Identification of Structural Features from Mass Spectrometry Using a Neural Network Approach: Application to Trimethylsilyl Derivatives Used for Medical Diagnosis

Ahmad Eghbaldar,[†] Thomas P. Forrest,[§] Daniel Cabrol-Bass,[*,†] Aimé Cambon,[‡] and Jean-Marie Guigonis[‡]

LARTIC and Chimie organique du fluor, Université de Nice-Sophia Antipolis, Parc Valrose,
06108 Nice Cedex 2, France, and Department of Chemistry, Dalhousie University,
Halifax, Nova Scotia, Canada

Received September 14, 1995[⊗]

An artificial neural network (ANN) has been trained to recognize the presence or absence of specific structural features (SF) in trimethylsilyl derivatives of organic acids from their mass spectra. The input vector is constructed without knowledge of the molecular ion, which is generally not observed in the spectra of these compounds. The results are used in conjunction with a classical search in a spectral library to identify organic acids in biological fluids for rapid acidemias diagnosis.

INTRODUCTION

Gas chromatography/mass spectrometry (GC/MS) is a powerful analytical technique to detect abnormal compounds in physiological fluids providing valuable information for diagnosis of rare, often critical metabolic disorders.[1] In the case of organic acidemias, an important group of natural errors of metabolism, rapid identification of the metabolites is required for a prompt diagnosis, which is essential to prevent death or irreversible retardation.[2]

Organic acidemias arise from enzymatic deficiencies which result in the abnormal accumulation of organic acids in physiological fluids. These acids originate from normal metabolism of carbohydrate and lipids, from pathological metabolism of amino acids and biological amines, and also from the metabolism of drugs. As far as metabolic dysfunction is concerned, we are only interested in organic acids in physiological fluids. To date, 250 organic acids have been identified for diagnosis of diseases, and the number of identified diseases is increasing rapidly. These acidic metabolites are mostly fatty acids, dicarboxylic acids, and phenolic acids. The study of these organic acids by GC/MS has permitted the identification of a great number of acidemias because it has facilitated the elucidation of the structure of the metabolites.[3]

Automatic identification by searching computerized mass spectral databases is routinely used, but in some cases manual interpretation remains necessary. In this context, Artificial Neural Network (ANN) can be of great help to the analyst by providing reliable information about the presence or absence of specific structural feature (SF) in the unknown compounds. Such an ANN has been trained to recognize 12 SFs from the mass spectra. The results of the training and testing of the ANN are critically evaluated and an example of practical use is presented here.

METHOD

Chemical Preparation of the Sample. In the case of metabolic dysfunction, organic acids accumulate in the plasma and are excreted in abnormal quantities in the urine. Before analysis by CG/MS several steps are necessary to convert the acids into a suitable form for separation by gas chromatography. The organic acids must be extracted from the aqueous phase because their conversion into volatile compounds must be completed in an anhydrous phase. The urine sample is acidified, and organic compounds are extracted by ether. The extracts are dried by sodium sulfate under a stream of nitrogen. As organic acids are highly polar compounds with low volatility, derivatization is necessary before separation by gas chromatography. Trimethylsilyl derivatives are easily prepared using bistrimethylsilylacetamide (BSTFA) as the reagent, resulting in the silylation of the carboxyl, hydroxyl, and phenolic functional groups.

CG/MS Analyses. Gas chromatographic separation is achieved on apolar fused-silica (DB1 or DB5) capillary column ($L = 30$ m, $\Phi = 0.25$ mm). Mass spectra are recorded on a Finnigan MAT Incos 500E GC/MS system. The use of Kovats indices alone do not allow reliable diagnosis. It is therefore necessary to make the structural identification of the trimethylsilylated organic acids present in the sample with the aid of mass spectrometry.

Analysis of the Mass Spectra by ANN. The first step in the identification of each compound consists of a direct search of specific compounds in the spectral library. Although simple and attractive, this approach may fail. Obviously it is not applicable if the compound has not been identified in advance and its spectra is not recorded in the library. Unfortunately this situation is not always easy to recognize since the indices used by the searching algorithm to rank the similarity between the spectra of reference compounds and of unknown compounds may lead to very small differences, even for spectra of different compounds. In some cases, the similarity indices calculated for the spectra of different unknown compounds and one reference spectrum in the library are so close that no reliable discrimination can be achieved. As a consequence, careful examination of the hit list produced by the library search algorithm is absolutely essential, and in many cases it remains necessary for the analyst to interpret the mass spectra in order to elucidate an unknown structure. In these circumstances the use of ANN, combined with other taxonomic methods, can provide

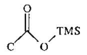
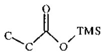
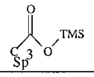
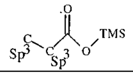
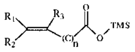
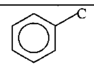
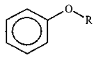
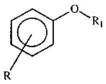
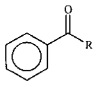
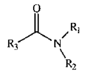
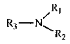
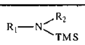
[*] Author to whom inquiries about the paper should be addressed.
[†] LARTIC, Université de Nice-Sophia Antipolis.
[‡] Chimie organique du fluor, Université de Nice-Sophia Antipolis.
[§] Dalhousie University.
[⊗] Abstract published in *Advance ACS Abstracts,* April 1, 1996.

