academicJournals

Vol. 8(7), pp.102-109 July, 2014
DOI:10.5897/AJPAC2014.0569
Article Number:C1B386B46244
ISSN 1996 - 0840
Copyright © 2014
Author(s) retain the copyright of this article
http://www.academicjournals.org/AJPAC

African Journal of Pure and Applied Chemistry

Full Length Research Paper

Prediction of substituent types and positions on skeleton of eudesmane-type sesquiterpenes using generalized regression neural network (GRNN)

Alawode T. T.[1]* and Alawode K. O.[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.

Received 22 April, 2014; Accepted 15 July, 2014

Sesquiterpenes are formed from countless biogenetic pathways and are therefore a constant challenge to spectroscopists in structure elucidation. In this study, we explore the ability of generalized regression neural network (GRNN), an architecture of artificial neural networks (ANNs), to predict the substituent types on eudesmanes, one of the most representative skeletons of sesquiterpenes. Carbon-13 (^{13}C) nuclear magnetic resonance (NMR) chemical shift values of skeletons of 291 eudesmane sesquiterpenes 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. After training, the network was simulated using 34 test compounds. The results showed that the GRNN had between 73.33 to 100% recognition rates of the test compounds. GRNN could therefore be a powerful aid in the structural elucidation of organic compounds.

Key words: Artificial neural networks (ANNs), generalized regression neural network (GRNN), eudesmane skeleton, sesquiterpenes, structural elucidation.

INTRODUCTION

Many phytochemical research efforts are directed at isolation of the compounds responsible for the activities displayed by plants. Elucidation of structures of the isolated compounds from their proton nuclear magnetic resonance (^{1}H NMR) and Carbon-13 (^{13}C) NMR spectra is often a difficult task. Computer-assisted structure elucidation (CASE) methods have been developed to help in this regard. CASE seeks to find, within a given solution space, the single structure that best fits a set of chemical and spectral boundary conditions. Structural elucidation involves finding, from structural information of an unknown compound derived from chemical and/or spectra evidence, the fittest structural formula that satisfies all the constraints (Yongquan, 2003). The input information consists of molecular formula derived from mass spectrometry or elemental analysis, and routine ^{1}D and ^{2}D NMR spectra.

The starting point for structure elucidation is molecular formula derived from Mass Spectrometry (MS), ^{1}D and ^{2}D NMR spectra. The collective spectral information is interpreted as a set of substructures predicted to be present or absent in the unknown. The deduced

*Corresponding author. E-mail: onatop2003@yahoo.com.

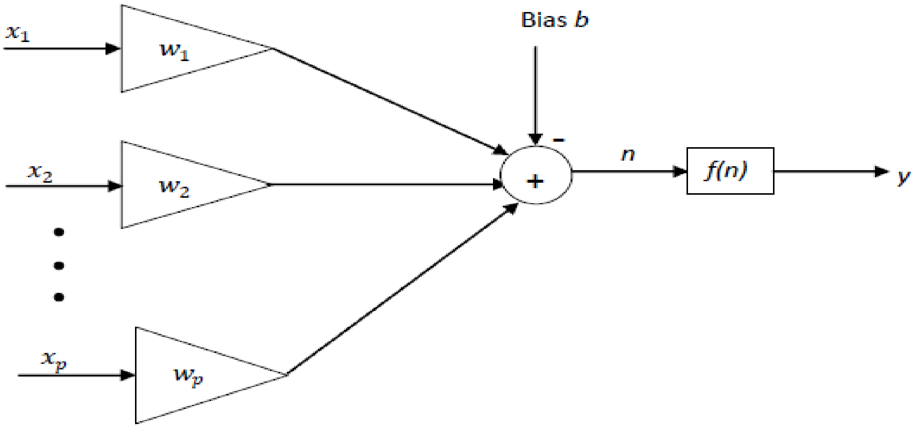

Figure 1. Single neuron model.

information, together with its molecular formula, is the usual input in structure generation. 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 (Elyashberg et al., 2002; Strokov and Lebedev, 1999). 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 (Meiler and Kock, 2004).

