Internet Electronic Journal of
Molecular Design

January 2005, Volume 4, Number 1, Pages 9–16

Editor: Ovidiu Ivanciuc

Proceedings of the Internet Electronic Conference of Molecular Design, IECMD 2003
November 23 – December 6, 2003
Part 7

Neural Networks for Secondary Metabolites Prediction in *Artemisia* Genus (Asteraceae)

Tanja Schwabe,[1] Marcelo J. P. Ferreira,[1] Sandra A. V. Alvarenga,[2] and Vicente P. Emerenciano [1]

[1] Instituto de Química, Universidade de São Paulo, Caixa Postal 26077, 05513–970, São Paulo, SP, Brazil
[2] Faculdade de Engenharia de Guaratinguetá, Universidade Estadual Paulista, 12516–410, Guaratinguetá, SP, Brazil

Received: November 1, 2003; Revised: August 10, 2004; Accepted: September 7, 2004; Published: January 31, 2005

Citation of the article:

T. Schwabe, M. J. P. Ferreira, S. A. V. Alvarenga, and V. P. Emerenciano, Neural Networks for Secondary Metabolites Prediction in *Artemisia* Genus (Asteraceae), *Internet Electron. J. Mol. Des.* **2005**, *4*, 9–16, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

Neural Networks for Secondary Metabolites Prediction in *Artemisia* Genus (Asteraceae) [#]

Tanja Schwabe,[1] Marcelo J. P. Ferreira,[1] Sandra A. V. Alvarenga,[2] and Vicente P. Emerenciano [1,]*

[1] Instituto de Química, Universidade de São Paulo, Caixa Postal 26077, 05513–970, São Paulo, SP, Brazil
[2] Faculdade de Engenharia de Guaratinguetá, Universidade Estadual Paulista, 12516–410, Guaratinguetá, SP, Brazil

Received: November 1, 2003; Revised: August 10, 2004; Accepted: September 7, 2004; Published: January 31, 2005

Internet Electron. J. Mol. Des. **2005**, *4* (1), 9–16

Abstract

Motivation. The chemistry of secondary metabolites is a peculiar field of study due to its complexity and the interest it raises in other fields of pharmacology. The plants of the Asteraceae, one of the largest families of plants, have been intensely studied for this reason and have been resulted in the identification of around 28000 occurrences of substances in the species of the family. The chemistry of the Asteraceae is extremely complex and the great problem with databases compiled from the literature is the lack of knowledge about the precision of the data. Thus, the imprecision of the data leads us to use specific techniques to work with this kind of incomplete data. So, the use of artificial neural networks is very adequate. In the present study we focus attention at the genus *Artemisia* and work at the infra genus level in order to try to predict the occurrence of chemical substances present in the genus.

Method. The methodology applied starts by taking all the information on the genus *Artemisia* from the database. An entry matrix was assembled with the occurrences of the six most representative chemical classes in the genus: flavonoids, monoterpenes, sesquiterpenes, sesquiterpene lactones, polyacetylenes and coumarins. The training of the network was performed with the statistical package Statsoft using the backpropagation algorithm.

Results. The secondary metabolites most frequently present in the genus *Artemisia* are monoterpenes and sesquiterpene lactones. Since monoterpenes are present in almost all species, this variable is highly correlated to the variable corresponding of the number total of occurrences. Analyzing the variables corresponding to the sesquiterpene lactones, flavonoids and coumarins show that the two previous ones have similar test set and range errors (*c.a.* 0.20) while for coumarins, the error is the same, but range falls to half of that.

Conclusions. The results presented show that the mechanism of the neural networks may be effective to predict the occurrence of secondary metabolites in plant genera if an adequate network is used. In this study we show too the application of the artificial neural networks in the chemistry of natural products, a field in which the numerical precision is very small.

Keywords. *Artemisia*; asteraceae; artificial neural networks; secondary metabolites; occurrence prediction.

[#] Presented in part at the Internet Electronic Conference of Molecular Design, IECMD 2003.

* Correspondence author; phone: 55–11–3091–2056; fax: 55–11–3815–5579; E–mail: vdpemere@iq.usp.br.

