CODEN IEJMAT *Internet Electronic Journal of Molecular Design* **2006**, *5*, 213–223 ISSN 1538–6414
BioChem Press http://www.biochempress.com

Inter*net* Electronic Journal of Molecular Design

April 2006, Volume 5, Number 4, Pages 213–223

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Lemont B. Kier on the occasion of the 75th birthday

Diterpene Skeletal Type Classification and Recognition using Self–Organizing Maps

Vicente de Paulo Emerenciano,[1] Marcus Tullius Scotti,[1] Ricardo Stefani,[2] Sandra A. V. Alvarenga,[3] Jean Marc Nuzillard,[4] and Gilberto V. Rodrigues [5]

[1] Instituto de Química, Universidade de São Paulo, Caixa Postal 26077, 05513–970 São Paulo – SP Brazil
[2] Departamento De Química, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo. Av. Bandeirante, 3900, 14040–901 Ribeirão Preto – SP Brazil
[3] Faculdade de Engenharia de Guaratinguetá, Universidade Estadual Paulista, 12516–410 Guaratinguetá – SP Brazil
[4] FRE2715, University of Reims, Moulin de la Housse, BP, 1039, 51687 Reims Cedex 2 France
[5] Departamentode Química, ICEX, UFGM, Belo Horizonte, MG, Brazil

Received: February 22, 2006; Accepted: March 8, 2006; Published: April 30, 2006

Citation of the article:
V. P. Emerenciano, M. T. Scotti, R. Stefani, S. A. V. Alvarenga, J. M. Nuzillard, and G. V. Rodrigues, Diterpene Skeletal Type Classification and Recognition using Self–Organizing Maps, *Internet Electron. J. Mol. Des.* **2006**, *5*, 213–223, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

Diterpene Skeletal Type Classification and Recognition using Self–Organizing Maps

Vicente de Paulo Emerenciano,[1,*] Marcus Tullius Scotti,[1] Ricardo Stefani,[2] Sandra A. V. Alvarenga,[3] Jean Marc Nuzillard,[4] and Gilberto V. Rodrigues [5]

[1] Instituto de Química, Universidade de São Paulo, Caixa Postal 26077, 05513–970 São Paulo – SP Brazil
[2] Departamento De Química, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo. Av. Bandeirante, 3900, 14040–901 Ribeirão Preto – SP Brazil
[3] Faculdade de Engenharia de Guaratinguetá, Universidade Estadual Paulista, 12516–410 Guaratinguetá – SP Brazil
[4] FRE2715, University of Reims, Moulin de la Housse, BP, 1039, 51687 Reims Cedex 2 France
[5] Departamentode Química, ICEX, UFGM, Belo Horizonte, MG, Brazil

Received: February 22, 2006; Accepted: March 8, 2006; Published: April 30, 2006

Internet Electron. J. Mol. Des. **2006**, *5* (4), 213–223

Abstract

Motivation. Kohonen Self–Organizing Feature Map (SOM Kohonen map) is a technique used for pattern classification. The method can be applied to classify different classes of organic compounds based on ^{13}C NMR chemical shift data. This can be a very useful tool in structure validation, which is one of the steps of automated structure elucidation process. In this paper we present the use of Kohonen ANN to predict and classify different skeletal types of diterpenes.

Method. The Kohonen neural network was trained using Matlab version 6.5 with the package Somtoolbox 2.0. A total of 957 cases belonging to 12 different skeletal types of diterpenes were used to train the network.

Results. During the training phase, 91.12% of the patterns were highly correctly classified, while for the testing phase, 75.22% of the input data were correctly classified by the Kohonen neural network.

Conclusions. As demonstrated by these results, SOM Kohonen neural network can be a reliable tool to predict diterpene skeletal types from ^{13}C NMR spectrum data.

Keywords. Diterpenes; neural networks; Kohonen self–organizing feature map; structure elucidation.

