviewed.[4] Neural networks are attractive as a means of spectral interpretation because the relationship between spectral properties and structural features in a molecule need not be specified, or even known, in advance. Instead, the network in effect deduces the relationship during the process of training. This is important, since the rules relating the features of a compound's spectrum to specific

† A portion of this work was presented at the 203rd American Chemical Society National Meeting, San Francisco, April 9, 1992.

⊗ Abstract published in *Advance ACS Abstracts,* February 1, 1996.

0095-2338/96/1636-0231$12.00/0

structural groupings are often so complicated or so poorly understood that construction of a rule-based interpretation system is impractical.

A number of successful applications of neural networks to spectral interpretation have recently been reported. These include studies on the use of infrared spectra[5−13] and mass spectral data[14,15] to identify substructural groupings in organic compounds.

In structure elucidation, data derived from several spectroscopic methods must be combined in making inferences about the presence of substructural groupings in a compound. Information from these disparate sources can in principle be very readily combined in a neural network. Numerical data, irrespective of its type or source, is simply conjoined, after suitable scaling, in a single input vector. The potential ease with which different types of spectral data can be combined provides additional motivation for investigating the use of neural networks in spectral interpretation.

In this paper, we report a pilot study in which information from the infrared spectra, carbon-13 NMR spectra, and the molecular formulas of organic compounds are combined for input to a network trained to recognize substructural groupings.

METHOD

Network Architecture. This study employed a two-layer (one layer of hidden units) fully connected feedforward network.[16] The input vector **X** was a coded representation of data from the infrared and carbon-13 NMR spectra of a compound, along with information from the compound's molecular formula. The output vector **Y** represented coded information about the presence or absence of predefined substructural features in the compound, features that are expressed as substructures. The input to the hidden units $\mathbf{H}_j$ and to the output units is the matrix product of the units in the preceding layer and the connecting coefficients **C**. This input was made nonlinear by passing it through a logistic "squashing" function. For maximum generality, two adjustable parameters were inserted into the logistic function for each layer: a β to alter the stiffness of the sigmoid function and a θ to set its threshold. Values for these parameters were determined during optimization (see below). The initial ranges of the coefficients and the values of η, the training step sizes in the correction formulas, were also determined during optimization. The connecting coefficients were initialized with random values in their optimized range. The equations used for propagation in the forward direction were

$$\mathbf{H}_j = \sum^{i} \mathrm{X}_i \mathrm{C}_{1ij} \tag{1}$$

$$\mathbf{H}_j \leftarrow [1 + \mathrm{e}^{-\beta_1(\mathbf{H}_j - \theta_1)}]^{-1} \tag{2}$$

$$\mathbf{Y}_k = \sum^{i} \mathbf{H}_j \mathrm{C}_{2jk} \tag{3}$$

$$\mathbf{Y}_k \leftarrow [1 + \mathrm{e}^{-\beta_2(\mathbf{Y}_k - \theta_2)}]^{-1} \tag{4}$$

The difference between **Y** and the target value **T** was used to correct the coefficients by backpropagation, using the generalized delta rule.[17] The equations used were

$$\delta_{2k} = \beta_2 \mathbf{Y}_k (1 - \mathbf{Y}_k)(\mathbf{T}_k - \mathbf{Y}_k) \tag{5}$$

$$\mathbf{C}_{2jk} \leftarrow \mathbf{C}_{2jk} + \eta_2 \delta_{2k} \mathbf{H}_j \tag{6}$$

$$\delta_{1j} = \beta_1 \mathbf{H}_j (1 - \mathbf{H}_j) \sum^{k} \delta_{2k} \mathbf{C}_{2jk} \tag{7}$$

$$\mathbf{C}_{1ij} \leftarrow \mathbf{C}_{1ij} + \eta_1 \delta_{ij} \mathbf{X}_i \tag{8}$$

The number of hidden units can have a profound effect on the performance of a neural network. If too few are used, the network is ineffective in making discriminations. When more than the optimum number is used, the effectiveness of the network does not improve, while the required training time increases rapidly and the danger of overtraining increases. In the final network, which had 512 input units and 85 output units, the optimum number of hidden units, derived from experience[6] and experimentation, was typically found to be about 50.

Training Database. A database containing the machine-readable infrared and carbon-13 NMR spectra of organic compounds, along with machine manipulable connection table representations of the molecular structures of the compounds, was needed to prepare the input vectors and target vectors for training the network and testing its output. Since no such database was available at the outset of this work, we took the infrared database of about 6500 compounds used in our previous studies[5,6] and a C-13 NMR database of about 10 000 compounds and executed a search for common molecular structures. In this way, a combined infrared/carbon-13 NMR database containing 1560 compounds was obtained. While small, this number was adequate for the present pilot study. The compounds in the database contained from 1 to 40 carbon atoms, with an average molecular weight of 145.5, and represented a broad range of structural features.

Spectral Representation. A total of 512 input units was used: 128 for infrared data, 294 for carbon-13 NMR data, and 90 for information pertaining to the molecular formula. The relative number of units used for each of the three classes of input information was based on our estimate of the richness of the information content of each class.