A structure elucidation problem is equivalent to a combinatorial optimization problem if the spectra-based structural information of the unknown is treated as constraints to be satisfied. The central task is thus to prune the size of the search space to a computationally acceptable extent. The methods mentioned above attempt to reduce the size of the search by taking advantage of problem-specific information. Nevertheless, pruning heuristics are not always enough because the incompleteness of chemical and/or spectroscopic evidence as the existence of vague information makes the actual search space expand drastically (Yongquan, 2003).

Artificial Neural Networks (ANNs) are defined as computational models with structures that are derived from a simplified concept of the brain, in which a number of nodes, called neurons, are interconnected in a network-like structure (Scotti et al., 2012). Due to its parallel nature, ANNs, could speed up the process of structural elucidation 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 (Rufino et al., 2005). ANNs are employed in pattern recognition problems, especially those associated with prediction, classification or control. The technique has been applied to the prediction of biological activity of natural products or congeneric compounds (Wrede et al., 1998; Fernandes et al., 2008), the identification, distribution and recognition of patterns of chemical shifts from ^{1}H-NMR spectra (Aires-de-Sousa et al., 2002; Binev and Aires-de-Sousa, 2004) and identification of chemical classes through ^{13}C-NMR spectra (Fraser and Mulholland, 1999).

Neural networks are nonlinear processes that perform learning and classification. ANNs 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. Figure 1 depicts a single neuron model. Such a neuron first forms weighted sum of the inputs.

$$n = (\sum_{i=1}^{P} w_i x_i) + b$$

where *P* and w_i are the number of elements and the interconnection weight of the input vector x_i, respectively, and *b* is the bias for the neuron. The knowledge is stored as a set of connection weights and biases. The sum of the weighted inputs with a bias is processed through an activation function, represented by *f*, and the output that it computes is as follows:

$$f(n) = f\left[\sum_{i=1}^{P} w_i x_i) + b\right]$$

There are many ways to define the activation function such as the threshold function, sigmoid function, and the hyperbolic tangent function. The type of activation function depends on the type of the neural network to be designed. A neural network can be trained to perform a particular function by adjusting the values of connections that is, weighting coefficients, between the processing elements.

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. This type of learning is known as supervised learning. 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. Each unit can send its output to the units on the higher layer only and receive its input from the lower layer. This structure is known as multilayer perceptron.

Rufino et al. (2005) showed that ANNs 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, which 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.

In the present work, we show that where the skeleton of a class of compounds has been identified, the substituents positions and types on the skeleton can be predicted using generalized regression neural network (GRNN), one of the architectures of ANNs. We focus on eudesmane-type compounds, one of the most representative skeletons of sesquiterpenes. Sesquiterpenes are formed from countless biogenetic pathways and therefore produce several types of carbon skeletons (Oliveira et al., 2000; Ferreira et al., 2004). This makes elucidation of their structure very challenging. In a previous work, Olievera et al. (2000) described the use of the expert system, SISTEMAT, as an auxiliary tool in the process of structure elucidation of eudesmanes. Eudesmane-type sesquiterpenoids and their biological activities have been the focus of numerous phytochemical, pharmacological and synthetic studies. Since sesquiterpenes exhibit a wide range of biological activities, and include compounds that are plant growth regulators, insect antifeedants, antifungals, anti-tumour compounds and antibacterials, much efforts has been directed at relating their structures to function (Wu et al., 2006).

A GRNN is based on kernel regression networks (Celikoglu and Cigizoglu, 2007; Cigizoglu and Alp, 2005; Kim et al., 2004; Hannan et al., 2010). A GRNN does not require an iterative training procedure. It approximates any arbitrary function between input and output vectors, drawing the function estimate directly from the training data. In addition, it is consistent that as the training set size becomes large, the estimation error approaches zero, with only mild restrictions on the function (Kim et al., 2004; Hannan et al., 2010).

