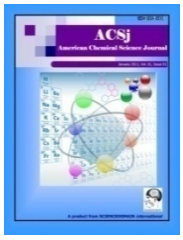

American Chemical Science Journal
4(6): 890-900, 2014

SCIENCEDOMAIN *international*
www.sciencedomain.org

Prediction of Substituent Types and Positions on Skeleton of Myrcane-Type Monoterpenoids using Generalized Regression Neural Network

Taye T. Alawode[1*] and Kehinde O. 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.

Authors' contributions

This work was carried out in collaboration between both authors. Both authors read and approved the final manuscript.

Original Research Article

Received 15th May 2014
Accepted 11th July 2014
Published 24th July 2014

ABSTRACT

Aim: To explore the ability of GRNN as a tool of structural elucidation in predicting the substituent types on myrcane, one of the representative skeletons of monoterpenoids.
Methodology: Generalized regression neural network (GRNN) was used in the study. Carbon-13 (^{13}C) NMR chemical shift values of skeletons of 104 myrcane monoterpenoids were used as the input data used for the network. Each substituent type on the skeleton of the different compounds were coded and used as the output data for the network. These data were used to train the network while the spread constant of the GRNN was varied. After training, the network was simulated using 15 test compounds.
Results: GRNN at a spread constant of 1.0 gave the best result. The network had between 80 to 90% recognition rates in 14 of the 15 test compounds. The network could not predict correctly the substitution pattern on 'compound 11' as all the positions was predicted to be unsubstituted. This could be due to the non-existence of precise rules for the compound.
Conclusion: GRNN, one of the architectures of Artificial Neural Networks (ANNs), could be a powerful aid in the structural elucidation of organic compounds.

**Corresponding author: Email: alawodett@fuotuoke.edu.ng, onatop2003@yahoo.com;*

Keywords: ANNs; GRNN; myrcane skeleton; monoterpenoids; structural elucidation.

1. INTRODUCTION

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. Structural elucidation (using CASE methods) involves finding, from structural information of an unknown compound derived from chemical and/or spectra evidence, the fittest structural formula that satisfies a set of chemical and spectral boundary conditions [2]. An invaluable component of the CASE system is 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 [3,4]. 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 [5].

The structure of any natural product is conventionally divisible into three sub-units: (i) the skeletal atoms; (ii) heteroatoms directly bonded to the skeletal atoms or unsaturations between them; and (iii) secondary carbon chains, usually bound to a skeletal atom through an ester or ether linkage [6]. A procedure that utilizes ^{13}C NMR for terpenoid skeleton identification has been described previously. The program REGRAS, developed for the expert system, SISTEMAT, could recognize the substructures and the skeleton present in a compound [7]. When REGRAS was tested on skeleton elucidation of 35 terpenoid compounds, excellent results were obtained [8]. Another program, MACRONO (also written for SISTEMAT) could expunge chemical shifts not due to the skeletal carbons from the initial dataset, which can then be input into SISTEMAT for skeletal identification. A new version of the program was successfully tested in the identification of the substituents and skeletons of 60 compounds [6,9]. SISTEMAT has been found to achieve both high reliability and good performance when applied to structural elucidations and chemical shift evaluations of monoterpenoids. It has also been applied successfully to other classes of natural products like diterpenes, triterpenes and flavonoids [10]. More recently, it was established that ANN methods give fast and accurate results for identification of skeletons and for assigning unknown compounds among distinct fingerprints (skeletons) of aporphine alkaloids [11]. In the present work, we show that Generalized Regression Neural Networks (GRNNs), one of the architectures of Artificial Neural Networks, can predict substituents' positions and types on the Myrcane-type monoterpenoid skeleton.

ANNs have been applied to the prediction of biological activity of natural products or congeneric compounds [12,13], the identification, distribution and recognition of patterns of chemical shifts from ^{1}H-NMR spectra [14,15] and identification of chemical classes through ^{13}C-NMR spectra [16]. 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 [17]. Fig. 1 shows a single neuron model.