S0095-2338(95)00129-6 CCC: $12.00

valuable assistance in the structure determination.[4−7] Although the application of ANN to identify specific SFs, fragments, or substructures from spectroscopic information can be routinely carried out on a microcomputer, subject to the availability of specific software,[8,9] the prior construction, training, and testing of ANN requires more powerful hardware and software resources. Furthermore presently there is no well established methodology for ANN development in real practical problems, consequently one has to adopt an empirical approach to choose the architecture, learning method, and input and output vectors which best fit the problem at hand.

Architecture of the Neural Network. A classical multilayered feed-forward (MLF) neural network[10−12] is constructed using the ASPIRIN software[13] from the MITRE Signal Processing Laboratory, running on an IBM RS 6000 workstation. The usual logistic function $1/(1 + e^{-x})$ is used as the activation function. The input layer is made of 199 units which receive information extracted from the mass spectra (see infra: input vector). The number of units in the hidden layer has been varied from 15 to 45, and the number is retained at 36 as a good compromise between the quality of adjustment during training and the network complexity. The size of the output layer is established by the number of SFs to be recognized by the ANN. In the present study a set of 12 SFs (see Table 1) has been selected on the basis of (a) their significance for the identification of trimethylsilylated organic acids, (b) their statistical distribution in the collection of compounds used in training and testing, and (c) the fact that known fragmentation patterns suggest that they could be recognized from the mass spectra.

Table 1. Definition and Occurrence of the Structural Features in the Training and Test Sets

Output	SF	Remarks	Training Set (464)	Test Set (298)
SF1		TMS ester on carbon atom having any hybridization, possibly bearing a functional group (with the exception of ethylenic bond)	**200**	**139**
SF2		TMS ester on a chain of at least two carbon atoms having any hybridization, possibly bearing a functional group (with the exception of ethylenic bond)	**117**	**101**
SF3		TMS ester on a carbon atom hybridized sp3 bearing no functional group (with the exception of TMS amine and TMS oxide and TMS ester)	**157**	**114**
SF4		TMS ester on a ethane unit bearing no functional group (with the exception of TMS amine and TMS oxide and TMS ester)	**100**	**85**
SF5		Ethylenic bond in a TMS ester (n≥0)	**96**	**24**
SF6			**211**	**94**
SF7		Unsubstituted phenoxy group	**137**	**68**
SF8		Substituted phenoxy group	**71**	**26**
SF9			**84**	**21**
SF10		All types of amide groups	**104**	**50**
SF11		All types of amine groups	**75**	**15**
SF12		All types of TMS amines	**91**	**70**

Input Vector. Common, low-mass ions below 40 are not very diagnostic, so the input vector is composed of the intensities of peaks with $40 \leq m/z \leq 220$ augmented by a series of intensities corresponding to mass-losses. One difficulty encountered in the extraction of significant information to calculate these losses originates from the fact that the molecular ion is rarely observed in the case of trimethylsilylated organic acids. Because we are interested only in recognizing parts of compounds, the information from the fragmentation comes at the lower end in the charged fragments and from the rest of the spectrum in the form of neutral losses. Dealing only with high energy electron impact spectra, the majority of the neutral fragments that are lost are even-electron species, except for those lost from the molecular ion. These may be odd-electron species and less commonly even-electron species. If the molecular ion were known, then one could handle the losses from the molecular and the daughter ions. In our case, the following scheme for counting the neutral losses ignoring this information has been used: (a) ignore peaks with intensity below 1%; (b) ignore $M + n$ peaks that do not have an intensity greater than the expected isotopic intensity; (c) for all remaining peaks, find the difference with all possible daughter ions, ignoring losses that come from peaks than are very different in intensity (16×) and those for which the value m/z of the daughter is greater than half of the parent; (d) sum the pair intensities; retain the most intense and increment by 25% for each additional loss of the same mass; (e) the following differences are retained to be incorporated as elements of the input vector: [16, 18, 26, 27, 28, 30, 42, 43, 44, 60, 71, 72, 73, 90, 116, 117, 132, 144, 216]. The above algorithm is somewhat arbitrary, but it is to some extent chosen on the basis of examination of the pattern produced when applied to our cases. It gives a greater difference in the pattern as compared to many others we have tried, but no real optimization has been undertaken. Thus, the resulting input vector is made of 180 peaks plus 19 losses giving 199 units in the input layer.

Output Vector. The output vector is generated from the known structure by giving the real value of 0.0 to each vector element when the associated SF is not present in the structure, and the value of 1.0 when present.