A GRNN consists of four layers: input layer, pattern layer, summation layer and output layer as shown in Figure 2. The number of input units in 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 (Jang et al., 1997; Hannan et al., 2010):

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

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

where 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, and σ is the spread parameter, whose optimal value is determined experimentally.

Compared to other ANN models such as the backpropagation (BP) neural network model, the GRNN needs only a fraction of the training samples a BP 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 (Specht, 1991). 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

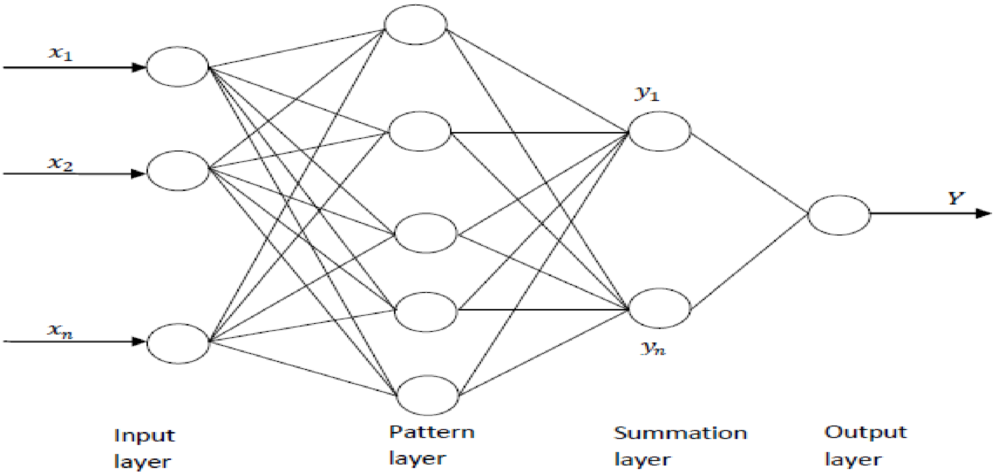

Figure 2. General structure of GRNN.

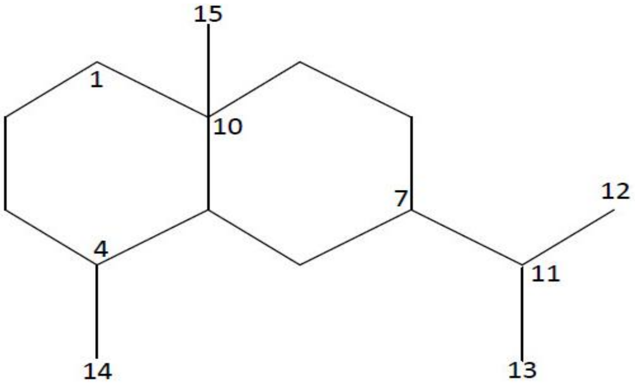

Figure 3. The eudesmane skeleton.

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 BP 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 (Sun et al., 2008; Mahesh et al., 2014).

MATERIALS AND METHODS

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 (Rodrigues et al., 1997). 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 325 (out of the total 350) eudesmane compounds published by Olievera et al. (2000). This information can be extracted from data of eudesmane sesquiterpenes published in literature by isolating ^{13}C values of the skeletal (carbon) from those of the substituents. The compounds left out were those whose substituents were not stated explicitly due to structural complexity. 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 325 compounds used, 34 were reserved for use as test cases (these were not used in training the neural network). The structure of the eudesmane skeleton with the numbering of each carbon atom is shown in Figure 3.

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 1-291. 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 1-15. The ^{13}C chemical shift data for each Carbon at each of the 15 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 1-34). The ^{13}C chemical shift data for skeletons of the test compounds are presented in Table 1. Since ANNs learn through examples, the test compounds were selected based on the representativeness of their substitution patterns in the table of structural information published by Oliveira et al. (2000). This was done largely by visual inspection. These represent the input data for the test compounds.

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 (MATLAB and Statistics Toolbox Release, 2009a). 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

Table 1. ^{13}C NMR chemical shift data for test compounds.

