Carbohydrate Research, 207 (1990) C1–C5
Elsevier Science Publishers B.V., Amsterdam

Preliminary communication

Identification of the mass spectra of partially methylated alditol acetates by artificial neural networks

Jeffery Sellers, William York, Peter Albersheim, Alan Darvill, and Bernd Meyer*
Complex Carbohydrate Research Center and Department of Biochemistry, University of Georgia, 220 Riverbend Road, Athens, GA 30602 (U.S.A.)

(Received May 31st, 1990; accepted in revised form July 20th, 1990)

Partially methylated alditol acetates (PMAAs) are used as key glycosyl-residue derivatives in the structural characterization of oligo- and poly-saccharides. These derivatives are most readily separated by gas-liquid chromatography (g.l.c.) and identified by electron-impact mass spectroscopy (e.i.-m.s.). Artificial neural networks were originally created to model biological neural networks. Continued development of the underlying mathematical processes have produced networks that can be trained to respond to specific inputs by activating the corresponding output neurons. We have used the mass spectra of the PMAAs of xylitol, arabinitol, rhamnitol, and fucitol to determine whether artificial neural networks can identify mass spectra (*cf.* Fig. 1 and compounds **1–22**). Previous work in this laboratory has shown that artificial neural networks are capable of recognizing one-dimensional ^{1}H-n.m.r. spectra of alditols and of complex carbohydrates[1,2]. We now report the successful application of artificial neural networks to the recognition of the e.i.-mass spectra of PMAAs.

We have been working with feed-forward neural networks with back-propagation of errors[3] to recognize mass spectra (Fig. 2). The network was trained by presenting to the input neuron layer the patterns (mass spectra) to be distinghuished and by defining the desired response of the output layer, that is, by assigning each of the 22 output neurons to one of the 22 PMAAs used in the training set. Connection strengths between the input and hidden layer neurons and between the hidden and output layer neurons were modified by the network as it used the back-propagation learning algorithm to minimize the output errors[4]. This iterative process continued until the network activated the correct output neuron for each PMAA. An output neuron activation level of 0.9 ± 0.1 was defined to represent the "on" state which identified the e.i.-mass spectrum (input pattern) of a specific PMAA. An output neuron activation level of 0.1 ± 0.1 was defined to represent the "off" state. Following the training phase, network responses to new spectra which gave activation levels between 0.2 and 0.8 were considered to be indicative of related PMAAs but did not represent a definitive

* Author to whom inquires should be directed.