1 INTRODUCTION

In natural products chemistry, NMR spectroscopy is one of the most important techniques available for structure determination of isolated compounds. Nevertheless, spectra interpretation is a task for a well–trained chemist and is a very time consuming step in the elucidation process. This

Dedicated on the occasion of the 75th birthday to Professor Lemont B. Kier.
* Correspondence author; E–mail: vdpemere@iq.usp.br.

fact has inspired the development of expert systems for automatic structure elucidation [1–4]. Two revisions of expert systems in structural determination report the most important results of the area on these last years [5,6]. The computer–assisted methods on which these systems are based involve an artificial intelligence approach to structure elucidation. Problems can be efficiently be solved by the introduction of skeletal constraints that avoid combinatorial explosion to occur [7]. The first step during the interpretation of NMR data is to try to recognize the class the studied compound belongs to. We define here class as the major groups of secondary compounds as steroids, terpenoids, flavonoids, etc. The second step, used mainly in manual interpretation of spectra is the assignment of a skeletal type. Skeletons can be identified from ^{13}C NMR [7], ^{1}H NMR [8] or botanical data [9]. However, it is difficult to associate a spectral pattern to a group of substances of a same skeleton. The main reason for this is that the definition of skeletons or structure types is not standardized and the classification of skeletal types and subtypes goes beyond carbon connectivity and numbering. Thus, skeletons can be classified by configuration (E/Z, or R/S), ring junction geometry (clearodanes and *cis*–clerodanes), inversion of a sterogenic center (labdanes and *ent*–labdanes) or even oxidation state changes. This leads to a great variety of skeletal and substructural types and even though they may cause variations in spectroscopy, it can be very hard to make a good identification of the compound skeletal type at a glance, since the appearance of spectra can be very similar for different skeletons. Due to this difficulty, a computer is a useful tool to solve this problem, as it can treat a large amount of data at the same time.

Artificial neural networks (ANN) are one of the most used approaches to achieve computer classification and pattern recognition, since they are robust and able to detect groupings and patterns that may be unclear even by a trained human expert. Since ANNs are not restricted to linear correlations and can also take into account non–linear data correlations, they can be efficiently applied for modeling, prediction and classification. Hence, ANNs can be trained to aid the classification of compound skeleton types based on NMR data, since skeleton type identification is a pattern recognition and classification problem.

The most used ANN architecture for pattern recognition and classification is the SOM [10]. A SOM can map multivariate data onto a two dimensional grid, grouping similar patterns near each other. Each neuron in the grid is associated to a weight and similar patterns stimulate neurons with similar weight, so that similar patterns are mapped near each other. In the field of natural product spectroscopy, supervised neural networks were used to classify oxidized terpenes (limonoids) [11]. Other specific applications of neural networks in ^{13}C NMR are frequent [12–15]. Many efforts have been made towards classification ^{1}H NMR data using SOMs, with success [16,17]. However, the

same did occur with ^{13}C NMR data, as ^{13}C NMR spectra provide much more information about the chemical structure than ^{1}H NMR, classification of the former spectrum types is preferable than classification of the later type.

Our research group has been developed an expert system named SISTEMAT to aid natural product chemists in structure determination [15] and chemotaxonomy [18] tasks. In our development of SISTEMAT, a Kohonen ANN was trained to automatically determine diterpene skeletal types from ^{13}C NMR data. Diterpenes are one of the most widespread natural product class in the plant kingdom [19]. Many diterpenes present useful biological activities, such as antimicrobial [20] and anticancer [21]. This makes diterpenes a class of great interest for the pharmaceutical industries (Taxol [22], for example is a diterpene) and for researchers interested in new active compounds. Hence, SOM can be used to select compounds that may require further investigation, instead of performing detailed investigation on their spectroscopic data.

2 MATERIALS AND METHODS

All structure and ^{13}C NMR data were extracted from the SISTEMAT database. An in–house program for data extraction was written in Java and subsequently used to select 2600 diterpene compounds. A computer worksheet was created, each line of which corresponds to a spectrum and contains the chemical shifts, multiplicities and carbon types as integers ranging from 1 to 11. The skeleton name each compound belongs to is also reported. Our definition of a skeleton is based on the connectivity between carbon atoms. However, the nature of some stereogenic centers is also considered as characteristic. The Figure 1 shows the skeletons we were interested in. The skeleton naming conventions follow the current usage that prevails in natural product journals.

