IOSR Journal of Applied Chemistry (IOSR-JAC)
e-ISSN: 2278-5736.Volume 8, Issue 1 Ver. II. (Jan. 2015), PP 11-19
www.iosrjournals.org

Identification of Skeleton of Monoterpenoids from ^{13}CNMR Data Using Generalized Regression Neural Network (GRNN)

Taye Temitope Alawode[1], Kehinde Olukunmi Alawode[2]
[1]*Department of Chemical Sciences, Federal University Otuoke, Bayelsa State, Nigeria*
[2]*Department of Electrical and Electronic Engineering, Osun State University, Osogbo, Osun State, Nigeria*

Abstract: *This paper describes the use of Generalized Regression Neural Network (GRNN) in the identification of various skeletons of monoterpenoid compounds from their ^{13}C NMR chemical shift data. Towards this end, ^{13}C NMR chemical shift data of skeletons of 328 compounds belonging to various classes of monoterpenoids were used as input for the network. To generate the output data for the network, each compound belonging to a skeletal class was assigned a code of 1while every other possible skeleton types were given codes of 0. These data were used to train the network at varying spread constant values. After training, the network was simulated using 113 test compounds. At a spread constant of 15, the network had between 99.98 and 100% recognition of Myrcane skeleton, 100% recognition of the Santoline skeleton and 87.63 - 100% recognition of the Menthane skeleton. The network, however, could not identify successfully the Bornane and Pinane skeletons. To correct this anomaly, the training data for these classes of compounds were increased and the data re-trained. The results obtained improved considerably with between 68.25% and 99.95% recognition of the Bornane skeleton and 83.86% to 100% recognition of the Pinane skeleton. GRNN could be a powerful complimentary tool in the elucidation of structures of monoterpenoids.*

Keywords: *^{13}C NMR, GRNN, Monoterpenoid, Simulation, Skeleton.*

I. Introduction

Structural determination of natural products usually requires vast experience in spectral analysis. The fundamental stage in the process of structural elucidation is the determination of the compound carbon skeleton as this forms the basic unit to which the substance belongs. However, this is often difficult owing to high structure variety and diversity encountered in natural products chemistry. Studies in structural elucidation of monoterpenoids are of importance because this class of naturally occurring compounds possesses important pharmacological activities [1]. The advent of Computer Assisted Structural Elucidation (CASE) methods has simplified the process of interpretation of complex organic compounds, especially in the field of natural products. A high-quality reference library containing both structures and complete spectra or substructures and subspectra being representative of the types of compounds encountered in the laboratory, is an invaluable component for a CASE system [2, 3]. The premise implicit in the spectrum interpretation is that if the spectrum of the unknown and a reference library spectrum have a subspectrum in common, then the corresponding reference substructure is also present in the unknown. The components generated by spectra interpretation are fed into the structure generator, which will exhaustively generate all possible structures from these components. Examples of structure generators include MOLGEN, GENIUS and COCON. Their applications are described elsewhere [4]. Procedures that utilize ^{13}C NMR for skeleton identification have been previously developed and utilized with excellent results [5, 6, 7, 8].

Rufino et al [9] applied Artificial Neural Networks in the identification of skeletons of Aporphine alkaloids from ^{13}C NMR data asserted that ANNs because of their parallel nature can speed up the process of structural elucidation. ANNs have been applied to the prediction of biological activity of natural products or congeneric compounds [10, 11], the identification, distribution and recognition of patterns of chemical shifts from ^{1}H-NMR spectra [12,13] and identification of chemical classes through ^{13}C-NMR spectra [14]. ANNs are computational models derived from a simplified concept of the brain, in which a number of nodes, called neurons, are interconnected in a network-like structure [15]. Fig.1 shows a single neuron model.

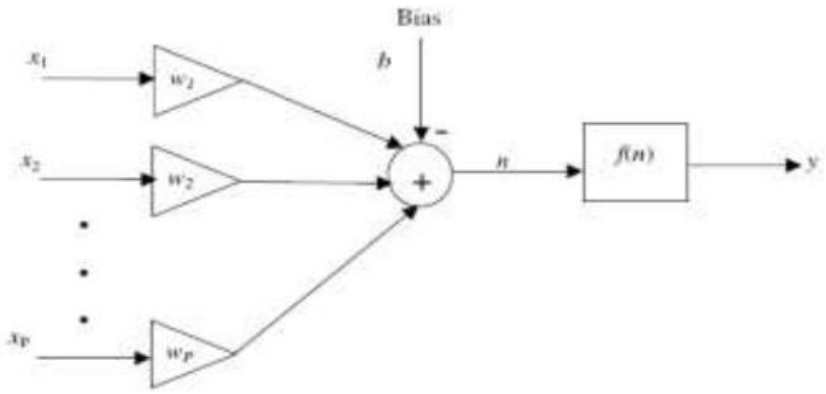

Figure 1: Single Neuron Model

Neural networks are nonlinear processes that perform learning and classification. Artificial neural networks consist of a large number of interconnected processing elements known as neurons that act as microprocessors. Each neuron accepts a weighted set of inputs and responds with an output. In general, neural networks are adjusted/ trained to reach, from a particular input, a specific target output until the network output matches the target. Hence the neural network can learn the system. The learning ability of a neural network depends on its architecture and applied algorithmic method during the training. A neural network is usually divided into three parts: the input layer, the hidden layer and the output layer. The information contained in the input layer is mapped to the output layers through the hidden layers.

In the present work, we show that Generalized Regression Neural Networks (GRNNs), one of the architectures of Artificial Neural networks can identify the skeletons of unknown monoterpenoid compounds among different (monoterpenoid) skeletons-Myrcane and Santoline (alicyclic monoterpenoids), Menthane (monocyclic monoterpenoids), Thujane, Bornane, Isocamphane and Fenchane (bicyclic monoterpenoids), and Pinane (a tricyclic monoterpenoid). Generalized Regression Neural Networks consists of four layers: input layer, pattern layer, summation layer and output layer as shown in Fig. 2. The theory of Generalized Regression Neural Networks has been described elsewhere [16].