Site	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
C-1	78.9	84.5	76.1	79	70.7	72.6	71.6	75	70	76.2	73.4	75.3	68.4	41.6	40.5	41.4	76.7	76.8
C-2	26.7	23.4	71.3	23.4	71.3	67.7	70.4	67.3	67.9	70.7	24	25.1	68.2	19.2	20.8	20.1	32.1	26.7
C-3	45.2	43	32.4	26.7	30.9	44.1	31.5	41.1	42.2	31.7	38.1	37.9	41.9	41.6	39.5	44.5	121.2	32.1
C-4	75.4	82.5	33.5	39.9	39.2	70.2	39.8	72.1	69.6	34.2	70.4	70.5	69.6	71.9	71.8	77	133.5	139.2
C-5	55.3	57.4	91.4	88.5	87	91.7	87.6	91.5	91.2	92.4	92.6	92.1	91.1	51	53	48.5	50.8	136.4
C-6	69.7	69.4	75.1	32	35.9	69.2	36.1	76.9	78.1	75.5	78.1	72.6	71.7	77.9	70.9	20.5	71.4	206.8
C-7	49.9	49.8	53	48	43.7	54.1	44	53.6	49.2	65.8	52.1	53.2	49	44.4	49.8	41.7	49.3	57.5
C-8	21.2	23.8	72	70	31.1	77.3	31.3	73.8	34.5	198.7	77.2	78.1	34.6	23.2	26.7	21.3	20.3	21.7
C-9	41	33.1	75.7	74.3	74.3	72.3	73.8	75.3	69.8	80.4	76.5	70.3	78.1	39.1	80.5	41.5	35.4	37
C-10	34.8	48.5	49	49	47	50.1	47.4	50.6	55.1	52.7	47.9	52.4	55.2	37.3	39.4	34.2	37.7	43
C-11	28.9	29.6	81.3	80.5	82.3	84.8	82.6	84.4	84.6	84.1	84.1	82.7	84.5	24	28.7	74.6	28.6	25.8
C-12	21.2	21.8	24.1	22.9	24	26.7	19.3	30	25.7	25.6	29.7	24.3	25.5	22.3	21.3	29.9	22.2	18.2
C-13	20.8	21	30.7	29.9	30.2	30.3	20.1	26.7	29.4	31.2	25.5	29.5	25.1	25.3	20.4	29.5	20.1	21
C-14	21.6	17.8	18.7	16.1	19.4	25.5	24.5	24.2	25.1	18.6	23.7	22.7	29.2	23.4	29.7	21.8	20.7	20.7
C-15	15.3	22.8	13.3	61.2	20	20.7	30.4	61.7	65.9	61	13.3	60.5	65.2	19.5	13.7	18.4	12.2	18.3

Table 1. Contd.

Site	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
C-1	37.4	79.2	80.7	31.9	36.5	81.5	78.7	33.4	76.1	75.5	42.9	40.9	44.3	37.8	30.7	41.9
C-2	33.8	32.5	28	25.1	25.6	40.6	30	23.1	29.6	27.8	19.5	22.7	75.3	23	18.5	26.7
C-3	199	35.2	33.8	73.4	72.9	39.4	39.7	72.3	39.1	35.6	43.6	39.7	121.6	121.1	32.2	37.8
C-4	128.9	146.4	148.2	75.4	85.7	70.9	71	83.4	65.9	81.3	73.5	79.7	139	134.8	143.4	143
C-5	162.6	56.2	48.7	48.9	48.6	46.8	47.6	45.6	60.6	56.7	57.9	47.3	47.2	46.9	129	57.9
C-6	28.8	67.2	24.2	143.2	140.1	23.5	26.8	26.1	66.3	71.7	73.3	19.6	28.9	29.4	32.9	69.3
C-7	49.7	49.6	47.6	145.4	145.1	142.1	129.9	130.5	56.3	50.7	50.3	39.4	40.1	40.1	37.1	48.1
C-8	22.6	18.5	21.9	201.3	200.3	116.1	202	210.7	67.5	25.2	26.8	23.3	26.7	27.4	35.7	23.9
C-9	42	36.5	36.6	57.7	57.6	23.1	55.4	60.4	44.5	39.9	42.6	40.6	39.8	40.1	79.9	40.4
C-10	35.9	41.8	39.1	39.2	40	36.9	40.3	36	42.4	40.8	36.3	35	35.2	32.3	39.1	37.4
C-11	72.4	26.3	72.7	72	71.7	35.1	146.4	146.1	137.7	143.7	142.3	146.6	145.1	145.3	131.6	147.3
C-12	26.8	21.1	27	29.3	28.9	21.9	23.1	23.7	128.8	125.2	125.8	110.8	125.1	172.4	125.5	124.6
C-13	27.5	16.4	27.2	28.8	29.1	21.3	23.8	23.1	167.4	167.9	168.3	22.7	172.3	125	170.8	168.1
C-14	10.9	107.9	107.2	22.4	18.7	29.9	25.9	19.3	63.7	74.8	23.8	18.1	21	21.1	19.8	106.9
C-15	22.6	11.7	11.2	17.7	18.5	12.9	12.7	18.3	12.9	15.3	19.7	18.8	16.4	15.7	19	17.6