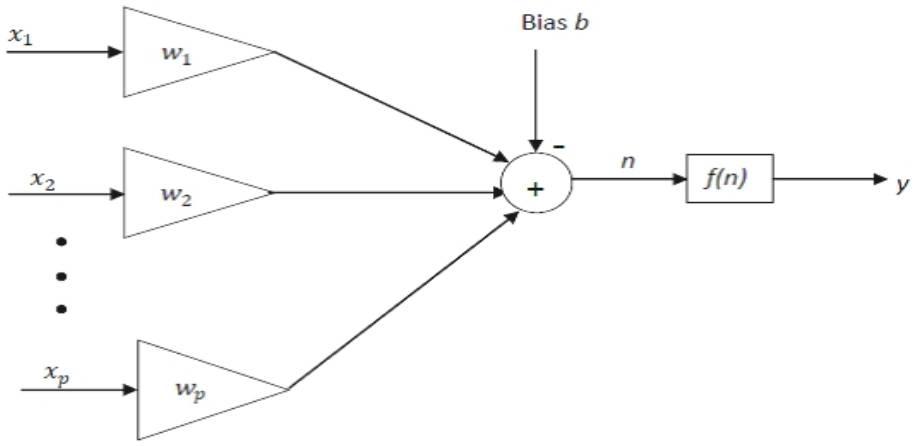

Fig. 1. Single neuron model [18]

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. Training procedure ceases if the difference between the network output and desired/actual output is less than a certain tolerance value. Thereafter, the network is ready to produce outputs based on the new input parameters that are not used during the learning procedure. 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.

A GRNN consists of four layers: input layer, pattern layer, summation layer and output layer as shown in Fig. 2. The number of input units in the input layer depends on the total number of the observation parameters. The first layer is connected to the pattern layer and in this layer each neuron presents a training pattern and its output. The pattern layer is connected to the summation layer. The summation layer has two different types of summation, which are a single division unit and summation units. The summation and output layer together perform a normalization of output set. In training of network, radial basis and linear activation functions are used in hidden and output layers. Each pattern layer unit is connected to the two neurons in the summation layer, S and D summation neurons. S summation neuron computes the sum of weighted responses of the pattern layer. On the other hand, D summation neuron is used to calculate un-weighted outputs of pattern neurons. The output layer merely divides the output of each S-summation neuron by that of each D-summation neuron, yielding the predicted value Y'i to an unknown input vector x as [18,19];

$$Y_i' = \frac{\sum_{i=1}^{n} y_i . exp - D(x, x_i)}{\sum_{i=1}^{n} exp - D(x, x_i)}$$

$$D(x, x_i) = \sum_{k=1}^{m} \left(\frac{x_i - x_{ik}}{\sigma}\right)^2$$

y_i is the weight connection between the i^{th} neuron in the pattern layer and the S-summation neuron, n is the number of the training patterns, D is the Gaussian function, m is the number of elements of an input vector, x_k and x_{ik} are the j^{th} element of x and x_i, respectively, σ is the spread parameter, whose optimal value is determined experimentally.