1 INTRODUCTION

The chemistry of secondary metabolites is a peculiar field of study due to its complexity and the interest it raises in other fields of pharmacology. This is because of the continuous search for substances with biological activity. On the border between chemistry and biochemistry, the study of new plant species is related to the preservation of the planet's biodiversity, especially in developing countries where the destruction of the forest takes place at alarming rates.

Compiling data on secondary metabolites already isolated from plants, thus "transforming" these into a database, is a hard task that our research group has undertaken. We have centered our efforts on the Asteraceae, one of the one of the largest families of plants [1,2].

The chemistry of the Asteraceae is extremely complex, in accordance with of its biological and anatomical complexity, based on morphological criteria [3] and DNA variation [4]. Several research groups have dedicated efforts to the study of Asteraceae plants in the laboratory. Among these, the group of Bohlmann [5] stands out. At that time, about 7,000 compounds had been isolated from plants of this family. Large reviews have been produced on some classes of secondary metabolites in the Asteraceae. Sesquiterpene lactones were reviewed by Seaman [6], diterpenes were reviewed by Bohlmann *et al*. [7], flavonoids by Emerenciano [8] and Bohm [9], triterpenes by Macari [10] polyacetylenes by Bohlmann [11], benzofuranes and benzopiranes by Proksck [12]. All these data were updated and used in an analysis of the subgroups of the family, using PCA techniques [13].

The great problem with databases compiled from the literature is the lack of knowledge about the precision of the data. This is because these data are subjected to laboratory errors, studies directed to the search for a specific chemical class, etc. One has to consider also the implications of geographical and ecological factors, as well as the phenotypic pressures all influencing the production of secondary metabolites. Thus, the imprecision of the data leads us to use specific techniques to work with this kind of incomplete data.

The use of neural networks has become a routine in chemistry. A review by Zupan and Gasteiger [14] describes some of the applications of neural networks in chemistry and later the examples presented in this review were developed in a book [15]. Virtually all areas of chemistry have scientific studies published using neural networks. It is interesting to note the applications on ^{13}C and ^{1}H NMR [16,17], in IR and mass spectra [18,19]. Applications of neural networks in natural products chemistry are somewhat less frequent. The study of oxidized triterpenes or limonoids by Fraser *et al*. [20] provides an example. In chemical taxonomy our groups is working with a large database for the prediction of the occurrence of substances in genus of Asteraceae [21]. In the study mentioned we were able to predict the presence of chemical classes in *ca*. 500 genera of Asteraceae.

In the present study we focus attention at the genus *Artemisia*. We work at the infra genus level in order to try to predict the occurrence of chemical substances present in the genus.

2 METHODS

The methodology applied starts by taking all the information on the genus *Artemisia* from the database. According to Ling [22,23] there are *ca*. 400 described species of *Artemisia* in the world and our database contains chemical data on about half of these, which is fundamental for the training of the network.

An entry matrix was assembled with the occurrences of the six most representative chemical classes in the genus: monoterpenes, sesquiterpenes, sesquiterpene lactones, flavonoids, polyacetylenes and coumarins. Figure 1 describes a typical substance from each class of compounds analyzed in this study [24–27]. We define occurrences as the number of times a substance of a certain class was present in a taxon. For example, if the database contains for two different species of the same tribe i n_1 compounds belonging to the chemical class j for the first species and n_2 compounds of the same chemical class for the second species, the number of occurrences $O_{i,j}$ in the tribe i is the sum corresponding to $n_1 + n_2$. If a unique compound is isolated from two different species of the same tribe, it is counted twice.