Table 1. Skeleton type selected for training

ID	Skeleton Name	ID	Skeleton Name
1	12,13–*sec*–13–*nor*–totarane	**7**	clerodane
2	13–*epi*–rosane	**8**	*ent*–atisane
3	abietane	**9**	*ent*–labdane
4	andromedane	**10**	phytane
5	cembrane	**11**	isopimarane
6	cyathane	**12**	labdane

After data extraction, a Perl script [23] was run to build the training set. It retained 957 spectra of compounds belonging to 12 different skeletal types (Table 1). The script also checked whether each chemical shift value was compatible with the number of attached number of hydrogens and oxygen atoms and the aliphatic, olefinic, aromatic or acetylenic nature of the corresponding carbon atom, using appropriate chemical shifts range. Acetylenic carbons atoms (index 8 to 11) were not

considered as the corresponding molecules are under–represented within the database Code 10 becomes 6 and 11 becomes 7. Only compounds with very high–quality spectroscopic data were selected for the training set. For the test set, 113 occurrences spectra were selected. Table 2 describes the training and test set. A Kohonen ANN was then trained and tested for the recognition of diterpene types from ^{13}C NMR spectra.

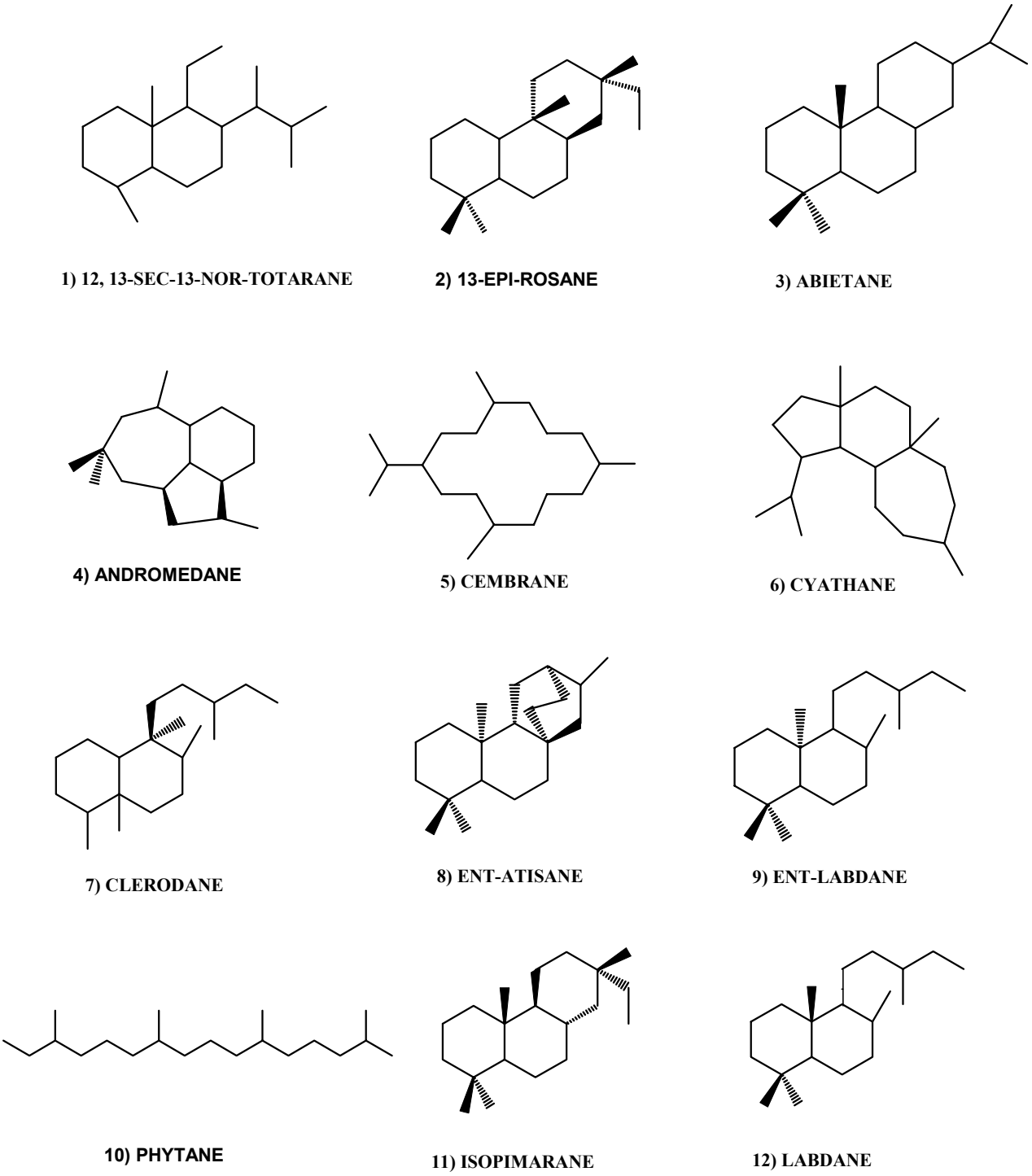

Figure 1. Examples of diterpenes skeletal types used to train the ANN.