Training and Testing. A collection of 762 spectra of trimethylsilyl compounds has been extracted from the NBS mass spectral library, from which 464 have been chosen randomly to constitute the training set, the 298 remaining being used for testing. The occurrences of the retained SFs in the two sets are given in Table 1.

The training is monitored by following the decrease of the root mean square residuals Sr calculated globally for all SFs (Sr_g) and individually for each one (Sr_f)

$$Sr_{\mathrm{g}} = \sqrt{\frac{\sum_{i=1}^{nt}\sum_{f=1}^{nf}(O_{if} - T_{if})^2}{nf \cdot nt}}$$

$$Sr_{\mathrm{f}} = \sqrt{\frac{\sum_{i=1}^{nt}(O_{if} - T_{if})^2}{nt}}$$

where nt, nf, O_{if}, and T_{if} are, respectively, the number of

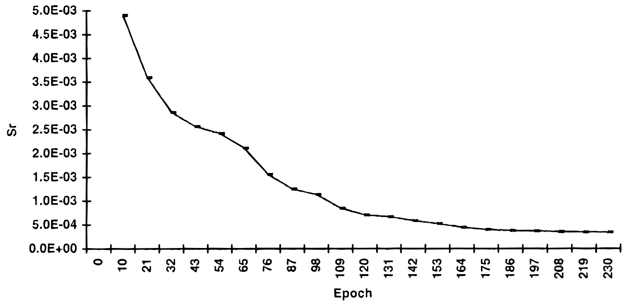


Figure 1. Variation of *Sr* during the learning phase.

examples in the set, the number of SFs, the response of network, and the desired response. Indice *i* refer to the compound and *f* to the SF.

For each 5000 iterations (one iteration corresponds to one presentation of a pattern to the network and a correction of the weights), the weight matrix is dumped into a file for future use in the testing phase. The classical backpropagation method[14] gives an oscillatory behavior during the learning phase, while a regular decrease of the global error is observed using the conjugate gradient method of optimization with line search[15] (Figure 1). The initial learning rate and inertia was set to 0.1 but is adapted after an entire presentation of the training set (epoch) by the line search method. Training was stopped after a fixed number of epochs, chosen large enough (230) to be sure that convergence is achieved.

It is well-known that neural networks can lead to overfitting, resulting in poor performance in prediction when applied to examples which are not included in the training set. Therefore the weight matrices stored during the learning phase are used with the test set to evaluate the variation of *Sr* on such "new" examples. The resulting graphs (see Figure 2) show that for each SF the values of *Sr* decreases rapidly to reach a minimum at approximately 50 epochs, the exact number varying with the SF; afterward they increase and stabilize for 85 epochs.

Thus, one could conclude, as many authors did in similar studies, that the best performances in prediction was obtained by using the network for which training was stopped at the minimum of *Sr* (50 epochs in the present study). In fact, things are not so simple, because *Sr* being a global index does not make any distinction between residuals obtained for compounds with (p) and without (a) the specific SF. For a better view of the network outputs it is judicious to calculate separately *Srp* and *Sra* for the two classes of compounds globally and for each SF_f

$$Srp_f = \sqrt{\frac{1}{np_f} \cdot \sum_{i=1}^{nt} (T_{if})(O_{if} - T_{if})^2}$$

$$Sra_f = \sqrt{\frac{1}{na_f} \cdot \sum_{i=1}^{nt} (1 - T_{if})(O_{if} - T_{if})^2}$$

with n_{pf} and n_{af} being, respectively, the number of presents and absents for SF_f.

The resulting curves (see Figure 3) show that the network learns differently to recognize the presence or the absence of each SF. Sometimes it adjusts its outputs on compounds with the SF, relaxing simultaneously its outputs for compounds without the SF, sometimes the opposite. The sole observation of the smooth decrease of *Sr* during training hides the oscillatory behavior of the network before a stabilization phase observed after 150 epochs approximately.

In order to provide a better basis for choosing the best state of network training for differentiating between compounds with or without a SF, we have introduced a discrimination index *DI*[16]

$$DI = \frac{(m_p - m_a)}{\sqrt{\frac{(n_p \cdot v_p + n_a \cdot v_a)}{nt}}}$$

with m_p = mean of neural output for presents; m_a = mean of neural output for absents; v_p = variance of neural output for presents; v_a = variance of neural output for absents; n_p = number of presents and n_a = number of absents. The larger this index is, the better the discrimination.

