Reports

Identification of the ^{1}H-NMR Spectra of Complex Oligosaccharides with Artificial Neural Networks

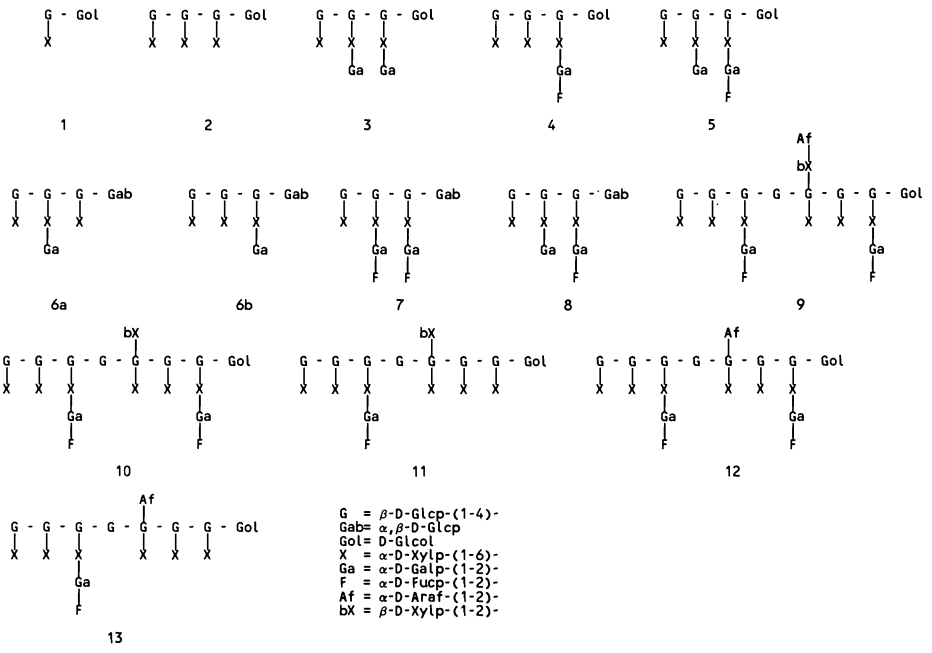
BERND MEYER,* TORBEN HANSEN, DONALD NUTE, PETER ALBERSHEIM, ALAN DARVILL, WILLIAM YORK, JEFFERY SELLERS

Artificial neural networks can be used to identify hydrogen nuclear magnetic resonance (^{1}H-NMR) spectra of complex oligosaccharides. Feed-forward neural networks with back-propagation of errors can distinguish between spectra of oligosaccharides that differ by only one glycosyl residue in twenty. The artificial neural networks use features of the strongly overlapping region of the spectra (hump region) as well as features of the resolved regions of the spectra (structural reporter groups) to recognize spectra and efficiently recognized ^{1}H-NMR spectra even when the spectra were perturbed by minor variations in their chemical shifts. Identification of spectra by neural network–based pattern recognition techniques required less than 0.1 second. It is anticipated that artificial neural networks can be used to identify the structures of any complex carbohydrate that has been previously characterized and for which a ^{1}H-NMR spectrum is available.

THE ELUCIDATION OF COMPLEX CARbohydrate structures often relies heavily on ^{1}H-NMR spectra, which provide a great deal of structural information from moderate amounts of sample (~100 nmol). Pattern recognition techniques could simplify the procedure of extracting structural information from ^{1}H-NMR spectra. Thomsen and Meyer have shown that artificial neural networks can recognize ^{1}H-NMR spectra of simple alditols (*1*). We have now extended the application of neural networks to the recognition of complex ^{1}H-NMR spectra of large, structurally repetitive oligosaccharides (Scheme 1). Only a small percentage of the signals of such complex ^{1}H-NMR spectra are separated into individually resolved NMR multiplets: most of the signals are contained in a region of strong overlap, called the hump region (Fig. 1). This type of NMR spectrum is typical of the spectra of many other biologically important molecules including DNA, RNA, and proteins. These molecules are composed of many closely related building blocks (nucleotides, amino acids, and glycosyl residues) that give rise to similar resonances in NMR spectra (*2*).