Table 2. Summary of training and test set

Diterpene types		Training set		Test set		Total	
Name	ID	Train	% total	Test	% total	Total	% total
12,13–*sec*–13–*nor*–totarane	**1**	38	90.48	4	9.52	42	100.00
13–*epi*–rosane	**2**	17	85.00	3	15.00	20	100.00
abietane	**3**	94	91.26	9	8.74	103	100.00
andromedane	**4**	28	77.78	8	22.22	36	100.00
cembrane	**5**	150	92.59	12	7.41	162	100.00
cyathane	**6**	12	80.00	3	20.00	15	100.00
clerodane	**7**	186	89.86	21	10.14	207	100.00
ent–atisane	**8**	29	82.86	6	17.14	35	100.00
ent–labdane	**9**	47	92.16	4	7.84	51	100.00
phytane	**10**	51	89.47	6	10.53	57	100.00
isopimarane	**11**	97	90.65	10	9.35	107	100.00
labdane	**12**	208	88.51	27	11.49	235	100.00
Total		957	89.44	113	10.56	1070	100.00

2.1 Neural Network Input Descriptors

A worksheet containing 957 rows and 25 columns codifies the 957 diterpenes and their respective ^{13}C NMR chemical shifts. In the last column we give the ID of the skeleton type corresponding to each diterpene, ID that serves only as a label for the SOM. The ^{13}C NMR was coded from the first to the 25th column, 26th column with the skeleton ID. Each line of the worksheet contained ^{13}C chemical shift for each compound. Another Perl script was used to create the input vector for the SOM training. For each compound, chemical shifts are grouped by carbon atom type, as defined in Table 3. Chemical shift values are sorted in ascending order within each group. Each carbon type is associated to a maximum number of retained chemical shifts (Table 3). If the number of shifts within a group is greater than the allowed maximum, the highest shift values are discarded. If the number of shifts within a group is smaller than the allowed maximum, fictitious shifts are added, whose value is first set as "missing". According to Table 3, SOM input variables $\mathbf{x}_1$ to $\mathbf{x}_4$ correspond to chemical shifts of methyl group. If a compound has only three methyl groups, the value of $\mathbf{x}_4$, that is initially "missing" is finally replaced by the mean of all "non missing" $\mathbf{x}_4$ values. Of course the same processing is applied to all input variable values.

Table 3. Diterpene carbon types and number of columns for each type in the input vector

Carbon type	^{13}C signal multiplicity	# of columns
–CH_3	4	4
–CH_2–	3	6
–CH–	2	4
–C–	1	3
=CH2	3	2
=CH	2	3
=C–	1	3

The input vector structure was designed according to the frequency of each carbon type in diterpenes. For example, we found heuristically that the maximum number of carbon atoms

belonging to type 1 in diterpenes is 4, because a real structure often has functionalized positions. An overall number of 25 variables were found to be adequate for classification of diterpenes. Table 3 summarizes each type of carbon and the respective number of columns in the worksheet.

2.2 Neural Network Architecture

A Kohonen ANN was trained using Matlab 6.5 computing environment by Mathworks [24] and SOM Toolbox [25]. Matlab is a powerful and easy to use scientific computing language and is the choice for most scientific simulation and data analysis. SOM Toolbox is a set of Matlab functions that can be used to develop and implement SOM neural networks. A SOM grid with square geometry and 31×20 dimension size was created and trained. The training was conducted through the Batch–training algorithm. In this algorithm, the whole dataset is presented to the network before any adjustment is made. In each training step, the dataset is partitioned according to the regions of the map weight vectors. After this, the weights are calculated as stated by Eq. (1):

$$m_i(t+1) = \frac{\sum_{j=1}^{n} h_{ic}(t)\,\mathrm{X}_j}{\sum_{j=1}^{n} h_{ic}(t)} \tag{1}$$

where **X***j* is an input vector randomly chosen from the input dataset at time *t*, $h_{ic}(t)$ is the neighborhood around the winner neuron and *m* is the *i*–th weight value. Within this algorithm, the new weight vector is simple averages. This feature allows missing values to be ignored by the net. The number of epochs is automatically chosen by the Toolbox, *i.e.*, the neural network is trained until its convergence to minimal error.