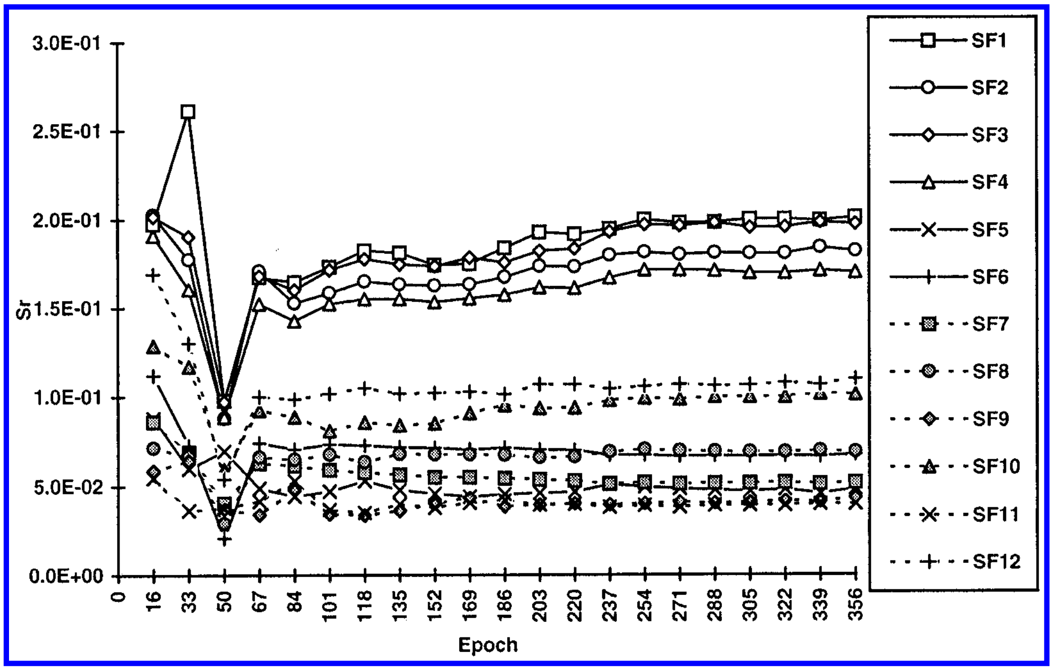
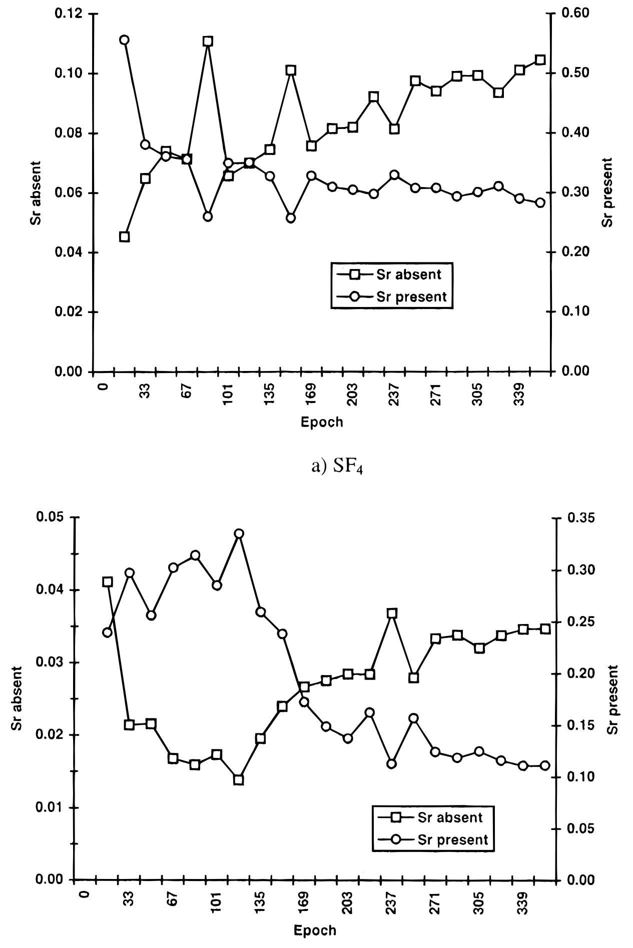


Figure 2. Variation of *Sr* on test for each structural feature with the number of learning cycles.



a) SF_4

b) SF_7

Figure 3. Variation of *Sra* and *Srp* for two structural features on test with the number of learning cycles: (a) SF_4 and (b) SF_7.

Plotting the value of *DI* calculated on the test set for each SF_f against the number of epochs reveals that the best discrimination is obtained for a training of 185 epochs; afterward a slight deterioration of the performance is noticeable. In the following we present the results obtained with the network in this optimal training state (see Figure 4).