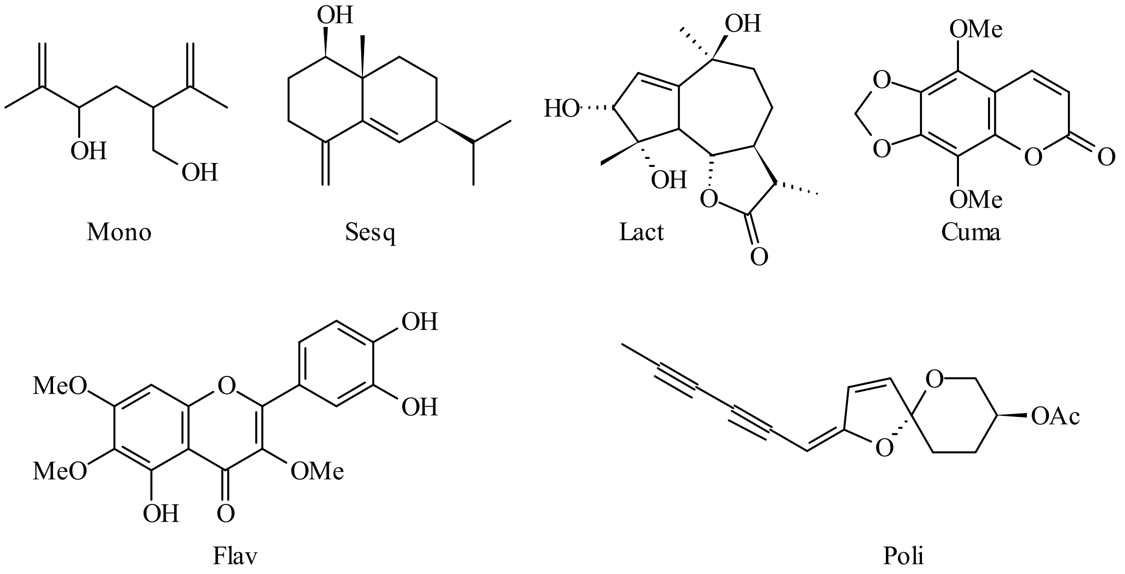

Figure 1. Typical compounds from each class of secondary metabolites analyzed in this study.

As the number of occurrences may depend on the factors mentioned above, we introduced other variables in the matrix to describe the presence or absence of the chemical classes in each case analyzed. These variables, referred to as heuristic, describe the probability of a certain species to present or not a certain chemical class. We defined arbitrarily the presence of a chemical class in a species as positive (1.0) if the value of the continuous variable of that species is greater than or equal to the average of all cases. Otherwise the heuristic variable equals zero. The twelve variables used as inputs for the network are described in Table 1.

The neural networks training for each variable was done using the "Neural Networks – Custom" option from the Statsoft for Windows. All variables were trained with the same parameters below described. Using a "custom" option, all nets were elaborated with one hidden layer. The multilayer perceptron (MLP) model was chosen with epochs number = 10000, learning rate = 0.01, momentum = 0.3 and the backprogation algorithm with a linear regression output function.

Table 1. Variables Used as Inputs for the Network

Mono =	Monoterpenes
Sesq =	Sesquiterpenes
Lact =	Sesquiterpene lactones
Cuma =	Coumarins
Poli =	Polyacetylenes
Flav =	Flavonoids
CLM =	Presence or absence of Monoterpenes
CLS =	Presence or absence of Sesquiterpenes
CLL =	Presence or absence of Sesquiterpene lactones
CLC =	Presence or absence of Coumarins
CLP =	Presence or absence of Polyacetylenes
CLF =	Presence or absence of Flavonoids
Total =	Sum of occurrences in the taxon
NCLA =	Sum of the number of classes in the taxon

The results of the network were compared with the descriptive statistics and a correlation matrix for the variables (Table 2). When ST Neural Networks reports statistics on training or network performance, separate statistics are calculated for the training, verification and test sets. Without cross verification, a network with a large number of weights can overfit the training data – learning as it were the noise present in the data rather than the underlying structure. The ability of a network not only to learn the training data, but also to perform well on previously–unseen data, is known as generalization. We can check that a network is generalizing properly in ST Neural Networks by observing whether the verification error is reasonably low. In some circumstances, we might run an iterative training algorithm and find that, although the training error decreases almost to zero, the verification error first decreases and then begins to rise again. This is a sure sign that over–learning is occurring, and one should stop training once deterioration in the verification error is observed. The ST Neural Networks software can also perform this error checking, and stop training automatically. When one trains a network with verification cases defined, ST Neural Networks plots two error lines on the graph: one for the training set and one for the verification set. The network is trained using the training set, but is also tested after each epoch using the verification set [28]. This graph was traced for the MONO variable (Figure 2). By analysing this graph, one can verify that the verification set curve (in red colour) is constant until the curve is abruptly interrupted indicating that the learning was not better. The ideal number of epochs is 100000.