the training of test data include perceptron, feed-forward BP and GRNNs. 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 each 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 eudesmane skeleton. After trying several neural network types and network parameters, the GRNN at a spread constant of 1.0 was found to give the best results.

RESULTS AND DISCUSSION

Eudesmanes may or may not be oxygenated. Oxygenated eudesmanes may be alcohols, ethers, epoxides, peroxides, aldehydes, ketones, carboxylic acids and lactones. These different functional group substituents are important in determining the individual biological activities of the various sesquiterpenoids, hence the need to correctly predict the substituent types and their positions on the skeleton.

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 (15). This ranged between 73.33 and 100% except for test compounds 8 and 12 where 33.33 and 40%, respectively were obtained.

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

Site	1		2		3		4		5		6		7	
	Exp.	Pred.	Exp.	Pred.	Exp.	Pred.	Exp.	Pred.	Exp.	Pred.	Exp.	Pred.	Exp.	Pred.
C-1	β-OH	β-OH	β-Oxy	α -Oxy	β-OH	α-OGly	OAc	OBzt	OAc	OAc	β -OAc	α -OAc	β -OAc	β –Oac
C-2	-	-	-	-	β-OH	α-OGly	-	-	OBzt	OBzt	β-OH	OEpcin	β-OiBu	α-OBut
C-3	-	-	-	-	-	-	-	-	-	-	-	-	-	-
C-4	α-OH	β-OH	β-Oxy	α -Oxy	-	-	-	-	-	-	α-OH	α-OH	-	-
C-5	-	-	-	-	α -Oxy	α -Oxy	α -Oxy	α-Oxy	α-Oxy	α-Oxy	α-Oxy	α-Oxy	α-Oxy	α-Oxy
C-6	OCin	α- OCin	β-OCin	α-OGly (OAc)$_4$	α-OAc	α-OAc	-	-	-	-	α-OAc	α-OAc	-	-
C-7	-	-	-	-	-	-	-	-	-	-	-	-	-	-
C-8	-	-	-	-	β-OBzt	α -OBzt	β -OAc	α-OAc	-	-	α-OBzt	α-OBzt	-	-
C-9	-	-	-	-	β-OBzt	β-OBzt	OBzt	OBzt	α-OEpcin	α-OEpcin	α-OBzt	α-OBzt	α- OCin	α-O-trans(3'-OAc-2-butenoate)
C-10	-	-	-	-	-	-	-	-	-	-	-	-	-	-
C-11	β	β	β	β	Oxy, α	Oxy, α	Oxy, α	Oxy, α	Oxy, α	Oxy, α	Oxy, α	Oxy, α	Oxy, α	Oxy, α
C-12	-	-	-	-	-	-	-	-	-	-	-	-	-	-
C-13	-	-	-	-	-	-	-	-	-	-	-	-	-	-
C-14	β	α	α	β	β	β	β	β	β	β	β	β	β	β
C-15	β	β	β	β	β	β	OAc	OAc	β	β	β	β	β	β
% Recognition	80		73.33		80		86.67		100		86.67		**86.67**	

Table 2. Contd.