0008-6215/90/$ 03.50

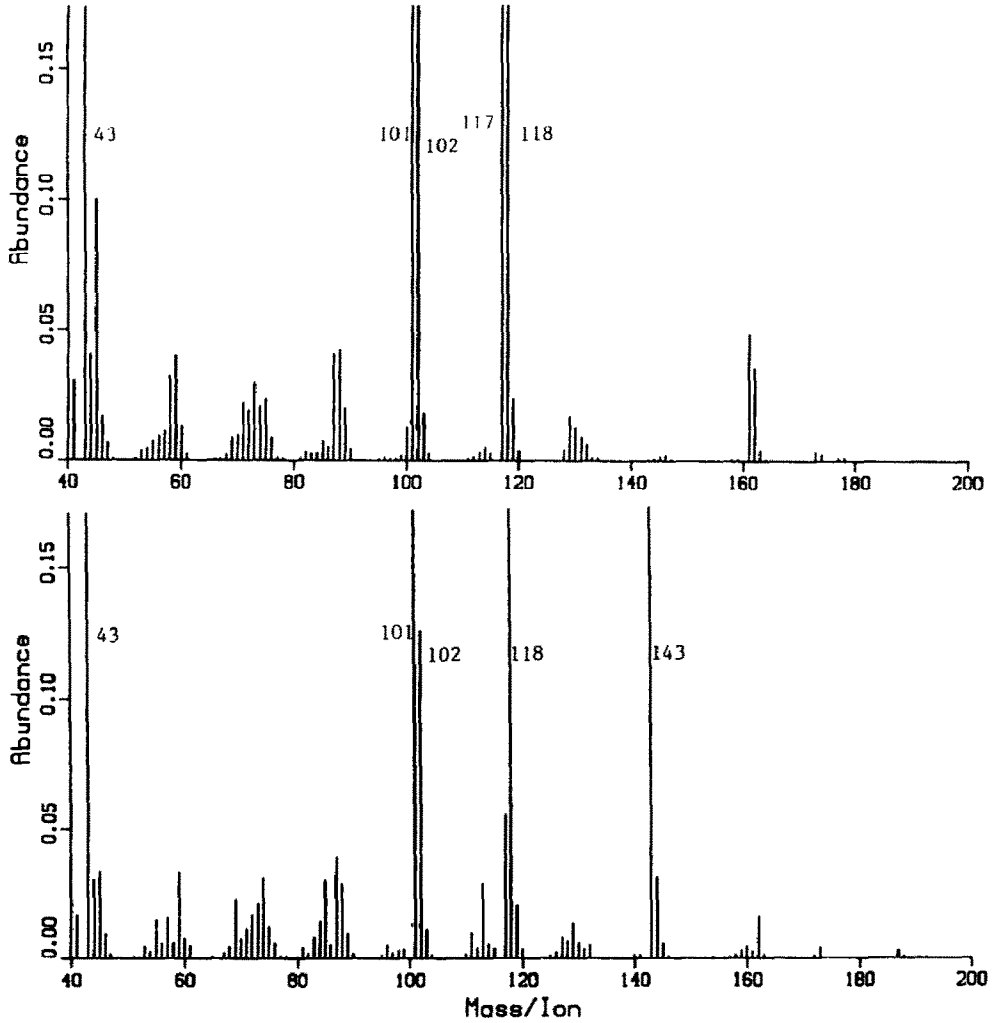
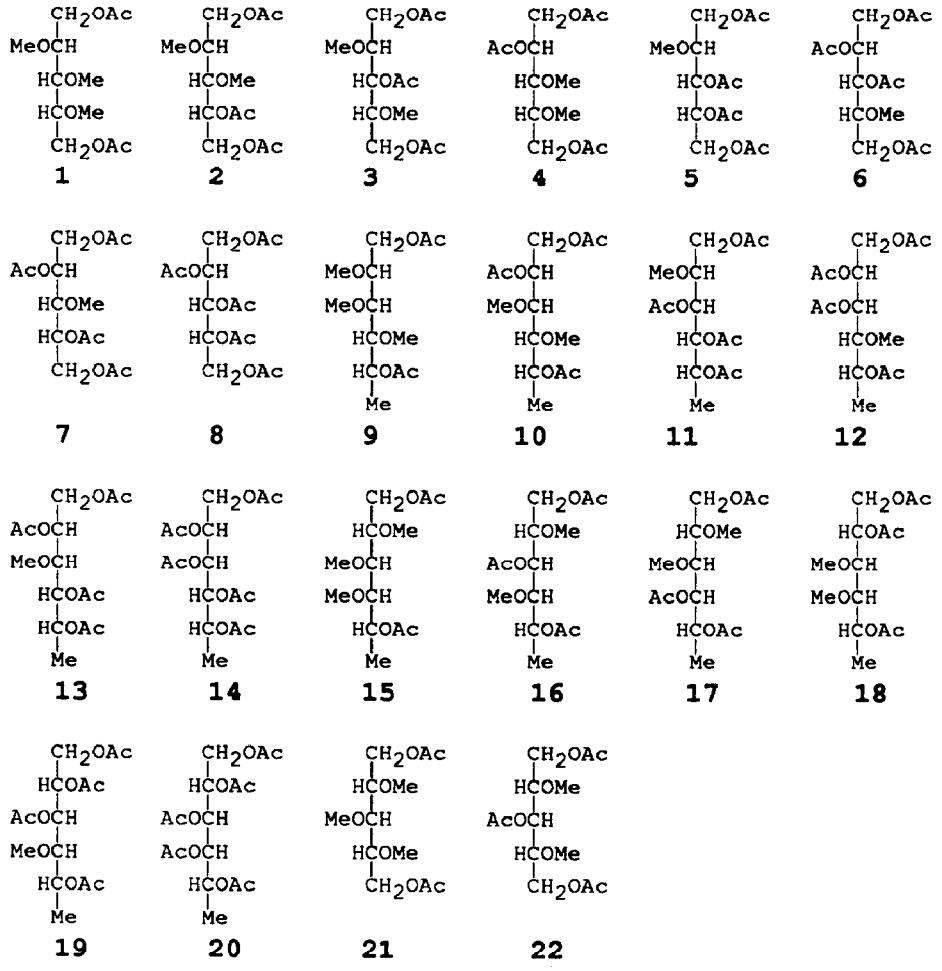