We chose the one-dimensional ^{1}H-NMR spectra of oligosaccharides derived from xyloglucan, a plant cell wall hemicellulose, to explore how well artificial neural networks can distinguish between such spectra (*3–5*). The structurally related molecules in the test set contained from 3 to 20 glycosyl residues (Scheme 1 and Fig. 1). Several residues in each xyloglucan oligosaccharide are identical; for example, **10** through **13** each contain eight β-(1-4)–linked glucosyl and six α-(1-6)–linked xylosyl residues. Thus, major portions of the NMR spectra of these compounds contain signals of repetitive residues that result in high degeneracies in their ^{1}H-NMR spectra. This type of spectrum should present the greatest challenge to pattern recognition techniques.

The 500-MHz ^{1}H-NMR spectra of compounds **1** through **13** used in this study had been recorded in our center (*3–5*). Thus, no special measures were taken when recording the spectra that were later used for training the artificial neural network as this use of the spectra was not anticipated. The free induction decay (FID) files of the spectra were retrieved from tape and Fourier-transformed without any preprocessing. The resulting spectra were normalized to a digital resolution of 0.5 Hz per point by interpolation. Furthermore, the chemical shifts of all spectra were normalized to the same standard (acetone at 2.225 ppm). Three spectral regions (1.15 to 1.34 ppm, 3.23 to 4.68 ppm, and 4.90 to 5.37 ppm) covering all of the signals in the ^{1}H-NMR spectra of **1** through **13** were combined into one pattern for each spectrum and presented to the input neurons of the neural network. The total width of these combined regions was 1056 Hz. The residual water signal at 4.75 ppm was replaced by zeros to avoid problems due to its greatly varying intensity and width. The signals within the spectra were normalized such that the intensities of all lines belonging to one selected proton summed to 1. This process required the identification within each spectrum of one completely resolved resonance of known multiplicity, usually one of the anomeric proton signals. This approach produced patterns of approximately the same relative intensity for presentation to the neural network. Examples of the neural network input patterns that were obtained after preprocessing the ^{1}H-NMR spectra are displayed in Fig. 1.



Scheme 1. Symbolic representation of the structures of **1** through **13**.

B. Meyer, P. Albersheim, A. Darvill, W. York, J. Sellers, Complex Carbohydrate Research Center and Department of Biochemistry, University of Georgia, Athens, GA 30602.
T. Hansen and D. Nute, Artificial Intelligence Programs, University of Georgia, Athens, GA 30602.

We used the commercially available feed-forward, back-propagation neural network software of McClelland and Rumelhart (*6, 7*) implemented on a Silicon Graphics 4D/220 GTX computer for the analysis of the spectra of xyloglucan oligosaccharides **1** through **13** (Scheme 1). Several artificial neural networks with different numbers of input, hidden layer, and output neurons were trained. The "on" state of an output neuron was defined as any neuron with an activation of 0.9 ± 0.1. The "off" state of an output neuron was defined as any neuron with an activation of 0.1 ± 0.1. The neurons with activation values between 0.2 and 0.8 were considered to be in an undefined state. One training cycle required approximately 8 s. Typically, 300 training cycles were required to obtain a root-mean-square (rms) error of 0.05.