Table 2. Results Obtained from the Best Neural Network

Variable	Range	Mean	Standard Deviation	Variance	Train Error	Test Error	Train Perform.	Test Perform.
Mono	171	10.105	21.756	473.312	0.10800	0.12900	0.999	0.999
Sesq	13	0.405	1.693	2.877	0.05000	0.16000	0.945	0.980
Lact	44	3.642	7.180	51.564	0.00100	0.17600	0.969	0.940
Cuma	10	0.819	1.893	3.583	0.06000	0.16900	0.999	0.949
Flav	37	1.374	4.533	20.553	0.02200	0.20100	0.999	0.948
Poli	13	0.879	2.187	4.784	0.07400	0.14900	0.976	0.964
CLM	1	0.368	0.484	0.234	0.00027	0.18373	0.929	0.920
CLS	1	0.084	0.278	0.077	0.00007	0.02418	0.939	0.917
CLL	1	0.310	0.464	0.215	0.00016	0.00080	0.836	0.786
CLC	1	0.242	0.429	0.185	0.00001	0.00002	0.919	0.840
CLF	1	0.126	0.333	0.111	0.00769	0.00702	0.895	0.855
CLP	1	0.226	0.419	0.176	0.00205	0.01992	0.956	0.807
NCLA	5	1.798	1.022	1.022	0.17100	0.23300	0.809	0.862
TOTAL	227	17.216	29.028	29.028	0.37400	0.47600	0.999	0.999