Site	8		9		11		12		13		20		21		22		23	
	Exp.	Pred.	Exp.	Pred.	Exp.	Pred.	Exp.	Pred.	Exp.	Pred.	Exp.	Pred.	Exp.	Pred.	Exp.	Pred.	Exp.	Pred.
C-1	α-OAc	OAc	β-OBut	β-OBut	α-OBzt	α-OBzt	β-OCin	β-OCin	α-ONic	α-ONic	β -OH	β -OH	β -OAc	β -OH	-	-	-	-
C-2	α-OAc	OiBu	β-OBut	β-OBut	-	-	-	-	α-OAc	α-OAc	-	-	-	-	-	-	-	-
C-3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	α-OH	α-OH	β-OAng	Δ^1
C-4	β-OH	α-OH	α-OH	α-OH	β-OH	β-OH	α-OH	β -OH	β -OH	β -OH	$\Delta^{4(14)}$	$\Delta^{4(14)}$	$\Delta^{4(14)}$	$\Delta^{4(14)}$	α-OH	β -OH	β -OAc	β -OAc
C-5	β-Oxy	α-Oxy	α-Oxy	α-Oxy	β-Oxy	β-Oxy	α-Oxy	β -Oxy	Oxy	Oxy	-	-	-	-	-	-	-	-
C-6	β-OH	α-OH	α-OAc	α-OAc	β -OAc	β-OAc	α-OAc	β -OAc	OAc	OAc	α-OH	α-OH	-	-	Δ^6	Δ^6	Δ^6	Δ^6
C-7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
C-8	β-OiBu	α -OAc	-	-	β -OAc	β-OH	β -OAc	O-Cis-(3'-OAc-2-butenoate)	-	-	-	-	-	-	Oxo	Oxo	Oxo	Oxo
C-9	α-OBzt	OBzt	α-OBzt	α-OFur	α-OAc	α-OAc	β-OBzt	α-OBzt	OFur	α-OEpcin	-	-	-	-	-	-	-	-
C-10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
C-11	Oxy, β	Oxy, α	Oxy, α	Oxy, α	Oxy, β	Oxy, β	Oxy, α	Oxy, β	Oxy	Oxy	**-**	β	OH, β	OH, β	OH	OH	OH	OH
C-12	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
C-13	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
C-14	α	β	β	B	A	α	β	α	α	A	-	-	-	-	β	α	α	α
C-15	β-OiBu, α	OiBu	OAc, β	OAc, β	A	α	OAc, β	OAc, α	OAc, α	OiBu, α	β	β	β	β	β	β	β	β
% Recognition	**33.33**		93.33		93.33		40		86.67		93.33		93.33		80		**93.33**	

Table 2. Contd.