For representing the infrared spectrum of a compound, the infrared range from 393 to 3906 cm^{-1} was divided into 128 intervals, and each interval was assigned to an input unit (units 0−127). As in our earlier studies[5,6] the units were not of equal width but decreased in width as the wavenumber decreased. This reflects our view that the low wave number region—the so-called fingerprint region—contains more structural information per wavenumber than does the high wavenumber region. If the spectrum of a compound had at least one significant peak in an interval, the input value for the unit assigned to that interval was a number between 0 and 1 in proportion to the intensity of the most significant peak. If no peak appeared in an interval, the value of 0 was given to the unit.

Carbon-13 NMR data was first grouped according to the signal multiplicity originating from one-bond carbon−hydrogen coupling. The chemical shift regions in which signals of each multiplicity are normally found were then divided into equal intervals 2 ppm in width, and each interval was assigned to an input unit. For quartets, the chemical

Table 1. Assignment of Data from Molecular Formula to Input Units

feature	no. of features present	no. of units assigned	assignment
C	1...21, 22−26, 27+	23	422...444
N	0...3, 4+	5	445...449
O	0...6, 7+	8	450...457
F	0...2, 3+	4	458...461
Cl	0...4, 5+	6	462...467
Br	0...2, 3+	4	468...471
I	0...1, 2+	3	472...474
P	0, 1+	2	475...476
S	0...2, 3+	4	477...480
unsat. equiv.	0...13, 14+	15	481...495
MW mod 14	0...13	14	496...509
MW mod 2	0...1	2	510...511

shift range 0−64 ppm was divided into 32 intervals (units 128−159); for triplets, the range 0−124 ppm was divided into 62 intervals (units 160−221); for doublets, the range 8−203 ppm was divided into 98 intervals (units 222−319); and for singlets, the range 24−228 ppm was divided into 102 intervals (units 320−421). If at least one signal occurred in one of the 2 ppm intervals, a 1 was entered in the corresponding input unit; in the absence of a signal, a 0 was entered for that unit.

Since, in general, only signal position and not signal intensity is significant, no attempt was made to represent intensities. Similarly, there was no representation of the number of equivalent carbon atoms giving rise to a signal, since in the presence of symmetry in the compound of unknown structure this information would not be available.

The scheme devised for entering information related to the molecular formula is summarized in Table 1. The input unit corresponding to the number of atoms of an element was assigned a value of 1 if it characterized the compound and a 0 otherwise. For example, a 1 would be assigned to unit 431 for a compound with 10 carbon atoms; for a compound with 30 carbon atoms a 1 would be entered for unit 444. In addition to atom numbers, certain other information derived from the molecular formula was entered: the number of unsaturation equivalents in units 481−495; the mass number mod 14, commonly used as a characterizing feature in mass spectrum interpretation; and the odd-even parity of the mass number, a form of the mass spectroscopists' so-called Nitrogen Rule.

In devising the system for representing information from the spectra and molecular formula and for assigning it to input units, an overall criterion was that each input unit should be activated by at least five of the compounds in the database.

Structural Representation. A large number of substructural features common in organic compounds was examined for suitability as output descriptors. Some corresponded to the traditional functional groups and substituent groups of organic chemistry. Others were larger assemblages of atoms or were of common structural skeletons such as "terpene" or "steroid". From these, a final set of 85 features, listed in Table 2, was selected for development and evaluation of the multispectral network. The final selection was made with several criteria in mind: (1) each feature should occur in at least 20 compounds of the database; (2) the list should include features both easy and difficult to identify; (3) the features should represent a range of sizes, from a single atom, for example an *sp* hybridized carbon atom, to an extended group comprising many atoms; and (4) the features should include some that would make an interpreter based on the network of value to the SESAMI system.

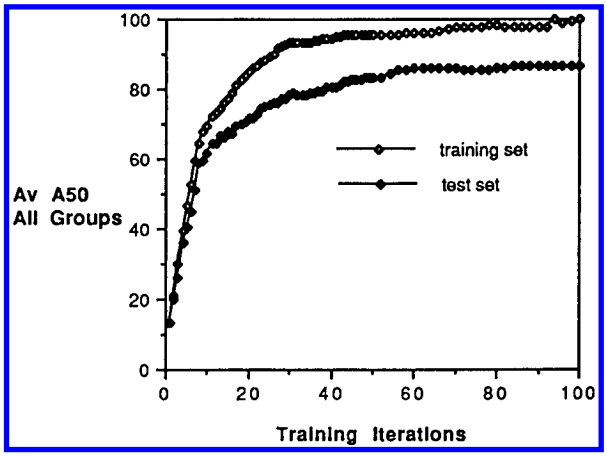

Figure 1. Typical learning curve: Average A_{50} for all output groups for training and test sets versus the number of training iterations.

Each of the 85 substructural features selected was assigned arbitrarily to an output unit. If a compound contained a given feature, the target value at the output unit assigned to that feature was set to a value of 1; if the feature was not present in the compound, the target value at that unit was set to 0. A substructure searching program was used to search the connection tables of the database to prepare a target vector for each compound.

Training Method and Cross-Validation. Valid measures of network performance can be obtained only by evaluating results from input examples which are not used in training the network. We therefore set aside one-tenth of the database (156 compounds) as a test set and used the remainder as training examples. Input from each training compound was presented to the network in turn, followed by correction of the coefficients by backpropagation. After all of the training examples had been presented, the compounds of the test set were presented, without backpropagation, and the output results were tabulated. This constituted one training iteration.