2.3 Computer Software and Hardware

The neural network was trained on a Pentium IV HT 3.0 GHz with 1.0 GB of RAM running Windows XP. Matlab R13 (6.5) from Matworks Inc. The Self–Organizing Map Toolbox was employed to train the network.

3 RESULTS AND DISCUSSION

A summary of the obtained results are in Table 5. Figure 2 shows the obtained Kohonen map as clusters after training, while Figure 3 shows the same map in a PCA–like 3D fashion. In Figure 3, the data is projected according to the Euclidian distances among the clusters and contents of each neuron within the map. Table 5 presents the results of each classification both as absolute values and as percentages. The percentage representation is useful from an analytical perspective, since one error on a dataset of size 10 is much more significant than 4 errors on a dataset of size 50.

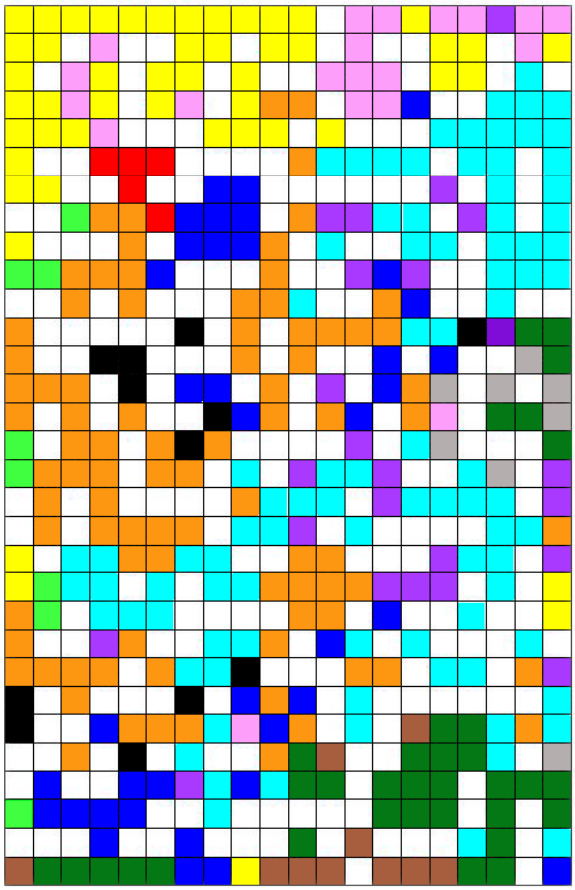

Figure 2. Kohonen map obtained after the training phase. The skeleton type and the respective color are shown in Table 4.

Table 4. Types of skeletons and the respective colors of the regions on the Kohonen map in Figure 2

ID	Name	Color
1	12,13–*sec*–13–*nor*–totarane	black
2	13–*epi*–rosane	gray
3	abietane	blue
4	andromedane	green
5	cembrane	yellow
6	cyathane	red
7	clerodane	orange
8	*ent*–atisane	brown
9	*ent*–labdane	purple
10	phytane	pink
11	isopimarane	dark green
12	labdane	cyan

In order to test the ability of the network to generalize, the 113 test cases randomly selected from SISTEMAT's database were applied to the trained network as new data, taking care that the test set contains representative samples of all trained skeletons.

After training, the neural network was able to classify each skeletal type based on the ^{13}C NMR of the compounds. During the training phase, 91.12% of the data set was successfully classified and

during the test phase 75.22% of the samples were correctly classified. Although the performance of the network during the test phase could be considered unsatisfactory by some chemometricians (only 75% accuracy), these results are still significant, considering the nature of the network (unsupervised) and the difficulty of differentiating diterpenoids types.