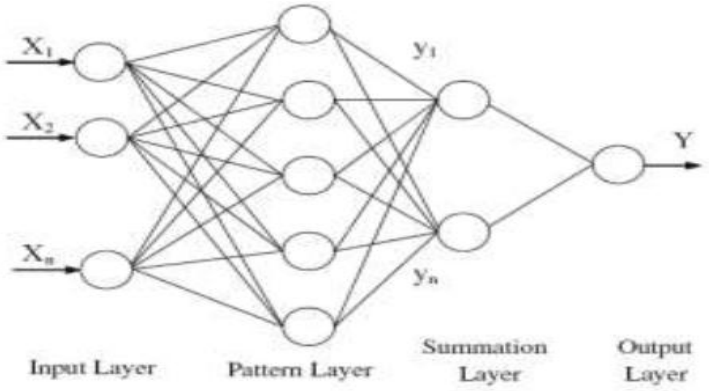

Figure 2: General Structure of GRNN

Compared to other ANN models such as the backpropagation neural network model, the GRNN needs only a fraction of the training samples a backpropagation neural network would need. Therefore it has the advantage that it is able to converge to the underlying function of the data with only few training samples available [17]. Furthermore, since the task of determining the best values for the several network parameters is difficult and often involves some trial and error methods, GRNN models require only one parameter (the spread constant) to be adjusted experimentally. This makes GRNN a very useful tool to perform predictions and comparisons of system performance in practice. Previous works relating the predictive capability of GRNN to backpropagation neural network and other nonlinear regression techniques highlighted the advantages of GRNN to include excellent approximation ability, fast training time, and exceptional stability during the prediction stage [18,19].

II. Materials And Methods

For identification purposes and for structural elucidation of new compounds, it is necessary to have access to extensive list of their structural data. In the present study, we made use of structural (skeletal) ^{13}C data of compounds reviewed and published by [20]. This information can be extracted from data of monoterpenoids published in literature by isolating ^{13}C values of the skeletal (carbon) from those of the substituents. ANNs work through learning method, their training must, therefore, be done with the use of detailed and correct data to avoid an erroneous learning process. A total of 441 compounds were employed in this study. Of these, 113 were reserved for use as test cases (these were not used in training the neural network). These included 33 Mrycane, 3 Santoline, 38 Menthane, 5 Thujane, 12 Bornane, 3 isocampahane, 15 Pinane and 4 Fenchane monoterpenoid compounds. ANNs learn through examples and the test compounds are selected based on the representativeness of their skeletons among data used for training. The skeletons of the compounds and the numbering of the carbon atoms are shown in Fig. 3.

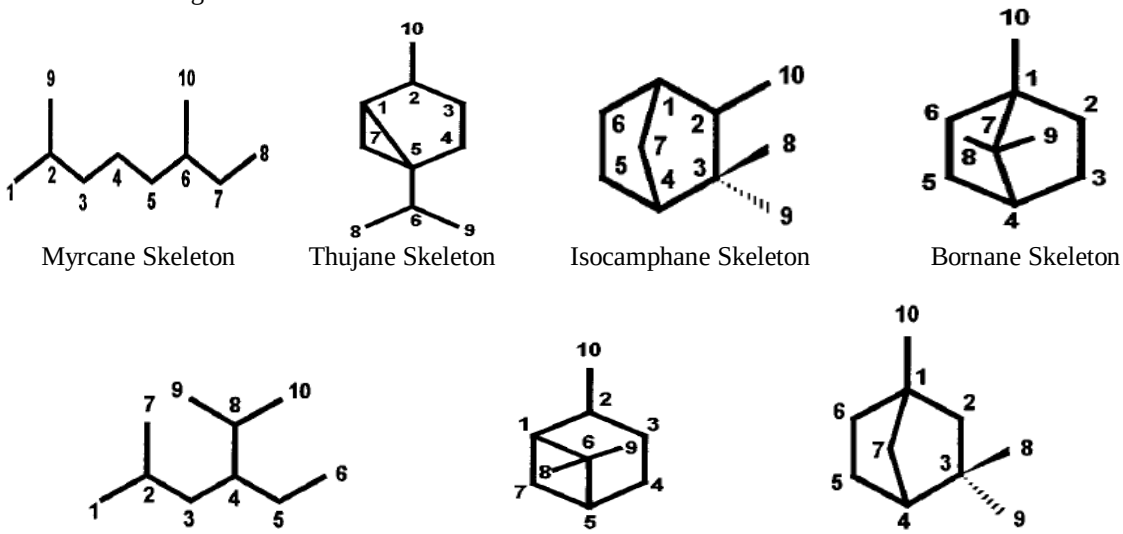

Figure 3: Skeletons of Monoterpenoid compounds used

Three Excel worksheets containing coded information on the input and target data for the training and test compounds were prepared. On the first row of the first sheet, the compounds were assigned codes 1-328. In the first column of the same sheet, the positions of each carbon atoms on the skeleton (as shown in Figure 3) were coded as 1-10. The ^{13}C chemical shift data for each Carbon at each of the 10 positions was recorded for each compound. These represent the input data subsequently used in training of the net. Another excel sheet in the format just described was prepared except that it contained ^{13}C chemical shift data for the test compounds (coded 1-113). The ^{13}C chemical shift data for skeletons of the test compounds are presented in Table 1. The target data were prepared on the third excel sheet. The compounds were assigned codes 1-328 as previously described. In the first column of the excel sheet, the eight different skeletons were listed vertically. Each compound is identified as belonging to a particular skeleton by assigning it a code 1 or 0. A compound belonging to a particular skeleton type is assigned a code of 1 while all the other compounds are assigned 0 for that skeleton type.