The results from a typical training session are shown in Figure 1 (A_{50}, the measure of performance used, is discussed below). The performance of a neural network increases at first with the number of training iterations. After extended training, the network typically exhibits perfect performance when measured by the results from the training set. It is able to do this because it eventually learns to recognize the training examples by using information particular to individual compounds, small fingerprint bands or noise peaks in the infrared spectrum, for example. In so doing, it makes less use of features of general significance in the input examples and performance as measured against a test set suffers. This phenomenon is known as overtraining. Thus, there is an optimum number of training iterations. In this work, the optimum number of iterations in a complete training session was determined by experience[6] and experimentation to be 50. Because of the relatively small size of the database, the number of test examples—156—was too small to allow a statistically meaningful evaluation of the performance of the network for all of the 85 output groups. We therefore made use of a cross-validation technique.[18] Nine additional networks were trained, each using a different one-tenth of the database as a test set. The results from all

Table 2. Output Performance for 85 Substructural Features with Combined Infrared, Carbon-13 NMR and Molecular Formula Input

	substructural feature	N^a	$A_{50}{}^b$	$R_{90}{}^c$	acc.^{d}	rec.^{e}	Y(+)^f	Y(−)^g	useable^h
1.	sp carbon	87	91.6	59.5	82.9	66.7	0.676	0.013	√
2.	acetylene	28	100.0	57.1	100.0	53.6	0.486	0.006	√
3.	nitrile	48	96.0	65.6	86.5	66.7	0.649	0.005	√
4.	hydroxy	457	97.8	79.9	87.0	85.1	0.840	0.064	√
5.	alcohol	226	98.7	85.3	89.8	85.4	0.840	0.020	√
6.	CH_2OH	96	96.9	72.1	88.1	77.1	0.751	0.010	√
7.	sec. alcohol	124	98.1	83.1	92.0	83.1	0.792	0.014	√
8.	CH_3-CHOH	31	100.0	74.8	91.7	71.0	0.660	0.004	√
9.	tert. alcohol	27	10.2	0.0	54.5	22.0	0.257	0.007	
10.	phenol	96	96.5	61.4	74.0	77.1	0.749	0.023	√
11.	C−Cl	162	99.4	100.0	94.7	99.8	0.948	0.007	√
12.	R−Cl	73	97.8	70.9	85.3	79.5	0.749	0.011	√
13.	CH_2Cl	34	100.0	67.6	85.2	67.6	0.643	0.005	√
14.	unsat. Cl	31	88.6	47.2	89.5	54.8	0.520	0.005	
15.	aryl-Cl	65	96.2	95.4	91.0	93.8	0.878	0.007	√
16.	carbonyl	594	99.8	99.7	97.0	97.5	0.965	0.025	√
17.	acetyl	101	92.0	66.2	82.0	81.2	0.764	0.014	√
18.	CH_2-CO	196	93.1	63.3	83.7	76.0	0.739	0.029	√
19.	unsat. Carbonyl	107	68.2	21.5	70.2	55.1	0.529	0.021	
20.	aryl CO	138	91.9	64.4	78.8	78.3	0.772	0.024	√
21.	aldehyde	45	61.6	31.1	80.0	44.4	0.433	0.009	
22.	Ar-CHO	20	90.9	50.0	90.0	45.0	0.420	0.004	√
23.	ester	142	99.0	79.2	87.4	83.1	0.807	0.016	√
24.	aliph. ester	85	100.0	88.2	89.4	89.4	0.841	0.010	√
25.	acetoxy	18	100.0	81.3	92.9	72.2	0.663	0.003	√
26.	$CH_2C(=O)O$	41	93.2	63.4	84.4	65.9	0.673	0.007	√
27.	unsat. ester	21	38.5	4.8	55.4	23.8	0.244	0.006	
28.	arom. ester	33	66.0	0.0	65.4	51.5	0.471	0.009	
29.	methyl ester	46	100.0	78.3	88.4	82.6	0.778	0.005	√
30.	$CH_2OC=O$	65	94.2	81.4	92.7	78.5	0.741	0.006	√
31.	ethyl ester	46	97.2	84.8	90.7	84.8	0.783	0.004	√
32.	ketone	182	99.2	80.8	87.1	81.9	0.802	0.022	√
33.	methyl ketone	67	94.6	67.2	86.9	79.1	0.721	0.008	√
34.	methylene ketone	84	92.7	64.3	87.5	75.0	0.721	0.011	√
35.	CH_2-CO-CH_2	29	72.5	0.0	80.0	41.4	0.480	0.006	
36.	unsat. ketone	51	60.7	29.4	67.7	41.2	0.463	0.013	
37.	arom. ketone	70	92.7	74.6	89.8	75.7	0.723	0.010	√
38.	carboxyl	166	99.4	85.2	89.9	85.5	0.840	0.019	√
39.	aliph. COOH	114	99.6	84.0	90.5	83.3	0.818	0.014	√
40.	CH_2COOH	53	80.3	29.4	73.2	56.6	0.522	0.010	
41.	unsat. COOH	35	83.3	46.9	81.8	51.4	0.523	0.006	
42.	arom. COOH	19	82.6	35.5	76.9	52.6	0.630	0.006	
43.	Amide	61	81.3	37.2	76.5	63.9	0.739	0.013	
44.	ether	116	96.8	70.7	87.9	75.0	0.722	0.016	√
45.	CH_2-O	135	97.3	66.5	84.3	71.9	0.497	0.017	√
46.	CH_2-O-CH_2	22	64.7	45.5	62.5	45.5	0.953	0.004	
47.	methyl	892	99.9	99.1	98.3	96.9	0.807	0.039	√
48.	*O*-methyl	111	99.9	86.4	97.9	83.0	0.726	0.007	√
49.	methyl ether	60	96.8	81.0	93.6	73.3	0.794	0.004	√
50.	Ar-OCH3	44	100.0	93.4	92.7	86.4	0.494	0.004	√
51.	*N*-methyl	44	81.5	29.5	82.1	52.3	0.933	0.008	
52.	*C*-methyl	756	99.7	98.6	97.2	94.7	0.605	0.037	√
53.	allylic methyl	99	87.0	45.8	75.9	60.6	0.752	0.016	
54.	Ar-CH_3	71	97.3	74.5	83.8	80.3	0.757	0.012	√
55.	quatern. methyl	34	100.0	82.4	90.0	79.4	0.651	0.004	√
56.	gem. dimethyl	96	88.6	46.9	71.1	66.7	0.900	0.017	
57.	ethyl	304	99.2	93.7	92.3	90.8	0.953	0.023	√
58.	*n*-propyl	151	99.3	88.0	89.9	88.1	0.616	0.012	√
59.	isopropyl	73	85.6	47.5	76.7	63.0	0.616	0.015	
60.	CH_2-CH_2-CH_2	307	99.2	85.2	89.2	85.7	0.838	0.028	√
61.	allylic CH_2	117	92.9	51.9	85.0	58.1	0.583	0.013	√
62.	Ar-CH_2	82	70.7	24.4	69.7	56.1	0.562	0.021	
63.	*n*-butyl	113	99.6	96.5	92.9	92.9	0.897	0.008	√
64.	isobutyl	30	100.0	67.5	95.2	66.7	0.651	0.004	√
65.	*sec*-butyl	21	84.0	14.7	70.0	66.7	0.632	0.007	
66.	*tert*-butyl	45	81.8	44.4	77.4	53.3	0.535	0.010	
67.	double bond	311	92.1	52.9	81.6	71.4	0.689	0.047	√
68.	$CH_2=CH-$	43	99.2	72.6	91.2	72.1	0.667	0.006	√
69.	$CH_2=CH$-CH_2-	18	100.0	72.2	100.0	66.7	0.607	0.001	√
70.	*gem* disubst C=C	25	73.5	24.0	87.5	32.0	0.368	0.003	
71.	isopropenyl	18	68.2	27.8	83.3	38.5	0.344	0.002	
72.	*vic* disubst C=C	95	60.8	22.6	67.8	42.1	0.432	0.002	
73.	trisubst C=C	62	79.5	32.9	78.6	53.2	0.501	0.011	