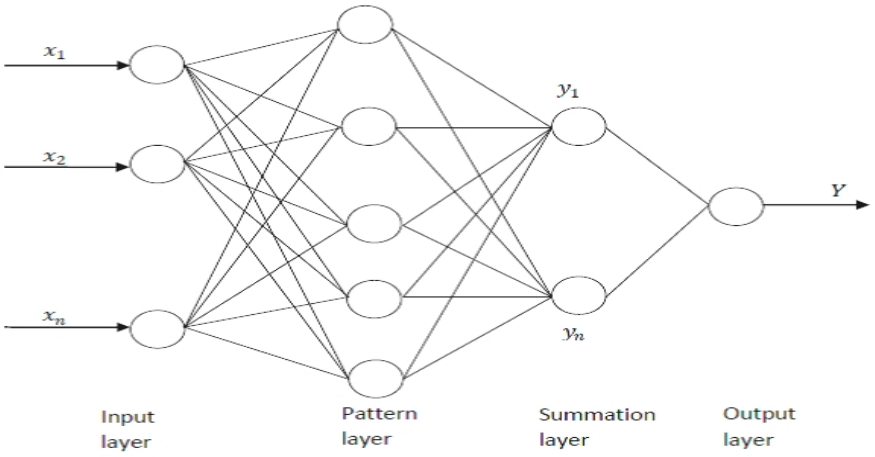

Fig. 2. General structure of GRNN [18]

Compared to other ANN models such as the back propagation neural network model, the GRNN needs only a fraction of the training samples a back propagation 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 [20]. 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 [21,22].

2. METHODOLOGY

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, substituents and stereochemical information of 119 Myrcane compounds published by [1].This information can be extracted from data of Myrcane 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 well detailed and correct data to avoid an erroneous learning process. Of the total of 119 compounds used in this study, only 15 were reserved for use as test cases (these were not used in training the neural network). This is because ANNs learn through examples and the test compounds can only be selected based on the representativeness of their substitution patterns among the test compounds. Selection of test compounds was largely done by visual inspection. The structure of the myrcane skeleton

with the numbering of each carbon atom is shown in Fig. 3 while Fig. 4 shows the fifteen selected test compounds.