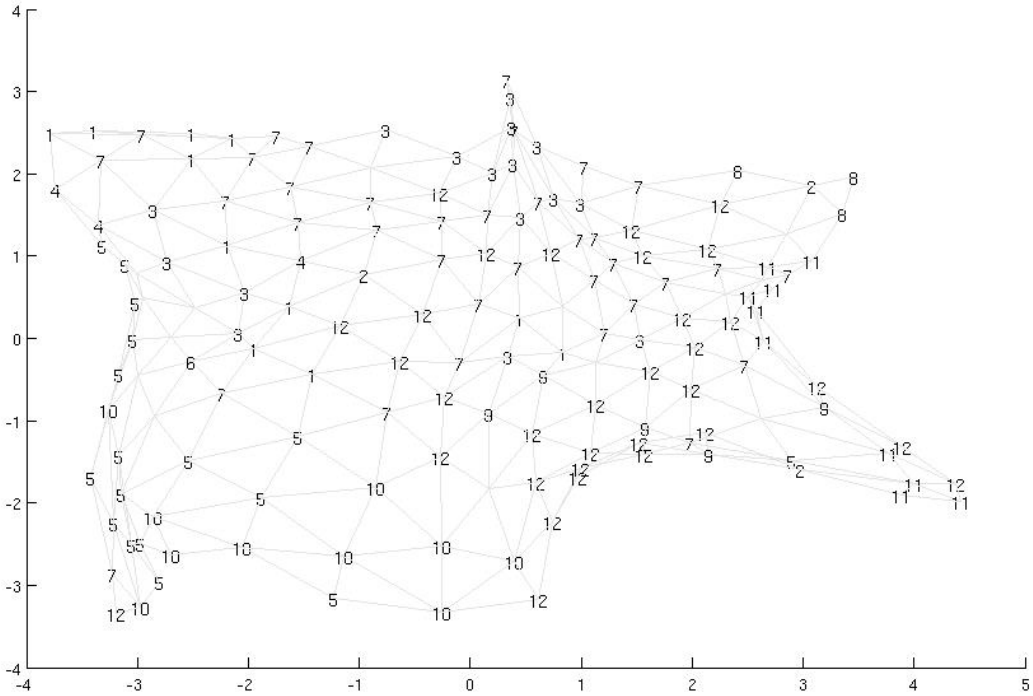

Figure 3. 3D PCA–like projection of the Kohonen map after training. The distances among the clusters are easier to visualize in this mode. The skeleton type and the respective number are shown in Table 1.

Tabel 5. Results obtained with test and training set

	Traininig set		Test set	
ID	Match	%	Match	%
1	30	78.95	4	100.00
2	14	82.35	3	100.00
3	89	94.68	8	88.89
4	24	85.71	5	62.50
5	142	94.67	10	83.33
6	12	100.00	2	66.67
7	169	90.86	15	71.43
8	27	93.10	5	83.33
9	36	76.60	3	75.00
10	40	78.43	4	66.67
11	94	96.91	6	60.00
12	195	93.75	20	74.07
Total	872	91.12	85	75.22

A detailed inspection of the Kohonen map shows that the neural network had correctly classified similar skeletons near each other similar and far from unrelated skeletons. An inspection of both maps clearly shows that the distribution of the diterpenes on the Kohonen map is influenced by the structural properties, which reflects on ^{13}C NMR chemical shifts. The Kohonen map expresses this

fact without having knowledge about structural features of diterpenes, by placing diterpenes with similar structures in close vicinity on the map or even on the same neuron.

Taking for example the percentage of success for very similar skeletons, such as *ent*–labdanes (**9**) and labdanes (**12**) where the difference among skeletons is only the orientation of a methyl group attached to C–10, even with the lower accuracy of the neural network to classify *ent*–labdanes, the performance of the neural network to classify it during the test set is as good as the overall performance of the net. It is also interesting to note that on the 3D projection, the distance between the neurons representing both skeletons is not high and in both diagrams, *ent*–labdane cluster is neighbor of isopimares (**11**), even though, structurally speaking, *ent*–labdanes is much more similar to labdanes. On the other hand, *ent*–labdanes are stereochemicaly similar to isopimaranes. Moreover, this means that the neural network is sensitive to minor variations in the spectral data, which reflects the structural feature and diversity in this specific case: a difference in stereochemistry.

Another interesting feature of the maps is that skeletons that are similar from the biogenetic point of view form vicinal clusters. For example, the two simplest skeletons, from the biogenetic point of view, are the acyclic phytane (**10**) and cembrane (**5**). They appear as two neighbor clusters in the upper part of the map (Figure 2).