Fig. 1. The electron-impact mass spectra of **1** (top) and **17** (bottom), which are typical of the spectra used in training the neural network. Only m/z from 40–200 are shown. The y-scale has been expanded to emphasize the smaller signals.

identification. The back-propagation program of McClelland and Rumelhart[5] was used to train a neural network consisting of an input layer of 400 neurons, a hidden layer of 25 neurons, and an output layer of 22 neurons.

A Hewlett–Packard model 5890 g.l.c. instrument with an H.–P. 5970 mass-selective detector was used for separation and e.i.-m.s. analysis of the PMAAs. The spectra were recorded, stored on an H.–P. 9000/200 workstation, and subsequently transferred to a DECstation 3100 for further processing. In order to use the spectra as input ot the neural network, all data within each spectrum were normalized relative to the largest mass/charge (m/z) peak in the spectrum. Each m/z ratio was rounded to an



Scheme 1. Structures of partially methylated alditol acetates: arabinitols **1–8**, rhamnitols **9–14**, fucitols **15–20**, and xylitols **21–22**.

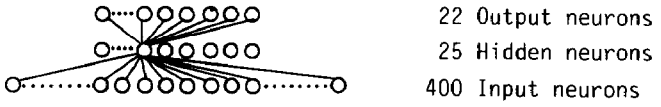


Fig. 2. Diagram of a feed-forward neural network model. The connections of a single hidden neuron are illustrated. Each of the 25 hidden layer neurons are connected to every neuron in both the input and output layers. Signals are fed from the input neurons through the hidden layer and on to the output neurons. During training, error signals based on the difference between the desired output and the actual output are propagated from the output neurons back through the hidden layer to the input neurons. These signals are used, in an iterative process, to modify the connection strengths so that each spectrum evokes a unique output neuron activation pattern. After training, the recognition of a spectrum involves presenting the spectrum to the input neurons, feeding the signal through the network using the connection strengths, and calulating the output neurons' response.

integer number. An input pattern was then generated by mapping the normalized abundance for each mass-to-charge ratio to the corresponding input neuron, *i.e.*, *m/z* 1–400 for the 400 input neurons. Spectra typical of those used in this study are shown in Fig. 1.

Four solutions, each containing PMAA derivatives of either xylitol, arabinitol, rhamnitol, or fucitol, were used in this study. Twenty-two well-resolved PMAAs were selected from the chromatographic profiles of the four PMAA solutions. Several mass spectra were obtained as each PMAA eluted from the capillary column. Mass spectra with a total-ion current of at least 30% of the maximum total-ion current observed for that PMAA were used to train the neural network. This resulted in a training set of 66 input patterns for the 22 different PMAAs.