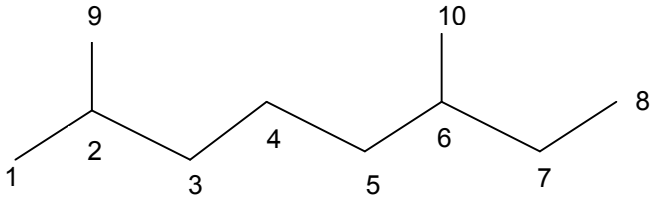

Fig. 3. The myrcane skeleton

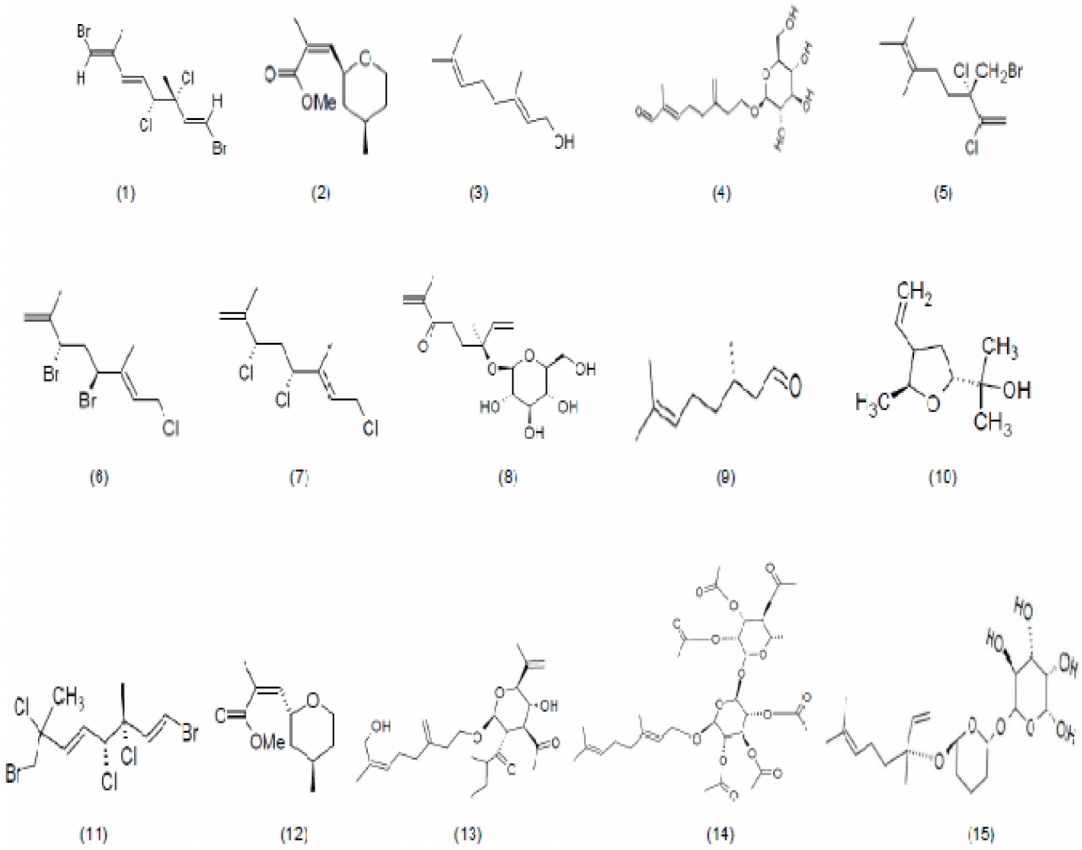

Fig. 4. The selected 15 test compounds

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 1-104. In the first column of the same sheet, the positions of each carbon atoms on the skeleton (as shown in Fig. 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 1-15). The ^{13}C chemical shift data for skeletons of the test compounds are presented in Table 1.

In preparing the target data, each substituent type (on first encounter) was assigned 3 number codes. These codes serve to identify the substituent while also taking into account its possible stereochemistry (α or β) in various positions of the skeletons in other compounds. Carbon positions without substituents were assigned a code of 0 while α and β positions without substituent(s) were assigned codes of 1 and 2 respectively. For example, OH group was given a code of 3, an α-OH is given a code of 4 while a β-OH was assigned a code of 5.

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' and to select the appropriate network for training. The network types employed in the training of data include perceptron, feed-forward back propagation (BP) and Generalized Regression Neural Networks. Several network parameters including number of layers, training function, adaptation learning function, performance function, number of neurons, were varied for feed-forward BP and perceptron neural networks while for GRNN, only the spread constant was varied. The effectiveness of training 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 predict correctly the substituents and their positions on the myrcane skeleton. After trying several neural network types and network parameters, the Generalized Regression Neural Network (GRNN) at a spread constant of 1.0 was found to give the best results.

3. RESULTS AND DISCUSSION

The results obtained after training of the neural network and simulating with the test data using GRNN are presented in Table 2. Percentage (%) recognition of the compounds was calculated from the number of correctly predicted points relative to the total number of positions on each compound (10). This ranged between 80% and 90%. Results for test compounds 11 is not shown because the network presented all the positions on the skeleton as un-substituted. This may be due to the non-existence of precise rules for the compound. From the results presented in Table 2, the un-substituted positions (designated as '-') on the myrcane skeleton in all the compounds tested were correctly predicted except for the un-substitued C-3 position on test compound 5 which was predicted to hold a –Cl group. The results obtained when perceptron and feed-forward backpropagation neural networks (employing varying network parameters) were used are not presented since the substituents predicted to be on the myrcane skeleton for all the test compounds, are largely inaccurate.

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. Most research efforts are directed at developing expert systems to help in this regard. Despite the progress, identifying the types, positions and stereochemistry of substituents on the skeleton of an unknown compound can be a daunting task to the unskilled. The present work seeks to simplify this task.