In both map representations, the most disperse clusters are that of abietane (**3**, blue) and clerodane (**7**, orange). Even with the non–connectivity of both clusters, the accuracy of neural network to classify both skeletons is high for training set (90.86% for clerodanes, 94.68% for abietanes) and for the test set (71.43% for clerodanes, 88.89% for abietanes). It is remarkable to see that these skeletons are close to the clusters of cyclic skeletons [isopimarane (**11**), labdane (**12**), *ent*–labdane (**9**), 13–*epi*–rosane (**2**)] and share similar, but not identical, features with all these skeletons.

A noticeable feature is that although the accuracy for 12,13–*sec*–13–*nor*–totarane (**1**) and 13–epi–rosane (**2**) during the training phase was low, they were the only two skeletons with 100% of match during the test phase. A careful look at both representation of SOM (2D and 3D PCA–like projection) shows that these skeletons are neighbors of skeletons that have some common features with them.

Nevertheless, the unsupervised training is powerful enough to classify both skeletons, despite of the small number of spectra. In other words, the efficiency of the neural network to classify the skeleton type may depend on the number of occurrences of the skeleton present in the data set, but the relationship is not linear. Our SOM is able to detect features in a NMR spectrum that belong to a specific skeleton and to use this information to group clusters of a skeleton together or near a similar skeleton.

4 CONCLUSIONS

Self–organizing feature maps can be a useful tool for structure elucidation process. The method can be a reliable tool to identify different types of skeletons with good accuracy. Hence, SOM can be used to choose spectra that may require further investigation. NMR spectra analysis is a laborious task and any system that may help the chemist in this task is welcomed. Analysis of spectra with the use of SOMs can also help the rapid classification of which compound that may require further investigation regarding biological activity or chemotaxonomy studies.

Acknowledgment

Vicente P. Emerenciano, Marcus T. Scotti, Sandra A. V. Alvarenga and Gilberto V. Rodrigues acknowledge the financial support of this research by the CNPq and FAPESP Foundations. Ricardo Stefani and Marcus Tullius Scotti acknowledge a doctorate fellowship from CAPES. Jean M. Nuzillard acknowledges grants from CNRS.

5 REFERENCES

[1] C. Djerassi, D. H. Smith, C. W. Crandell, N. A. B. Gray, J. G. Nourse and M. R. Lindley, The Dendral Project: Computational Aids, to Natural Products Structure Elucidation, *Pure and Appl. Chem* **1982**, *54*, 2425–2442.

[2] J. E. Dubois, A. Panaye and R. Attias, DARC System: Notions of Defined and Generic Substructures: Filiation and Coding of frel Substructure (SS) Classes, *J. Chem. Inf. Comput. Sci* **1987**, *27*, 74–82.

[3] B. D. Christie and M. E. Munk,The role of two–dimensional nuclear magnetic resonance spectroscopy in computer–enhanced structure elucidation, *J. Am. Chem. Soc*. **1991**, *113*, 3750–3757.

[4] K. Funatsu and S. Sasaki, Recent advances in the automated structure elucidation system, CHEMICS. Utilization of two–dimensional NMR spectral information and development of peripheral functions for examination of candidates, *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 190–204.

[5] M. Jaspars, Computer assisted structure elucidation of natural products using two–dimensional NMR spectroscopy, *Nat. Prod. Rep*. **1999**, *16*, 241–248.

[6] C. Steinbeck, Recent developments in automated structure elucidation of natural products, *Nat. Prod. Rep*. **2004**, *21*, 512–518.

[7] M. J. P. Ferreira, G. V. Rodrigues and V.P. Emerenciano, MONOREG – An expert system for structural elucidation of monoterpenes. *Can. J. Chem*. **2001**, *79*, 1915–1925.

[8] M. J. P. Ferreira, M. B. Constantin, G. V. Rodrigues and V. P. Emerenciano, A new program for skeleton prediction of iridoids through ^{1}H NMR data, *Spectrosc. Lett*. **2004**, *37*, 587–605.

[9] M. J. P. Ferreira, S. A. V. Alvarenga, P. A. T. Macari, G. V. Rodrigues and V. P. Emerenciano, A program for skeleton prediction of natural products based on botanical information. *Biochem. Syst. Ecol*. **2003**, *31*, 25–43.