Table 1: ^{13}C NMR Chemical Shift data for test compounds

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
C-1	114.8	115.5	114.8	111.3	113.9	114.1	110.9	114	112.2	109.9	110.0	124.7	143.8	130.3	108.1
C-2	143.6	142.6	142.6	147.1	144.1	143.6	147.2	144	148.4	145.8	145.3	144.3	137.3	137.6	135.3
C-3	57.5	63.9	64.4	75.3	85.0	69.4	28.1	85.6	77.0	81.7	81.8	201.8	122.5	126.5	137.6
C-4	43.4	42.7	43.0	32.7	31.5	34.1	33.2	26.7	30.3	30.4	30.7	31.9	134.0	119.0	133.1
C-5	56.2	62.7	63.0	31.1	31.6	85.0	75.0	29.7	39.6	36.8	36.2	35.6	69.5	68.6	69.0
C-6	139.8	142.6	139.8	139.1	138.8	138.6	142.2	78.4	73.5	82.9	85.3	80.0	71.5	71.6	71.9
C-7	124.5	125.3	124.6	127.0	129.1	123.1	119.5	138.3	146.1	143.6	142.1	141.2	139.5	139.3	128.5
C-8	39.7	39.5	40.4	68.1	58.7	60.6	61.1	117.5	111.5	111.1	111.1	116.0	116.3	116.4	110.5
C-9	18.3	16.8	17.7	17.6	16.2	18.1	17.8	16.2	17.9	17.8	18.0	17.6	189.3	65.5	19.1
C-10	12.6	11.5	11.8	41.7	41.5	18.1	23.4	47.0	28.0	26.6	26.1	23.8	24.6	25.1	28.3

Table 1 (continues): ^{13}C NMR Chemical Shift data for test compounds

	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
C-1	108.0	132.3	131.8	121.7	115.6	130.2	170.3	170.3	169.5	25.7	25.3	195.1	25.7	25.7	24.2
C-2	135.3	137.0	136.0	140.7	141.8	120.5	128.4	128.4	128.7	131.5	131.6	153.6	130.6	131.1	133.3
C-3	138.6	124.0	123.5	133.6	137.0	110.8	144.3	144.2	144.3	124.3	124.0	139.6	125.0	124.8	122.5
C-4	133.0	123.4	124.1	126.6	124.9	151.8	27.1	27.1	27.1	25.6	25.6	26.3	25.6	25.5	25.6
C-5	69.0	69.4	69.4	69.1	45.8	41.2	37.6	37.0	36.8	37.0	36.9	35.2	37.4	37.2	33.5
C-6	71.5	71.6	71.6	72.2	72.8	72.6	30.7	30.5	30.5	27.8	27.7	27.7	29.3	29.6	64.6
C-7	127.9	139.5	139.3	137.3	144.8	144.2	36.9	40.6	37.5	51.0	51.0	50.7	39.7	35.7	64.5
C-8	110.0	116.5	116.5	110.5	112.0	112.0	68.9	61.0	68.5	203.8	202.8	202.0	60.2	62.9	198.8
C-9	19.5	36.7	69.4	43.9	18.6	9.8	12.5	12.4	12.4	19.8	17.6	19.6	17.6	17.6	17.6
C-10	25.3	25.1	25.1	28.0	27.3	27.4	19.8	19.8	19.7	17.6	19.6	9.1	19.6	19.5	22.2

Table 1 (continues): ^{13}C NMR Chemical Shift data for test compounds

	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
C-1	25.7	18.3	67.8	26.1	21.0	21.0	73.1	74.5	70.1	73.6	70.7	70.9	71.7	73.3	69.7
C-2	131.0	135.0	137.8	134.2	135.8	131.1	64.9	56.3	43.0	49.1	44.3	44.1	62.3	59.6	33.1
C-3	125.1	126.4	126.1	122.2	125.8	129.0	58.6	52.6	70.5	213.1	57.5	56.5	36.4	36.2	34.6
C-4	22.9	74.6	74.3	51.2	49.2	49.2	43.3	43.5	40.6	51.9	41.3	41.0	34.0	34.5	68.9
C-5	41.4	40.8	40.4	137.4	139.3	139.1	16.9	16	21.3	18.3	23.8	15.5	22.1	21.8	34.6
C-6	72.8	30.3	30.1	115.8	114.5	114.8	31.0	25.8	30.0	30.4	29.7	31.0	31.1	25.3	33.1
C-7	34.4	34.4	34.2	18.2	68.6	69.8	26.7	26.9	26.8	26.3	26.7	26.7	25.1	25.1	26.7
C-8	8.4	67.8	67.8	84.3	146.8	146.5	75.1	74.8	73.4	73.6	74	74.2	74.1	74	77.6
C-9	17.6	25.6	14	23.6	110.5	110.8	28.3	28.4	30.4	26.9	30.6	29.2	28.8	28.8	25.7
C-10	26.3	23.3	22.2	24.2	13.8	14.1	29	28.5	30.7	30.6	30.5	28.4	28.4	28.3	25.7