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
C1	108.1	168.3	25.5	195.2	17.7	114.8	114.8	124.7	25.3	24.6	41.5	168.3	11.3	25.5	25.7
C2	135.3	141.7	131.1	146.6	133.3	143.6	142.6	144.4	131.6	71.1	66.9	141.0	134.9	131.8	131.0
C3	137.6	128.8	124.9	154.5	122.1	57.5	64.4	202.4	124	85.6	137.8	128.0	127.4	122.3	125.5
C4	133.1	74.8	26.6	27.7	25.7	43.4	43.0	32.7	25.6	26.4	137.1	69.5	26.0	26.2	23.1
C5	69.0	39.1	39.7	36.6	38.3	56.2	63.0	34.5	36.9	37.5	67.8	37.3	35.7	39.5	40.9
C6	71.9	30	137.1	140	74.2	139.8	139.8	79.5	27.7	83.0	71.3	24.9	145.5	142.5	80.0
C7	128.5	34.1	124.5	35.3	140.3	124.5	124.6	144.7	51.0	143.9	127.5	32.5	36.3	119.3	144.3
C8	110.5	67.8	58.7	68.7	117.9	39.7	40.4	114.2	202.8	119.2	110.5	62.3	68.2	65.2	114.0
C9	19.1	12.9	17.3	9.1	23.5	18.3	17.7	17.8	17.6	27.0	27.7	12.8	61.2	17.3	24.2
C10	28.3	22.2	16.0	111.7	38.5	12.6	11.8	25.1	19.6	26.8	28.0	19.5	111.1	16.2	17.7

Table 2. Expected (Exp.) and predicted (Pred.) substituents on myrcane skeleton

SITE	1		2		3		4		5		6		7		8	
	Exp.	Pred.	Exp.	Pred.	Exp.	Pred.	Exp.	Pred.	Exp.	Pred.	Exp.	Pred.	Exp.	Pred.	Exp.	Pred.
C-1	Br,Δ^1	Br,Δ^1	OMe, Oxo	Oxo, OH	-	-	Oxo	Oxo	-	-	Δ^1	Δ^1	Δ^1	Δ^1	Δ^1	Δ^1
C-2	-	-	Δ^2	Δ^2	Δ^2	Δ^2	Δ^2	Δ^2	Δ^2	Δ^2	-	-	-	-	-	-
C-3	Δ^3	Δ^3	β	β	-	-	-	-	-	Cl	α-Br	α-Cl	α-Cl	α-Cl	3-oxo	3-oxo
C-4	-	-	Oxy	Oxy	-	-	-	-	-	-	-	-	-	-	-	-
C-5	α-Cl	β-Cl	-	-	-	-	-	-	-	-	β-Br	β-Cl	α-Cl	β-Cl	-	-
C-6	α-Cl	α-Cl	-	-	Δ^6	Δ^6	$\Delta^{6,10}$	$\Delta^{6,10}$	α-Cl	α-Cl	Δ^6	Δ^6	Δ^6	Δ^6	β -OGly	β -OGly
C-7	Δ^7	Δ^7	-	-	-	-	-	-	Cl, Δ^7	Cl, Δ^7	-	-	-	-	Δ^7	Δ^7
C-8	Br	Br	Oxy	Oxy	OH	OAc	OGly	OGly-(2',-OAc)	-	-	Cl	Cl	Cl	Cl	-	-
C-9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
C-10	B	β	β	β	-	-	-	-	Br, β	Br, β	-	-	-	-	α	β
% Recognition	90.00		90.00		90.00		90.00		90.00		80.00		90.00		90.00	