[10] T. Kohonen, *Self–Organization and Associative Memory* Third Edition, Springer Verlag, Berlin, 1989.

[11] L. A. Fraser, D. A. Mulholland and D. D. Fraser, Classification of Limonoids and Protolimonoids using Neural Networks, *Phytochem. Anal*. **1997**, *8*, 301–311.

[12] J. W. Ball and P. C. Jurs, Automated selection of regression models using neural networks from ^{13}C NMR spectral predictions, *Anal. Chem*. **1993**, *65*, 505–512.

[13] J. P. Doucet, A. Panaye, E. Feuilleaubois and P. Ladd, Neural networks and ^{13}C shift prediction, *J. Chem. Inf. Comp. Sci*. **1993**, *33*, 320–324.

[14] J. Meiler and M. Will, Automated structure delucidation of organic molecules from ^{13}C NMR spectra using genetic algorithms and neural networks, *J. Chem. Inf. Comp. Sci*. **2001**, *41*, 1535–1546.

[15] A. R. Rufino, A. J. C. Brant, J. O. B. 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*. **2005**, *45*, 645–651.

[16] S. Kalelkar, E. R. Dow, J. Grimes, M. Clapham and H. Hu, Automated Analysis of proton NMR spectra from combinatorial rapid parallel synthesis using self–organizing maps, *J. Comb. Chem*. **2002**, *4*, 622–629.

[17] B. K. Lavine, C. E. Davidson and D. J. Westover, Spectral pattern recognition using Self–organizing maps, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1056–1064.
[18] S. A. V. Alvarenga, M. J. P. Ferreira, G. V. Rodrigues and V. P. Emerenciano, A general survey and some taxonomic implications of diterpenes in Asteraceae, *Bot. J. Linn. Soc.* **2005**, *147*, 291–308.
[19] F. Seaman, F. Bohlmann, C. Zdero and T. J. Mabry, *Diterpenes of flowering plants Compositae* (*Asteraceae*), Springer–Verlang, Berlin, 1990.
[20] H. R. El–Seedi, N. Sata, K. B. G. Torssell and S. Nishiayma, New labdane diterpenes from *Eupatorium glutinosum*, *J. Nat. Prod.* **2002**, *65*, 728–729.
[21] M. Kondoh, F. Nagashima, I. Suzuki, M. Harada, M. Fujii, Y. Asakawa and Y. Wanatabe, Induction of Apoptosis by new *ent*–kaurene type diterpenoids isolated from New Zealand livework *Jungermannia* species, *Planta Med.* **2005**, *71*, 1005–1009.
[22] S. J. Stohs, Taxol in cancer treatment and chemoprevention; in: *Phytopharmaceuticals in Cancer Chemoprevention*, Eds. D. Bagchi, H and G. Preuss, CRC Press LLC, New York, 2005, pp 519–524.
[23] L. Wall, T. Christiansen, J. Orwant, *Programming Perl*, Third Edition, O' Reilly, 2000.
[24] Mathworks Inc. **2004**, http://www.mathworks.com.
[25] J. Vesanto, J. Himberg, E. Alhoniemi and J. Parhankangas, *SOM Toolbox for Matlab 5*, **2005**, http://www.cis.hut.fi/projects/somtoolbox.

Biographies

Vicente de Paulo Emerenciano is professor of chemistry at the University of São Paulo. Ph.D. degree in organic chemistry at the University of São Paulo, Dr. Emerenciano undertook post doctoral research with Professor Jaques–Emile Dubois at Paris VII University. More recently, Prof. Emerenciano has collaborated on projects with Professor Daniel Cabrol–Bass and Dr. Jean–Marc Nuzillard at Nice and Reims Universities as invited Professor and "Chercheur associé" respectively.

Marcus Tullius Scotti and **Ricardo Stefani** are doctorate students at University of São Paulo, Brazil.

Sandra A.V. Alvarenga is Professor of Chemistry as Uinverisidade Estadual Paulista, Sao Paulo, Brazil.

Jean–Marc Nuzillard is Research Director at the Faculty of Chemistry, Reims, France. His research project includes structural determination of organic compounds and computer–assisted structure determination. He developed the LSD program, an structure generator avaliable at www.univ–reims.fr/LSD.

Gilberto V. Rodrigues is Professor of Chemistry as UIniversidade Federal de Minas Gerais, Belo Horizonte, Brazil.