Table 1 (continues): ^{13}C NMR Chemical Shift data for test compounds

	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
C-1	71.8	70.7	69.9	72.8	72.8	73.3	71.5	74.3	75.4	72.0	74.0	77.2	75.4	68.9	73.3
C-2	27.6	31.5	31.6	42.7	43.8	71.3	43	37.3	56.7	57.1	38.5	92.6	57.0	83.2	82.6
C-3	22.2	22.1	22.4	55.2	54.8	65.8	55.2	54.8	35.4	91.2	54.9	50.3	36.0	35.0	31.9
C-4	33.0	28.2	29.2	42.3	41.6	42.5	49.2	42.5	35.2	49.9	42.8	36.2	35.6	41.2	41.3
C-5	22.2	22.2	22.6	36.8	29.1	15.2	50.6	28.5	35.6	21.6	36.9	21.1	36.0	24.5	24.4
C-6	27.6	31.7	31.7	60.0	60.4	31.1	43.7	58.5	58.9	28.6	57.2	28.4	57.0	36.8	33.1
C-7	51.6	27.2	27.1	24.4	24.2	24.5	25.8	24.3	23.0	26.1	24.2	22.0	23.1	32.7	28.8
C-8	74.4	76.1	75.7	74.6	74.7	74.7	74.9	74.5	74.5	74.9	74.3	74.5	74.5	83.6	82.4
C-9	28.5	50.4	68.7	30.3	28.8	28.9	30.7	29.3	29.1	32.1	29.9	28.1	28.4	22.7	23.4
C-10	28.5	24.4	23.5	30.0	28.3	27.9	29.7	27.8	27.7	31.9	29.7	27.8	28.4	30.0	30.3

Table 1 (continues): ^{13}C NMR Chemical Shift data for test compounds

	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
C-1	68.0	65.4	135.8	133.6	133.3	133.7	162.8	134.1	133.1	132.9	133.3	139.1	136.1	140.4	25.2
C-2	83.8	80.9	122.5	118.6	120.9	120.5	123.2	149.0	126.4	127.0	125.8	127.5	125.3	123.0	181.3
C-3	33.5	56.0	24.2	34.5	27.0	26.5	202.6	69.5	28.6	28.7	26.6	76.5	69.8	66.0	121.4
C-4	41.0	45.1	44.9	71.7	45.0	42.7	76.3	52.7	39.8	39.9	44.8	40.8	54.1	46.7	207.8
C-5	24.3	18.1	41.7	31.1	24.0	24.0	30.0	38.7	25.8	25.8	23.4	26.1	24.2	17.5	40.0
C-6	35.5	34.3	39.0	27.2	31.1	31.0	32.3	198.3	26.3	26.2	26.6	68.3	30.8	31.5	26.2
C-7	31.0	33.5	22.2	23.3	23.3	22.3	23.3	15.0	73.6	71.6	73.5	21.1	22.8	23.2	38.5
C-8	83.0	80.5	29.3	36.9	72.2	84.7	30.1	74.8	32.1	32.1	72.5	30.4	74.8	72.4	19.2
C-9	23.0	24.2	21.8	17.0	27.2	23.3	16.2	24.0	19.8	19.9	27.3	20.5	24.1	28.1	19.8
C-10	30.1	30.8	18.3	16.9	26.0	23.3	16.5	30.0	19.5	19.6	26.3	17.0	30.1	29.1	61.7

Table 1 (continues): ^{13}C NMR Chemical Shift data for test compounds

	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90
C-1	25.2	29.1	29.4	28.9	46.7	47	44.7	44.8	47.6	46.7	47.1	47.1	48.2	50.5	50.5
C-2	173.6	177.5	152.2	156.5	79.8	79.4	73.9	74.8	86.1	83.5	86.9	83.9	82.0	83.0	82.5
C-3	122.6	124.0	76.1	74.7	76	76.7	68.1	69.8	84.2	82.3	80.7	80.6	43.0	34.2	34.2
C-4	205.8	208.1	35.9	37.2	51.8	49.1	51	48.3	52.7	49.6	51.0	48.3	82.0	53.5	53.5
C-5	39.7	40.7	37.1	37.5	24.5	23.4	18.8	18.5	26.1	24.5	18.6	18.3	35.0	75.0	75.0
C-6	26.2	26.3	32.4	32.5	33.6	32.5	26.4	26.3	26.1	25.8	35.0	33.7	27.0	40.2	40.2
C-7	37.5	38.0	18.6	20.0	49.2	48.1	49.8	48.4	50.8	49.2	50.4	49.3	49.4	48.2	48.2
C-8	19.0	19.3	19.5	19.6	21.6	19.9	18.6	18.3	20.1	18.7	20.2	19.3	17.0	20.3	20.7
C-9	19.6	20.2	19.5	19.6	22.1	20.5	20.4	19.1	21.6	19.7	21.3	19.8	17.2	21.6	21.5
C-10	62.0	18.7	109.7	106.3	11.6	10.4	14.8	13.4	13.6	12.3	12.0	10.7	14.3	13.8	13.8

Table 1 (continues): ^{13}C NMR Chemical Shift data for test compounds

	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105
C-1	58.4	55.1	40.6	41.1	47.2	47.9	43.8	48.0	47.5	47.5	47.7	49.7	43.3	39.1	43.4
C-2	217.1	85.5	56.5	52.5	144.4	147	151.1	148.3	149.8	150.1	150.4	170.1	147.8	147.0	147.8
C-3	40.2	45.2	39.7	40.2	116.1	119.6	117.6	118.9	115.5	115.3	115.3	121.1	117.5	118.6	117.0
C-4	49.5	50.6	49.2	49.3	31.3	73.3	69.6	70.3	79.0	79.0	73.8	203.9	31.6	30.6	80.8
C-5	80.8	23.9	23.9	23.9	40.9	48.2	47.1	47.0	45.7	45.9	44.4	57.6	41.0	36.5	41.0
C-6	38.7	23.8	29.5	29.4	38.0	39.0	46.0	46.1	46.0	46.0	46.2	54.0	37.9	43.4	38.0
C-7	46.7	35.8	35.8	36.0	31.5	35.5	28.4	28.6	29.0	29.1	29.4	40.8	31.1	31.6	31.2
C-8	20.6	23.4	28.0	28.0	26.4	27.0	20.3	26.6	20.5	20.5	26.5	26.6	26.2	16.1	21.1
C-9	20.7	25.5	23.7	23.9	20.8	22.6	26.3	20.4	26.6	26.6	20.5	22.0	21.1	69.0	26.2
C-10	9.3	30.8	63.4	68.0	23.0	22.6	64.2	22.6	22.8	22.7	22.8	23.5	65.6	65.7	65.8