Table 2 (Continued)

	substructural feature	N[a]	A_{50}[b]	R_{90}[c]	acc.[d]	rec.[e]	Y(+)[f]	Y(−)[g]	useable[h]
74.	aromatic	693	99.1	99.7	97.0	98.9	0.974	0.034	√
75.	benzene	481	98.1	92.6	89.4	93.1	0.922	0.052	√
76.	phenyl	165	99.1	77.7	85.4	81.8	0.790	0.023	√
77.	benzyl	35	89.7	54.4	75.0	60.0	0.535	0.007	√
78.	*o*-disubst.	172	88.8	<47.1	97.5	74.4	0.738	0.031	
79.	*m*-disubst.	47	51.4	19.1	56.8	94.7	0.463	0.014	
80.	*p*-disubst.	104	92.0	53.2	71.7	68.3	0.668	0.028	√
81.	pyridine	52	100.0	68.2	85.7	69.2	0.685	0.007	√
82.	naphthalene	29	98.3	58.6	94.4	58.6	0.547	0.004	√
83.	indole	20	90.9	60.0	85.7	60.0	0.567	0.003	√
84.	steroid	23	86.8	97.8	90.9	87.0	0.820	0.002	√
85.	terpene	55	94.2	70.9	82.0	74.5	0.731	0.007	√
	mean, all groups								
	weighted		94.7	79.3	88.5	82.3	0.767	0.025	
	unweighted		89.1	61.8	84.1	70.1	0.678	0.014	

[a] Number of compounds having the substructural feature. [b] Accuracy with cutoff set for 50% recovery. [c] Recovery with cutoff set for 90% accuracy. [d] Accuracy with cutoff at Y = 0.500. [e] Recovery with cutoff set at Y = 0.500. [f] Mean output value Y for compounds having the substructural group. [g] Mean output value for compounds lacking the substructural group. [h] Either $A_{50} \geq 90\%$ or $R_{90} \geq 50\%$.

ten test sets could then be pooled, since in no case was a network tested against a compound used in its training and since each of the ten networks was trained on very nearly the same set of training examples.