The initial training set (TS-1) containing the NMR spectra of oligosaccharides **1** through **13** was used to train a neural network with 2113 input, 10 hidden layer, and 13 output neurons (NN-1). Each input neuron represented 0.5 Hz of the ^{1}H-NMR spectra (a total of 1056 Hz). After training the artificial neural network, we used plots of the weights that connect the input layer to the hidden layer to analyze the convergence properties of the neural network. Signals presented to the input layer of the neural network combine with positive weights to increase the activation of hidden layer neurons, and signals combine with negative weights to decrease the activation of hidden layer neurons. A weight plot obtained from the trained network (Fig. 2) represents the connection strengths between all input neurons and one of the ten hidden layer neurons. It is apparent that both the hump region (3.23 to 4.21 ppm) and the structural reporter group regions (4.47 to 5.37 ppm and 1.15 to 1.34 ppm) contribute significantly to the recognition capabilities of the neural network. Comparisons of weight plots of all hidden layer neurons with the actual spectra revealed that the neural network utilized most of the signals in the original ^{1}H-NMR spectra of **1** through **13** to activate or deactivate hidden layer neurons.

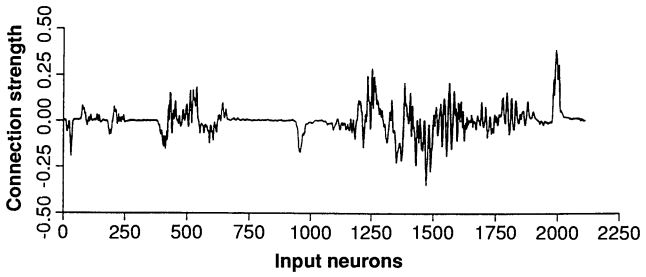


Fig. 2. Weight plot of one hidden layer neuron connected to all input layer neurons of NN-1 (see text). Signals from the input patterns are emphasized by positive weights and deemphasized by negative weights. Weight plots for other hidden layer neurons show different combinations of the input pattern signals.

^{1}H-NMR spectra of different preparations of the same compound typically show slight variations in chemical shifts and linewidths. We expected that training the neural network with deliberately imperfect data would increase the ability of the neural network to correctly recognize spectra that inherently contain similar imperfections (*1*). To test this hypothesis, we intentionally included "fuzziness" (that is, minor variations in the input data) in the training set by generating four additional copies of each spectrum. Each of the four copies contained all the signals of the original spectrum. In the first copy the entire spectrum was shifted 0.5 Hz to the right, and in the second copy the entire spectrum was shifted 1.0 Hz to the right. Similarly, the third and fourth copies contained the entire spectrum shifted 0.5 Hz and 1.0 Hz to the left, respectively. We expanded the initial training set of 13 spectra to 65 by including the spectra modified by fuzziness. This 65-spectra data set was then used to train NN-1. Once again, the neural network converged and recognized the spectra of **1** through **13**. The rms error of the trained neural network was 0.03, indicating excellent agreement between target and actual output patterns. We tested the trained neural network with spectra of **2** through **5** that originated from different preparations of those molecules. (We had recorded the spectra without the intention of using them in the neural network approach. Thus no special precautions had been taken to record the spectra under standardized conditions.) All eight spectra were recognized correctly (three instances of spectra of **2**, two of **3**, one of **4**, and two of **5**). Thus, the neural network was trained to tolerate variations in the input patterns, an important characteristic for the wide applicability of this technique. Furthermore, spectra recorded under normal conditions are recognized easily.

We know from the structural reporter group concept that signals with chemical shifts outside the poorly resolved hump region can be used to recognize spectra of a variety of oligosaccharides (*2*). We wanted to test how well neural networks recognize spectra, using either the structural reporter group signals or the hump region signals. We split the ^{1}H-NMR spectra of oligosaccharides **1** through **5** into two sets. One set contained only the signals of the structural reporter groups. The second set of partial spectra contained only the signals in the hump region. The spectral resolution was maintained at 0.5 Hz per input neuron. We used a neural network with 1003 input, 5 hidden layer, and 5 output neurons for the structural reporter group region and a neural network with 981 input, 5 hidden layer, and 5 output neurons for the hump region. Even with only these partial data sets, both neural nets converged and were able to recognize each of the spectra. This was expected for the structural reporter group region; the result with the hump region was less intuitive. We interpret this result to mean that artificial neural networks will be able to discriminate between spectra whose differences are not apparent to the human observer. Although it was not obvious that artificial neural networks could utilize the poorly resolved signals in the hump region to recognize spectra, we did know that the hump region contains the information necessary to distinguish between oligosaccharides. The artificial neural network easily extracted from the hump region the information needed to discriminate between the spectra.