Site	24		25		26		27		28		30		33	
	Exp.	**Pred.**	**Exp.**	**Pred.**	**Exp.**	**Pred.**	**Exp.**	**Pred.**	**Exp.**	**Pred.**	**Exp.**	**Pred.**	**Exp.**	**Pred.**
C-1	β-OAc	β-OH	α-OH	α-OH	-	-	β -OH	β -OH	βO(αOH-dihydroCou)	βO(α-OH-iVa)	-	-	-	-
C-2	-	-	-	-	-	-	-	-	-	-	-	-	-	-
C-3	-	-	-	-	β-OAng	α-OEpang	-	-	-	-	-	-	-	-
C-4	β-OH	β-OH	α-OH	α-OH	α-OAc	α-OAc	β -OH	β -OH	β -OH	β -OH	β-OFuc(3’4’Oisopropylidene)	β-OTig	Δ^{4}	Δ^{4}
C-5	-	-	-	-	-	-	-	-	-	-	-	-	-	-
C-6	-	-	-	-	-	-	α-OH	α-OH	α-OH	α-OH	-	-	-	-
C-7	Δ^{7}	Δ^{7}	$\Delta^{7(11)}$	$\Delta^{7(11)}$	$\Delta^{7(11)}$	$\Delta^{7(11)}$	-	-	-	-	-	-	-	-
C-8	-	-	Oxo	Oxo	Oxo	Oxo	α-OH	α-OMe Acr(4’OH)	-	-	-	-	-	-
C-9	-	-	-	-	-	-	-	-	-	-	-	-	β-OAc	β-OAc
C-10	-	-	-	-	-	-	-	-	-	-	-	-	-	-
C-11	-	-	-	-	-	-	Δ^{11}, β	Δ^{11}, β	Δ^{11}, β	Δ^{11}, β	Δ^{11}, β	Δ^{11}, β	Δ^{11}	Δ^{11}
C-12	-	-	-	-	-	-	-	-	-	-	-	-	-	-
C-13	-	-	-	-	-	-	Oxo, OMe	Oxo, OMe	Oxo, OMe	Oxo, OMe	-	-	OH, Oxo	OH, Oxo
C-14	A	α	β	β	β	β	OH, α	OH, α	O-Gly, α	O-Gly, α	α	α	-	-
C-15	B	β	α	α	β	β	β	β	β	β	α	α	β	β
% Recognition	93.33		100		93.33		93.33		93.33		93.33		**100**	

Results for test compounds 10, 14, 15, 16, 17, 18, 19, 31, 32 and 34 are 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 these compounds. From the results presented in Table 2, there is 100% recognition of the un-substituted positions (designated as ‘-’) on the eudesmane skeleton in all the compounds tested The results obtained when perceptron and feed-forward BP neural networks (employing varying network parameters) were used are not presented since the substituents predicted to be on the eudesmane skeleton for all the test compounds, are largely inaccurate.

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 (Schneider and Wrede, 1998). 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. This is despite the large variations in the nature of substituents on the eudesmane skeleton of the various compounds used in the study. 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 eudesmane skeletons. Moreover, it would also be possible to perform the training of the networks interactively, so that every researcher dealing with the identification of substituents on skeletons of natural products could create a network specialized in groups of such complex substances.

Conflict of interest

The authors declare that no conflict of interest have influenced this study.

REFERENCES

Aires-de-Sousa J, Hemmer M, Gasteiger J (2002). “Prediction of ^{1}H NMR Chemical Shifts Using Neural Networks”. Anal. Chem. 74(1):80-90. http://dx.doi.org/10.1021/ac010737m

Binev Y, Aires-de-Sousa J (2004). "Structure-Based Predictions of 1H NMR Chemical Shifts Using Feed-Forward Neural Networks". Chem. Inf. Comput. Sci. 44: 940-945. http://dx.doi.org/10.1021/ci034228s

Celikoglu HB, Cigizoglu HK (2007). Public transportation trip flow modeling with generalized regression neural networks.

Adv. Eng Softw. 38:71-79. http://dx.doi.org/10.1016/j.advengsoft.2006.08.003

Cigizoglu HK, Alp M (2005). Generalized regression neural network in modelling river sediment yield. Adv. Eng. Softw. 37:63-8. http://dx.doi.org/10.1016/j.advengsoft.2005.05.002

Elyashberg ME, Blinov KA, Williams A J, Martirosian ER, Molodtsov SG (2002). Application of a new expert system for the structure elucidation of natural products from their ^{1}D and ^{2}D NMR data. J. Nat. Prod. 65: 693-703. http://dx.doi.org/10.1021/np0103315

Fernandes MB, Scotti MT, Ferreira MJP, Emerenciano VP (2008). Use of self-organizing maps and molecular descriptors to predict the cytotoxic activity of sesquiterpene lactones. Eur. J. Med. Chem. 43:2197-2205. http://dx.doi.org/10.1016/j.ejmech.2008.01.003

Ferreira MJP, Oliveira FC, Rodrigues GV, Emerenciano VP (2004). ^{13}C NMR Pattern Recognition of Guaiane Sesquiterpenes. Internet Electr. J. Mol. Des. 3(11):737-749.