Optimization. There are a number of procedural parameters in training that affect network performance; these include the stiffness and threshold of the squashing functions (β_1, β_2 and θ_1, θ_2, respectively), the step sizes for correction in backpropagation (η_1, η_2) and, since there are two layers, two parameters for each of the two initial ranges of random values of the starting coefficients. These parameters are best regarded as an interacting set, since the best value of one will depend on the values of the others. We have found that the final performance of a network can be enhanced by careful optimization of this set of parameters. It is not practical to do this by trial and error, since certain combinations of values give networks that perform very poorly. A robust optimization method is needed for exploring what is, in effect, a highly irregular response surface in ten dimensions.

The simplex method is admirably suited to this type of optimization. The initial vertices of a 10-dimensional simplex were prepared by selecting 11 sets of parameters and using each set to train a network with full cross-validation. Using the mean A_{50} values for all 85 output groups as a measure of performance, the variable-size sequential simplex algorithm[19] was then implemented. Each step consisted of training a new network with a new set of parameters. Convergence was achieved after about 200 simplex steps.

The results of two such optimizations are shown in Table 3. The values of the parameters in the two sets obtained in separate optimizations are somewhat different, but the final optimized network performance is closely comparable in the two runs.

The final results from a fully optimized network for the 85 substructural groupings surveyed are given in Table 2.

Computational Methods. Programs for preparing the database, substructure searching, preparing input and target vectors, training the neural networks and evaluating the results, cross-validation, and simplex optimization were written in Pascal and C. Developmental work was carried out on a VAX 3500 system. Production runs were carried out on a MasPar MP-1 system or on a Hewlett-Packard 9000 series RISC workstation. A complete run, including cross-validation and optimization, required about 1 h of CPU time on the Hewlett-Packard workstation.

Table 3. Typical Network Parameters and Network Performance After Simplex Optimization

	run 1	run 2
β_1	1.107	1.202
β_2	1.108	1.113
η_1	0.120	0.140
η_2	0.122	0.173
θ_1	0.622	0.351
θ_2	0.571	0.551
$\mathbf{C}_1$ (initial)	−0.262 to 0.081	−0.245 to 0.105
$\mathbf{C}_2$ (initial)	−0.156 to 0.181	−0.205 to 0.138
A_{50}, av, all groups	80.6	82.3

RESULTS AND DISCUSSION

Measures of Effectiveness. If the neural network were perfectly capable of recognizing the presence of an output group, the carboxyl group, for example, then each of the compounds in the database having the group would give an output value Y of 1.00 at the output unit assigned to the group, while each of the compounds lacking the group would give an output value of 0.00 at that unit. An actual result, that for the carboxyl group, is shown in Figure 2. Most of the 166 compounds having a carboxyl group do give output values of 1.00 or nearly so, but a few give smaller values of Y (solid curve). The broken curve shows the distribution of Y values for the 1394 compounds which have no carboxyl group. Most give values of 0.00 or nearly so, but a few give higher values. The feature to note is that there is some overlap between the two curves, so that perfect accuracy in identifying this group cannot be achieved by the network. A value for Y of 0.50 from an unknown compound at this output unit, for example, would be ambiguous, since there are compounds both with and without carboxyl groups that give such an output value.

In inferring structural features in a compound of unknown structure, a choice must be made for each output group as to the cutoff value of Y that is used in deciding whether or not the group is present in the compound. If a high cutoff value is chosen, then predictions that the group is present

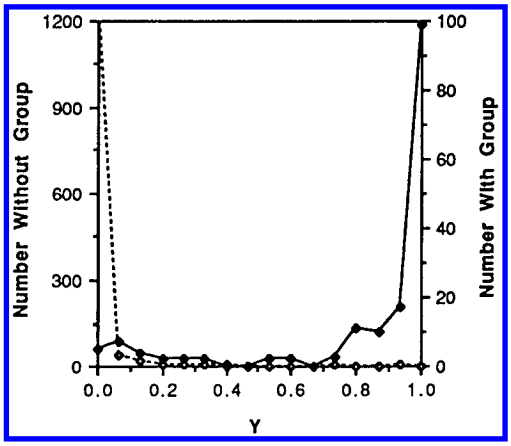

Figure 2. Distribution of test set results at output unit 38 (carboxyl group). Solid curve: compounds with carboxyl group; broken curve: compounds without carboxyl group.

will be of high accuracy. This greater accuracy is achieved, however, at the expense of a lower recovery of valid information, since many compounds having the group will give output values less than the cutoff value and so will not be detected. Alternatively, the recovery of valid information could be increased by choosing a lower cutoff value of Y, at the expense of decreased accuracy. In a particular application, practical considerations will govern the decision as to whether accuracy or recovery is the more important. The point here is that, because different cutoff values can be chosen, it is difficult to summarize the effectiveness of the identification of an output group with a single numerical measure.