SITE	9		10		12		13		14		15	
	Exp.	**Pred.**	**Exp.**	**Pred.**	**Exp.**	**Pred.**	**Exp.**	**Pred.**	**Exp.**	**Pred.**	**Exp.**	**Pred.**
C-1	-	-	-	-	OMe, Oxo	Oxo, OH	-	-	-	-	-	-
C-2	Δ^2	Δ^2	OH, α	OH, β	Δ^2	Δ^2	Δ^2	Δ^2	Δ^2	Δ^2	Δ^2	Δ^2
C-3	-	-	Oxy	Oxy	α	α	-	-	-	-	-	-
C-4	-	-	-	-	Oxy	Oxy	-	-	-	-	-	-
C-5	-	-	-	-	-	-	-	-	-	-	-	-
C-6	-	-	Oxy	Oxy	-	-	$\Delta^{6,10}$	$\Delta^{6,10}$	Δ^6	Δ^6	α- OGly-(6'-OFuc)	OGly-(6'-OAra)
C-7	-	-	Δ^7 , β	Δ^7 , β	-	-	-	-	-	-	Δ^7	Δ^7
C-8	Oxo	Oxo	-	-	Oxy	Oxy	OGly-(3'MeBu-4',6'-Ac)	OGly-(3'MeBu-4',6'-Ac)	OGly-$(OAc)_3$-[6'-ORha-$(OAc)_3$]	OGly-$(OAc)_3$-[6'-OAra-$(OAc)_3$]	-	-
C-9	-	-	-	-	-	-	OH	OAc	-	-	-	-
C-10	α	β	α	α	β	β	-	-	-	-	β	B
% Recognition	90.00		90.00		90.00		90.00		90.00		90.00	

Programs have previously been written which could identify substituents and their positions on the skeleton of natural products, though they are applied in the identification of skeletons of unknown compounds. For example, the program REGRAS, by analysis of ^{13}C NMR from a given compound and, from ranges of chemical shifts, could identify the chemical functions existing on specific positions of carbon skeletons and at the end of the procedure match the types of carbon atoms obtained against a database, displaying as results the likely skeletons of the question substance. This process was tagged 'disfunctionalization'. The program MACRONO could identify substituent groups attached to any of the atoms in the conventional skeleton of a natural product. The program was developed for finding the subspectra due to the carbons in the said substituent groups among the raw ^{13}C NMR spectroscopic data from any given natural product (by means of comparisons of all possible subsets of all the observed chemical shifts with those contained in an apposite database, built with literature of ^{13}C NMR spectroscopic data regarding those groups). The chemical shifts due to the skeletal carbons from the initial dataset, is then used as input into the expert system SISTEMAT, for skeletal identification.

Similarly, a practical use for the current work may be realized by creating a database of ^{13}C NMR spectroscopic data of skeletons of several classes of natural products and developing a program which can identify the ^{13}C NMR data due to the skeleton of an unknown compound (by comparison of the raw 13C NMR data of the compound with those contained in the database). The output (^{13}C NMR data of the skeleton of the unknown) can be fed as input into the Generalized Regression Neural Network (GRNN) for prediction of the substituents and their positions on the skeleton. This would further simplify the process of structural elucidation of organic compounds.

4. CONCLUSION

Neural networks learn from examples and acquire their 'knowledge' by induction. They can generalize, provide flexible non-linear models of input/output relationships can cope with noisy data and are fault-tolerant [19]. From this study, it could be seen that the predictions obtained using the GRNN were in good agreement with the actual substituents on the skeletons of the test compounds. Where the skeleton type of a natural product has been ascertained by sequential comparison of unknown target spectrum with a set of library spectra or using ANNs, GRNN could be an excellent complimentary tool to use in predicting the nature of substituents attached to myrcane skeletons.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Ferreira MJP, Emerenciano VP, Linia GAR, Romoff P, Macari PAT, Rodrigues GV. 13C NMR spectroscopy of monoterpenoids. Progress in Nuclear Magnetic Resonance Spectroscopy. 1998;33:153.
2. Yongquan H. Evolutionary algorithm as an approach for computer assisted structure elucidation of organic and bioorganic compounds. Ph.D Thesis. Max-Planck-Institute for Chemical Ecology and Friedrich-Schiller-University: Germany; 2003. Accessed : February 21, 2014. Available : http://www.db-thueringen.de/servlets/DerivateServlet/Derivate-8506/thesis.pdf)