Table 1 (continues): ^{13}C NMR Chemical Shift data for test compounds

	106	107	108	109	110	111	112	113
C-1	43.4	38.2	43.4	67.9	53.9	53.5	60.4	52.5
C-2	147.8	151.6	147.8	138.7	222.1	221.6	221.6	218.4
C-3	117.0	147.5	117.0	118.4	47.2	45.3	47.1	54.6
C-4	80.8	31.3	31.6	33	45.3	50.3	44.6	41
C-5	41.0	40.7	41.0	62.6	25	77.8	36	24.8
C-6	38.0	37.6	38.0	30.1	31.8	41.8	76.7	32
C-7	31.2	33.0	31.2	205.8	41.6	38.1	38.5	41.4
C-8	21.1	25.7	21.1	27.3	23.3	23.8	23.8	49
C-9	26.2	20.9	26.2	14.7	21.7	21.5	21.6	18.2
C-10	65.8	191.0	65.8	23	14.6	14.6	12	14.4

After the construction of the worksheets, the data were transferred into the Neural Network toolbox of MATLAB 7.8.0. From the command window, the 'nntool' command was used to designate the imported data appropriately as 'input' or 'target'. The Generalized Regression Neural Network architecture was selected for training of the skeleton-identification system at spread constants of 0.5, 1, 2.5, 5, 7.5, 9, 10, 12, 15, 17.5, 20, 25, 30, 50 and 100. The effectiveness of training at each value of spread constant was assessed by simulation with the test data (not previously used for training and therefore unknown to the network). The aim was to ascertain whether the neural network would be able to identify correctly the skeleton type to which each test compound belong. The Generalized Neural Network (GRNN) at a spread constant of 15.0 was chosen as the baseline for results presentation as all classes of compounds give reasonably good results at this value.

When it was observed that for the network could not identify with high accuracy compounds having the Thujane, Bornane, Isocamphane, Pinane and Fenchane skeletons, the training data was increased. This was done by adding to the original training set randomly selected compounds from the previously used set of test compounds. The randomly selected compounds were from the classes of compounds whose skeletons were not correctly predicted. This reduced the total number of test compounds to 93 comprising of 33 compounds with the Myrcane skeleton, 3 with the Santoline skeleton, 38 compounds with the Menthane skeleton, 3 compounds with the Thujane skeleton, 5 compounds with the Bornane skeleton, 3 with Isocamphane skeleton, 5 with Pinane skeleton and 3 with the Fenchane skeleton. This procedure was carried out to ascertain whether the observed inaccuracies were due to insufficient training data. Graphs of observed errors in individual prediction against spread constant values for randomly selected compound(s) from each skeleton class are plotted to give an insight into the range of spread constant values where the best results may be obtained. For Bornane, Pinane and Fenchane, results obtained after re-training of the system were used. For these set of compounds, the GRNN was trained at spread constants of between 1 and 25. This is because from the previously trained data set (comprising of 113 compounds), it has been observed that least errors were obtained within this range.

III. Results And Discussion

The results obtained after training of the neural network and simulating with the original set of 113 test data using GRNN are presented in Table 2. The probability that a compound belongs to a particular skeletal type is expressed as percentages. (When a value of 1 is returned by the network for a particular skeletal type, there is 100% certainty that the unknown compound possess that skeleton while a value of 0 indicates a null probability). If correctly predicted, compounds 1-33 should be Myrcane; 34-36 Santoline; 37-74 Menthane; 75-79 Thujane; 80-91 Bornane; 92-94 Isocamphane; 95-109 Pinane; and 110- 113 Fenchane. The results showed that out of the 33 Myracane compounds used as test data, the network had 99.98% - 100% recognition rate of 30 compounds. A recognition rate of 71.7% and 78.58% was observed for compounds 28 and 29 (with 28.23% and 21.41% probability respectively that these compounds had Thujane skeleton). Compound 31 was wrongly predicted as Thujane skeleton (99.92%). The network had 100% recognition rate for the 3 compounds belonging to the Santoline skeleton and 87.63 - 100% recognition for compounds belong to the class of Menthane monoterpenoids. Of the 5 compounds with Thujane skeleton tested, 2 were erroneously predicted to have Pinane

skeleton. For compounds 80-88 (all belonging to the Bornane series), the network could not identify the compounds as belonging to any specific skeleton as the probabilities were almost evenly distributed between Menthane, Bornane, Pinane and Fenchane skeleton types. The network could only identify 2 of the 4 Isocamphane compounds at 85.99% and 86.47% and wrongly predicted most of the compounds belonging to the Pinane class as Thujane (with lesser probabilities as Myrcane). Also, only 2 of the 4 Fenchane compounds were recognized at 61.28% and 76.74%.