The 66 input patterns in this training set were used to train the neural network to distinguish each PMAA from the other PMAAs. The 22 PMAAs included six pairs of epimeric molecules (*i.e.*, compounds **1** and **21**, **3** and **22**, **9** and **15**, **10** and **18**, **12** and **19**, and **14** and **20**). Each presentation of the 66 input patterns required a computing time of 10 s to calculate the network response and to update the connection strengths. After completion of the training period, presentation of an individual pattern to the network resulted in the identification of the correct PMAA. After 670 presentations of the entire training set, the network had a root mean square (RMS) error of 0.03, indicating good agreement between the desired response and the actual output of the network. At this level of training, the maximum deviation of the network response from the target output was 0.19, which exceeded the maximum deviation of 0.1 defined earlier. This deviation occurred when an input pattern from the epimeric pair of **14** and **20** was presented to the network. By training the neural network to an RMS value of 0.0072 (approximately 2000 presentations), the maximum deviation from the target output was less than 0.04, indicating much better agreement between the target and actual outputs. That is, each PMAA activated its associated output neuron to a level between 0.86 and 0.94, while all other neurons were activated to a value between 0.06 and 0.14.

In testing the neural network with patterns from the training set, the correct answer was given for all mass spectra presented. The mass spectra of the six epimeric pairs were also correctly identified, indicating the neural network could extract from the mass spectra stereochemical information not routinely detected by scientists. Until now, scientists have utilized g.l.c. retention times to distinguish between the mass spectra of PMAAs that are stereoisomers. It is possible to incorporate retention-time data into the neural network analysis should that prove to be of value. Computing the response of the trained artificial neural network for the m.s. of any PMAA required less than 0.1 s.

The ability to identify PMAAs as well as to discriminate between PMAA stereoisomers from their e.i.-m.s. demonstrates the powerful capabilities of artificial neural networks in spectral analysis. Recent results have shown neural networks are able to recognize m.s. not included in the training set[6]. The effects of instrumental variations on the ability of neural networks to identify mass spectra of PMAAs, particularly of stereoisomers, are being studied. We conclude from our preliminary results that neural

networks can be trained to identify all possible partially methylated alditol acetates. While our efforts have been focused on PMAAs, neural networks capable of interpreting g.l.c.–m.s. data have broad applications including metabolite studies, environmental trace analyses, and assays of biological samples. The advantages of using a neural network for the analysis of data, rather than using a peak-matching library search, are the ease with which the neural network can be tailored to the needs of the researcher and the speed with which the knowledge base of the neural network will give an answer. An important characteristic of the neural approach is that only the training process is computationally intensive. Using an already trained network, a small personal computer can readily carry out the identification process in seconds. Furthermore, neural network analysis of spectra does not rely on human definition of the possible deviations within the data from one experiment to the next, but these differences are automatically extracted by the neural network from the training set.

ACKNOWLEDGMENTS

This work was supported in part by the U.S. Department of Energy grant DE-FG09-85-ER13424; by the USDA/DOE/NSF Plant Science Centers program with this project funded by the U.S. Department of Energy grant DE-FG09-87ER13810; by Digital Equipment Corporation (External Research Agreement 768); and by the Advanced Computational Methods Center of The University of Georgia.

REFERENCES

1 J. Thomsen and B. Meyer, *J. Magn. Reson.*, 84 (1989) 212–217.
2 B. Meyer, T. Hansen, D. Nute, P. Albersheim, A. Darvill, W. York, and J. Sellers, unpublished results.
3 D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, *Parallel Distributed Processing*, Vol. 1, MIT Press, Cambridge, MA, 1986, pp. 318–362.
4 D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, *Parallel Distributed Processing*, Vol. 1, MIT Press, Cambridge, MA, 1986, pp. 290–299.
5 J. L. McClelland and D. E. Rumelhart, *Explorations in Parallel Distributed Processing*, MIT Press, Cambridge, MA, 1988, pp. 121–159.
6 J. Sellers, W. York, P. Albersheim, A. Darvill, and B. Meyer, unpublished results.