Table 2. Statistical Results Obtained on the Test Set

	RL	AL	Rr (%)	Qr (%)	Qrp (%)	Qra(%)	EQr (%)
SF_1	0.377	0.377	100	77	72	83	54
SF_2	0.478	0.643	97	79	68	85	49
SF_3	0.770	0.841	98	81	76	85	58
SF_4	0.455	0.667	99	83	71	89	54
SF_5	0.479	0.643	97	95	70	97	56
SF_6	0.622	0.622	100	92	90	92	82
SF_7	0.437	0.602	98	94	89	96	80
SF_8	0.467	0.491	99	93	57	98	55
SF_9	0.577	0.789	97	96	75	98	56
SF_{10}	0.509	0.579	99	90	78	91	63
SF_{11}	0.397	0.562	98	96	64	98	57
SF_{12}	0.622	0.622	100	87	85	88	66

Classification. The purpose of the network is to help in the determination of the presence or absence of specific SF_f. Thus, for each SF_f the calculated output real values O_f ranging from 0.0 to 1.0 have to be converted to a logical answer. For this purpose, for each SF_f we define two threshold levels: RL_f (reject level) and AL_f (acceptance level), such that when $O_f \leq RL_f$ the corresponding SF_f is reported as absent in the compound, when $AL_f \leq O_f$ it is reported as present, and when $RL_f < O_f < AL_f$ it is reported as unclassified. The values of RL_f and AL_f are determined by minimization of a cost function taking into account the number of correct minus the number of false classifications calculated on the test set.[17] Thus it is valuable to evaluate the network's performance in classification, using indicators calculated on the basis of the logical answers obtained after application of the above rules to the calculated outputs. The values of the following indicators calculated for each SF_f for the network after 185 epochs of training are reported in Table 2. *Rr* (response ratio) is the ratio of the number of classified examples to the total number of examples. *Qr* (global response quality) is the ratio of the number of correct responses to the number of responses given by the network excluding the unclassified cases. *Qrp* and *Qra* (quality of present and absent responses) is the same indicator as *Qr* but applied only to the absent (respectively present) responses. It is clear that these indices are highly dependent on the composition of the studied population. Overoptimistic conclusions could easily be drawn if one does not take into

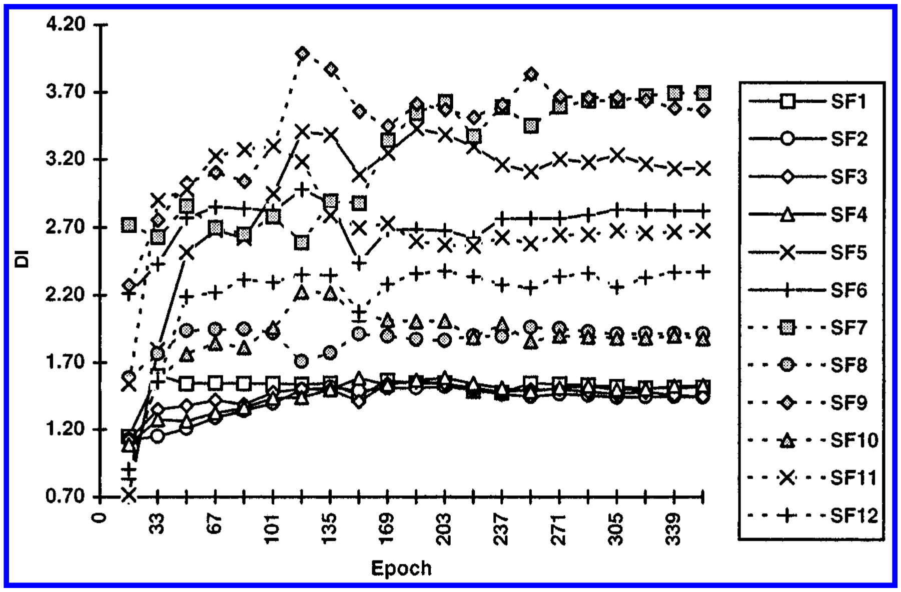


Figure 4. Variation of *DI* on test for each structural feature with the number of learning cycles.