Other workers, faced with this difficulty, have proposed a variety of measures for assessing performance. In our previous work[5,6] we used the accuracy at 50% recovery (A_{50}) as a means of comparing results for different output groups and also as a general measure of performance. This measure is calculated by choosing a cutoff value of Y which results in 50% recovery of the valid predictions. The cutoff is the median of the output values for compounds having the group. The A_{50} value for a given structural feature is then the accuracy of the predictions with Y values equal to or greater than the cutoff, i.e., the ratio of correct predictions to all predictions (including false positions), expressed as a percentage. In the data for the carboxyl group shown in Figure 2, for example, the median for the solid curve occurs at Y = 0.84; with this as a cutoff, predictions that a carboxyl group is present in the test compounds have an accuracy of 99.4%. The A_{50} values for the 85 substructural groupings surveyed in this study are shown in the second column of Table 2.

In a computer-enhanced structure elucidation system such as SESAMI, it is important to utilize inferences of prediction accuracy of 90% or greater, even if this is achieved at the cost of a lower recovery of information. The recovery at 90% accuracy (R_{90}), which emphasizes this, is a measure of performance which we have found to be useful. The R_{90} measure is calculated by selecting a value of Y such that 90% of the output values equal to or greater than this cutoff are due to compounds having the structural feature in question (correct identifications), the remaining 10% being false positives. The recovery then is the ratio of correct identifications to the total number of compounds having the structural feature in question, expressed as a percentage. In

Table 4. Effect of Combined Input on Network Performance

	A_{50}, av, all groups	
input	with mol formula	without mol formula
IR only	56.3	45.2
C-13 NMR only	77.1	64.8
IR + C-13 NMR	81.7	75.4

Figure 2, for example, 90% accuracy occurs when the cutoff value of Y is 0.57. Of all the carboxylic acids in the database, 85.2% (R_{90}) gave Y values equal to or greater than this. The R_{90} measure is quite a conservative measure of performance, and a value of more than 50% is considered to be excellent. R_{90} values for all of the 85 output groups studied are in the third column of Table 2.

In the case of many of the substructural groups surveyed, the distribution curves showed little overlap. In that case, changing the cutoff value of Y has little effect on either the accuracy or recovery, and it is simpler to set the cutoff value arbitrarily at 0.500 in calculating accuracy and recovery. The fourth and fifth columns of Table 2 show accuracies and recoveries calculated in this way.

Also of interest are the mean values of Y for each of the groups studied, both for the set of compounds with and without the group. These are given in the sixth and seventh columns of Table 2.

General Performance. The overall performance of the neural network, using input from infrared spectra, carbon-13 NMR spectra, and molecular formulas, was excellent. Most of the functional groups and substituent groups common in organic compounds—triple bond, nitrile, alcohol, phenol, carbonyl, ester, ketone, carboxyl, ether, methyl, ethyl, and phenyl, for example (Table 2, entries 2, 3, 5, 10, 16, 23, 32, 38, 44, 47, 57, 76)—displayed A_{50} values rating close to 100% accuracy. The mean for all of the groups in all of the compounds in the database was A_{50} = 94.7% and R_{90} = 79.3%.

Identifications with A_{50} values in excess of 90% or R_{90} values greater than 50% are of potential value in the SESAMI computer-enhanced structure elucidation system. Of the 85 substructural groups surveyed, 59 met this criterion. These are indicated in the last column of Table 2.

Detection of the traditional functional groups of organic chemistry is important in structure elucidation, but of greater utility in reducing the number of compatible molecular structures is the detection of larger assemblages of atoms. For example, the hydroxyl group is a component of many functional groups. Alcohols comprise a subset of the hydroxyl group. Primary and secondary alcohols are two classes of alcohols. The methyl carbinols represent a more limited group of secondary alcohols. As we go deeper into this nested arrangement of substructures, the information content of the retrieved substructure increases, as applied to structure elucidation problems. Yet, as shown by the data for these substructures in Table 2 (entries 4−8), there is little change in the quality of the predictions as measured by either the A_{50} or R_{90} values. Comparison of the data for the carbonyl, ester, methyl ester, ethyl ester, acetyl, and acetoxy substructures (Table 2, entries 16, 22, 29, 31, 17, 25) shows a similar result.

We anticipated that the inclusion of carbon-13 NMR data in the network's input would enhance its ability to provide carbon skeletal information by identifying extended hydro-

carbon groupings not necessarily near a polar functional group. The results for the methyl, ethyl, *n*-propyl, *n*-butyl, $-(CH_2)_3-$, and isobutyl groupings (Table 2, entries 47, 57, 58, 63, 60, 64) bear this out. Again, both A_{50} and R_{90} values suggest excellent performance.

The multispectral network was also able to identify with high accuracy the presence of major molecular skeletons—benzene, naphthalene, pyridine, indole (Table 2, entries 75, 81−83)). The compounds representing these skeletons in the database differ not only in their substituents but also in their substitution patterns. The identification of the classes which these molecular skeletons represent is significant information in elucidating structure.

Included in the list of output groups were some rather general structural classes (steroid, terpene—Table 2, entries 84, 85). The identification of database compounds as belonging to these categories was very good. This result suggests that preliminary classification of unknown natural products as to skeletal type might be possible, providing that enough examples of the types were present in the training database.