Table 2: Probability of the Test Compound To Belong to the Skeletons Researched (σ =15)

Tested Skeletons	Tested compounds (%)														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Myrcane	100	100	100	100	100	100	100	100	100	100	100	100	0	100	100
Santoline	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Menthane	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Thujane	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bornane	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Isocamphane	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pinane	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Fenchane	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 2 (continues): Probability of the Test Compound To Belong to the Skeletons Researched(σ=15)

Tested Skeletons	Tested compounds (%)														
	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Myrcane	100	100	100	100	100	100	99.98	99.99	99.98	100	100	100	71.70	78.58	100
Santoline	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Menthane	0	0	0	0	0	0	0.02	0.01	0.02	0	0	0	0	0	0
Thujane	0	0	0	0	0	0	0	0	0	0	0	0	28.23	21.41	0
Bornane	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Isocamphane	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pinane	0	0	0	0	0	0	0	0	0	0	0	0	0.07	0.01	0
Fenchane	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 2(continues): Probability of the Test Compound To Belong to the Skeletons Researched(σ=15)

Tested Skeletons	Tested compounds (%)														
	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
Myrcane	0	100	100	0	0	0	0.01	0.03	0.01	0	0.12	0.06	0.12	0.12	0.30
Santoline	0	0	0	100	100	100	0	0	0	0	0	0	0	0	0
Menthane	0	0	0	0	0	0	99.99	99.97	99.99	99.71	99.88	99.94	99.88	99.88	99.55
Thujane	99.92	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bornane	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Isocamphane	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pinane	0.08	0	0	0	0	0	0	0	0	0.29	0	0	0	0	0.15
Fenchane	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 2(continues): Probability of the Test Compound To Belong to the Skeletons Researched (σ=15)

Tested Skeletons	Tested compounds (%)														
	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
Myrcane	12.35	0.65	0.19	1.77	0.90	0	2.61	0.97	2.76	0	2.08	0	2.50	0	0
Santoline	0	0	0	0	0	0		0	0	0	0	0	0	0	0
Menthane	87.63	99.34	99.81	98.20	99.09	99.98	97.296	99.03	97.22	100	97.90	100	97.48	100	100
Thujane	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bornane	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Isocamphane	0.11	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pinane	0	0	0	0.03	0.01	0.02	0.1	0	0.02	0	0.02	0	0.02	0	0
Fenchane	0	0	0	0	0	0		0	0	0	0	0	0	0	0

Table 2(continues): Probability of the Test Compound To Belong to the Skeletons Researched (σ=15)

Tested Skeletons	Tested compounds (%)														
	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
Myrcane	0	0	0	0.03	0	0	0	0	5.82	5.46	0.66	0	0.02	0	0
Santoline	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Menthane	100	100	100	99.97	100	100	100	100	94.18	94.54	99.34	100	99.98	100	0
Thujane	0	0	0	0	0	0	0	0	0	0	0	0	0	0	99.84
Bornane	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Isocamphane	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pinane	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.16
Fenchane	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 2(continues): Probability of the Test Compound To Belong to the Skeletons Researched(σ=15)

Tested Skeletons	Tested compounds (%)														
	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90
Myrcane	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Santoline	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Menthane	0	0	0	0	0.23	25.41	13.30	16.71	31.54	33.04	31.28	34.13	0.81	0	0
Thujane	99.78	99.96	0.20	0.31	0.05	0.09	0.05	8.63	0.026	00.06	0.02	0.05	0	0	0
Bornane	0	0	0	0	35.02	34.52	52.76	50.08	20.39	24.21	26.23	27.13	69.17	99.96	99.95
Isocamphane	0	0	0.08	0.02	9.05	8.37	15.93	14.79	8.73	8.74	7.75	7.15	12.84	0	0
Pinane	0.22	0.04	99.72	99.47	16.73	17.81	3.29	4.66	26.35	22.47	15.23	16.65	0.13	0.02	0.02
Fenchane	0	0	0	0	16.26	13.81	14.66	13.69	12.96	11.48	19.49	14.89	17.05	0.02	0.02

Table 2(continues): Probability of the Test Compound To Belong to the Skeletons Researched (σ=15)

Tested Skeletons	Tested compounds (%)														
	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105
Myrcane	0	0	0	0	0.02	63.19	7.01	27.49	23.60	23.52	37.68	0.13	0.03	0	58.98
Santoline	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Menthane	0	0.05	12.35	11.92	0	0	0	0	0	0	0	0	0	0	0
Thujane	0	0.10	0	0	99.80	36.64	92	72.48	76.37	76.45	62.30	60.19	97.05	96.36	40.54
Bornane	68.25	26.03	0.02	0	0	0	0	0	0	0	0	0	0	0	0
Isocamphane	0	42.73	85.99	86.47	0	0	0	0	0	0	0	0	0	0	0
Pinane	0	1.17	1.62	1.61	0.17	0.17	0.99	0.03	0.03	0.03	0.02	39.69	2.92	3.64	0.50
Fenchane	31.75	25.47	0.02	0	0	0	0	0	0	0	0	0	0	0	0

Table 2(continues): Probability of the Test Compound To Belong to the Skeletons Researched (σ=15)

Tested Skeleton	Tested Compounds (%)							
	106	107	108	109	110	111	112	113
Myrcane	58.98	0	0	16.14	0	0	0	0
Santoline	0	0	0	0	0	0	0	0
Menthane	0	0	0	0	0	0	0	0
Thujane	40.54	0	96.52	0	0	0	0	0
Bornane	0	0	0	0	87.86	38.72	23.26	85.80
Isocamphane	0	0	0	0	0	0	0	0
Pinane	0.50	1	3.47	83.86	0	0	0	0
Fenchane	0	0	0	0	12.14	61.28	76.74	14.20