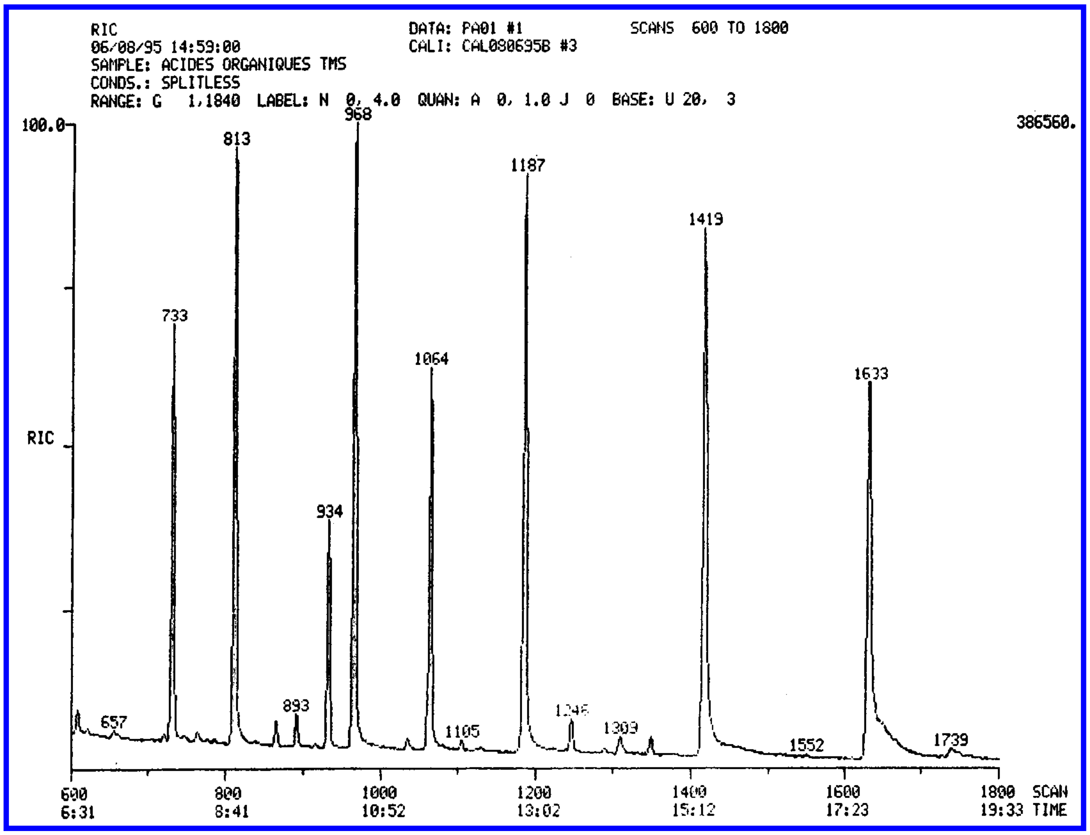


Figure 5. A typical chromatogram.

consideration the probability of correct classification by chance. To avoid this pitfall we have introduced a new index EQr_f (extra-statistical quality) representing a measure of the improvement in quality of the classification achieved by the network over chance-correct classification[16] for SF_f.

$$EQr_f = \frac{(GQ - St_f)}{(1 - St_f)}$$

$$St_f = 1 - 2 \cdot P_f + 2 \cdot P_f^2$$

$$P_f = \frac{np_f}{np_f + na_f}$$

It is defined as the ratio of the observed improvement to the maximum possible improvement which is itself calculated on the basis of the representation (P_f) of the SF_f in the test set. Statistical results obtained for the training set show a near perfect classification (over 99.9% correctly classified). The results that were obtained for the test set are presented in Table 2.

Examination of the values reported in Table 2 shows that the response ratio and the quality of the responses are very high for all investigated SF. In all cases the improvement over chance by the ANN is significant as indicated by the values of *EQr*. However, as a general trend, the responses for the recognition of the absence of any SF are more reliable than those for the presence (e.g., $Qra > Qrp$).

APPLICATION

In order to exemplify the potential of ANN in metabolite identification we report a typical study. A chromatogram (see Figure 5) obtained from a urine sample consists of dozens of peaks from which the eight most intense have to be identified. The mass spectral library search allows automatic identification of all these peaks (see Table 3); however, a problem remains because two of them (scan 733 and 968) are both assigned to the same TMS diester, i.e., methyl propanedioic bis(trimethylsilyl) ester, by the search.

Table 3. Assignment by Library Search of the Eight Most Intense Chromatographic Peaks

scan	intensity	library assignment	fit
733	72	methyl propanedioic acid, bis(TMS) ester	994
813	95	3,6,9-trioxa-2,10-disilaundecane,2,2,10,10-tetramethyl	989
934	40	3,7-dioxa-2,8-disilanonane,2,2,8,8-tetramethyl-5-(trimethylsilyl)ox	993
968	100	methyl propanedioic acid, bis(TMS) ester	957
1064	62	2-butenedioic acid (Z)-, bis(TMS) ester	960
1187	85	pentanedioic acid, bis(TMS) ester	959
1419	80	hexanedioic acid, bis(TMS) ester	998
1633	56	heptanedioic acid, bis(TMS) ester	997

As a matter of fact, the two mass spectra (see Figures 6 and 7) are very similar; however, the experienced spectroscopist can confirm the assignment of spectra 733 to methyl propanedioic bis(trimethylsilyl) ester and elucidate the spectra 968 as being succinic bis(trimethylsilyl) ester by comparison of small peaks at $m/z = 129$, 133, 172, and 218. Application of our neural network to these spectra lead to the results presented in Table 4.