The predictive ability of the network was not uniformly excellent. There were a number of substructural features for which either the A_{50} or the R_{90} values were low. In most of these cases, the number of examples of the feature in the compounds of the database was too small for adequate training. Examples include the furan ring (11 examples, $A_{50} = 30.0\%$, $R_{90} = 10.2\%$), the epoxide grouping (11 examples, $A_{50} = 27.2\%$, $R_{90} = 36.4\%$), and the cumulene double bond system (8 examples, $A_{50} = 0.7\%$, $R_{90} = 0.0\%$). These and other groups of rare occurrence were excluded from the final set of 85 features surveyed. Our experience suggests that a substructural feature should be present in at least 1.5% of the members of the training set (about 21 compounds in this study). Otherwise the network will not learn to identify the substructure, no matter how salient its spectral features are.

There were, however, some features for which poor performance was observed even when that condition was met. The tertiary alcohol, unsaturated ester, disubstituted double bond, and meta-disubstituted benzene groups are among those which may be discerned in Table 2 (entries 9, 27, 70, 72, 79). Apparently these groups lack sufficiently distinctive spectral features that can be recognized and utilized by the network.

Some control experiments were carried out in which fictitious "groups" lacking any common features were studied. For example, all of the compounds in the database whose I.D. numbers were evenly divisible by 29 were assigned an output unit representing the "1/29th group". After extensive training—actually overtraining—the network learned to correctly identify these compounds in the training set. However the network completely failed to recognize the "1/29th group" in the test set (48 examples, $A_{50} = 6.3\%$, $R_{90} = 0.0\%$).

Effect of Multispectral Input. A primary goal of this study was to evaluate not only the feasibility but also the effectiveness of combining data from different types of spectroscopy as input to a neural network. We therefore carried out some experiments designed to assess the relative effectiveness of each of the types of spectroscopy in identifying substructural features and to determine if enhanced effectiveness resulted from their combination. This was done by training networks, using the same method and the same set of output groups, but restricting the input by omitting either the infrared data or the NMR data (with appropriate truncation of input units). The effect of omitting data from the molecular formula was assessed in the same way. The results of these experiments are summarized in Table 4. The average of the A_{50} values for all of the 85 substructural groups was taken as the measure of performance.

Table 5. Response of Individual Groups to Restricted Input

		mean A_{50} values for group			
	group[a]	IR alone	NMR alone	IR + NMR	IR + NMR + mol formula
2.	acetylene	27.3	5.1	100.0	100.0
3.	nitrile	15.8	50.0	96.0	96.0
6.	$-CH_2OH$	87.5	63.6	87.5	96.9
8.	$CH_3-CHOH-$	40.0	19.0	66.7	100.0
10.	phenol	28.6	48.3	77.8	96.5
15.	aryl Cl	17.9	50.0	45.5	96.2
16.	carbonyl	100.0	98.3	100.0	100.0
17.	acetyl	100.0	92.9	100.0	92.0
21.	aldehyde	18.0	50.0	63.6	61.6
31.	ethyl ester	37.5	75.0	97.0	97.2
32.	ketone	56.5	97.3	93.1	99.2
48.	OCH_3	88.8	91.3	100.0	100.0
50.	$ArOCH_3$	80.0	66.7	100.0	100.0
68.	vinyl	76.9	55.6	100.0	99.2
67.	double bond	46.8	73.3	95.5	92.1
73.	isopropyl	24.1	63.6	87.5	85.6
75.	benzene	85.0	99.3	96.5	98.1
76.	phenyl	100.0	83.0	98.1	99.1
81.	pyridine	12.3	100.0	100.0	100.0

[a] Entry numbers correspond to those in Table 2.

From these experiments we draw a number of conclusions. First, carbon-13 NMR data alone lead to a better performing network than one based only on infrared data. In this connection, however, it should be noted that many of the output groups are merely carbon skeleton fragments. It is to be expected that carbon-13 NMR would be a more powerful probe for such structural features. Second, the inclusion of data from the molecular formula improves the identification of substructural groups when combined with infrared data alone, with carbon-13 NMR data alone, or with a combination of infrared and carbon-13 NMR data. Finally, and most noteworthy, the combined use of infrared and carbon-13 NMR data as input improves network performance over networks using either type of data alone.

These conclusions were confirmed by noting, for each of the networks trained with restricted input, how many individual output groups were identified at the $A_{50} \geq 90\%$ level of accuracy. Without use of molecular formula information, infrared data alone trained 6 groups to this level; 19 groups with carbon-13 NMR data alone, and 35 groups when both infrared and NMR data were used. When combined with information from the molecular formula, the corresponding figures were with infrared only, 15 groups, with NMR only, 32 groups; and with both infrared and NMR, 45 groups.

It is of interest to note which type of spectral data is most effective in training particular individual substructural features. Data for some selected groups are presented in Table 5. For most groups, carbon-13 NMR data were more effective than infrared data. This was most notably the case for the hydrocarbon groups (e.g., isopropyl, entry 73). NMR data were also more effective in distinguishing between

different types of carbonyl groups (e.g., aldehyde, entry 21; ethyl ester, entry 31; ketone, entry 32). A few groups, however, were more readily identified with infrared data alone than with NMR data alone. Examples include the vinyl group (entry 68), the phenyl group (entry 76), the primary alcohol group (entry 6), and aromatic methyl ethers (entry 50). Some groups were easily identified by either form of spectral data (carbonyl, entry 16; acetyl, entry 17; *O*-methyl, entry 48; benzene, entry 75). Interestingly for this study, there were features not well identified by either infrared or NMR data alone but which were accurately identified by a combination of the two (e.g., acetylene, entry 2; nitrile, entry 3; methyl carbinol, entry 8; phenol, entry 10). In many cases the information from the molecular formula did not appear to assist in the group identification; in others, it led to a modest increase in accuracy. Molecular formula information appeared to be of greatest utility in identification of groups with heteroatoms (e.g., aryl chloride, entry 15) and highly unsaturated and aromatic groupings.