To ascertain whether the inadequacies observed especially in results involving Thujane, Bornane, Isocampahne, Pinane and Fenchane compounds were due to insufficient training data, the number of the training data were increased as previously described. After training and simulating with the 93 compounds whose ^{13}C NMR values are used as the test data, at the baseline spread constant value of 15, the Bornane skeletons are now recognized at 68.25%, 70.82%, 99.10%, 99.95% and 99.95% for the 5 test compounds used. The network also had between 83.86 and 100% recognition of the 5 compounds with Pinane skeleton used for the test. No significant improvement was obtained for Isocamphane, Thujane and Fenchane skeletons. This is expected since out of a total of 20 compounds added to the original training set (of 328 compounds), 7, 10, 2, 1and 0 belong to the Bornane, Pinane, Thujane, Fenchane and Isocamphane classes respectively. Fewer numbers of compounds from the Thujane, Fenchane and Isocamphane classes were used for the re-training because only 5, 4 and 3 respectively from these classes were present in the original set of test data. The predictive ability of GRNN might have been affected by the size of the learning database. In a previous work, Ferreira et al (1998) showed that the expert system SISTEMAT could only predict the pinane skeleton types with only 0.714 accuracy, implying that other skeletons also appear but with low statistical significance. In their pioneering work, Rufino et al (2005) showed that ANN methods give fast and accurate results for identification of skeletons and for assigning unknown compounds among distinct fingerprints (skeletons) of aporphine alkaloids. The computation method is much faster than the utilization of traditional methods for skeleton prediction as the time-consuming sequential search (especially for large spectra library) and matching procedures (sequential comparison of an unknown target spectrum with the set of library spectra) employed by the conventional databases is avoided. This makes neural networks ideal for selecting results for structure generators or checking the entries of a database. If a large number of skeletons have to be predicted or a fast and easy check of a structure is necessary, this approach is advantageous. Moreover, the large amount of the disk space for saving the database or long time for loading data from external computers will no longer be necessary.

From Fig 4 (below), it could be observed that the spread constant ranges over which excellent prediction results were obtained seems to be specific for each skeleton class. Best prediction of the Myrcane skeleton, was within spread constant range of 10 – 30; and for Menthane skeleton best results were obtained between 5 and 30.

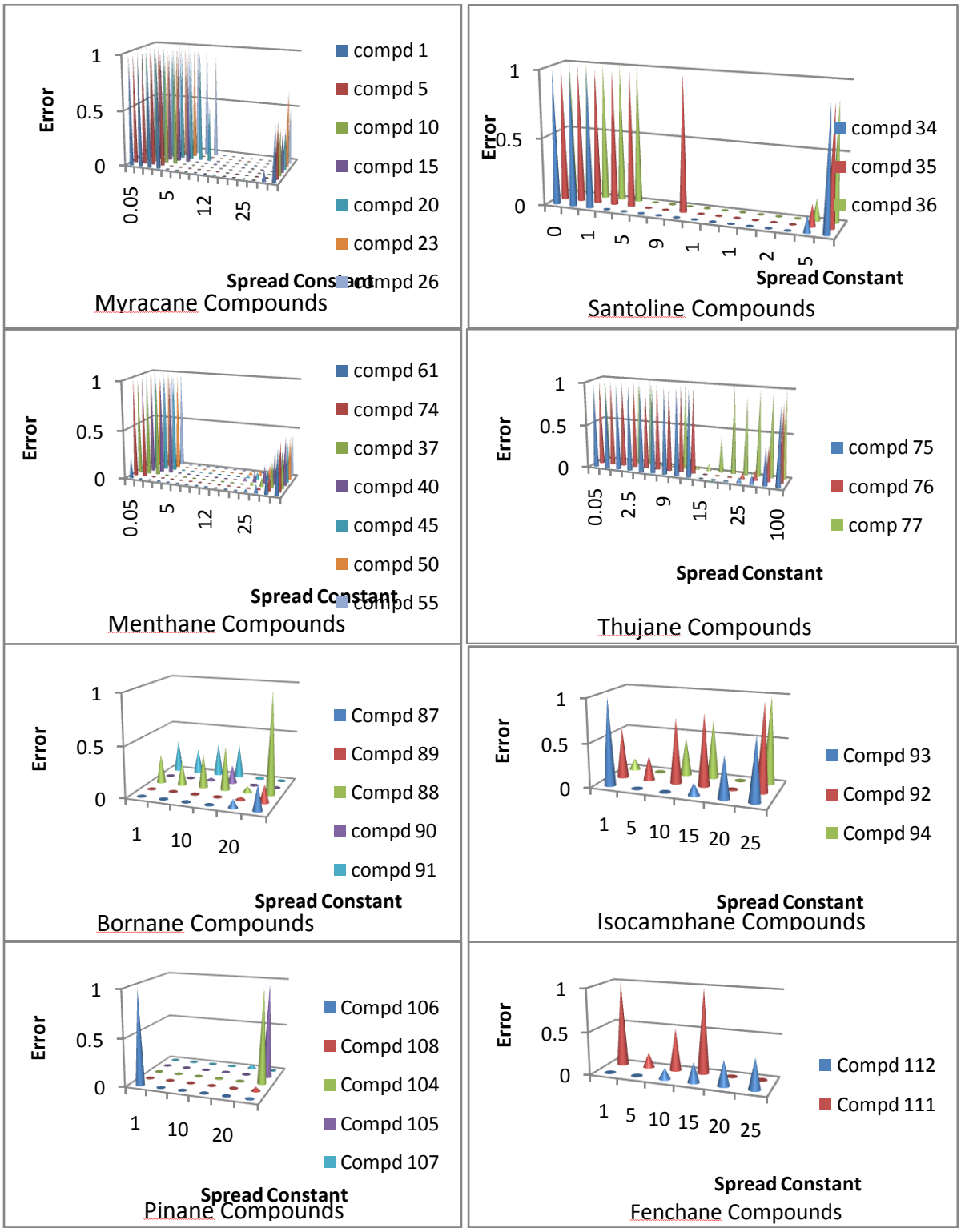

Figure 4: Graphs of observed errors in individual prediction against spread constant values

Though available test data are few, one can cautiously infer that the best predictions appears to be obtained within the spread constant range of 1 - 20 for Bornane skeleton, 7.5 -20 for Santoline skeleton and 5-20 for Pinane skeleton. Within these broad ranges of values, errors in prediction were zero in most cases. The variation of the generalized error with change in spread constant is an important parameter to access the efficacy of any GRNN. A network that gives a constant error for a broad range of spread constant is considered better since designers can choose from a wide range of spread constant values for their network.