With the exception of SF_4, small differences in the output numerical values are obtained for the various SF_f; concerning the logical answers the only difference is observed for SF_4 which is indicated as absent in spectra 733 and present in spectra 968. The latter answer indicates that the compound corresponding to spectra 968 must have at least two carbon

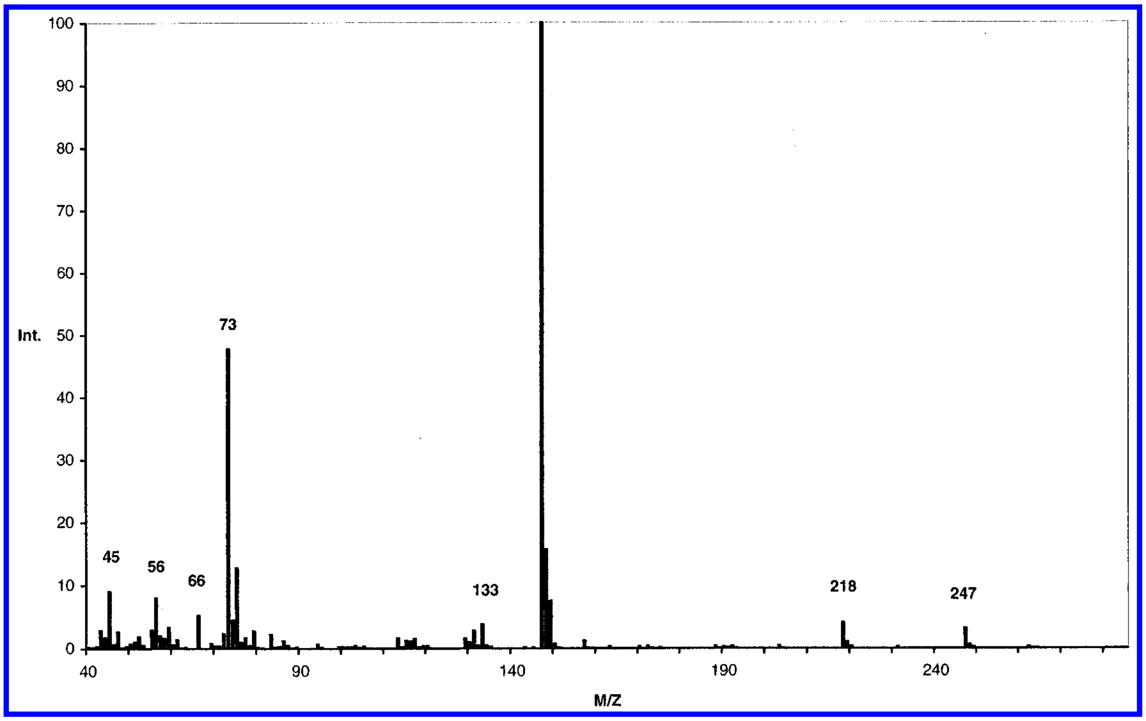


Figure 6. Mass spectrum number 733.

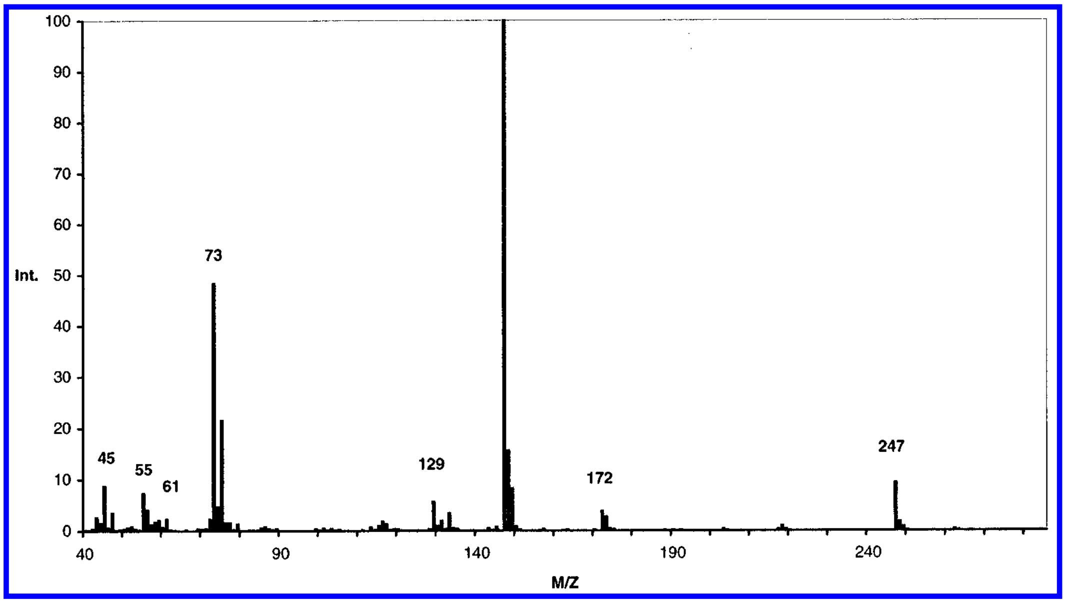


Figure 7. Mass spectrum number 968.