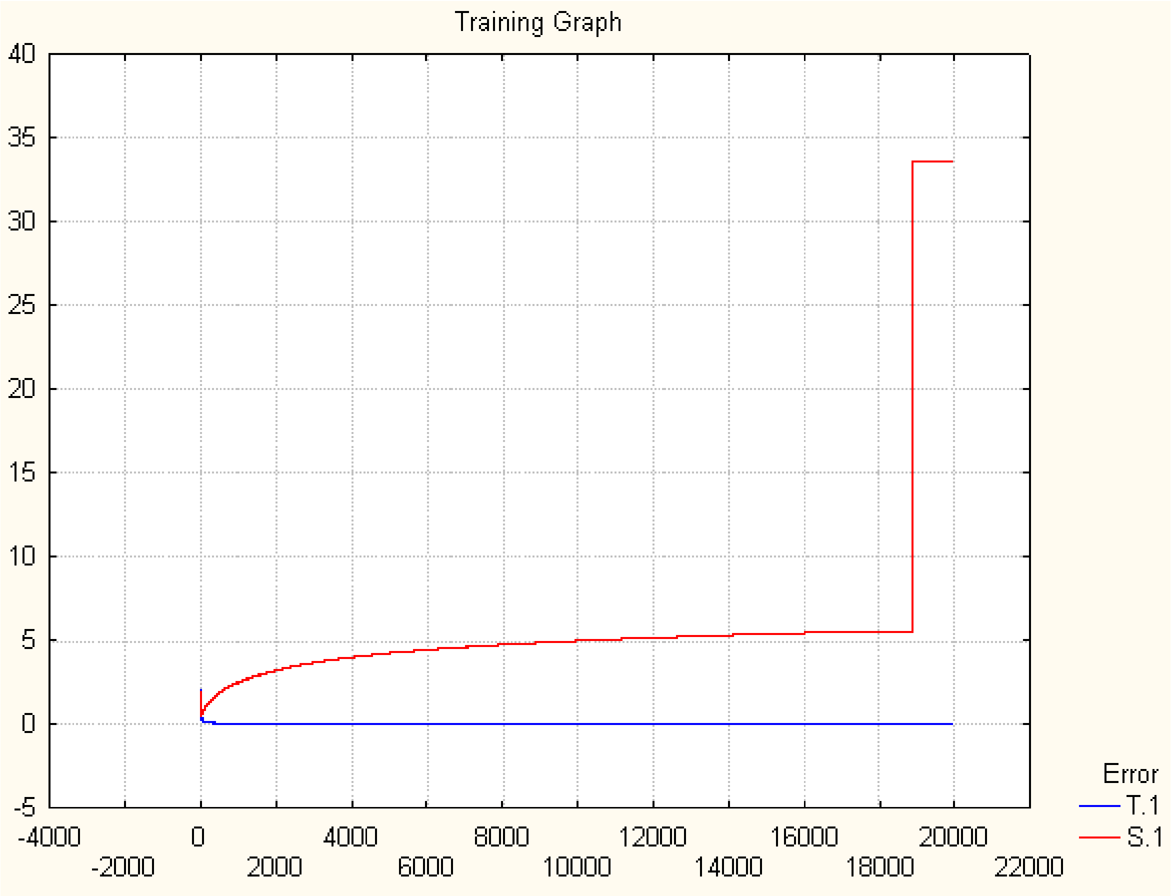

Figure 2. Training and verification sets for the MONO variable.

3 RESULTS AND DISCUSSION

It is adequate to inspect the data obtained both from the chemical and the computational points of view. The secondary metabolites most frequently present in the genus *Artemisia* are monoterpenes and sesquiterpene lactones featuring 171 and 35 occurrences, and flavonoids (range = 37). The data in Table 3 show a great variance for these data when compared with the other variables. These are also the variables where the errors of the network were the smallest both in the test set and in the training set if we take them as percentages. Obviously, the total number of substances in a certain species may also be large, if there are more studies on it in the literature. Since monoterpenes are present in almost all species, this variable (MONO) is highly correlated to the variable TOTAL.

Table 3. Correlation Matrix of the Variables in *Artemisia* Genus

	Mono	Sesq	Lact	Cuma	Flav	Poli	CLM	CLS	CLL	CLC	CLF	CLP	NCLA	TOTAL
Mono	1.00	0.32	0.21	0.34	0.50	0.41	0.53	0.23	0.10	0.11	0.21	0.25	0.49	0.95
Sesq	0.32	1.00	0.22	0.07	0.33	–0.01	0.31	0.79	0.05	–0.05	0.13	0.09	0.41	0.40
Lact	0.21	0.22	1.00	–0.13	0.21	0.11	0.13	0.32	0.65	–0.16	0.08	0.11	0.39	0.45
Cuma	0.34	0.07	–0.13	1.00	0.37	0.17	0.10	0.07	–0.20	0.71	0.10	0.23	0.35	0.36
Flav	0.50	0.33	0.21	0.37	1.00	0.15	0.17	0.23	0.05	0.19	0.65	0.10	0.42	0.63
Poli	0.41	–0.01	0.11	0.17	0.15	1.00	0.26	0.02	0.12	0.14	0.18	0.74	0.57	0.44
CLM	0.53	0.31	0.13	0.10	0.17	0.26	1.00	0.40	–0.04	–0.02	0.04	0.21	0.43	0.50
CLS	0.23	0.79	0.32	0.07	0.23	0.02	0.40	1.00	0.12	–0.04	0.11	0.06	0.42	0.34
CLL	0.10	0.05	0.65	–0.20	0.05	0.12	–0.04	0.12	1.00	–0.25	0.02	0.13	0.28	0.24
CLC	0.11	–0.05	–0.16	0.71	0.19	0.14	–0.02	–0.04	–0.25	1.00	0.04	0.19	0.27	0.13
CLF	0.21	0.13	–0.08	0.10	0.65	0.18	0.04	0.11	0.02	0.04	1.00	0.02	0.30	0.31
CLP	0.25	0.09	0.11	0.23	0.10	0.74	0.21	0.06	0.13	0.19	0.02	1.00	0.67	0.31
NCLA	0.49	0.41	0.39	0.35	0.42	0.57	0.43	0.42	0.28	0.27	0.30	0.67	1.00	0.62
TOTAL	0.95	0.40	0.45	0.36	0.63	0.44	0.50	0.34	0.24	0.13	0.31	0.31	0.62	1.00