Fraser L, Mulholland DA (1999). A robust technique for group classification of the C-13 NMR spectra of natural products from Meliaceae. Fresenius J. Anal Chem. 365:631-634. http://dx.doi.org/10.1007/s002160051535

Hannan SA, Manza RR, Ramteke RJ (2010). Generalized Regression Neural Network and Radial Basis Function for Heart Disease Diagnosis. Int. J. Comput. Appl. 7(13):7-13.

Jang JSR. Sun CT, Mizutani E (1997). Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence, Prentice Hall, Upper Saddle River, New Jersey, USA; [Chapter 9]

Kim B, Lee DW, Parka KY, Choi SR, Choi S (2004). Prediction of plasma etching using a randomized generalized regression neural network. Vacuum 76:37-43. http://dx.doi.org/10.1016/j.vacuum.2004.05.018

Mahesh C, Kannan E, Saravanan MS (2014). Generalized regression neural network based expert system for hepatitis b diagnosis. J. Comput. Sci. 10(4):563-569. http://dx.doi.org/10.3844/jcssp.2014.563.569

MATLAB and Statistics Toolbox Release (2009a). The MathWorks, Inc., Natick, Massachusetts, United States.

Meiler J, Kock M (2004). Novel Methods of Automated Structure Elucidation based on ^{13}C NMR Spectroscopy. Magn. Reson. Chem. 42:1042-1045. http://dx.doi.org/10.1002/mrc.1424

Oliveira FC, Ferreira MJP, Nu´n˜ez CV, Rodriguez GV, Emerenciano VP (2000). ^{13C}NMR spectroscopy of eudesmane sesquiterpenes. Prog. Nucl. Magn. Reson. Spectrosc. 37:1-45.

Rodrigues GV, Campos IPA, Emerenciano VP (1997). 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 pp. 191-200.

Rufino AA, Brant AJC, Santos JBO, Ferreira MJP, Emerenciano VP (2005). Simple Method for Identification of Aporphine Alkaloids from ^{13}C NMR Data Using Artificial Neural Networks. J. Chem. Inf. Model. 45:645-651. http://dx.doi.org/10.1021/ci0498416

Schneider G, Wrede P (1998). Artificial neural networks for computer-based molecular design. Prog. Biophys. Mol. Biol. 70:175-222. http://dx.doi.org/10.1016/S0079-6107(98)00026-1

Scotti MT, Emerenciano V, Ferreira MJP, Scotti L, Stefani R, da Silva MS, Mendonça Junior FJB (2012). Self-Organizing Maps of Molecular Descriptors for Sesquiterpene Lactones and Their Application to the Chemotaxonomy of the Asteraceae Family. Molecules 17, 4684-4702. http://dx.doi.org/10.3390/molecules17044684

Specht DF (1991). A General Regression Neural Network. IEEE Trans. Neural Netw. 2(6):568-576. http://dx.doi.org/10.1109/72.97934

Strokov II, Lebedev KS (1999). Computer aided method for chemical structure elucidation using spectral databases and ^{13}C NMR correlation tables. J. Chem. Inf. Comput. Sci. 39:659-665. http://dx.doi.org/10.1021/ci980184p

Sun G, Hoff SJ, Zelle BC, Nelson MA (2008). 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. Am. Soc. Agric. Biol. Eng. 51(2):685-694.

Wrede P, Landt O, Klages S, Faterni A, Hahn U, Schneider G (1998). Peptidase design aided by neural networks: biological activity of artificial signal peptidase I cleavage sites. Biochemistry 37:3588-3593. http://dx.doi.org/10.1021/bi9726032

Wu Q, Shi Y, Jia Z (2006). Eudesmane sesquiterpenoids from the Asteraceae family. Nat. Prod. Rep. 23:699-7134. http://dx.doi.org/10.1039/b606168k

Yongquan H (2003). 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. (Available online: http://www.db-thueringen.de/servlets/DerivateServlet/Derivate-8506/thesis.pdf)