IV. Conclusion

From this study, it could be seen that the predictions obtained using GRNN are in good agreement with the actual skeletons of the compounds tested. The network was tested with compounds belonging to diverse skeleton types and good results were obtained in almost all the cases. The quality of predictions of the network, however, depends on the availability of sufficiently diverse training data (covering adequately all the classes of monoterpenoid compounds) for the network. GRNN, could therefore be a powerful complimentary tool in structural elucidation of monoterpenoids.

References

[1]. N.H. Fischer, Plant terpenoids as allelopathic agents in Harbone J.B. and Tomas-Barberan F.A. (Eds.) Ecological Chemistry and Biochemistry of Plant Terpenoids,. Clarendon Press, Oxford, 1991, 377.

[2]. M. E. Elyashberg, K.A. Blinov, A. J. Williams, E. R. Martirosian and S. G. Molodtsov, Application of a new expert system for the structure elucidation of natural products from their 1D and 2D NMR data. Journal of Natural Products. 65, 2002, 693-703.

[3]. I.I. Strokov and K. S. Lebedev, Computer aided method for chemical structure elucidation using spectral databases and C-13 NMR correlation tables. Journal of Chemical Information & Computer Sciences. 39, 1999, 659-665.

[4]. J. Meiler and M. Kock, Novel Methods of Automated Structure Elucidation based on 13C NMR Spectroscopy. Magn. Reson. Chem. 42, 2004, 1042–1045.

[5]. M.J.P. Ferreira, G.V. Rodrigues, A.J.C. Brant and V.P. Emerenciano, REGRAS: an auxiliary program for pattern recognition and substructure elucidation of monoterpenes. Spectroscopy. 15, 2000, 65–98.

[6]. M.J.P. Ferreira, A. J.C Brant., G.V. Rodrigues and V.P. Emerenciano, Automatic identification of terpenoid skeletons through ^{13}C nuclear magnetic resonance data disfunctionalization. Analytica Chimica Acta. 429, 2001, 151–170.

[7]. M.J.P.Ferreira, F.C.Oliveira, S.A.V.Alvarenga, P.A.T. Macari, G.V.Rodrigues and V.P. Emerenciano, Automatic identification by ^{13}C NMR of substituent groups bonded in natural product skeletons. Computers & Chemistry. 26, 2002, 601–632.

[8]. G.V. Rodrigues, I.P.A. Campos and V.P Emerenciano, Applications of artificial intelligence to structure determination of organic compounds **. Determination of groups attached to skeleton of natural products using ^{13}C Nuclear Magnetic Resonance Spectroscopy. Spectroscopy, 1997,191-200.

[9]. A.A. Rufino, A .J. C. Brant, J. B. O. Santos, M.J.P. Ferreira and V.P. Emerenciano, Simple Method for Identification of Aporphine Alkaloids from ^{13}C NMR Data Using Artificial Neural Networks. J. Chem. Inf. Model. 45, 2005, 645-651.

[10]. P. Wrede, O. Landt, S. Klages, A. Faterni, U. Hahn and G. Schneider, Peptidase design aided by neural networks: biological activity of artificial signal peptidase I cleavage sites. Biochemistry, 37, 1998, 3588-3593.

[11]. M.B. Fernandes, M.T. Scotti, M.J.P. Ferreira and V.P. Emerenciano, Use of self-organizing maps and molecular descriptors to predict the cytotoxic activity of sesquiterpene lactones. European Journal of Medicinal Chemistry. 43, 2008, 2197-2205.

[12]. J. Aires-de-Sousa, M. Hemmer and J. Gasteiger, "Prediction of 1H NMR Chemical Shifts Using Neural Networks". Analytical Chemistry, 74(1), 2002, 80-90.

[13]. Y. Binev and J. Aires-de-Sousa, "Structure-Based Predictions of 1H NMR Chemical Shifts Using Feed-Forward Neural Networks". Chem. Inf. Comput. Sci., 44, 2004, 940-945.

[14]. L. Fraser and D. A. Mulholland , A robust technique for group classification of the C-13 NMR spectra of natural products from Meliaceae. Fresenius J Anal Chem.,. 365, 1999, 631-634.

[15]. M.T. Scott, V. Emerenciano, M.J.P. Ferreira, L. Scotti, R. Stefani, M.S. da Silva and F.J.B. Mendonça Junior, Self-Organizing Maps of Molecular Descriptors for Sesquiterpene Lactones and Their Application to the Chemotaxonomy of the Asteraceae Family. Molecules. 17, 2012, 4684-4702.

[16]. S.A. Hannan, R.R. Manza, and R.J. Ramteke, Generalized regression neural network and radial basis function for heart disease diagnosis. International Journal of Computer Applications, 7(13), 2010, 7-13.

[17]. Specht, A General Regression Neural Network. IEEE Transactions on Neural Networks. 2(6), 1991,568-576.

[18]. G. Sun, S. J. Hoff, B. C. Zelle and M. A. Nelson, Development and comparison of Backpropagation and Generalized regression neural network models to predict diurnal and seasonal gas and pm10 Concentrations and emissions from swine buildings. American Society of Agricultural and Biological Engineers. 51(2), 2008, 685-694.

[19]. C. Mahesh, E. Kannan and M. S. Saravanan (2014). Generalized regression neural network based expert system for hepatitis b diagnosis. Journal of Computer Science. 10 (4), 2014, 563-569.

[20]. M. J. P. Ferreiraa, V. P. Emerenciano, G. A. R. Linia, P. Romoff,P. A. T. Macarib and G. V. Rodrigues, ^{13}C NMR spectroscopy of monoterpenoids. Progress in Nuclear Magnetic Resonance Spectroscopy, 33, 1998, 153–206.