Obviously, the most frequent secondary metabolites in the database are those most isolated by researchers and this number is debatable from several points of view. From the laboratory point of view, we can mention the throwing away of extract fractions *a priori* considered unproductive by the researcher. From the analytical point of view there are the difficulties of interpretation of spectral data (every time more improbable, since the appearance of 2D NMR techniques). From the phenotypic point of view, it is accepted among phytochemists that the production of secondary metabolites is connected to the attraction or repulsion due to interactions among plants, among plants and insects, the soil, etc, and differences among genotypes. Although we work in a low hierarchical level (genus) the taxonomic units not always constitute natural groups of species. These are also subjected to erroneous classifications and it is thus impossible to warrant the presence of a species in the same genus until studies of molecular biology [29] corroborate the morphological data. Lastly, the database is not complete because some isolated compounds may not have been published or the collection of data in the literature may not have been precise. Therefore, a small database such as the one cited here, with imprecise information becomes adequate to an inspection *via* neural networks.

The errors observed when one works with large number of occurrences must be regarded carefully in function of the factors mentioned. As an example of this, a simple inspection of the variables LACT, FLAV CUMA and TOTAL in Tables 1 and 2 show that the two previous ones have similar test set and range errors (*c.a.* 0.20) while for CUMA, the error is the same, but the range decreases to half of that. For the TOTAL variable the error for the test set is 0.13 and it includes all other variables.

The use of heuristic variables (presence/absence) based on the average of every variable was adopted here to try to introduce variables that escape the above–mentioned sources of error.

4 CONCLUSIONS

The results presented in Table 2 show that the mechanism of the neural networks may be effective to predict the occurrence of secondary metabolites in plant genera if an adequate network is used. In this study we show the application of the ANN in the Chemistry of Natural Products, a field in which the numerical precision is very small.

It was also demonstrated that the problem of the predictability may be treated through the ANN in a vast universe such as phytochemistry. As one works in very low hierarchical levels in the botanical point of view, the errors in the data may be even greater. But even so, the mechanism of neural networks, and, more precisely, the adequate training with the best cases of the available sample were able to produce results that may be considered good. It was also observed that, in spite of the strong ability to deal with imprecise data, the skill of the chemist to analyze the first training sets was paramount for the performance of the ANN.

The use of mathematical models in biology is not new, but models that are general and can produce good predictions will be useful to direct, or at least orient, laboratory work in search of new active principles of plants.

Acknowledgment

This work was supported by grants from the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) and by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

5 REFERENCES

[1] K. Bremer, *Asteraceae. Cladistics and Classification*, Timber Press, Portland, Oregon, 1994.

[2] K. Bremer, Major clades and grades of the Asteraceae; in: Proceedings of the International Compositae Conference, Eds. D. J. N. Hind and H. J. Beentje, Royal Botanic Gardens, Kew, 1996, *1*, Systematics, pp. 1–7.

[3] K. R. Sporne, Ovule as an indicator of evolutionary status in angiosperms, *New Phytol*. **1969**, *68*, 555–566.

[4] R. K. Jansen, K. E. Holsinger, H. J. Michaels and J. D. Palmer, Phylogenetic analysis of chloroplast DNA restriction site data at higher taxonomic levels: an example from the Asteraceae, *Evolution* **1990**, *44*, 2089–2105.

[5] C. Zdero and F. Bohlmann, Systematics and evolution within the Compositae, seen with the eyes of a chemist, *Pl. Syst. Evol*. **1990**, *171*, 1–14.

[6] F. Seaman, Sesquiterpene lactones as taxonomic markers in the Asteraceae, *Bot. Rev.* **1982**, *48*, 121–595.

[7] F. Seaman, F. Bohlmann, C. Zdero and T. J. Mabry, *Diterpenes in flowering plants: Compositae* (*Asteraceae*), Springer, New York, 1990.

[8] V. P. Emerenciano, M. A. C. Kaplan, O. R. Gottlieb, Evolution of sesquiterpene lactones in angiosperms, *Biochem. Syst. Ecol.* **1985**, *13*, 145–166.