Table 4. Outputs and Answers given by the ANN for Spectra 733 and 968[a]

scan		SF_1	SF_2	SF_3	SF_4	SF_5	SF_6	SF_7	SF_8	SF_9	SF_{10}	SF_{11}	SF_{12}
733	output	0.999	0.821	0.886	0.200	0.006	0.035	0.000	0.000	0.000	0.051	0.001	0.018
	answer	P	P	P	A	A	A	A	A	A	A	A	A
970	output	0.972	0.913	0.975	0.781	0.007	0.001	0.000	0.000	0.001	0.004	0.001	0.029
	answer	P	P	P	P	A	A	A	A	A	A	A	A

[a] P = present and A = absent.

atoms between the two TMS ester groups giving a valuable hint for the structure determination.

It is to be noted again that although building, training, and testing ANNs requires lengthy computations and cautions, their applications in the analytical laboratory can be done routinely on commonly available microcomputers. The spectra exported as JCAMP files are automatically processed by a specialized program ANNE which makes use of description file to construct the input vector from it.[9] The ANN answers are displayed in a multiwindow environment which allows the analyst to take decisions by combining several sources of information when available.

CONCLUSION

Rapid diagnosis of acidemias by identification of organic acids in physiological fluids can be achieved routinely using

GC/MS and conventional search in spectral databases. However, there are cases for which manual structure elucidation from the spectra remains necessary. Artificial neural networks can be trained to recognize the presence or absence of specific SF in an unknown metabolite from the mass spectra. Once trained these networks can be used directly in the analytical laboratory providing valuable help to the spectroscopist.

REFERENCES AND NOTES

(1) Jellum, E.; Kvittingen, A. A.; Stokke, O. Mass Spectrometry in Diagnosis of Metabolic Disorders. *Biomed. Environ. Mass Spec*. **1988**, *16*, 57−62.

(2) Williams, J. C. Organic Acidurias in Children. *Spectra* **1985**, *10,* 2, 4−6.

(3) Jellum, E. Profiling of human body fluids in healthy and diseased states using gas chromatography and mass spectrometry with special reference to organic acids. *J. Chromatogr.* **1977**, *143*, 427−462.

(4) Curry, B.; Rumelhart, D. E. MSnet: A Neural Network That Classifies Mass Spectra, *Tetrahedron Comput. Method.* **1990**, *3*, 3/4, 213−237.

(5) Lohninger, H. Classification of Mass Spectral Data Using Neural Networks *Software development in Chemistry*; Gmehling J., Ed.; Springer Verlag: Berlin Heidelberg, 1991; Vol. 5, pp 159−168.

(6) Lohninger, H.; Stanel 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**, *334*, 186−189.

(7) Werther, W.; Lohninger, H.; Stanlcl, F.; Varmuza, K. Classification of Mass Spectra—A Comparison of Yes/No Classification Methods for the Recognition of Simple Structural Properties. *Chemom. Intell. Lab. Syst*. **1994**, *22*, 1, 63−76.

(8) Van Est, Q. C.; Schoenmakers, P. J.; Smits, J. R. M.; Nijssen, W. P. M. Practical Implementation of Neural Networks for the interpretation of Infrared Spectra. *Vib. Spectrosc.* **1993**, *4*, 263−272.

(9) Cabrol-Bass, D.; Cachet, C.; Cleva, C.; Eghbaldar, A.; Forrest, T. P. Application pratique des réseaux neuro mimétiques aux données spectroscopiques (infrarouge et masse) en vue de l'élucidation structurale. *Can. J. Chem.* **1995**, *73*, 1412−1426.

(10) Zupan, A.; Gasteiger, J. Neural networks: A new method for solving chemical problems or just a passing phase. *Anal. Chim. Acta* **1991**, *248*, 1−30.

(11) Lacy, M. E. Neural Network technology and its Application in Chemical Research. *Tetrahedron Comput. Method.* **1990**, *3*,3/4, 119−128.

(12) Jansson, P. A. Neural Networks: An Overview. *Anal. Chem.* **1991**, *63*,6, 357A−361A.

(13) Leighton R. R. The Aspirin/Migraines Software Tools. Technical Report MP-91 W00050; The MITRE Corp.

(14) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning Representations by Back-Propagating Errors. *Nature* **1986**, *323*, 533−536.

(15) Leonard, J.; Kramer, M. A. Improvement of the Backpropagation Algorithm for Training Neural Networks. *Comput. Chem. Eng.* **1990**, *14*, 3, 337−341.

(16) Sbirrazzuoli, N.; Cachet, C.; Cabrol-Bass, D.; Forrest T. P. Indices for the Evaluation of Neural Network Performance as Classifier: Application to Structural Elucidation in Infrared Spectroscopy. *Neural Comput., Applic.* **1993**, *1*, 229−239.

(17) Ricard, D.; Cachet, C.; Cabrol-Bass, D.; Forrest, T. P. Neural Network Approach to Structural Feature Recognition form Infrared Spectra. *J. Chem. Inf. Comput. Sci.* **1993**, 33, *2*, 202−210.

CI9501296