The results reported here can be used to develop an artificial neural network–based pattern recognition system for identifying structurally diverse complex carbohydrates.

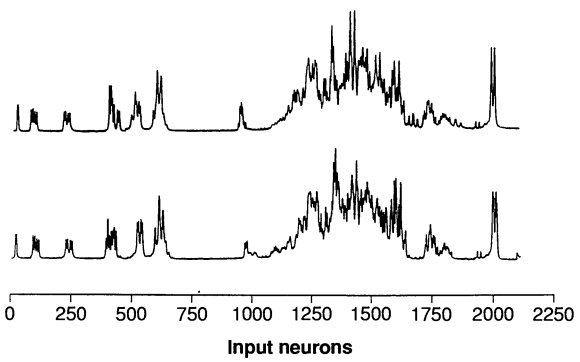


Fig. 1. Input patterns generated from the ^{1}H-NMR spectra of **9** and **12**.

Implementation of such a neural network–based system will require additional studies to determine the most efficient neural network architecture and the best training set to obtain a system with good discrimination capabilities while maintaining tolerance toward experimental variations.

The powerful recognition capabilities of artificial neural networks reported here suggest that neural networks are adaptable to other spectroscopic techniques in addition to NMR spectroscopy. We have already extended this approach to the identification of mass spectra (*8*). Neural networks have a great advantage over conventional spectroscopic library searches in that neural networks do not require definition of rules by scientists but rather extract the characteristic differences between the spectra during the training process. Furthermore, the retrieval of information from a neural network system is significantly faster than conventional library searches, which take many seconds up to minutes to retrieve the information. The response time of our neural networks is less than 0.1 s and is virtually independent of the number of spectra that are contained in its knowledge set. Other pattern recognition techniques that have been applied to the NMR spectra of carbohydrates have not been able to fully recognize all members of a family of molecules and required a complete assignment of the spectra before they could be used in the pattern recognition (*9*).

REFERENCES AND NOTES

1. J. Thomsen and B. Meyer, *J. Magn. Reson.* **84**, 212 (1989).
2. J. F. G. Vliegenthart, L. Dorland, H. van Halbeek, *Adv. Carbohydr. Chem. Biochem.* **41**, 209 (1983).
3. W. S. York, H. van Halbeek, P. Albersheim, A. Darvill, *Carbohydr. Res.* **200**, 9 (1990).
4. M. Hisamatsu, W. S. York, P. Albersheim, A. Darvill, unpublished results.
5. L. L. Kiefer, W. S. York, P. Albersheim, A. G. Darvill, *Carbohydr. Res.* **197**, 139 (1990).
6. D. E. Rumelhart and J. L. McClelland, *Parallel Distributed Processing* (MIT Press, Cambridge, MA, 1986), vol. 1.
7. J. L. McClelland and D. E. Rumelhart, *Explorations in Parallel Distributed Processing* (MIT Press, Cambridge, MA, 1988).
8. J. Sellers, W. S. York, P. Albersheim, A. Darvill, B. Meyer, *Carbohydr. Res.*, in press.
9. W. J. Goux, *J. Magn. Reson.* **85**, 457 (1989).

10. This work was supported in part by U.S. Department of Energy grant DE-FG09-85ER13424; by the U.S. Department of Agriculture–U.S. Department of Energy–National Science Foundation Plant Science Centers program with this project funded by 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.