[9] B. A. Bohm and T. F. Stuessy, *Flavonoids of the sunflower family*, Springer–Wien, New York, 2001

[10] P. A. T. Macari, J. P. Gastmans, G. V. Rodriguez and V. P. Emerenciano, An expert system for structure elucidation of triterpenes, *Spectroscopy– An International Journal* **1994/1995**, *12*, 139–166.

[11] F. Bohlmann, T. Burkhardt and C. Zdero, *Naturally occurring acetylenes*, Academic Press, London, 1973.

[12] P. Proksch and E. Rodriguez, Chromenes and benzofuranes of the Asteraceae, their chemistry and biological significance, *Phytochemistry* **1983**, *22*, 2335–2355.

[13] S. A. V. Alvarenga, M. J. P. Ferreira, V. P. Emerenciano and D. Cabrol–Bass, Chemosystematic studies of natural compounds isolated from Asteraceae. Characterization of tribes by principal component analysis, *Chem. Intell. Lab. Syst.* **2001**, *56*, 27–37.

[14] J. Gasteiger and E. Zupan, Neural networks in Chemistry. *Angew. Chem. Int. Ed. Engl.* **1993**, *32*, 503–527.

[15] E. Zupan and J. Gasteiger, *Neural Networks for Chemists – An Introduction*, VHC–Verlag, Weinhein, 1993.

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

[17] J. Aires–de–Souza, M. C. Hemmer and J. Gasteiger, Prediction of ^{1}H NMR chemical shifts using neural networks, *Anal. Chem.* **2002**, *74*, 80–90.

[18] C. Cleva, C. Cachet and D. Cabrol–Bass, Clustering of infrared spectra with Kohonen networks. *Analusis* **1999**, *27*, 81–90.

[19] H. Lohninger and F. Stanci, Comparing the performance of neural networks to well–established methods of multivariate data analysis: the classification of mass spectral data, *Fresenius J. Anal. Chem.* **1992**, *344*, 188–189.

[20] L. Fraser, D. A. Mulholland and D. D. Fraser, Classification of limonoids and protolimonoids using neural networks. *Phytochem. Anal.* **1997**, *8*, 301–311.

[21] M. J. P. Ferreira, A. J. C. Brant, S. A. V. Alvarenga, F. M. M. Magri, A. R. Rufino and V. P. Emerenciano, Prediction of occurrences of diverse chemical classes in Asteraceae through neural networks, *Phytochem. Anal.* **2003**, in press.

[22] Y. R. Ling, The Old World *Artemisia* Linn. (Compositae). *Bull. Bot. Res. Harbin* **1992**, *12*, 1–108.

[23] Y. R. Ling, The New World *Artemisia*; in: *Advances in Compositae Systematics*, Eds. D. J. N. Hind, C. Jeffrey and G. V. Pope, Royal Botanic Gardens, Kew, 1995, pp. 255–281.

[24] G. D. Brown, G. –Y. Liang and L. –K. Sy, Terpenoids from the seeds of *Artemisia annua*, *Phytochemistry* **2003**, *64*, 303–323.

[25] R. X. Tan, H. Q. Tang, J. Hu and B. Shuai, Lignans and sesquiterpene lactones from *Artemisia sieversiana* and *Inula racemosa*, *Phytochemistry* **1998**, *49*, 157–161.

[26] T.–S. Wu, Z.–J. Tsang, P.–L. Wu, F.–W. Lin, C.–Y. Li, C.–M. Teng and K.–H. Lee, New constituents and antiplatelet aggregation and anti–HIV principles of *Artemisia capillaris*, *Bioorg. Med. Chem.* **2001**, *9*, 77–83.

[27] J. A. Marco, J. F. Sanz–Cervera, F. Sancenón, M. Arnó and J. Vallès–Xirau, Sesquiterpene lactones and acetylenes from *Artemisia reptans*, *Phytochemistry* **1994**, *37*, 1095–1099.

[28] Statistica Neural Networks, Statsoft, Inc. Tulsa, OK, EUA, 2000.

[29] E. E. Schilling and J. L. Panero, A revised classification of subtribe Helianthinae (Asteraceae: Heliantheae). I. Basal lineages, *Bot. J. Linn. Soc.* **2002**, *140*, 65–76.