3. Elyashberg ME, Blinov KA, Williams AJ, Martirosian ER, Molodtsov SG. Application of a new expert system for the structure elucidation of natural products from their 1D and 2D NMR data. Journal of Natural Products. 2002;65:693-703.
4. Strokov II, Lebedev KS. Computer aided method for chemical structure elucidation using spectral databases and C-13 NMR correlation tables. Journal of Chemical Information & Computer Sciences. 1999;39:659-665.
5. Meiler J, Kock M. Novel methods of automated structure elucidation based on 13C NMR spectroscopy. Magn. Reson. Chem. 2004;42:1042–1045.
6. Rodrigues GV, Campos IPA, Emerenciano VP. 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.
7. Ferreira MJP, Rodrigues GV, Brant AJC, Emerenciano VP. REGRAS: An auxiliary program for pattern recognition and substructure elucidation of monoterpenes. Spectroscopy. 2000;15:65–98.
8. Ferreira MJP, Brant AJC, Rodrigues GV, Emerenciano VP. Automatic identification of terpenoid skeletons through 13C nuclear magnetic resonance data disfunctionalization. Analytica Chimica Acta. 2001;429:151–170.
9. Ferreira MJP, Oliveira FC, Alvarenga SAV, Macari PAT, Rodrigues GV, Emerenciano VP. Automatic identification by 13C NMR of substituent groups bonded in natural product skeletons. Computers & Chemistry. 2002;26:601–632.
10. Ferreiraa MJP, Emerenciano VP , Linia GAR, Romoff P, Macarib PAT, Rodrigues GV. ^{13}C NMR spectroscopy of monoterpenoids. Progress in Nuclear Magnetic Resonance Spectroscopy.1998;33:153–206.
11. Rufino AA, Brant AJC, Santos JBO, Ferreira MJP, Emerenciano VP. Simple method for identification of aporphine alkaloids from 13C NMR data using artificial neural networks. J. Chem. Inf. Model. 2005;45:645-651.
12. Wrede P, Landt O, Klages S, Faterni A, Hahn U, Schneider G. Peptidase design aided by neural networks: Biological activity of artificial signal peptidase I cleavage sites. Biochemistry. 1998;37:3588-3593.
13. Fernandes MB, Scotti MT, Ferreira MJP, Emerenciano VP. Use of self-organizing maps and molecular descriptors to predict the cytotoxic activity of sesquiterpene lactones. European Journal of Medicinal Chemistry. 2008;43:2197-2205.
14. Aires-de-Sousa J, Hemmer M, Gasteiger J. Prediction of 1H NMR chemical shifts using neural networks. Analytical Chemistry. 2002;74(1):80-90.
15. Fraser L, Mulholland DA. A robust technique for group classification of the C-13 NMR spectra of natural products from *Meliaceae*. Fresenius J Anal Chem. 1999;365:631-634.
16. Binev Y, Aires-de-Sousa J. Structure-based predictions of 1H NMR chemical shifts using feed-forward neural networks. Chem. Inf. Comput. Sci. 2004;44:940-945.
17. Jang JSR, Sun CT, Mizutani E. Neuro-fuzzy and soft computing: A computational approach to learning and machine intelligence, Prentice Hall, Upper Saddle River, New Jersey, USA; 1997. [Chapter 9].
18. Hannan SA, Manza RR, Ramteke RJ (2010). Generalized regression neural network and radial basis function for heart disease diagnosis. International Journal of Computer Applications. 2010;7(13):7-13.
19. Specht. A general regression neural network. IEEE Transactions on Neural Networks.1999;2(6):568-576.

20. Sun G, Hoff SJ, Zelle BC, Nelson MA. 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. 2008;51(2):685-694.
21. Mahesh C, Kannan E, Saravanan MS. Generalized regression neural network based expert system for hepatitis b diagnosis. Journal of Computer Science. 2014;10(4):563-569.
22. Schneider G, Wrede P. Artificial neural networks for computer-based molecular Design. Progress in Biophysics & Molecular Biology. 1998;70:175-222.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sciencedomain.org/review-history.php?iid=528&id=16&aid=5443