4 June 1990; accepted 5 November 1990

Rapid Changes in the Range Limits of Scots Pine 4000 Years Ago

ANNABEL J. GEAR AND BRIAN HUNTLEY

Paleoecological data provide estimates of response rates to past climate changes. Fossil *Pinus sylvestris* stumps in far northern Scotland demonstrate former presence of pine trees where conventional pollen evidence of pine forests is lacking. Radiocarbon, dendrochronological, and fine temporal-resolution palynological data show that pine forests were present for about four centuries some 4000 years ago; the forests expanded and then retreated rapidly some 70 to 80 kilometers. Despite the rapidity of this response to climate change, it occurred at rates slower by an order of magnitude than those necessary to maintain equilibrium with forecast climate changes attributed to the greenhouse effect.

MUCH CONCERN IS FOCUSED upon the responses of organisms and ecosystems to global climate change (*1*). The paleoecological record provides evidence of both past distributions (*2–4*) and migration rates in response to past climate changes (*3, 4*). For example, subfossil stumps demonstrate that trees were formerly present in areas that are today treeless. Scots pine (*Pinus sylvestris* L.) stumps are recorded from peat deposits across northern Europe (*5–11*). Most occur where pollen analyses have recorded regional pine forests (*5–7, 11*). Far northern Scotland, however, is exceptional. Although stumps are present (*5*), conventional palynological studies have not provided evidence of pine forests (*12*). Two hypotheses have been advanced. First, pine trees were sparse for several millennia, growing only in particularly favorable sites and perhaps producing little pollen because of harsh climatic conditions; the stumps represent trees on mire surfaces during periods favorable for preservation (*5, 13*). Second, regional pine forests developed for an interval brief enough to be overlooked by conventional pollen studies. We have tested these hypotheses by making a regional survey of the distribution of stumps in conjunction with radiocarbon and dendrochronological analyses and a detailed palynological study.

The systematic mapping of subfossil pine stumps preserved in blanket peats in far northern Scotland showed that pine was formerly present throughout most of the region; only in a small area of the extreme northwest and a somewhat larger area of the northeast have pine stumps not been located (Fig. 1). Mapping involved systematic searches for stumps exposed in peat faces and ditches. The mapping was greatly facilitated by widespread afforestation and associated road building and drainage operations. Radiocarbon dating of 22 samples of subfossil pinewood from 11 localities (Fig. 1) gave ages ranging between 4405 and 3815 years B.P. (before present) (Table 1), a time span of 590 ± 71 years. Time spans are shorter at individual sites, ranging up to only 350 ± 85 years. A ^{14}C measurement on subfossil pinecones from blanket peat in 10-km National Grid square 29/84 gave an age comparable to those for nearby pine stumps (SRR-3563 4450 ± 65 years B.P. (Table 1). These ages are comparable with previously published ages for stumps from the region, which range between 4393 and 3976 years B.P. (*5*). They contrast, however, with the age of 6980 ± 100 years B.P. (Q-887) for a stump from Coire Bog (National Grid Reference 28/582857; 57°50′10″N 4°23′50″W) in the extreme south of the region studied (Fig. 1); the age of this stump is comparable with ages determined from many localities in the region of the Highlands that lies to the south of our study area (*5, 6*). High pine-pollen values are also recorded at Coire Bog between ~7000 and 5000 years B.P. (*6*), as at many palynologically studied sites throughout the Highlands. In addition, some of the most northerly of the relict stands of native pine forest are found today at scattered localities near Coire Bog (Fig. 1).

At one site with stumps in northern Scotland, Lochstrathy (National Grid Reference 29/796491; 58°24′51″N 4°3′40″W), dendrochronological data were also collected. Ring widths were measured on 42 stumps, and the transformed sequences were cross matched by minimizing the multiple proba-

Environmental Research Centre, University of Durham, Department of Biological Sciences, South Road, Durham DH1 3LE, United Kingdom.