Conclusions. This study demonstrates that disparate spectral data can be readily combined for input to a single neural network. The network parameters are conveniently optimized by the simplex method. The resulting network is capable of identifying substructural features with high accuracy, with improved performance over networks employing data from a single spectroscopic source. The broad range of features that can be detected include those with significance in solving real-world structure elucidation problems.

ACKNOWLEDGMENT

The financial support of this research by the National Institutes of Health (Grant GM 37963), Sterling Drug, Inc., and Allied Signal, Inc. is gratefully acknowledged. Zafer Kadi was very helpful in assisting the porting of the neural network program to the MasPar computer. We thank the Nicolet Company for the infrared database and Prof. Wolfgang Robien of the University of Vienna for most entries of the carbon-13 NMR database.

REFERENCES AND NOTES

(1) Munk, M. E.; Velu, V. K.; Madison, M. S.; Robb, E. W.; Badertscher, M; Christie, B. D.; Razinger, M. Chemical Information Processing in Structure Elucidation. *Recent Advances in Chemical Information II*; Collier, H., Ed.; Royal Society of Chemistry: Cambridge, U.K., 1992; pp 247−263.

(2) Christie, B. D.; Munk, M. E. Structure Generation by Reduction: A New Strategy for Computer-Assisted Structure Elucidation. *J. Chem. Inf. Comput. Sci*. **1988**, *28*, 87−93.

(3) Haykin, S. *Neural Networks. A Comprehensive Foundation*; MacMillan: New York, 1994.

(4) Gasteiger, J.; Zupan, J. Neural Networks in Chemistry. *Angew. Chem., Int. Ed. Engl*. **1993**, *32*, 503−527.

(5) Robb, E. W.; Munk, M. E. A Neural Network Approach to Infrared Spectrum Interpretation. *Mikrochim. Acta* **1990** I, 131−155.

(6) Robb, E. W.; Madison, M. S.; Munk, M. E. Neural Network Models for Infrared Spectrum Interpretation. *Mikrochim. Acta* **1991**, II, 505−514.

(7) Fessenden, R. J.; Györgyi, L. Identifying Functional Groups in Infrared Spectra Using Neural Networks. *J. Chem. Soc., Perkin Trans. 2* **1991**, 1755−1762.

(8) Meyer, M.; Weigelt, T. Interpretation of Infrared Spectra by Artificial Neural Networks. *Anal. Chim. Acta* **1992**, *265*, 183−190.

(9) Weigel, U.-M.; Herges, R., Automatic Interpretation of Infrared Spectra: Recognition of Aromatic Substitution Patterns Using Neural Networks. *J. Chem. Inf. Comput. Sci*. **1992**, *32*, 723−731.

(10) Meyer, M.; Meyer, K.; Hobert, H. Neural Networks for Interpretation of Infrared Spectra Using Extremely Reduced Data. *Anal. Chim. Acta* **1993**, *282*, 407−415.

(11) Ricard, D.; Cachet, C.; Cabrol-Bass, D. Forrest, T. P. Neural Network Approach to Structural Feature Recognition from Infrared Spectra. *J. Chem. Inf. Comput. Sci*. **1993**, *33*, 202−210.

(12) Klawun, C.; Wilkins, C. L. A Novel Algorithm for Local Minimum Escape in Back-Propagation Neural Networks: Application to Interpretation of Matrix Isolation Infrared Spectra. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 984−993.

(13) Nović, M.; Zupan, J. Investigation of Infrared Spectra-Structure Correlations Using Kohonen and Counterpropagation Neural Networks. *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 454−466.

(14) Curry, B.; Rumelhart, D. E. MSnet: A Neural Network Which Classifies Mass Spectra. *Tetrahedron Comput. Methodol* . **1990**, *3*, 213−238.

(15) Lohninger, H.; Stancl, F. Comparing the Performance of Neural Networks to Well-Established Methods of Multivariate Data Analysis: The Classification of Mass Spectral Data. *Fresenius' J. Anal. Chem*. **1992**, *344*, 186−189.

(16) Wasserman, P. D.; *Neural Computing; Theory and Practice*; Van Nostrand Reinhold: New York; 1989; pp 43−59.

(17) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. In *Parallel Distributed Processing, Vol. I*; Rumelhart, D. E.; McClelland, J. L., and the PDP Research Group, Eds.; MIT Press: Cambridge MA, 1986; pp 318−362.

(18) Efron, B. *The Jackknife, the Bootstrap, and Other Resampling Plans*; Society for Industrial and Applied Mathematics: Philadelphia, PA, 1982.

(19) Walters, F. H.; Parker, L. R., Jr.; Morgan, S. L.; Deming, S. N. *Sequential Simplex Optimization*; CRC Press: Boca Raton, FL, 1991; pp 76−95.

CI950094+